Large-scale Inference in the Focally Damaged Human Brain

TIANBO XU

Queen Square Institute of Neurology

University College London

Submitted in partial fulfillment of the requirements for a degree of Doctor of Philosophy

Declaration

I, Tianbo Xu confirm that the work presented in this thesis is my own. It has been carried out at the UCL Queen Square Institute of Neurology under the supervision of Prof Parashkev Nachev and Prof Geraint Rees between April 2014 and September 2019. The partial work has been presented at national and international conferences and is published in peer-reviewed journals. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

London, September 2019

Acknowledgements

I have been extremely fortunate to pursue my PhD under the supervisions of Prof Parashkev Nachev and Prof Geraint Rees at Queen Square Institute of Neurology, UCL. None of the work in this thesis would have been accomplished without their support, advice and guidance.

My heartfelt thanks go to Prof Parashkev Nachev who was an outstanding and generous mentor and supervisor. I am especially indebted to him for igniting my interest in neuroimaging and giving me the chance to be a part of his team. The work presented in this thesis was in the context of his program of research funded by Wellcome Trust and the National Institute for Health Research University College London Hospitals Biomedical Research Center.

Prof Geraint Rees's support was immeasurable, and I owe him a huge debt of gratitude for supervising me.

Prof Rolf Jager generously provided the full radiologist reports recorded in National Hospital for Neurology and Neurosurgery which can be traced back to 2004. Without his invaluable support, we would not have been able to collect a large-scale smoothly to conduct our studies.

The consistent support from my wife, father and parents-in-law strengthened me. I am immensely grateful to them.

Dedicated to the memory of G Sun, “Mum” (1952 - 2009).

"Stop trying to change reality by eliminating complexity."

– David Whyte

Abstract

Clinical outcomes in focal brain injury reflect the interactions between two distinct anatomically distributed patterns: the functional organisation of the brain and the structural distribution of injury. The challenge of understanding the functional architecture of the brain is familiar; that of understanding the lesion architecture is barely acknowledged. Yet, models of the functional consequences of focal injury are critically dependent on our knowledge of both.

The studies described in this thesis seek to show how machine learning-enabled high-dimensional multivariate analysis powered by large-scale data can enhance our ability to model the relation between focal brain injury and clinical outcomes across an array of modelling applications. All studies are conducted on internationally the largest available set of MR imaging data of focal brain injury in the context of acute stroke (*N=1333*) and employ kernel machines at the principal modelling architecture.

First, I examine lesion-deficit prediction, quantifying the ceiling on achievable predictive fidelity for high-dimensional and low-dimensional models, demonstrating the former to be substantially higher than the latter. Second, I determine the marginal value of adding unlabelled imaging data to predictive models within a semi-supervised framework, quantifying the benefit of assembling unlabelled collections of clinical imaging. Third, I compare high- and low-dimensional approaches to modelling response to therapy in two contexts: quantifying the effect of treatment at the population level (therapeutic inference) and predicting the optimal treatment in an individual patient (prescriptive inference). I demonstrate the superiority of the high-dimensional approach in both settings.

Impact Statement

Fundamentally, the series of studies carried in this PhD project is to propose a high-dimensional multivariate approach methodologically and conceptually so as to improve our understanding on focally damaged human brain, including but not limited to the contexts of lesion prediction, therapeutic and prescriptive inference. The complex distributed structural architecture of human brain lesions is the principal reason for adopting this high-dimensional multivariate analysis to revolutionise the stereotyped conventional mass-univariate methods. To parameterise the intrinsic complexity of brain lesions, we applied machine learning techniques to provide feasible solution for the corresponding computational complexity. In this way, the inextricably intertwined voxels of brain lesions were modelled as a whole to reveal the hidden patterns and interactions that could be responsible to the neurological deficits.

The results derived from the series of experiments dedicated to this PhD program showed their impacts in a few aspects. Firstly, they revealed the reality that the complexities of the architecture of lesions is comparable to that of the macroscopic functional organisation of the brain. Explicitly, to characterise lesion architecture with high-dimensional modelling becomes the crucial prerequisite to capture the relationships between lesion and outcome not only for clinical prediction but functional mapping. This compels us to conclude the conventional studies of inferences about human brain lesion-deficit mapping should be re-evaluated with the approaches involving the complexity of lesion architecture sufficiently. Secondly, the results shown in our therapeutic inference study have substantial implications for translational research, providing one remediable explanation for the common failure of interventional studies in humans involving agents shown to be effective in simpler animals. In other words, by applying machine learning techniques, to model the full complexity of human brain with the proposed high-dimensional multivariate approach is expected to uncover the treatment effects that would otherwise be missed. We hope our methods can be widely used among both researchers and clinicians when they run a clinical trial. Thirdly, our study of prescriptive inference demonstrated the complexity of the human brain interacts with the complexity of the pathological process determining the intrinsic dimensionality of brain lesions, which would be applicable to a wide range of prediction and inference in medical domain to contribute to the realm of personalised medicine.

In the domain of neuroscience, the real value of machine learning is to formalise the complex decision, which perfectly matched the intrinsic complexity of human brain. Inevitably, we are motivated to apply the state-of-art computational techniques combined with a sustainably growing large-scale data set. Thus, we are encouraged to apply for further research grants to continuously emphasise and amplify the contribution of high-dimensional modelling in understanding of human brain. The positive findings from this PhD project will definitely enhance our position to do this. Meanwhile, based on the above contexts, our profile is strongly capable to establish a wide range of collaborations with both academic and industrial organisations nationally and internationally.

Contents

DECLARATION ... 1

ACKNOWLEDGEMENTS ... 2

ABSTRACT ... 5

IMPACT STATEMENT ... 6

CONTENTS ... 8

LIST OF FIGURES ... 15

LIST OF TABLES ... 17

PUBLICATIONS ASSOCIATED WITH THIS THESIS ... 18

CHAPTER 1 INTRODUCTION ... 19

1.1 OVERVIEW ... 19

1.1.1 Transcranial magnetic stimulation and direct current stimulation ... 20

1.1.2 Functional and structural neuroimaging ... 22

1.1.3 Lesion studies ... 23

1.2 INDIVIDUALITY: A HIGH-DIMENSIONAL CHARACTERISTIC ... 25

1.3 LARGE-SCALE STUDIES OF THE FOCALLY DAMAGED HUMAN BRAIN ... 29

1.3.1 The rationale for “big data” ... 29

1.3.2 Obtain large-scale data ... 31

1.3.3 The characteristics of large-scale data ... 33

1.3.4 The challenge of a large-scale, multi-site, cross-sessional imaging studies ... 35

1.4 LESION-FUNCTION INFERENCE ... 38

1.4.1 Univariate modelling ... 38

1.4.1.1 Region of interest (ROI) based lesion-symptom mapping (overlap-subtraction method) 39

1.4.1.2 Voxel-based lesion-symptom mapping (VLSM) ... 41

1.4.2 Multivariate approaches ... 43

1.5 MULTIVARIATE BASED MACHINE LEARNING ALGORITHMS ... 46

1.5.1 Supervised learning algorithms ... 46

1.5.1.1 Datasets for learning algorithms ... 46

1.5.1.2 The choice of learning algorithms ... 47

1.5.1.2.1 Bias and variance ... 48

1.5.1.2.2 Properties of the input data ... 50

1.5.1.3 Support vector machines (SVMs) ... 51

1.5.1.3.1 The separating hyperplane and kernel trick ... 51

1.5.1.3.2 The hyperplane with maximum margin ... 56

1.5.1.3.3 Soft margins ... 58

1.5.1.3.4 Kernel functions and the selection of kernels ... 61

1.5.1.3.5 Extensions of SVMs ... 62

1.5.2 Semi-supervised learning algorithms ... 63

1.5.2.1 Inductive learning and transductive learning ... 65

1.5.2.2 Semi-supervised learning assumptions ... 66

1.5.2.3 Semi-supervised learning methods ... 67

1.5.2.3.1 Self-training ... 67

1.5.2.3.2 Co-training ... 69

1.5.2.3.3 Generative models ... 70

1.5.2.3.4 Discriminative models ... 73

1.5.2.3.4.1 Graph-based methods ... 74

1.5.2.3.4.1.1 Mincut ... 76

1.5.2.3.4.1.2 Harmonic function ... 77

1.5.2.3.4.1.3 Manifold regularisation ... 78

1.5.2.3.4.2 Transductive support vector machines (TSVMs) ... 79

1.5.3 Multi-label algorithms ... 84

1.5.3.1 The key concerns in multi-label classification ... 86

1.5.3.2 The overview of methods for multi-label classification ... 86

1.5.3.2.1 The problem transformation methods ... 87

1.5.3.2.1.1 Transforming to binary classification ... 87

1.5.3.2.1.2 Transforming to label ranking ... 89

1.5.3.2.1.3 Transforming to multi-class classification ... 90

1.5.3.2.2 Algorithm adaptation methods ... 91

1.5.3.2.2.1 Multi-label decision tree ... 91

1.5.3.2.2.2 Multi-label support vector machines ... 92

1.5.3.3 Evaluation metrics ... 95

1.5.3.3.1 Example-based metrics ... 95

1.5.3.3.2 Label-based metrics ... 97

1.5.4 Accuracy estimation and assessment of model generalizability ... 98

1.5.4.1 Cross-validation ... 101

1.5.4.1.1 Exhaustive cross-validation 101
1.5.4.1.2 Non-exhaustive cross-validation 102
1.5.4.1.3 Cross-validation methods and model selection 103
1.5.4.2 The bootstrap method 105
1.6 SUMMARY 107
CHAPTER 2 THE LIMITS OF BEHAVIOURAL OUTCOME PREDICTION FOLLOWING FOCAL BRAIN INJURY 109
2.1 INTRODUCTION 109
2.2 METHODS 114
2.2.1 Patients 114
2.2.2 Imaging 114
2.2.2.1 MR data acquisition 114
2.2.2.2 Image pre-processing 115
2.2.2.3 Lesion segmentation 116
2.2.3 Dimensionality reduction 118
2.2.4 Simulations 118
2.2.4.1 Simulation one: lesion-deficit prediction depending on single Brodmann region performed by a mass-univariate technique (Fisher’s exact test) against a multivariate technique (SVM) 120
2.2.4.1.1 Data preparation 120
2.2.4.1.2 Lesion-deficit models for clinical prediction 121
2.2.4.1.3 Mass-univariate (Fisher’s exact test) analysis 121
2.2.4.1.4 Multivariate (SVM) analysis 122
2.2.4.1.5 Comparison of behavioural outcome prediction with mass-univariate and multivariate analysis 123
2.2.4.2 Simulation two: lesion-deficit prediction depending on dual Brodmann regions performed by a mass-univariate technique (Fisher’s exact test) against a multivariate technique (SVM) 124
2.2.4.2.1 Data preparation 124
2.2.4.2.2 Lesion-deficit model for clinical prediction 124
2.2.4.2.3 Mass-univariate (Fisher’s exact test) analysis 125
2.2.4.2.4 Multivariate (SVM) analysis 126
2.2.4.2.5 Comparison of behavioural outcome prediction between mass-univariate and multivariate analysis 127
2.2.5 Software and hardware 127
2.3 RESULTS 128

2.3.1 Simulation one: lesion-deficit prediction dependent on single Brodmann regions performed by a mass-univariate technique (Fisher's exact test) versus a multivariate technique (SVM) 128
2.3.1.1 Comparison of Fisher's exact test against SVM 128
2.3.2 Simulation two: lesion-deficit prediction dependent on dual Brodmann regions performed by a mass-univariate technique (Fisher's exact test) versus a multivariate technique (SVM) 133
2.3.2.1 Comparison of Fisher's exact test against SVM 133
2.4 DISCUSSION 137
2.4.1 Simulation one: lesion-deficit prediction dependent on single Brodmann regions performed by a mass-univariate technique (Fisher's exact test) versus a multivariate technique (SVM) 137
2.4.2 Simulation two: lesion-deficit prediction dependent on dual Brodmann regions performed by a mass-univariate technique (Fisher's exact test) versus a multivariate technique (SVM) 139
2.5 CONCLUSION 141
CHAPTER 3 THE BENEFITS OF UNLABELLED DATA IN LESION-DEFICIT PREDICTION 142
3.1 INTRODUCTION 142
3.2 METHOD 146
3.2.1 Patients 146
3.2.2 Imaging 146
3.2.2.1 MR data acquisition 146
3.2.3.2 Image pre-processing 147
3.2.2.3 Lesion segmentation 148
3.2.3 Dimensionality reduction 150
3.2.4 Technique 150
3.2.5 Simulations 151
3.2.5.1 Simulation one: single Brodmann area based lesion-deficit prediction performed by a supervised learning technique (SVM) versus a semi-supervised learning technique (transductive SVM) 151
3.2.5.1.1 Data preparation 151
3.2.5.1.2 Lesion-deficit models for clinical prediction 152
3.2.5.1.3 Supervised learning models (SVM) 153
3.2.5.1.4 Semi-supervised learning models (TSVM) 154

3.2.5.1.5 Performance comparison between supervised learning and semi-supervised learning approaches 155
3.2.5.2 Simulation two: dual Brodmann area based lesion-deficit prediction performed by a supervised learning technique (SVM) versus a semi-supervised learning technique (transductive SVM)........ 158
3.2.5.2.1 Data preparation........ 158
3.2.5.2.2 Lesion-deficit model for clinical prediction........ 158
3.2.5.2.3 Supervised learning models (SVM)........ 159
3.2.5.2.4 Semi-supervised learning models (TSVM)........ 159
3.2.5.2.5 Performance comparison between supervised learning and semi-supervised learning approaches 160
3.2.6 Software and hardware........ 161
3.3 RESULTS........ 162
3.3.1 Simulation one: single Brodmann region based lesion-deficit prediction performed by a supervised approach (SVM) versus a semi-supervised approach (TSVM) 162
3.3.1.1 Performance comparison between supervised learning and semi-supervised learning models 162
3.3.2 Simulation two: dual Brodmann regions based lesion-deficit prediction performed by a supervised approach (SVM) versus a semi-supervised approach (TSVM) 171
3.3.2.1 Performance comparison between supervised learning and semi-supervised learning models 171
3.4 DISCUSSION........ 181
3.4.1 Simulation one: lesion-deficit prediction depending on single Brodmann region performed by a supervised learning method (SVM) versus a semi-supervised learning method (TSVM) 181
3.4.2 Simulation two: lesion-deficit prediction depending on a combinatorial pair of Brodmann regions performed by a supervised learning method (SVM) against a semi-supervised learning method (TSVM)........ 182
3.5 CONCLUSION 184

CHAPTER 4 THERAPEUTIC INFERENCE IN THE FOCALLY DAMAGED HUMAN BRAIN 185

4.1 INTRODUCTION........ 185
4.2 METHODS 188
4.2.1 Patients........ 188
4.2.2 Imaging 188
4.2.2.1 CT data acquisition 188

4.2.2.2 CT data analysis ... 189
4.2.2.2.1 Imaging pre-processing ... 189
4.2.2.2.2 Semi-automatic gaze segmentation ... 189
4.2.2.3 MR data acquisition ... 190
4.2.2.4 MR data analysis ... 192
4.2.2.4.1 DWI pre-processing ... 192
4.2.2.4.2 T2 pre-processing and semi-automatic gaze segmentation ... 193
4.2.2.5 Lesion segmentation ... 193
4.2.2.6 Behaviour: gaze deviation ... 195
4.2.3 Therapeutic inference ... 199
4.2.3.1 Data preparation ... 199
4.2.3.2 Lesion non-altering interventions ... 199
4.2.3.3 Lesion altering interventions ... 200
4.2.3.4 Low- and high-dimensional models ... 200
4.2.4 Software and hardware ... 204
4.3 RESULTS ... 205
4.3.1 Non-lesion altering intervention ... 207
4.3.2 Lesion-altering intervention ... 210
4.4 DISCUSSION ... 215
4.5 CONCLUSION ... 218
CHAPTER 5 THE IDENTIFICATION OF INDIVIDUAL SUSCEPTIBILITY TO TREATMENTS WITH PRESCRIPTIVE INFERENCE ... 219
5.1 INTRODUCTION ... 219
5.2 METHOD ... 222
5.2.1 Patients ... 222
5.2.2 Imaging ... 222
5.2.2.1 MR data acquisition ... 222
5.2.5.2 Image pre-processing ... 223
5.2.2.3 Lesion segmentation ... 224
5.2.3 Dimensionality reduction ... 226
5.2.4 Synthetic ground truths of lesion-deficit and susceptibility to treatments ... 227
5.2.5 Prescriptive inference ... 232
5.2.5.1 Prescriptive inference with a positive response to a given treatment ... 233
5.2.5.2 Prescriptive inference with a polarised response to a given treatment ... 233
5.2.5.3 Low- and high-dimensional models ... 234
5.2.5.3.1 Model implementation ... 235

5.2.5.3.2 Model evaluation......235
5.2.6 Software and hardware......237
5.3 RESULTS......238
5.3.1 Prescriptive inference with positive response to a given treatment......238
5.3.2 Prescriptive inference with a polarised response to a given treatment......240
5.3.3 Comparison between the two series of simulations.......242
5.4 DISCUSSION......250
5.5 CONCLUSION......253
CHAPTER 6 GENERAL DISCUSSION......254
6.1 THE LIMITS OF LESION-DEFICIT PREDICTION......256
6.2 ENHANCED DETECTABILITY OF THERAPEUTIC EFFECTS BY HIGH-DIMENSIONAL MODELLING 260
6.3 HIGH-DIMENSIONAL MODELLING OF HETEROGENEOUS TREATMENT EFFECTS......262
6.4 LIMITATIONS AND FUTURE WORK......264
CHAPTER 7 APPENDIX......267
7.1 APPENDIX A: THE DETAILS OF SCANNERS INVOLVED INTO THIS THESIS......267
7.2 APPENDIX B: DEMOGRAPHIC INFORMATION......269
CHAPTER 8 BIBLIOGRAPHY......271

List of figures

Figure 1.1 Absence of individuality..26

Figure 1.2 Reproduction of individual face 27

Figure 1.3 The difficulties of high-dimensional data..31

Figure 1.4 The systematic bias of template overlay method (Mah et al., 2014b)..45

Figure 1.5 Bias-variance trade-off 49

Figure 1.6 The complexity versus prediction error (Fortmann-Roe, 2012)..50

Figure 1.7 Separating planes of SVM..55

Figure 1.8 The maximum margin 58

Figure 1.9 Soft-margin of SVMs 61

Figure 1.10 The contribution of unlabelled data examples in semi-supervised learning..65

Figure 1.11 Basic concept of graph-based methods in semi-supervised learning 76

Figure 1.12 The comparison between TSVMs and inductive SVMs..80

Figure 1.13 underfitting, balanced and overfitting..100

Figure 2.1 Lesion clustering by lesion volume 119

Figure 2.2 Lesion distribution..120

Figure 2.3 The comparison of AUC between Fisher’s exact test and SVM 130

Figure 2.4 Comparison of ROC curve between Fisher’s exact test and SVM..132

Figure 2.5 Differences between Fisher’s exact test and LIBSVM across 819 combinatorial two Brodmann regions..135

Figure 2.6 Comparison of ROC curve between Fisher’s exact test and SVM..136

Figure 3.1 Lesion distribution..152

Figure 3.2 Balanced accuracy and AUC..157

Figure 3.3 Balanced accuracy 163

Figure 3.4 AUC scores..165

Figure 3.5 ROC curve (subset portion: 30%) 166

Figure 3.6 ROC curve (subset portion: 45%) 167

Figure 3.7 ROC curve (subset portion: 60%) 168

Figure 3.8 ROC curve (subset portion: 75%) 169

Figure 3.9 ROC curve (the whole lesion set)..170

Figure 3.10 balanced accuracy..173

Figure 3.11 AUC scores..174

Figure 3.12 ROC curve (subset portion: 30%) 176

Figure 3.13 ROC curve (subset portion: 45%) 177
Figure 3.14 ROC curve (subset portion: 60%) 178
Figure 3.15 ROC curve (subset portion: 75%) 179
Figure 3.16 ROC curve (the whole lesion set) 180
Figure 4.1 Gaze segmentation 190
Figure 4.2 Distribution of patient gaze on admission (CT) 197
Figure 4.3 Distribution of patient gaze at the time point of the MRI 198
Figure 4.4 Therapeutic function 203
Figure 4.5 Voxel-wise tuning of gaze deviation on admission 205
Figure 4.6 High-dimensional classifier derived from SVMLin 206
Figure 4.7 Empirical therapeutic inference function of non-lesion altering intervention 209
Figure 4.8 Empirical therapeutic inference function of non-lesion altering intervention (with additional co-factors) 210
Figure 4.9 Empirical therapeutic inference function of lesion-altering intervention 213
Figure 4.10 Empirical therapeutic inference function of lesion-altering intervention (with additional co-factors) 214
Figure 5.1 Two-dimensional clusters of lesion anatomy (Bonkhoff et al., 2019) 228
Figure 5.2 Ground truth: the separation of affected and unaffected lesions 229
Figure 5.3 Ground truth: the separation of susceptibility 231
Figure 5.4 The components of violin plot 237
Figure 5.5 Sensitivities in low- and high-dimensional models 239
Figure 5.6 Sensitivities in low- and high-dimensional models 241
Figure 5.7 Identification of patients with treatment (efficacy ratio: 10%) 244
Figure 5.8 Identification of patients with treatment (efficacy ratio: 30%) 245
Figure 5.9 Identification of patients with treatment (efficacy ratio: 50%) 246
Figure 5.10 Identification of patients with treatment (efficacy ratio: 70%) 247
Figure 5.12 Identification of patients with treatment (efficacy ratio: 90%) 248
Figure 5.12 Identification of patients with treatment (efficacy ratio: 100%) 249
Figure 7.1 Age and sex distribution 269
Figure 7.2 Ethnical group 270

List of tables

Table 1.1 Common kernel functions 62
Table 2.1 The necessity of high-dimensional model in lesion-deficit mapping (Xu et al., 2017a) 112
Table 2.2 The detailed performance between Fisher's exact test and SVM 131
Table 3.1 The details of balanced accuracy 164
Table 3.2 AUC scores 166
Table 3.3 balanced accuracy 174
Table 3.4 balanced accuracy 175
Table 5.1 The separation of lesions by the underlying functional deficit 230
Table 5.2 The separation of lesions by the susceptibility to a specific treatment 232
Table 5.3 Simulation one: sensitivity 240
Table 5.4 Simulation two: sensitivity 242
Table 7.1. The list of scanner details 268

Publications associated with this thesis

XU, T., JHA, A. & NACHEV, P. 2017a. The dimensionalities of lesion-deficit mapping. *Neuropsychologia*.

XU, T., ROLF JÄGER, H., HUSAIN, M., REES, G. & NACHEV, P. 2017b. High-dimensional therapeutic inference in the focally damaged human brain. *Brain*.

Chapter 1 Introduction

1.1 Overview

One of the major challenges of neuroscience is to identify the macroscopic functional architecture of human brain. Our current understanding of the relationships between brain anatomy and a specific function or deficit is derived from the studies conducted by a variety of techniques over the development of neuroscience. Lesion studies, usually referred as lesion-symptom mapping or lesion-behaviour mapping, represent the first technique, traceable back to the middle of the 19th century. For much of the history of neuroscience, it has been the most important means of localizing brain function by revealing the relation between brain lesions and clinical outcomes (Rorden et al., 2007, Timmann et al., 2016). Initially, lesion studies relied on post- mortem examination to investigate how the observed behaviours of patients could be linked with localised brain damage (Broca, 1861, Wernicke, 1874). Later, the invention of non-invasive brain imaging (typically, computed tomography (CT) and MRI) technology revolutionised our ability to visualise the human brain, facilitating group studies of both patients and control subjects, and significantly improving the spatial resolution for analysis (Bates et al., 2003). Nevertheless, in the last two or three decades, the dominance of lesion studies has faded owing to the emergence of more recent techniques such as transcranial magnetic stimulation (TMS) (Barker et al., 1985), transcranial direct current stimulation (tDCS) and especially functional and structural magnetic resonance imaging (sMRI and fMRI) (Rorden and Karnath, 2004, Rorden et al., 2007, Rorden et al., 2009). I wish to begin by evaluating the array of tools now available in order to conclude whether or not lesion-symptom mapping still retains its exclusive value for investigating the functional anatomy of the brain. Furthermore, I wish to examine whether or not its values can be enhanced by a methodologically and conceptually upgraded approach to analysis.

1.1.1 Transcranial magnetic stimulation and direct current stimulation

Founded on the principles of electromagnetic induction, TMS is a non-invasive technique widely applied to investigate the functional consequences of focal brain disruption. In practice, an electromagnetic coil held over the exterior surface of the cranium is discharged with a pulse of current, producing a transient magnetic field in the brain, which induces a small current within the tissue itself. As the living tissue is electrically conductive, the induced electric field causes the depolarisation and hyperpolarisation of the underlying neurons, which transiently disturbs brain function (Sack and Linden, 2003, Walsh and Cowey, 2000).

From a methodological perspective, TMS can be experimentally designed to address a wide of range of hypotheses in terms of localisation, timing and functional relevance. Multiple trains of repetitively delivered pulses (rTMS) can disrupt activity over longer time periods, and has been used to induce plasticity through after-effects (Klomjai et al., 2015). As a tool for anatomical inference, the induced focal depolarisation and hyperpolarisation during TMS sessions enables within-subject comparisons between two functional states, which can increase the inferential power dramatically (Mah, 2014b).

Nonetheless, TMS is constrained in anatomical range, and its effects are difficult to interpret. The technical limitation is the depth of penetration and spatial accuracy owing to the anatomy of the skull and the greater impedance of grey matter compared with white matter. The effects of TMS are consequently restricted to superficial cortex and cannot reach the tissue in deep medial and subcortical structures (e.g. basal ganglia and thalamus) (Klomjai et al., 2015, Walsh and Cowey, 2000). Although stimulation could be powered to penetrate subjacent white matter, the size of the induced current may result in epileptic seizures or harm to brain tissue (Wassermann and Lisanby, 2001). In addition, as the electric field is induced by an annulus current beneath the coil, the stimulation is not tightly focal. Hence, spatial accuracy will be affected (Ridding and Rothwell, 2007). The difficulty of interpretation arises from the transient and ill-defined nature of the disruption, typically confined to a single location, generated in the context of the extremely complex functional organisation of the underlying neuronal activity. Diaschisis of functional connectivity, for example, potentially allows for multiple interpretations (Ruff et al., 2009, Sack and Linden, 2003).

To counteract these limitations, “oversampling” with multiple stimulations applied to overlapping regions may improve spatial accuracy. Multi-site stimulations may also extend the spatial range, and to examine the impact of simultaneous and synchronous focal deactivation. Nonetheless, the maximum number of concurrent stimulations is a practical limitation to the effectiveness of such an approach. Temporal resolution is equally important. Multiple neurons will be activated simultaneously in the delivery of a TMS pulse, the highest signal-to-noise ratio theoretically achieved at the moment of maximal activation. There will likely be non-linear effects here, disruption potentially giving way to facilitation in a way that varies with time. In other words, the exact timing is difficult to control to ensure the best signal-to-noise ratio (Walsh and Cowey, 2000).

Transcranial direct current stimulation (tDCS) is another non-invasive technique to interfere with the human brain through an intact scalp. Direct current is delivered at low density between two surface electrodes in order to facilitate or inhibit neuronal activity (Brunoni et al., 2012). In contrast to TMS, which induces transient depolarisation and hyperpolarisation to disrupt the brain, tDCS is a neuro-modulatory intervention that modifies neuronal excitability (Nitsche et al., 2008). In particular, tDCS can induce antagonistic effects to modulate cortical excitability. This unique characteristic, coupled with its low-cost and portability, has made tDCS a common tool for exploring neuroplasticity and especially neuropsychiatric disorders (Wagner et al., 2007, Brunoni et al., 2012, Nitsche et al., 2002). Nonetheless, it is severely limited in its spatial resolution, lacking both *focality* and accuracy. The electric field yielded by tDCS is non-focal and the affected range can cover the whole area between the electrodes. Moreover, in order to assure a weak current through the scalp and minimise the sensation, the size of electrodes is usually designed to be relatively large – typically, approximately 2500 mm^2. Although smaller electrodes have become available in the recent years, their safety has not been sufficiently demonstrated (Priori et al., 2009). Therefore, the use of tDCS for anatomical inference is bound to be associated with poor precision of localisation and highly uncertain inference.

1.1.2 Functional and structural neuroimaging

Functional MRI (Ogawa et al., 1990a) has enabled us to visualise the functioning of human brain tissue in response to a variety of tasks, including vascular and metabolic reactions (Kropotov, 2016). Technically, blood oxygenation level dependent (BOLD) contrast is measured to identify the local changes in brain activity that are presumed to drive the vascular changes that generate it. BOLD reflects the complex associations and interactions between neuronal activity, oxygen saturation, and as well as blood flow and volume. Compared with brain stimulation, fMRI provides excellent spatial resolution and accuracy of localised activation (Huettel, 2017) over the entire extent of the brain.

Nonetheless, fMRI is limited in its temporal resolution because of the properties of task-related BOLD changes and their complex relation to the underlying neural physiology (Brunoni et al., 2012). Consequently, neither absolute nor resting function derived from the signal change can be confidently interpreted. This is an especially important limitation in clinical studies (Kropotov, 2016). For example, a functional deficit might be confounded by baseline or other task-unrelated effects. Moreover, the BOLD signal is prone to be affected by many factors that cannot be captured by the experimental design. Hence, whether or not, or to what extent, the BOLD signal can reliably reflect the neuronal activity to function is not easily confirmed.

Crucially, a correlation between neuronal activity and a cognitive function does not indicate the activity is critical or even relevant to its performance. A task whose execution involves multiple components need not allow us to distinguish between the marginal contribution of each. As a result, any anatomical inference will be limited because correlation here does not imply causation. The criticality of a specific region for a given function can only be established by a unidirectional causal relationship between region inactivation and loss of function (Sack and Linden, 2003, Mah, 2014a, Raichle, 1998).

In fact, the network of activations commonly identified by functional imaging during any reasonably ecologically-valid task tends to be so complex as to allow multiple possibilities of interpretation. For example, neuronal activity may increase not only owing to excitation, but because the region is being deliberately inhibited. Moreover, brain mechanisms widely enable competitive interactions between neural substrates. A successfully executed function may require on neural substrate to overcome the activity of a field of other, competing ones. In addition, incidental or epiphenomenal activation of a brain region cannot be excluded. It is possible a region is not directly related to a particular function, but is activated by some other region that *is* critical. Activation merely demonstrates a change in neuronal activity, and need not disclose its nature (Sack and Linden, 2003). For studying both the healthy and the focally injured human brain, functional imaging leaves of room for improvement.

1.1.3 Lesion studies

I have not introduced every available technique exhaustively, but the techniques taken into consideration are primary representatives that cover the major approaches in the current era. Neither transcranial stimulation nor functional imaging can sustain strong claims about the role of a specific brain region for a particular function. Each is inherently limited in its spatial accuracy and inferential power. By contrast, the loss or impairment of a brain function constitutive of lesion studies is theoretically capable of determining that a specific brain region is required by a particular function, which is indeed a much stronger assertion. This intrinsic causal strength of the lesion approach makes it more powerful than the other tools at our disposal.

Lesion-symptom mapping retains its value as potentially the most powerful technique to provide accurate functional localisation in human brain, which is vital to model the relationships between anatomical architecture and behavioural expression (Lomber, 1999, Rorden and Brett, 2000). Nevertheless, the performance of previous lesion studies has been unsatisfactory. One important constraint is the ethical impossibility of inducing human brain lesion experimentally, restricting study design and data collection. Another, more important constraint arises from the computational difficulties of analysis. Many studies have now shown that the distinct spatial distributions of lesions and the underlying functional anatomy are highly complex, and their interaction therefore doubly so. The causal relationships between focal damage and functional deficit will be determined by their interactions (Glasser et al., 2016a, Mah et al., 2014b, Mah et al., 2015, Nachev, 2015, Xu et al., 2017a, Xu et al., 2017b, Zhang et al., 2014). In the contemporary clinical domain, a wealth of neuroimaging scans and fully digitalised clinical records dramatically alleviate the difficulty to acquire a proper sample collection of focally damage human brain. Moreover, with the rapid progression of computing technology, computational complexity that is used to enforce the two sets of interactions mentioned above to be neglected is not an issue any more. Thereby, a crucial question we need to examine is how lesion data can be best modelled to reveal the anatomical interactions so as to improve lesion-function inference. With regards with the extremely complex underlying anatomical structure and high individual variability, a critical pre-requisite is the lesion data must be sufficiently parameterised to assure the detectability of the anatomical interactions and adequate characterisation of individual differences and similarities. To tackle a model including many variables, high-dimensional multivariate inferential tools are required.

I will now proceed to unpack the ideas introduced in the foregoing paragraph, in which is justified the inferential approach developed and illustrated in this thesis.

1.2 Individuality: a high-dimensional characteristic

In medicine, research is usually performed in the form of group studies. Typically, a cohort of many patients presumed to be more or less homogeneous is used to derive inferences about the population as a whole. By contrast, the subject of clinical interventions is almost always an individual. A question therefore arises here about how to tackle the differences between individuals and the group. Conventionally, most differences are treated as noise, tacitly assuming that they are either uninformative or too complex to be modellable. But if we treat individuals merely as accidental deviants from some kind of group mean, we naturally lose the information that defines their individuality. Consequently, any intervention will be optimal not for the individual but for a non-existent ideal average of the group. Figure 1.1 illustrates this point with an analogy to a part of the body whose information structure we are intuitively familiar: the face.

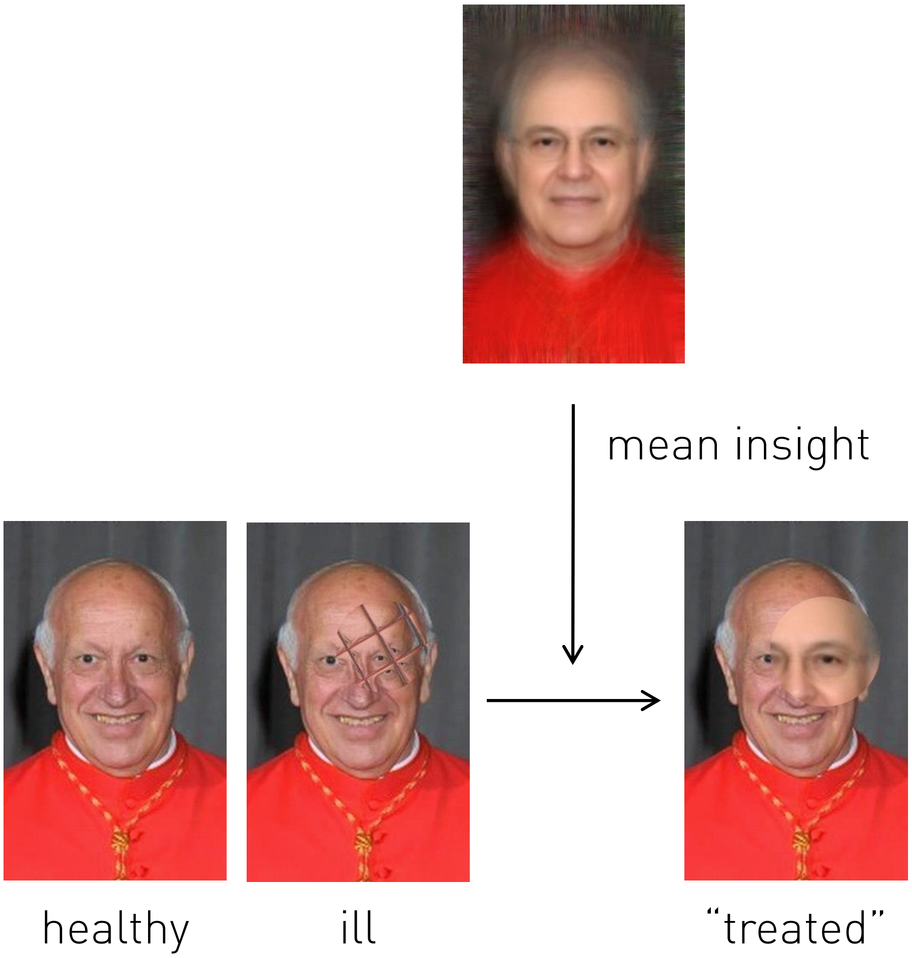


Figure 1.1 Absence of individuality

The face on the top was averaged from a plausibly homogenous group. The faces on the bottom line illustrated an individual from 'healthy' to 'ill' and how his face would be distorted after treatment with the crude group mean.

If we are to model individuality successfully, we need to understand how it is biologically constituted. Examined at a clinically-material scale, individuals vary a great deal even across what might be thought of as highly homogeneous groups, such as, for example, a set of male patients over 60 with stroke-induced aphasia. This variability arises from the fundamental nature of biology. First, the development of the human brain, along with the rest of the body, is guided by genetic plans deliberately shuffled at each generation, injecting diversity at the very point of conception. Second, the disparity between the maximal information content of the genome and the minimal plausible information content of the body is so great that much of the biology of an individual will have been dictated by environmental factors as variable as the world that surrounds us (Bartley et al., 1997, Baaré et al., 2001, Peper et al., 2007, Pol et al., 2006, Wright et al., 2002, Thompson et al., 2002, Thompson et al., 2001). It is implausible that this complexity should be purely accidental, and if it is not accidental, a high-dimensional parameterisation is necessary to apprehend it at the individual level.

The characteristics of the human face can be used as an illustration here. Generally, it is hardly possible to recognise an individual's face from the bland average of any collection of faces. In other words, a simple comparison of the difference between an individual face and any crude averaged face is rarely informative. Figure 1.2 illustrates how an individual face fails to be captured by the group mean or a single variable (the intraocular distance), no matter how great its precision of measurement. Far better individuation can be achieved by modelling a pixelated version of face with many more, even if less precise, parameters. Admittedly, in certain circumstances, an individual face may have a simple idiosyncratic feature, such as a scar, that could be captured by a univariate measure. But such circumstances will be exceptional. Analogically, to identify the individuality of a human brain we need to capture the underlying complex high-dimensional patterns as well. Human brains are high-dimensional constitutively: each individual brain represents a solution to the problem of creating a working brain, and no solution need me identical with any other.

The approach widely pursued in clinical research is to look for small sets of measures— “biomarkers”—rather than to construct high-dimensional models that involve all or many features. In a high-dimensional model, the inference will not only benefit from the entire set of features, but further information can be obtained from the interactions between them. Crucially, features that are uninformative in isolation may be powerfully illuminating in combination.

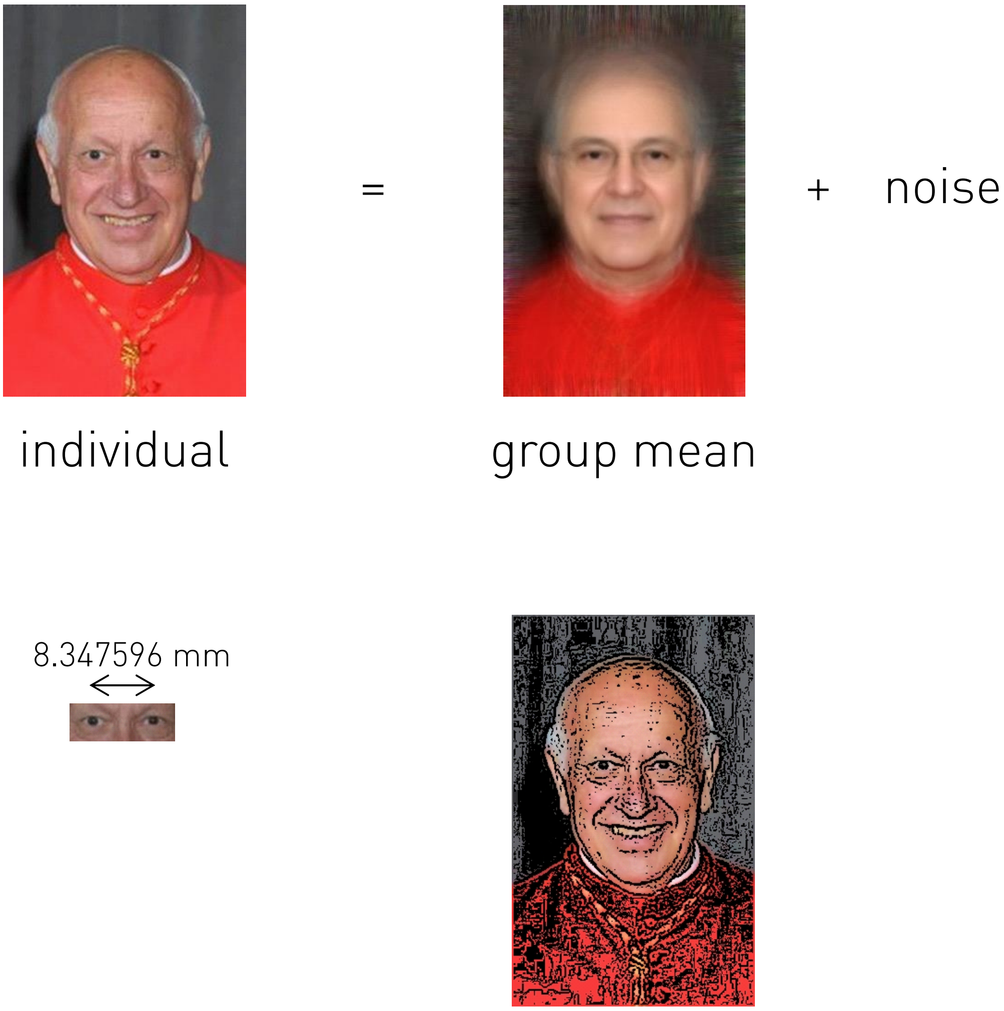


Figure 1.2 Reproduction of individual face

Conventionally, the analysis is highly relied on group mean. The difference between individuals and group mean is thus treated as noise. However, the presentation of human face apparently includes numerous features and can’t be described with any crude measurement though the measure is in extreme precision. To reproduce an individual face, sufficient parameterisation is inevitable, hence, high-dimensional analysis.

The importance, indeed the necessity, of modelling individuality within high-dimensional models is already obvious in the existing literature. One of the remarkable instances is in the field of translational medicine. Many of the positive interventional results observed in animal experiments have not been replicated in human studies (Brodie, 1962, Wang and Johnson, 2008, Bracken, 2009, Jucker, 2010), or found to be substantially less effective (Cummings et al., 2014, Sabbagh et al., 2013, Li et al., 2013, M Wilcock, 2010, Geerts, 2009). In the domain of biomedical research, audits have shown that only 8% of translational studies successfully pass Phase I. Although many factors are bound to contribute to this limited success rate—including molecular and physiological differences between man and experimental animals—the observed failure of translation may also be caused by insufficiently rich modelling of the underlying biology. A few studies have suggested the poor reproducibility of rodent research arises from insufficient similarity with human disease settings (Xu and Pan, 2013, Mak et al., 2014, Perel et al., 2007), but the importance of adequate model capacity has been widely neglected. It is plausible to expect that the discrepancy between animal and human studies could be ameliorated if the analysis is sufficiently high-dimensional. Precision medicine, involving individually tailored strategies of diagnosis, prognosis and therapy for each single patient is increasingly fashionable (Collins and Varmus, 2015, Mirnezami et al., 2012, Katsios and Roukos, 2010, Hamburg and Collins, 2010), but there is little recognition of the inevitable mathematical implications of the approach. The desire nonetheless reflects an awareness of the complexity arising from individuality and diversity, and the need to adopt novel approaches to understanding the disease and customizing individual interventions.

Over the history of neurology, one of the major obstacles to better understanding of brain mechanisms has been the complex and non-linear relationships between brain lesions and clinical outcomes. What I have said about the brain in general applies here with special force. The set of questions I would like to examine in this thesis agglomerate around the use of high-dimensional modelling for lesion-function inference.

1.3 Large-scale studies of the focally damaged human brain

Conceptually and methodologically, high-dimensional models are needed to characterise individuality with the precision needed to obtain optimal diagnosis and prognosis in the focally damaged human brain. To model high dimensionality, we need sufficient samples to cover the diversity of possible anatomical and pathological variations. Simply put, a highly parameterised model requires a commensurately large sample size. In this section, I shall enumerate the detailed reasons for large-scale data in this context, and what its characteristics should be; then, how to establish a large-scale dataset in the clinical domain and how to make its growth sustainable, not only for a specific research projects but a diversity of possible studies; lastly, what the challenges of managing a large-scale dataset containing sensitive clinical information are.

1.3.1 The rationale for “big data”

High-dimensional modelling and a large-scale data are indispensable and complementary. At the population level, the full diversity of complex patterns must be adequately sampled to characterise individuality. The more complex the patterns, the larger the necessary datasets.

We have proposed a novel approach to parameterizing the human brain with much greater richness than conventional approaches. For example, the most widely used approach—statistical parametric mapping (Friston et al., 1994, Ashburner and Friston, 2005)—models the brain with a very limited number of parameters, emphasizing regional specificity, and assuming an inherent smoothness of anatomical organisation. Under this assumption, simple univariate statistics, performed voxel-wise in a mass-univariate manner, are adequate for group inference (Friston et al., 1994). We have seen the observed complexity of the human brain, both functionally and anatomically, does not easily justify this assumption, though the exact level of complexity we must deal with is yet to be determined. To decipher this unknown complexity compels us to gather as wide an array of data samples as possible. Surveying large scale data can reveal clusters of relatively homogeneous instances, accessible to nearest neighbour modelling techniques that characterise each instance in relation to the centroid of its local neighbourhood rather than the group mean. The aim of such analysis—critically distinguishing it from other kinds of "big data" research, e.g. epidemiological research—is to use scale for better individuation rather than to power small effects into statistical significance. It requires us to collect not a fixed number of data samples, estimated from some kind of power calculation, but as much data as possibly can, for the number of clusters defining each individually informative centroid will generally be unknown.

If high-dimensional multi-parameter modelling is indeed necessary, we need to devise a solution for deal with the difficulty of high dimensionality. Generally, we may classify data into three categories by the number of samples (*N*) and dimensions (*P*): large *N* and small *P*; small *N* and large *P*; large *N* and large *P* (Lee and Yoon, 2017). Datasets with large *N* and small *P* tend to be tractable with classical statistical methods, at least with respect to simple properties of the data. Datasets with large *P* and small *N* are much harder to analyse, suffering from what is commonly described as the curse of dimensionality (Nasrabadi, 2007, Sinha et al., 2009, Lee and Yoon, 2017, Verleysen and François, 2005, Indyk and Motwani, 1998, Chen, 2009, Bellman, 2015). This is elegantly illustrated in the analysis presented in Figure 1.3. Random samples of synthetic data in the range -0.5 to 0.5 were drawn, divided into two equal groups of 1000 samples, with dimensionality varying from 1 to 1000. (https://www.biorankings.com/tech-reports/High-Dimensional_Data.pdf). The higher the dimensionality, the easier it was to separate the two samples, both from each other and from the origin, despite the fact that they are drawn from the same distribution centred on the origin. In short, with complex models the curse of dimensionality causes overfitting: learning of random patterns of the data with no plausible generalisation. The best protection against this problem is the acquisition of datasets in form of large P and large N: the larger the N, the better.

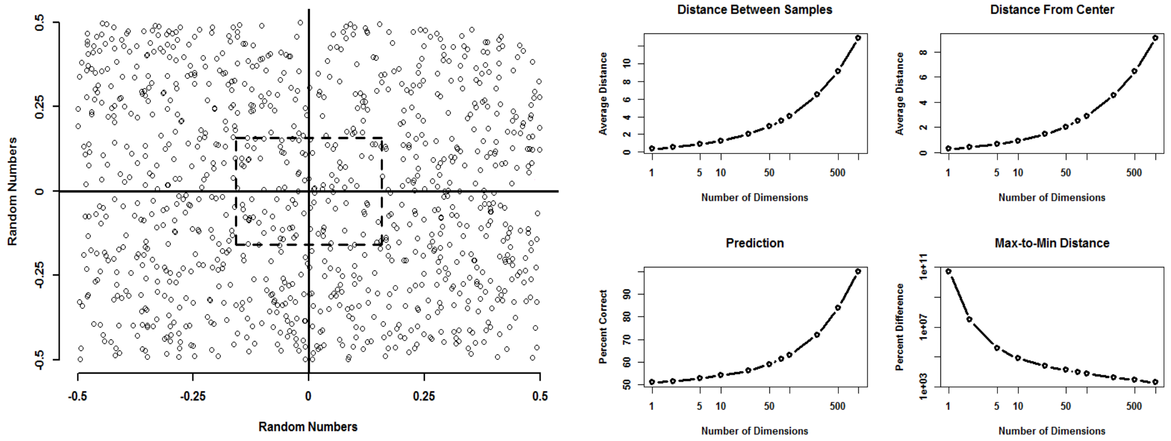


Figure 1.3 The difficulties of high-dimensional data

1000 data points in range of -0.5 and 0.5 were randomly generated for 100 iterations, 10th percentile was shown within the solid reference lines and 90% points were expected outside of it (left). The higher dimensions the data set had; the more complexity induced. The distance between samples and the distance from centre were both enlarged, which meant data points became more distributive and harder to tackle (right – upper left and right). With the increasing dimensions, the higher prediction rate indicated the wrong results induced due to complexity, and decreasing distance between maximum and minimum data points revealed they tended to be more difficult to be distinguished between each other (right – low left and right). The figures was sourced form: https://www.biorankings.com/tech-reports/High-Dimensional_Data.pdf.

1.3.2 Obtain large-scale data

To obtain large-scale data samples, we must first define the space of possible data sources. In the realm of human brain imaging, two general routes are open. The first and commonest route is to create a dedicated, externally-funded research project to collect and manage the data. Examples over the past decade include UK Biobank (http://www.ukbiobank.ac.uk/); the Human Connectome Project (http://www.humanconnectomeproject.org/) and the Human Brain Project (https://www.humanbrainproject.eu/en/). The number of human brain initiatives that generate neuroimaging associated with behavioural phenotype on various populations is rapidly rising. A non-exhaustive list includes Alzheimer's Disease Neuroimaging Initiative (http://www.adni-info.org/); the Thousand Functional Connectomes project and International Neuroimaging Data-sharing Initiative (http://fcon_1000.projects.nitrc.org/); the IMAGEN study of teenagers and mental health (http://www.imagen-europe.com); the AGES Reykjavik Study of Healthy Aging (http://www.hjarta.is/english/ages); and the Rotterdam study of aging (http://www.epib.nl/research/ergo.htm) (Van Essen et al., 2012). Predicated on large-scale financial support, such projects generally necessitate long-term planning and cooperation between many partners, accruing data over relatively long time-scales. This approach has the advantage of ensuring quality, homogeneity and completeness of the data, and the instruments and protocols used to acquire it. But there are disadvantages arising from the—typically high—cost, the length of time it takes to collect the data, and the organisational complexity of the necessary infrastructure. Where the data is sensitive, as tends to be true of brain imaging, the infrastructure is further complicated by security considerations which academic institutions are generally not optimally equipped to deal with. But perhaps the greatest problem is data selectivity. Obviously only data from those capable and willing to participate can be collected. This naturally excludes, or at least severely sub-samples, those incapacitated by disease, including many conditions of great interest to the neuroscience researcher.

An alternative approach is to make better use of data collected routinely for other purposes: in the context of our interest, clinical neuroimaging and corresponding clinical records. The volume of such routine clinical data is considerable if we can collect data retrospectively, enabling large datasets within relatively short time scales, and also continuous, as clinical activity naturally persists indefinitely. This approach then potentially yields large and securely growing datasets. Moreover, selection bias in the data is now introduced mostly by pathology, for a patient will generally not exercise the choice not to be treated. Crucially, drawing data from the clinical stream affords access to dynamic clinical contexts—such as acute focal brain injury—that leave little if any logistical space for research-specific investigation, and renders large-scale collections achievable at much lower cost, held within the secure digital environment of a hospital.

This approach is nonetheless far from easy to implement. Access to sensitive clinical data is administratively and technically complex, difficult to justify to the source—the hospital—on scientific grounds alone. Dual-use installations of joint operational value to the hospital are a solution, but it is one yet to be tested at scale.

Regarding to the data themselves, challenges arise from two aspects. The first is data integration. Data in clinical practice tends to be instrumentally heterogeneous, in the imaging domain derived from a mix of scanners employing a wide range of sequence types and acquisition parameters. The resultant variability complicates data aggregation, especially when it is non-random, biased by the clinical context. The second aspect concerns the administration and management of clinical data. Novel solutions for data retrieval, storage, and management are required. How to overcome these challenges and manage a large-scale clinical data warehouse is what this thesis in part seeks to determine.

1.3.3 The characteristics of large-scale data

Nowadays, big data is ubiquitous. Since the term big data was introduced, sometime in the late 1980s, infrastructural and analytic techniques have dramatically improved. Over the past decade in particular, mobile and wearable devices combined with commodity digital networks have vastly expanded the volumes of available data. Correspondingly, frameworks and algorithms have emerged to exploit the value inherent in large-scale data.

A widely-quoted description of big data was originally introduced by Gartner: high volume, high velocity, and high variety – Three Vs for short (Laney, 2001). Later, veracity was added as an additional characteristic to form a Four Vs model (Beyer and Laney, 2012); then, value and variability were appended to form a Six Vs model (Gandomi and Haider, 2015). These dimensions capture the essential features of big data. Yet, it is notable that they are associated and complementary rather than independent of each other (Gandomi and Haider, 2015). All the attributes are equally important to form a valuable dataset.

Besides these global attributes of a large scale dataset, medical big data has some additional distinctive features (Lee and Yoon, 2017). First, volumes in medicine are usually smaller and more variable in size than in other fields. For instance, datasets in epidemiology are usually much larger than those in neurology. The factors that limit the volume of a dataset in medicine are generally as follows: safety, expense and discomfort. To be specific, the context of safety in medical environment, particularly, in clinical domain, requires all the activities to be supported by a clinical need. Clinical data often involves expensive equipment and running costs. If the data is generated for clinical trials, the involvement of personnel is costly as well. Where an investigation causes the patient discomfort or presents a risk the volume of data derived from it will naturally be lower. Second, the high variety of data sources in medicine tend to make datasets extremely complex. For example, relevant data may include administrative records, clinical registries, patient recorded summary by physicians, imaging sessions, imaging reports by radiologist, biomarker data, genetic data, clinical trials and so on. It requires sophisticated management to match and link the variety of individual data efficiently, securely and precisely. Third, the sensitivity of clinical data imposes a requirement for secure storage and considerate usage. Sharing between hospitals and other institutions is typically difficult. Finally, clinical data is usually acquired and collated by inflexible, clinically-determined protocols. On one side, it is helpful to have the data in relatively structured formats; on the other, as clinical data is not reproducible in some situations, the certainty and validity of dataset can be affected by measurement errors or errors of inputting wrong codes (Lee and Yoon, 2017).

To draw benefit from high-dimensional modelling of large-scale datasets we need to solve an array of challenges. These span data collection and retrieval, storage, processing, an analysis (Wang and Krishnan, 2014).

A large-scale clinical database needs to be cost-effective, but it also needs to be stable and efficient. For instance, in terms of data collection and retrieval, data security and automation are high priorities. But also supremely important is minimizing the risk of disruption to clinical systems, avoiding data flow congestion and conflicts. In terms of data storage, we must consider security, access and sustainable extension. We also need a framework combining disparate types of data as necessitated by research objectives. Additionally, flexibility and extensibility are key to establishing and maintaining a scalable framework.

To be more specific, technically, the database needs to be supported by file server and backup server with sufficient storage and reliability. All communication should be limited to a secure intranet to ensure high security. For data retrieval, an automated pipeline must be built to access the clinical data stream safely and robustly. Once acquired, the data must be reorganised and integrated to form a comprehensive record for each patient. Meanwhile, the resources in clinical environment are extremely competitive. Such a system must operate without disturbing information flows necessary for clinical service delivery and maintenance.

In short, this is a complex task, requiring substantial human and technical resource. But it is a necessary task: the next question is how to maximise the value of the collated data.

Data itself is nothing if it is locked in a vacuum. It acquires value only when we extract intelligence from it, conditionally on the power of the methods we use. Conventional studies select small samples from the population and compare them within models with small numbers of parameters to examine a particular hypothesis, relying on simplicity to generalise to the entire population (Gandomi and Haider, 2015). Our interest here is not in such simple models, but the much more complex modelling large scale data potentially enables. Let us now consider what this entails.

1.3.4 The challenge of a large-scale, multi-site, cross-sessional imaging studies

The heterogeneity of clinical data is a product of the biological heterogeneity of the clinical population and the instrumental heterogeneity of data acquisition systems (Van Essen et al., 2012). In conventional research imaging studies, data tends to be collected on the same scanner, with a set of standardised acquisition parameters, ensuring consistency and reproducibility. Participants are specifically chosen, matched to the required demographic and clinical variables. In a real-life clinical environment, however, population characteristics will be widely distributed and outside the researcher’s control, introducing potentially great variability. A variety of scanner manufacturers, field strengths, and sequence characteristics are likely to have been used, especially when a dataset has accrued over a long period of time, introducing a great deal of instrumental variability. Even for a single scanner, the replacement of hardware components and software upgrades may alter the characteristics of the acquired image (Focke et al., 2011).

The impact of using instrumentally heterogeneous MR data has been previously explored (Han et al., 2006, Jovicich et al., 2009, Joshi et al., 2009, Chen et al., 2014, Rohrer et al., 2005), including comparison of the differences between inter-scanner and intra-scanner MRI sessions (Focke et al., 2011, Pardoe et al., 2008, Jovicich et al., 2013). In one study, a group of old participants (mean age 69.5), and two younger groups (mean age 34 and 36.6), were multiply scanned, both on the same scanner and across different scanners and field strengths. Regional volumes, such as hippocampal, thalamic and lateral ventricular volumes scanned by the same scanner on different days varied by less than 4.3% and 2.3% on older and younger groups, and comparable results were obtained across scanners (Jovicich et al., 2009). A similar study sought to identify the variability of cortical thickness with the manufacturer, field strength, and sequence characteristics of the image. The results showed within-scanner acquisitions to be negligibly variable; across scanners and field strengths, however, a slight bias was observed, reflected in thicker estimated cortex on 3 T scans. The impact by different acquisition sequences was larger; neither effects can be ignored in the design and analysis of cross-platform and longitudinal studies (Han et al., 2006).

Besides identifying the impact of multi-site and different acquisition parameters on anatomical quantification, studies have also suggested methods for correcting instrumental variability and otherwise mitigating its effects. One study showed the variability of phantom data could be reduced by 20% - 50% with high frequency correction and 20% - 25% with low frequency correction (Joshi et al., 2009). Another showed magnetic field strength, pulse sequence and radio-frequency receiving coil resulted in significant effects on variability and suggested a source-based morphometry model was effective to reduce the effects in some certain on multi-sites (Chen et al., 2014).

While such corrections may help, they cannot resolve the problem fundamentally, especially where scanner and biological variability are co-linear: any correction will consume both instrumental and biological variance. In voxel-based morphometry, for example, it is widely recommended to analyse data acquired on a single scanner with a specific sequence, even though multi-site studies will help to form a larger dataset and enhance statistic power (Focke et al., 2011, Pardoe et al., 2008). Voxel-based morphometry is a voxel-wised mass-univariate method, in which the voxel-wised comparison is limited by the variation of scan sessions derived from multi-site. This inspires us to seek a novel approach to overcome the heterogeneity stem from the instrument-related factors to have cross-sectional study possible.

One approach is to focus on the covariance between features, which then allow global effects such as the instrumental factors mentioned above to drop out of the consideration. To illustrate the proposed approach, we can return to the face example introduced earlier in this thesis. If human faces are measured by a few discrete features, recognizing and compare them will be seriously affected by measurement bias. For instance, if the precision of inter-ocular distance is biased across tests and the bias cannot be modelled, inter-ocular distance cannot be used as a cross-sectional measurement. But if the face is represented as a pixelated image, in which each pixel is an independent feature contributes to the measurement, information across all pixels will be less susceptible to distortion by the largely affine transforms instrumental variability will tend to introduce. High-dimensional modelling thus theoretically enables us to overcome instrumental effects, allowing us to extract biological signals from relatively low-quality, instrumentally heterogeneous data. We shall explore this idea further in subsequent sections.

1.4 Lesion-function inference

The major focus of this thesis is the combination of high-dimensional modelling with large-scale data to make inferences about the relation between focal brain injury and clinical outcomes, revealing the brain's functional architecture in the process. There are two broad approaches to modelling here: univariate analysis and multivariate analysis. Univariate methods analyse data one variable at a time for descriptive, predictive of inferential purposes. By contrast, multivariate methods model more than one variable, including their potentially complex interplay.

The human brain is arguably the most complex biological system we know, composed of multiple, densely connected, non-linearly interacting elements, at both the physical and functional level (Lessov-Schlaggar et al., 2016, Raichle et al., 2001, Power et al., 2011, Fox et al., 2005, Bassett and Gazzaniga, 2011, Sporns et al., 2005, Brett et al., 2002, Meyer-Lindenberg, 1996, Mah, 2014b). Yet most previous investigation of human brain, particularly those employing lesion-function mapping, tend to use univariate analysis and statistical approaches that are only appropriate for linear relationships within a relatively simple system. As we shall see, this may explain why many experimental studies succeed in simpler animals, but fail to replicate in humans (Van der Worp et al., 2010, Mak et al., 2014, Xu and Pan, 2013, Jucker, 2010).

In the following two sections, I will describe the univariate methods in widespread used through the evolution of lesion-symptom mapping studies; then introduce a few multivariate based approaches adopted in the empirical sections of this thesis.

1.4.1 Univariate modelling

Lesion-function mapping can be traced back to the 19th century, when Paul Broca inspected two patients who lost the ability to produce more than one syllable and subsequently conducted post-mortem examinations of their brains (Broca, 1861). He identified an association between injury to the posterior inferior frontal gyrus of the brain and language production. Although Broca is not the first person to propose a relationship between language production and the left hemisphere of the brain, his approach of using the anatomical location of brain injury together with the behavioural deficit to localise human brain function founded a tradition that has influenced neuropsychological research ever since (Broca, 1861, Rorden and Karnath, 2004, Dronkers et al., 2007). Broca's approach was rooted in examining numerous brain injuries and identifying the commonalities of lesion volume and location. He treated the brain components discretely, in other words, each unit location was treated as an independent region spatially correlated with the behavioural disorder. This reveals it as a univariate approach, essentially examining the relation between a single variable and a behavioural outcome. The studies to emerge in the following decades share the same fundamental inferential basis. For instance, Carl Wernicke published on the relationship between brain injury to the left posterior temporal cortex and language comprehension (Wernicke, 1874). In the mid-20th century, loss of memory was discovered to be related to bilateral damage to the medial temporal lobes (Scoville and Milner, 1957, Penfield and Milner, 1958). Studying split-brain patients revealed the dominance of the left hemisphere in language and calculation, and the right in spatial tasks (Vinken and Bruyn, 1969).

Lesion-symptom mapping studies began to flourish with the emergence of technologies for *in vivo* anatomical study, using Computed Tomography and Magnetic Resonance Imaging, to visualise, quantify and monitor the focally lesioned human brain in the living patient. The technical innovation has not been matched by conceptual innovation: univariate approaches still dominate the field. Two typical univariate methods that are most used are template overlay method and voxel-wise lesion-symptom mapping (Frey et al., 1987, Damasio and Damasio, 1989, Bates et al., 2003).

1.4.1.1 Region of interest (ROI) based lesion-symptom mapping (overlap-subtraction method)

The advent of neuroimaging enabled anatomical studies to be performed in advance of the patient's death, giving rise to group studies of the relationship between brain injury and behaviour. The first technique to make use of the enhanced data scale was template overlay (Frey et al., 1987). The technique involves transferring the lesions observed on individual imaging imaging onto a standard anatomical template. The extent of lesion overlap across the group can then be used to define the anatomical location common to those sharing the same behavioural deficit. Inferences drawn from this approach remain influential in the literature: for example, that language production and comprehension is related to the left hemisphere; and the spatial cognition is related to the right hemisphere (Rorden and Karnath, 2004, Rorden and Brett, 2000).

The approach nonetheless exhibits an array of obvious defects. The pattern of focal damage need not match the functional architecture of the underlying human brain. The underlying assumption of modularity is clearly naïve, for the reasons already discussed. There is an unwarranted assumption that the structure of functional modules is distributed in the same anatomical structure across individuals: intrinsic differences in brain anatomy exist between individuals (Brett et al., 2002, Hawrylycz et al., 2012), confounding group studies with template overlap method.

Brain lesions do not respect the boundaries of the functional networks reflected in the connectional structure of the human brain – the human brain *connectome* (Sporns et al., 2005, Bullmore and Bassett, 2011). When the brain is affected by a focal lesion, on the one hand, the inherent plasticity and adaptability of the human brain enables a mechanism of reconfiguration to mediate recovery, obscuring the causal connection between structure and function; on the other hand, structurally intact areas may be impaired owing to disconnection, as the patterns of anatomical connections are not visible from CT or MRI scans (Sporns et al., 2005, Rorden and Karnath, 2004). Therefore, the common region extracted from superimposed brains need not be representative of the underlying neural dependence.

Another critical concern is differential vulnerability. In stroke, the structure of the vascular supply makes some areas more prone to injury than others; similar effects render damage from other pathologies non-random. This not only alters the sensitivity for detecting lesion-function relationships across the brain, it also introduces spatial biases, as I discuss at length below.

A natural remedy is to introduce control populations (Rorden and Karnath, 2004, Karnath et al., 2001a, Weiller et al., 1993, Karnath et al., 2002), giving rise to the subtraction overlay method. A lesioned group expressing the behavioural disorder is here matched with another lesioned group without the disorder. Analogously to the template overlay method, two overlay plots are created by superimposing the brain lesions from the two groups. Subtraction of the overlays then yields three regional categories. First, regions commonly damaged in the test group but spared in the control group emerge with positive values, and are inferred to be critical in proportion to the degree of positivity. Regions commonly damaged in the control group but spared in the test group yield negative values, and are inferred to be irrelevant. Regions damaged equally or not at all are inferred to be irrelevant or untested respectively. Though better than the simple version, the subtraction overlay method still relies on the assumption of modularity and simple localisation in human brain.

1.4.1.2 Voxel-based lesion-symptom mapping (VLSM)

Overlap studies are performed at a given voxel resolution, and imply a set of voxel-wise operations. An explicitly voxel-wise approach to lesion mapping was nonetheless first applied to investigate verbal fluency and auditory comprehensive after brain damages on 101 chronic stroke patients with lesions in the left hemisphere (Bates et al., 2003). Each voxel was treated as an independent variable, and binarised as either damaged or intact. Behavioural performance—here fluency scores and auditory comprehension measures—was adopted as the dependent variable. A *t*-statistic was calculated for the strength of association of each voxel. Voxels located in the region of the insula and the deep parietal white matter (e.g. arcuate or superior longitudinal fasciculus) affected the performance of fluency most. For the auditory comprehension tasks, voxels within middle temporal gyrus showed the strongest association; in addition, regions of dorsolateral prefrontal cortex and parietal association cortex were involved. Previously identified areas within the inferior frontal gyrus—Broca's area—and superior temporal gyrus—Wernicke's area—where also shown. A set of analyses of covariance (ANCOVAs) were conducted among the relevant *a priori* regions of interest (ROIs) to clarify the relative contribution of each highlighted area.

Unlike overlap subtraction-based methods, which rely on binarised deficit scores, VLSM can model continuous behaviours. Most cognitive skills tend to be intrinsically continuous rather than binary. Operating voxel-wise is superior to operating across larger ROIs that may be comprised of multiple sub-regions unequally contributing to the deficit (Gleichgerrcht et al., 2017). VLSM generates a brain map indexed by a voxel-wise parametric test, such as the two-sample *t*-test, or the measurement of effect size, without requiring any *a priori* regional parcellation. More complex tests can be applied at each voxel, for example ANCOVAs that enable the removal of multiple confounds. VLSM thus addresses lesion-symptom relationships with greater finesse, but it still assumes a simple modularity of organisation essentially the identical with that assumed by ROI based methods.

Several critical issues remain unaddressed. First, dysfunction arising from disconnection of regions remote from the lesion (diaschisis) will remain undetected (Price et al., 2001, Feeney and Baron, 1986, Carrera and Tononi, 2014), unless the connectivity of white matter is taken into account outside the standard voxel-wise, mass-univariate statistical framework. Moreover, the complex patterns of differential vulnerability will distort inferred maps in response to lesion convariances driven by the vascular tree, at least in the case of vascular lesions (Rorden et al., 2007, Mah et al., 2014b, Gleichgerrcht et al., 2017, Zhao et al., 2017, Kimberg et al., 2007). Theoretically and conceptually, the inherent limitation of this mass univariate approach is to ignore the spatial dependencies between voxels (Ge et al., 2014).

Complex confounding dependencies might also arise from registration error, especially when the quality of registration co-varies with the parameters of the lesion, as is inevitable unless the patient has been scanned both before and after the appearance of the lesion. This is a problem not only with VLSM, but any massively univariate approach that assumes a simple covariance structure. Functional imaging and VBM escape it only because covariance patterns of functional activation and non-pathological structural variation happen to be well-behaved. SPM suggests Gaussian smoothing as a remediable solution for this problem. To be more specific, the smoothness is required to be greater than the voxel size to reach homology in term of functional structure between inter-subjects. However, this will not resolve the fundamental problem but merely reduce the impact of highly correlated locations (Kimberg et al., 2007, Ashburner and Friston, 2000, Good et al., 2001, Friston et al., 1994, Penny et al., 2011).

There have been attempts to remedy VLSM's defects (Kimberg et al., 2007, Kinkingnéhun et al., 2007, Medina et al., 2010, Rorden et al., 2009, Rorden et al., 2007, Groppe et al., 2011). For example, Rorden et al. suggested the use of different voxel-wise statistical tests—Liebermeister test and Brunner-Munzel test—with higher sensitivities to replace the traditionally used Yates-corrected chi-square test for binomial behavioural outcome and Student's t-test for continuous degree of behavioural symptom, respectively. The excessive conservatism of Bonferroni correction for multiple comparisons has been mollified with the use of permutation thresholding and false discovery rate (FDR) instead. But none of these modifications addresses the most fundamental problem: the failure to model the complex covariance structure of lesions.

1.4.2 Multivariate approaches

We need an approach that can deal robustly with the complexities of lesion data. This requires modelling architectures flexible enough to adapt to the peculiarities of the data, making the fewest assumptions about their distribution. In general, we need architectures that can handle multiple parameters, because the human brain is intrinsically high dimensional in its structure. Only multivariate methods can combine the information distributed across different dimensions, spatially and temporally, that describes the structure and function of the brain.

Multivariate analysis is a mature discipline (Badiru, 1992, Palmer, 1993, Sternberg et al., 1984). The established consensus on the complexity of the human brain justifies its adoption across neuroscience (Nandy and Cordes, 2003, McIntosh and Lobaugh, 2004, Norman et al., 2006, Mourao-Miranda et al., 2005, Friston et al., 2008, Friston et al., 2003). For example, canonical correlation analysis (CCA) reveals hippocampal activation on fMRI in a memory-related task better than conventional methods (Nandy and Cordes, 2003). Similarly, partial least squares

(PLS) can take advantage of spatial and temporal dependencies across voxels to identify the distributed patterns of task performance or functional connectivity (McIntosh and Lobaugh, 2004). Norman et al. (Norman et al., 2006) have surveyed multi-voxel pattern analysis (MVPA) analysis, including feature selection and pattern classification with a wide range of a linear or nonlinear classifier, demonstrating greater sensitivity compared with conventional methods. Though more powerful, multivariate approaches are constrained by the cost of computation, and the necessity for large scale data, at least where high-dimensional or otherwise complex models are employed.

The rapid rise in the computational power over the past decade has rendered large-scale studies tractable, transforming the applicability of multivariate methods in neuroscience as elsewhere (Mah et al., 2014b, Zhang et al., 2014, Zhao et al., 2017, Xu et al., 2017a, Xu et al., 2017b).

For example, Mah et al. studied a cohort of 581 acute ischemic lesions segmented from diffused-weighted imaging (DWI) and modelled at 2 x 2 x 2 mm resolution, yielding 90 469 within-brain voxels. The authors used a set of hypothetical lesion-deficit models to compare their anatomical fidelity of univariate or multivariate approaches. Assuming an asymptotically simple lesion-deficit relation—single voxel dependence—the maps derived from mass univariate analysis were displaced from the ground truth by a mean distance of 15.7 mm (SD = 9.15 mm), unsurprisingly following the stereotyped patterns of the vascular supply. A more physiologically plausible lesion-deficit model based on Brodmann areas showed a similar extent of mislocalisation but larger variation (15.9 mm; SD = 17.6 mm). By contrast, SVM models exhibited much less distortion (Mah et al., 2014b).

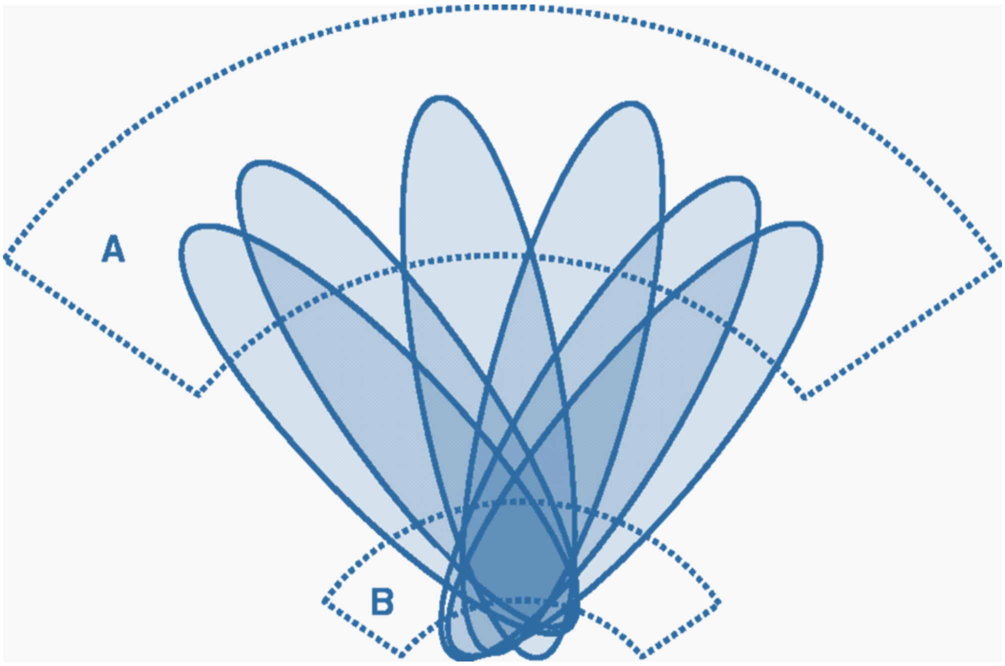


Figure 1.4 The systematic bias of template overlay method (Mah et al., 2014b)

Due to the bias of vascular structure existing in human brain, some areas in the brain are more vulnerable and tend to be affected collaterally. Thereby, the most overlapped area (B) extracted from a stereotyped pattern of brain injury was conventionally mislocalised as critical area. Whereas, the real critical area (A) that hidden behind complex anatomical patterns was thus ignored.

This study comprehensively demonstrated how the complexity of the lesion distribution results in mislocalisation when mass univariate methods are used. From a fundamental methodological standpoint, it firmly entrenched the necessity and superiority of high dimensional methods of lesion-function mapping. Naturally, brain lesions can rarely be described by a single voxel at any commonly used resolutions. But equally, damage to a set of voxels cannot be adequately modelled by a simple Gaussian field because the dependence between voxels is made complex by the inherent inhomogeneity of anatomical structures, particularly, the vascular tree. To map the lesion-symptom associations, it is crucial to capture the complexity of the structural architecture of lesions, which demands a high-dimensional multivariate framework implemented by machine learning techniques and powered by large scale data (Herbet et al., 2014, Nachev, 2015, Karnath and Smith, 2014, Mah et al., 2015).

In the following section, a set of multivariate machine learning techniques that fit high-dimensional dataset and related inferential tools involved into the studies in this thesis will be introduced though they are not exhaustive.

1.5 Multivariate based machine learning algorithms

Broadly speaking, machine learning can be grouped as supervised machine learning, unsupervised machine learning and reinforcement learning. Simply put, supervised learning can provide direct feedback as predicted outcome by learning labelled data sources. Unsupervised learning will not give any feedback but can reveal hidden structure behind data sources. Between supervised and unsupervised learning, the kind of problems with only a part of input data labelled is called semi-supervised learning. Reinforcement learning is used to learn a set of actions to gain optimal cumulative rewards. There are many different algorithms for each type of machine learning. The performance of learning algorithms is largely decided by the data size and structure: rather than simply judge one algorithm is superior to the others, it is necessary to introduce insight into the specific machine learning problem to choose the most appropriate algorithm.

1.5.1 Supervised learning algorithms

Supervised learning is to infer a function that can generalise as the transformation of a set of known inputs into a set of predicted outputs in the form of categories or real values (classification or regression). During the learning process, the algorithm iteratively learns and does prediction from training data until the performance is acceptable by some criterion of fidelity.

1.5.1.1 Datasets for learning algorithms

Essentially, a supervised learning algorithm is a data-driven process to fit multiple datasets at different stages. Initially, the model learns from a training dataset that is comprised of a collection of labelled examples pairing input features and desired outputs—also known as supervisory signals— to approximate a mapping function generalised from the training examples to determine the categories or values on new examples. During the training phase, the relations, associations and dependencies amongst the features of input objects are generalised as an inferred function whose parameters are iteratively adjusted until the function can predict the new input. For any learning algorithm, the quality of the training data ultimately determine performance, which demands thoughtfulness and care in collecting and labelling each instance. The difficulties of assembling a suitable training set of data derive from two aspects: one is the data quality; the other is dataset size. Data quality is crucial for training, as any inconsistencies between the pairs of input features and labels will be learnt through the training process, and propagate to the new, unseen instances so as to weaken the learning algorithm's predictive power. On the other aspect, dataset size is important for learning an optimal mapping function, particular for the high-dimensional inputs (e.g. brain imaging scans). The generalizability of a mapping function is sensitive to the scale of the training dataset; and improved by increasing training dataset volume. But there tends to be a quality-size tradeoff in the real world, making it difficult to assemble datasets of both high quality and large scale within reasonable time and resource.

A trained classifier model can be applied to predict the characteristics of new, unseen examples. The typical following stages are thus validation and testing. There are two aspects worth noting here: one is the necessity of separating validation and test sets; the other is the question of whether or not the test phase can be skipped. During the validation process, an independent dataset is used to evaluate the error function and estimate the model fit on the training dataset by tuning the hyperparameters so as to optimise the performance of prediction on new examples. Since, the validation dataset is used to determine the final model by minimizing the error rate, in order to avoid overfitting, the final validation of model performance needs to be derived from another independent dataset—the test dataset—quantifying performance by some standard measure (such as accuracy, sensitivity and specificity) so as to establish the generalizability of the final model. Validation is usually performed with the hold-out method or cross-validation (Kohavi, 1995, Whitney, 1971).

1.5.1.2 The choice of learning algorithms

As aforementioned, there are a wide range of options for supervised learning algorithms. Each algorithm may work well on a specific type of supervised learning problem; but worse than the others on other supervised learning problems. Besides improving the amount and quality of training dataset and conducting validation to avoid overfitting, in order to obtain proper performance, it is critical to choose the appropriate learning algorithm, matched to the specific problem. Broadly, the choice could be guided by considering the major aspects as follows: bias-variance tradeoff / dilemma (Platt, 1999); properties of the input data (Caruana et al., 2008); and presence of interaction or nonlinearity (Tu, 1996).

1.5.1.2.1 Bias and variance

Essentially, to estimate the performance of a learning model is to understand the prediction error of a learning model. Bias and variance are two essential types of prediction errors that quantify the performance of a supervised learning algorithm; and indicate a major difference in performance that distinguishes one algorithm from another. Theoretically, the error due to bias is related to the differences between predicted values and the correct answers, in other words: accuracy, which reflects the flexibility of an algorithm to fit new examples learnt from a training dataset. The error due to variance is related to the extent of difference between the predicted values learnt from different sets of training data in multiple times, in other words: consistency, which reflects the sensitivity of an algorithm to a specific training dataset. Bias and variance usually present a trade-off. This is because both bias and variance are highly sensitive to algorithmic complexity. If the structure of an algorithm tends to be simple and rigid, the algorithm will not be sufficiently flexible to learn new instances in training, leading to consistently incorrect prediction. This is knowns as high bias and low variance and can be illustrated with a “bull’s eye” diagram in Figure 1.5. By contrast, if the underlying structure of an algorithm is made very complex, for example, by adding more parameters to the model, these over flexible models will cause high sensitivities of each specific trained model, in other words, the models will treat different sets of training data very differently. This may result in the acceptable average accuracy; but varied and inconsistent prediction among the prediction models. This is known as low bias and high variance (Figure 1.5).

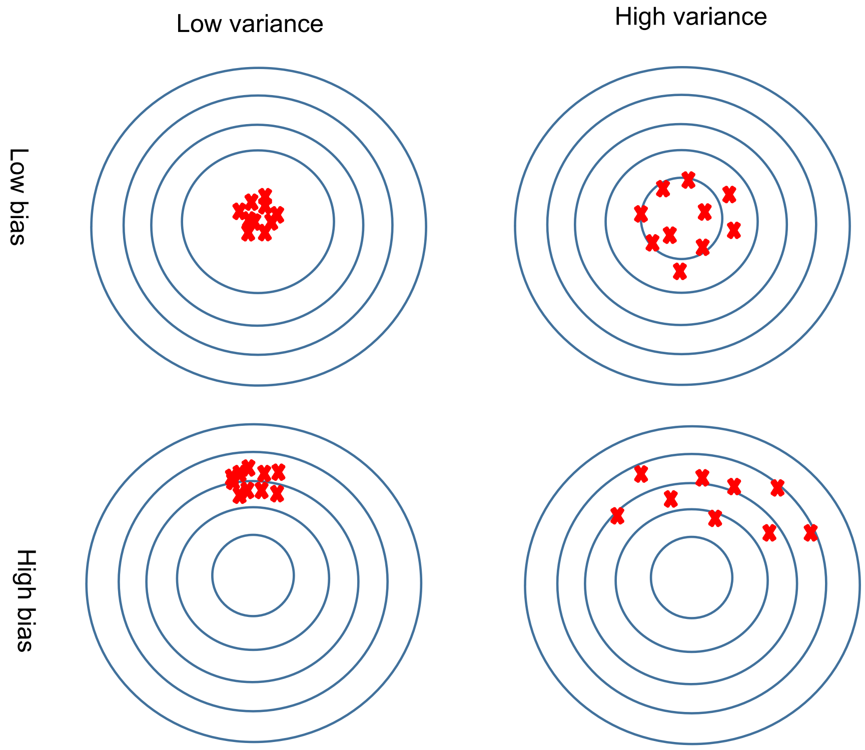


Figure 1.5 Bias-variance trade-off

This "bull's eye" diagram illustrates the relation of bias and variance. Theoretically, a predictive model in low bias and low variance would be ideal to any learning problem though it is hardly possible in practice. The trade-off between bias and variance determined the choice of learning algorithm needs to be adapted to the particular learning problem.

The trade-off between bias and variance corresponds to the contrast between under-fit and over-fit, which is critical for understanding the behaviours of prediction models. The total error of an algorithm can be broken down as follow:

$$Err(x) = Bias^2 + Variance + Irreducible\ Error$$

The irreducible error is a noise term capturing variability that is truly random. So, the critical concern is to balance bias and variance so as to obtain a prediction models with optimal fit. In figure 1.6, the relationships between model complexity and error indicates it is necessary to reconcile the errors due to bias and variance and look for minimizing the overall error rate of bias and variance.

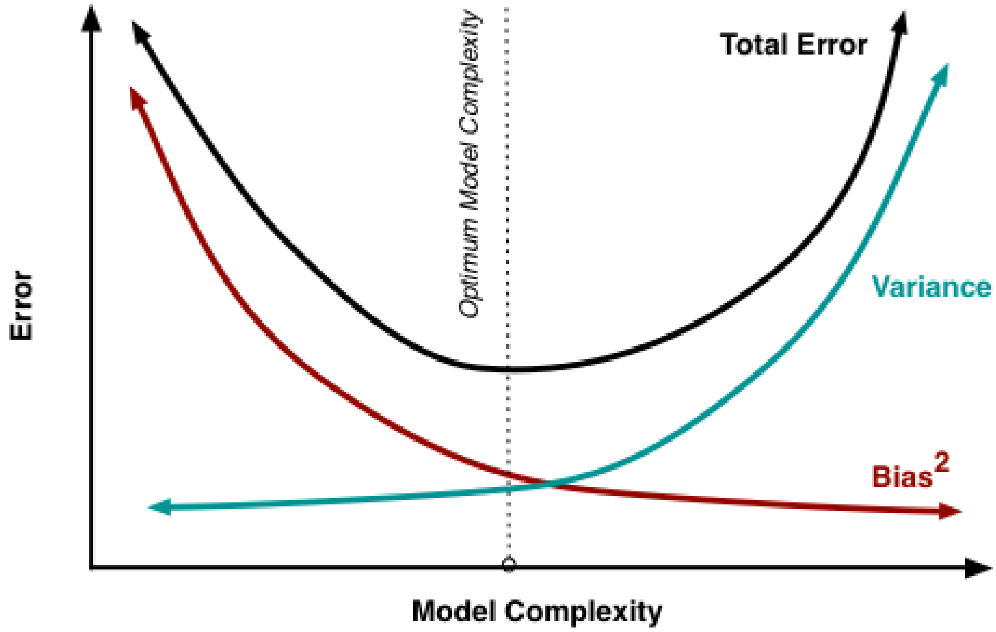


Figure 1.6 The complexity versus prediction error (Fortmann-Roe, 2012)

This figure shows the relation between total error and model complexity. Total error is composed of the errors arising from bias and variance. The trade-off between bias and variance is decided by the complexity of model. Specific to each particular problem, model complexity should be optimised by minimizing the total error.

1.5.1.2.2 Properties of the input data

Given the fundamental aspects of bias and variance arising in choosing an algorithm, some issues about training data need to be considered. Firstly, we need to consider the volume of training data and the number of features involved in each case. If the features are irreducibly many, distributed in a high-dimensional space, the true learning function will tend to be complex and require a prediction model to be tuned with low bias and high variance; hence, a large volume of training data is crucial to enable the prediction model to be learnable. If the features are few, or reducible to a few, we shall only require a less complex learning function with high bias and low variance to be learnt by a relatively small volume of data.

The dimensionality of the data aside, the interactions between the features also affect the complexity of the optimal learning function. For example, if each feature independently contributes to the desired output value, the model can be simple, incorporating only linear functions. If the features are complicated by interactions and dependencies, the optimal models will need to cope with non-linearity and complex parameter tuning is likely to be required.

It is clear within which category human brain data is likely to fall. As already discussed, we need large volumes of data, and algorithms capable of dealing with a high-dimensional input space involving the complex interactions. In the following sections, I review dominant, exemplar approaches to such modelling, including Support Vector Machine (Vapnik and Chervonenkis, 1968, Boser et al., 1992, Cortes and Vapnik, 1995) and decision trees (Quinlan, 1986, Utgoff, 1989).

1.5.1.3 Support vector machines (SVMs)

SVMs are a class of predominantly supervised learning models that learn labelled examples, then assign labels to new, unseen cases for either classification or regression purposes. Though they can be adapted to semi-supervised learning or unsupervised clustering (Bennett and Demiriz, 1999, Winters-Hilt and Merat, 2007), their main application is within a supervised context. As a classification approach, the Support Vector Machine (SVM) was initially proposed by Boser et, al., in which a high-dimensional decision plane, named a hyperplane, was introduced to separate the different class memberships of learnt instances with the maximum possible margin (Boser et al., 1992). The training instances are distributed in a multi-dimensional feature space, in which the learning algorithm seeks to find a hyperplane that separates into the target groups such that the distance between the hyperplane and nearest objects from each group is maximised. SVMs then use the hyperplane as a decision boundary to enable classification of unseen instances.

In essence, an SVM model can be treated as a mathematical entity which employs an algorithm to optimise a mathematical function with a particular maximum value with regard of a dataset. In order to understand the essential theory of SVMs, a few aspects of SVMs will be introduced in the following sections.

1.5.1.3.1 The separating hyperplane and kernel trick

We begin with the simplest case: two classes of data points represented as red points (negative class) and blue points (positive class) are continuously distributed in a line shown in Figure 1.7 (a). They can easily be separated by a single point on the single axis here. A more complex scenario is illustrated in Figure 1.7 (b), where two classes of values are linearly distributed in a two-dimensional feature space. The graphical illustration shows the data points can be classified by any of the straight lines depicted in the figure.

However, such a separation is not always straightforward. For example, to separate the elliptically distributed data points in Figure 1.7 (c - left), a non-linear solution has to be applied. SVMs solve such non-linearly separable problems by transforming the native input space to a higher dimensional feature space via a kernel. As shown in Figure 1.7 (c - right), by escalating the input two-dimensional space to three dimensions, the data points become linearly separable. Mathematically, the kernel employed in SVMs is essentially a mapping (transform) function to project the data points from a low dimensional space to a higher dimensional space while preserving their similarity.

As a demonstration, let us take two data points x = (x_1, x_2) and x' = (x'_1, x'_2) in the original space, and then apply a nonlinear function φ to map x = (x_1, x_2) and x' = (x'_1, x'_2) to the transformed higher dimensional feature space represented as (z_1, z_2, z_3) and (z'_1, z'_2, z'_3) by computing their inner product as the inner product is a measure of similarity.

$$\phi\ (x) \rightarrow \phi\ ((x_1, x_2) \rightarrow \phi\ (z_1, z_2, z_3) \rightarrow (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

$$\phi\ (x'\ ) \rightarrow \phi\ (x'_1, x'_2) \rightarrow \phi\ (z'_1, z'_2, z'_3) \rightarrow (x'^2_1, \sqrt{2}x'_1x'_2, x'^2_2)$$

We can formulate the mapping function - φ (x) in a three-dimensional space as below:

$$\beta_0 + \beta_1 x_1^2 + \beta_2\sqrt{2}x_1x_2 + \beta_3 x_2^2 = 0$$

As mentioned above, in SVMs, the dot products between the two data points are required to measure the similarity. First, we perform calculation below:

$$< \phi\ (x_1, x_2),\ \phi\ (x'_1, x'_2) > = < (x_1^2, \sqrt{2}x_1x_2, x_2^2), (x'^2_1, \sqrt{2}x'_1x'_2, x'^2_2) >$$

$$= x_1^2x'^2_1 + 2x_1x_2x'_1x'_2 + x_2^2x'^2_2$$

Then, we can also compute the dot product as follows:

$$<x, x'>^2 = <(x_1, x_2), (x'_1, x'_2)>^2 = (x_1x'_1 + x_2x'_2)^2 = x_1^2x'^2_1 + 2x_1x_2x'_1x'_2 + x_2^2x'^2_2$$

Another example illustrates how to transform the data points in the original space to a five-dimensional space.

Similarly, for the same data points: x = (x_1, x_2) and x' = (x'_1, x'_2), a mapping function - φ (x) can transform them from the original space to a five-dimensional space as follows:

$$\phi(x) \rightarrow \phi(x_1, x_2) \rightarrow \phi(z_1, z_2, z_3, z_4, z_5) \rightarrow (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

$$\phi(x') \rightarrow \phi(x'_1, x'_2) \rightarrow \phi(z'_1, z'_2, z'_3, z'_4, z'_5) \rightarrow (1, \sqrt{2}x'_1, \sqrt{2}x'_2, x'^2_1, \sqrt{2}x'_1x'_2, x'^2_2)$$

$$\beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1^2 + \beta_4\sqrt{2}x_1x_2 + \beta_5x_2^2 = 0$$

Again, the computations are conducted in the same two ways, respectively. First,

$$< \phi(x_1, x_2), \phi(x'_1, x'_2) >$$

$$= <(1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1x_2, x_2^2), (1, \sqrt{2}x'_1, \sqrt{2}x'_2, x'^2_1, \sqrt{2}x'_1x'_2, x'^2_2)>$$

$$= 1 + 2x_1x'_1 + 2x_2x'_2 + x_1^2x'^2_1 + 2x_1x_2x'_1x'_2 + x_2^2x'^2_2$$

Then,

$$(1 + <x, x'>)^2 = <(1 + (x_1, x_2), (x'_1, x'_2))>^2 = (1 + x_1x'_1 + x_2x'_2)^2$$

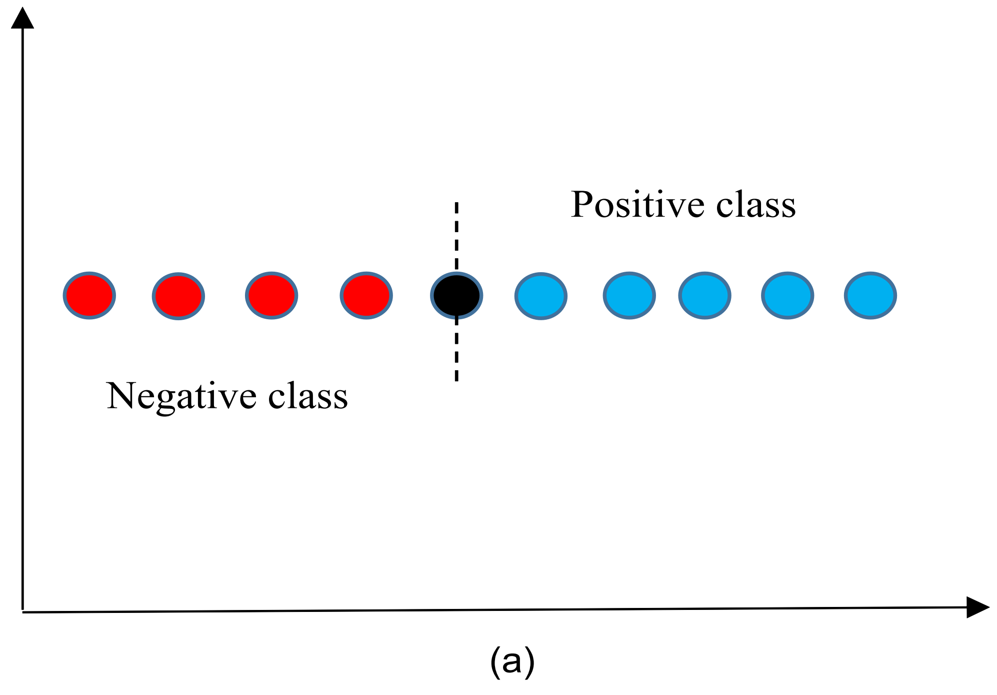
$$= 1 + 2x_1x'_1 + 2x_2x'_2 + x_1^2x'^2_1 + 2x_1x_2x'_1x'_2 + x_2^2x'^2_2$$

The kernel functions for three- and five-dimensional space can then be described as:

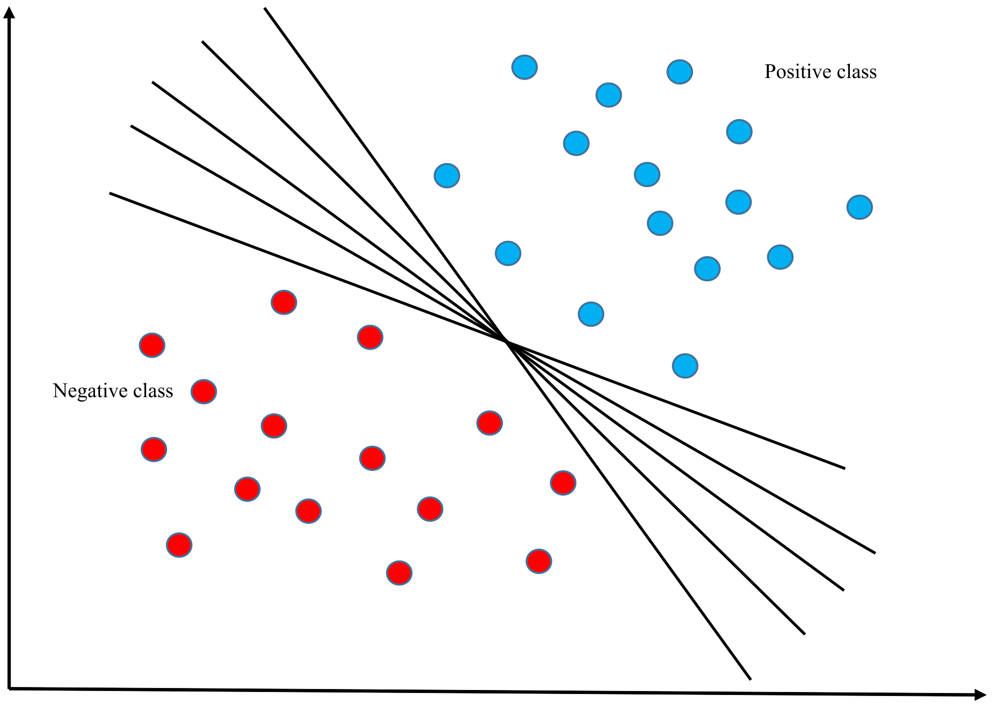
$$K(x, x') = < \phi(x), \phi(x') >^2$$

$$K(x, x') = (1 + < \phi(x), \phi(x') >)^2$$

This approach can be used to map data points in a space of arbitrary dimensionality. The kernel method provides a general solution to transforming data to a higher dimensional space infinitely wherever they are separable by a hyperplane. Surprisingly, we discover that we actually need not focus on the forms of the mapping functions other than maintain similarity by computing the inner product of two vectors in the original space then raising the result to a power for the corresponding higher dimensional space. In other words, we need not be concerned about the forms of φ but merely K. This surprisingly powerful tool is called the “kernel trick”. A kernel function defined as above can replace wherever an inner product is used. In applying the kernel function, it is not necessary for the algorithm to explicitly map the input data points into a higher dimensional space, which is particularly desirable if the transformed feature space is unfeasible to compute, such as an infinite dimensional space.

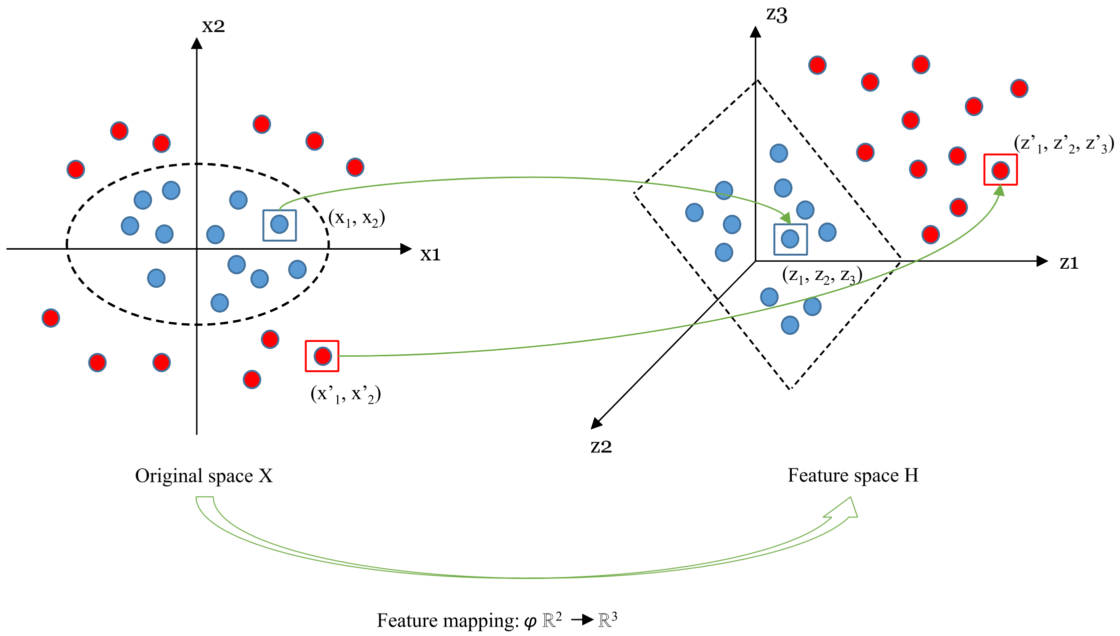


(a)



(b)

(b)



(c)

Figure 1.7 Separating planes of SVM

Occasionally, for some easy cases, the spatial distribution of a dataset can be separate linearly (a and b). But, as a more often case, a dataset tends to be not separable in a low-dimensional space linearly (c). A kernel function transforms the original distribution of a dataset to a higher dimensional space, which allows for a hyperplane to separate the dataset upon their membership.

Continuousness and symmetry are two essential requirements for a kernel function. In addition, based on Mercer's theorem, a kernel function is preferred to a positive semi-definite Gram matrix. This is not a necessary but a sufficient condition, which means the functions that are not satisfied Mercer's theorem can also be used as kernel function. The typical instances include Gaussian Kernel and Polynomial Kernel. Further details will be introduced in subsequent sections.

1.5.1.3.2 The hyperplane with maximum margin

In this section, we return to the separating hyperplane. We have seen that the classification of two sets of values may have multiple solutions when they are linearly separable. In order to choose the best amongst the potential solutions, SVM selects a plane that separates the points according to their class while ensuring the distance from closest points in each class is maximised and equalised. As shown in Figure 1.8, the chosen plane is defined as the hyperplane. The points closest to the hyperplane in each sub-group are referred as the Support Vectors (SVs) and the total distance between SVs in each sub-group and the hyperplane is referred as maximum margin of separation.

SVMs reasonably assume the test data set belong to the same distribution as the training data. In terms of statistical learning theory, SVMs select the hyperplane with the maximum margin as the decision boundary so as to maximise the ability to classify test examples correctly (Noble, 2006). In other words, SVMs adopt the hyperplane with maximum margin as a classifier that provides the most stable prediction against perturbations of the input data. To illuminate the choice of maximal margin, for a set of data points (x_i, y_i), i = 1, 2, …, n, where x_i is a input sample and $y_i \in (-\alpha, +\alpha)$, a formalised SVM with algebra for a given hyperplane is defined as below:

$$w^{T}\cdot x + b = 0$$

This represents the location of the hyperplane in feature space, in which w and b are both known as adjustable parameters: weight vector and bias. For a two-class classification problem, two flat planes along the outer margin of either class is defined as follows:

$$class\ 1: w^T \cdot x + b = \alpha \quad \forall x\ of\ class\ 1$$

$$class\ 2: w^T \cdot x + b = -\alpha \quad \forall x\ of\ class\ 2$$

We set the constant α = 1, by subtracting the outer margins of two classes, the margin with maximal width can be calculated as follows:

$$maximum\ margin\ w, b = maximise \frac{2}{\|w\|^2}$$

To obtain a classifier with the maximum margin, from the discriminant function above, is to maximise the geometric margin 1 / ||w||; equivalently, to minimise $||w||^2$. So, it can also be formulated in the other way:

$$maximum\ margin\ w, b = minimise \frac{1}{2} \| w \|^2$$

subject to

$$y_i\ (w^T \cdot x_i + b) \geqslant 1$$

By adjusting the parameters w and b, the equation above can be minimised to obtain an optimal classifier. If we assume the data is linearly separable, then, it is reasonable to expect each unseen test data point to be classified correctly. For a linearly separable data set, we have seen a larger margin yields better prediction. In reality, for some linearly separable problems, a greater margin could be obtained, and better performance overall, if some points are allowed to be misclassified (Ben-Hur and Weston, 2010). Moreover, in some sense, a trade off exists between the width of margin and the number of misclassified input data points. To investigate this aspect further, we need an understanding of the idea of a soft margin adopted by SVMs.

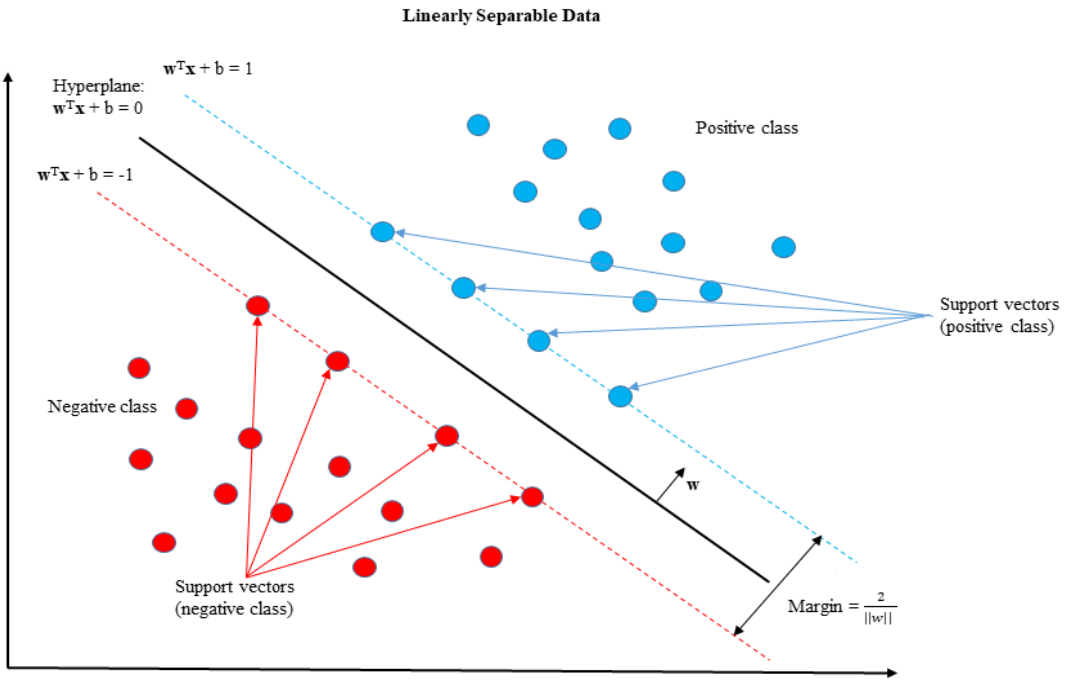


Figure 1.8 The maximum margin

A linearly separable data set can be divided into sub-groups by multiple flat planes. The optimal plane referred to as a hyperplane with a maximum margin is obtained by the trained classifier, which ensures the distance between either sub-group of support vectors and the plane is equally maximised.

1.5.1.3.3 Soft margins

We have seen that the optimisation problem for obtaining maximum separation is equivalent to minimizing ½ $||w||^2$. To solve this problem of primal minimisation, the Lagrange multipliers (α_i) can be introduced to obtain the dual formulation:

$$L\ (w, b,\ \alpha\ ) = \frac{1}{2} w^T\ w - \sum_{i=1}^{N} \alpha_i\, [y_i\ (w^T\ x_i + b) - 1]$$

In practice, to obtain an optimal classifier there is benefit in allowing some data points to go beyond the margin or to be misclassified so as to achieve a greater margin. Thus, the slack variable (ξ_i) is introduced and the problem of optimizing the classifier becomes:

$$maximum\ margin = minimise \frac{1}{2} \parallel w \parallel^2 + C\sum_i^N \xi_i$$

subject to

$$yi\,(w^T \cdot x + b) \geqslant 1 - \xi\, i\ for\ i = 1, 2, \cdots, N$$

As shown in Figure 1.9, where $\xi_i = 0$, the points on the margin are identified as support vectors. Where $\xi_i > 0$, points may be located outside the margin or hyperplane to some extent. In the case of $1 \geqslant \xi_i > 0$, they are between the margin and the hyperplane, which is named margin violation; in the case of $\xi_i > 1$, they are beyond the hyperplane and identified as misclassified. C is a regularisation parameter that sets the balance between maximum achieved margin and minimum permitted slack. When the value of C is small, the constraint is weak, allowing a greater margin; when the value of C is large, a narrow margin is achieved. If C is infinite, maximum constraint will be enforced, which is referred to as a hard margin. This parameter enables a trade-off between the size of the margin and the number of margin / hyperplane violations. The upper bound of a classifier's error rate is then determined by the extent of misclassification. The formulation above, initially proposed by Cortes and Vapnik, is known as soft-margin SVM (Cortes and Vapnik, 1995).

To be more specific, by applying the method of Lagrange multipliers, this optimisation problem can be solved with a convex quadratic transformed to a higher dimensional space:

$$L\,(w, b, \xi_i, \alpha_i, \beta_i) = \frac{1}{2} \parallel w \parallel^2 + C\sum_i^N \xi_i - \sum_{i=1}^N \alpha_i [y_i\,(w^T\ \phi\,(x_i) + b) - 1 + \xi_i] - \sum_{i=1}^N \beta_i\ \xi_i$$

Then, partial derivative is conducted on w, b, ξ_i to simplify the problem as follows:

$$\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^N \alpha_i y_i\ \phi\ (x_i)$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow 0 = \sum_{i=1}^N \alpha_i y_i$$

$$\frac{\partial L}{\partial \xi i} = 0 \rightarrow C - \alpha_i - \beta_i = 0$$

The optimisation problem thus becomes:

$$Minimise\ w, b,\ \xi_i L(w, b, \xi i, \alpha i, \beta i) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(x_i) \phi(x_j)$$

And the dual formulation becomes:

$$Minimise\ \alpha_i\ \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(x_i) \phi(x_j) - \sum_{i=1}^{N} \alpha_i$$

Subject to

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$C - \alpha_i - \beta_i = 0$$

$$0 \leqslant \alpha_i \leqslant C, i = 1, 2, \cdots, N$$

Exploiting the kernel method, we can replace the nonlinear mapping function φ(x_i) φ(x_j) with a kernel function K (x_i, x_j).

$$Q(\alpha) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{N} \alpha_i$$

If the value of α can be determined, w and b can be calculated as follows:

$$w^T = \sum_{i=1}^{N} \alpha_i yi\ \phi(x_i)$$

$$b = y_i - \sum_{i=1}^{N} \alpha_i y_i < \phi(x_i) \cdot \phi(x_i) >$$

Then, we can obtain the hyperplane and the decision function for classification as follows:

$$w \cdot x + b = 0$$

$$F(x) = sign\ (w \cdot x + b)$$

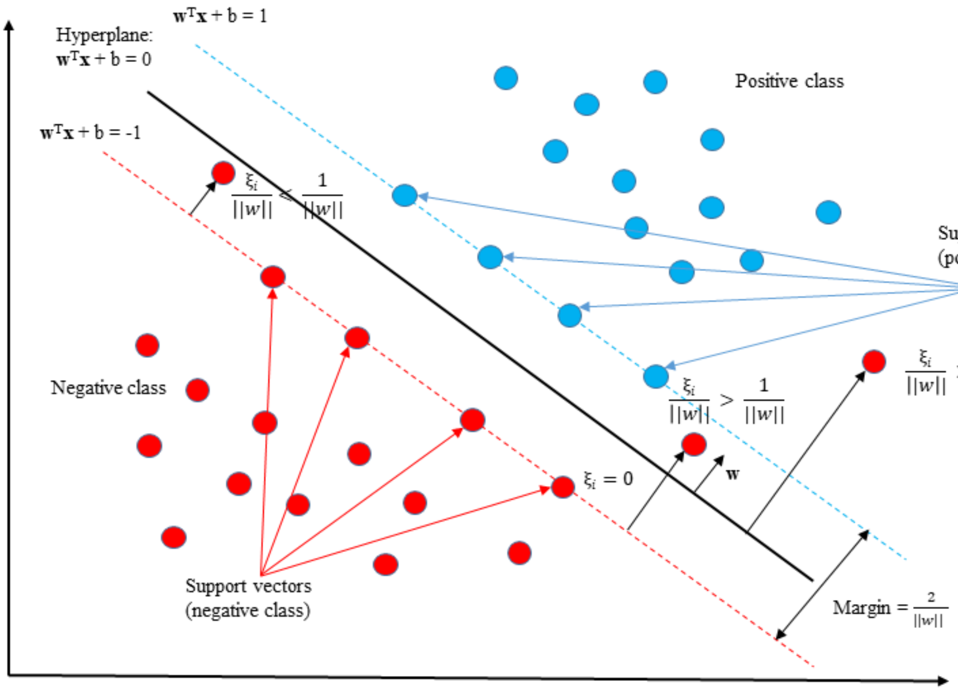

$\mathbf{w}^T\mathbf{x} + b = 1$
Hyperplane: $\mathbf{w}^T\mathbf{x} + b = 0$
$\mathbf{w}^T\mathbf{x} + b = -1$
Positive class
$\frac{\xi_i}{||w||} < \frac{1}{||w||}$
Support vectors (positive class)
$\frac{\xi_i}{||w||} > \frac{2}{||w||}$
Negative class
$\frac{\xi_i}{||w||} > \frac{1}{||w||}$
$\xi_i = 0$
$\mathbf{w}$
Support vectors (negative class)
Margin = $\frac{2}{||w||}$

Figure 1.9 Soft-margin of SVMs

In contrast with a hyperplane that completely separates the data points into two groups, a hyperplane with a soft margin allows for some slack for points positioned between the margin and the hyperplane (violation), or even beyond the hyperplane (misclassification). The number of slack points is adjustable with a regularisation parameter. There is a trade-off between the size of the margin and the number of slack points.

1.5.1.3.4 Kernel functions and the selection of kernels

We have seen that SVMs derive an optimal hyperplane, with maximum separation between the classes, based on methods of structural risk minimisation (Vapnik, 2013). Specifically, data sets that are not linearly separable can be made so by applying a kernel function on the input data, removing the need to conduct the computation in feature space. The classification performance of a classifier is highly dependent on the kernel function, as that determines the structure of the transformed feature space (Liu et al., 2005). It is critical to select a kernel function that maps the data set into a feature space neither too rich nor too impoverished, to avoid over-fitting or under-fitting respectively (Anthony et al., 2007).

A wide range of kernel functions has been proposed and evaluated in the literature. The four most widely adopted kernels are listed in the Table 1.1. Crudely, linear kernels tend to fit linear problems well; and non-linear problem tend to require non-linear kernels. A Gaussian kernel, commonly named Radial Basis Function (RBF) kernel, tends to suit linear problems as well as a linear kernel, because the latter is actually a special form of the former (Keerthi and Lin, 2003). Since the computational cost of using an RBF kernel, both during training and prediction, is higher, a linear kernel is nonetheless preferred for linear problems. For non-linear problems, RBF kernel is usually the first choice, as it is most versatile. The sigmoid kernel performs similarly in many settings (Lin and Lin, 2003) but is rarely better; a polynomial kernel requires more parameters than RBF, increasing the risk of over-fitting. Furthermore, there are numerical difficulties of using the polynomial and sigmoid kernels, limiting their usefulness (Hsu et al., 2003). Admittedly, the RBF kernel is not always ideal. When the number of attributes is very large, in particular when the number of attributes is far greater than the number of instances, a linear kernel will often provide better performance. In the case where both the number of instance and attributes are very large, linear and RBF kernels offer similar performance, but linear the kernel is much more efficient (Hsu et al., 2003) where the underlying data structure is suited to it.

When using the RBF kernel, optimal prediction performance can be achieved by tuning the adjustable parameter gamma appropriately(Souza, 2010).

Kernels	**Mathematical demonstration**
Linear	K (x, x’) = xT x’ + C
Polynomial	K (x, x’) = (γ xT x’ + C) d
RBF	K (x, x’) = exp (- γ ‖ x − x′ ‖2)
Sigmoid	K (x, x’) = tanh (γ xT x’ + C)

Table 1.1 Common kernel functions

1.5.1.3.5 Extensions of SVMs

SVM classifiers can be generalised to deal with a broader range of learning problems. For instance, Support Vector Clustering (SVC), which is also based on a kernel function, is designed to solve unsupervised learning problems (Ben-Hur et al., 2001). Multiclass SVMs learn the assignment of multiple labels to a set of data by breaking down what is a complicated multiclass problem into several simple two-class classification problems (Hsu and Lin, 2002). Transductive Support Vector Machines (TSVMs) can deal with semi-supervised learning problems where the data are partially labelled (Joachims, 2006). Regression can be solved by Support Vector Regression (SVR) (Drucker et al., 1997) and Least Squares Support Vector Machine (LS-SVM) regression (Suykens and Vandewalle, 1999) employing kernels suitable for predicting continuous outputs. The following sections provide more details on these extensions.

1.5.2 Semi-supervised learning algorithms

In supervised learning algorithms, a set of input data are paired with a corresponding set of labels during training as a precondition for creating a model that predicts the label from test data where the label is unknown. In practice, it not always possible to have each training input instance labelled with the correct answer. Unlabelled data is common for a variety of practical reasons, especially where the label is generated by a human expert and so often either not recorded or captured in a form, such as free text, that is not easy to parameterise. Semi-supervised learning can take advantages of partially labelled datasets for training, making it potentially valuable in real world practice.

Semi-supervised learning may be thought to lie between supervised and unsupervised learning and can be treated as an extension of either. However, most algorithm development here focuses on making use of unlabelled data in a discriminative setting. Its origins are to be found in the realm of self-training, developed in 1960s (Chapelle et al., 2009), thereafter further elaborated as transductive learning and inductive learning in 1970s (Vapnik and Chervonenkis, 1974, Pellegrino and Glaser, 1979). The combination of labelled and unlabelled information within semi-supervised models has been shown to improve predictive power compared with either purely supervised or unsupervised models across a range of tasks and model architectures (Zhu, 2006, Turian et al., 2010, Ando and Zhang, 2005, Bennett and Demiriz, 1999, Joachims, 2006). By drawing intelligence from both labelled and unlabelled instances, semi-supervised learning can improve reproducibility and increase stability (Sindhwani and Keerthi, 2006, Zhu and Goldberg, 2009). The constraints of modern clinical environments, where unlabelled data is far commoner than the labelled kind, make a semi-supervised approach important to consider for any healthcare application. This is not merely a resource issue: reliance on labelled data alone risks the entrenchment of biases arising from (typically human expert-derived) labels. Large-scale unlabelled data is therefore of significant potential value to clinical research.

Figure 1.10 illustrates the principle of deriving benefit from learning unlabelled data in addition to labelled samples, where the original decision boundaries are adjusted in response to the intrinsic structure of the data.

Analogously to supervised learning, semi-supervised learning can be grouped by semi-supervised classification, semi-supervised regression, semi-supervised clustering and semi-supervised dimensionality reduction. This thesis will focus on semi-supervised classification, though the principles apply more broadly.

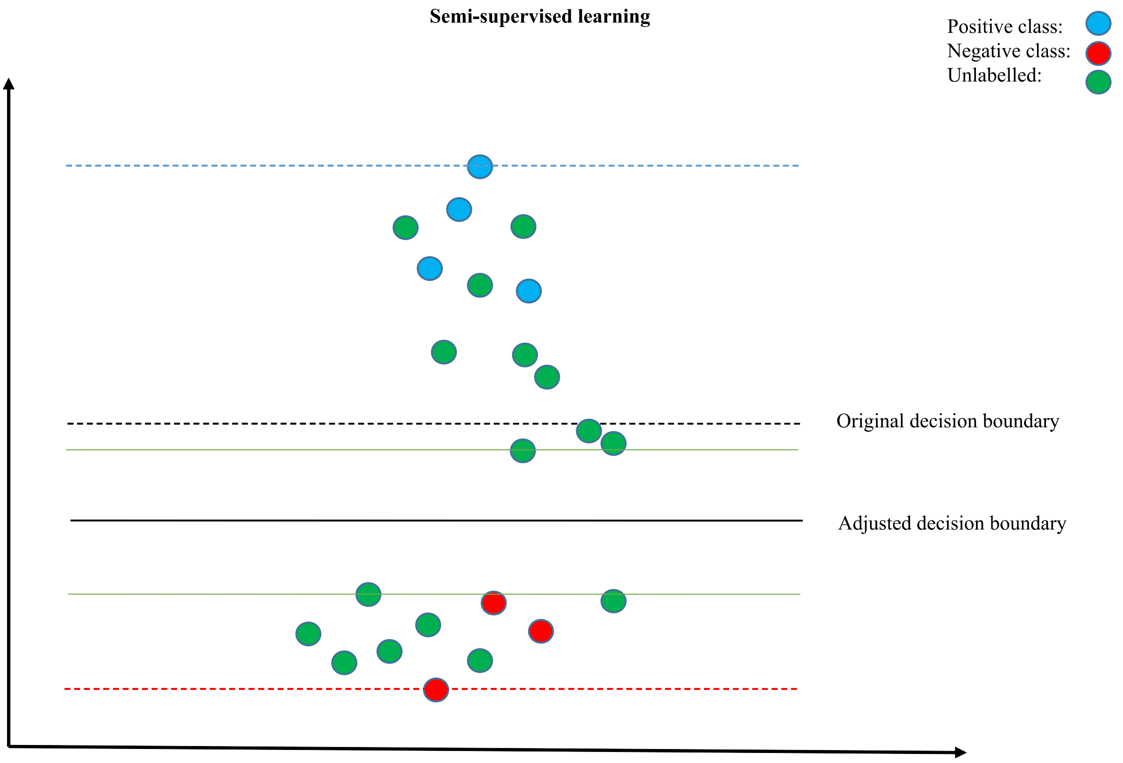


Figure 1.10 The contribution of unlabelled data examples in semi-supervised learning

The decision boundary showed in dotted line is determined by labelled data samples alone. When more data points without labelling are modelled, a better decision boundary can be found.

1.5.2.1 Inductive learning and transductive learning

In the realm of machine learning, the problem that yields the expected labels for a set of unlballed examples can be solved by two philosophies: induction and transduction.

Inductive learning is instantiated by algorithms whose goal is to learn a function for predicting unseen examples from observed instances. In other words, a universal model makes generalisations to understand or label the unseen instances. Concretely speaking, supervised learning can be generalised as a specific form of inductive learning. On the other hand, transductive learning algorithms are based on instance learning (Gammerman et al., 1998), which exploits unlabelled examples to extract information that can be transferred to labelled examples so as to infer the correct labels for unlabelled examples more easily.

The major difference between them is that inductive learning aims to induce a generalised function or rule to fit any new unseen examples. Whereas, for a specific data set, transductive learning can be thought as a procedure that learns the observed (labelled) training set, whist trying to predict the expected labels on the observed (unlabelled) test set. In the context of learning, transduction is in a less ambitious manner with respect to the induction (Bianchini et al., 2016). The characteristic of transduction is to avoid building a general model. Instead, it takes advantage of the information implicit in the observed unlabelled instances to understand them better. Admittedly, the obvious disadvantage is it is not appliable to the unseen examples as it is learnt for generalisations.

Overall, an extension of either supervised learning or unsupervised learning, semi-supervised learning is also based on a model driven by the data, but one where both labelled and unlabelled examples are used in the learning process. There are two ways for conducting semi-supervised learning: one approach is based on unsupervised learning associated with additional labelled data to improve the resultant representations such as characteristic clusters (induction); the other is based on supervised learning with extra unlabelled data to improve the accuracy of classification (transduction). In this thesis, we focus on the latter approach to optimise the prediction directly (Demiriz and Bennett, 2001, Sindhwani et al., 2005, Zhu and Goldberg, 2009, Van Engelen and Hoos, 2020).

1.5.2.2 Semi-supervised learning assumptions

The assumptions underlying semi-supervised learning algorithms are critical to the success and effectiveness of a semi-supervised learning model. Where a dataset is only partially labelled, correct assumptions about the underlying structure and distribution of the data can make unlabelled data more effectively learnable, thereby improving the reliability of estimates of the decision boundary (Zhu and Goldberg, 2009). Selecting the right set of assumptions is, however, an open question, and requires consideration in multiple aspects both empirically and theoretically. The assumptions commonly applied in semi-supervised learning methods include the smoothness assumption, the cluster assumption and the manifold assumption. The smoothness assumption encourages the model, on the one hand, to assign the same membership to data points that are in a high-density region and continuously close to each other; and, on the other, to place the decision boundary in low-density regions. The cluster assumption is actually a special occasion of smoothness assumption, where the data points located in each concentrated cluster tend to be of the same group. Compared to the smoothness and cluster assumptions, the manifold assumption is more widely employed in various semi-supervised learning methods (Belkin et al., 2006, Niyogi, 2013, Belkin and Niyogi, 2004, Goldberg et al., 2009), as it can be treated as a generalised extension of the smoothness assumption. In essence, the manifold assumption is the same as the smoothness and cluster assumptions in terms of similarity detection and comparison; but further assumes the data lie in a manifold of a lower dimensionality than the input space, where learning both labelled and unlabelled data can focus on the structure of the manifold so as to avoid the curse of dimensionality. We may understand the manifold assumption as an implication of the possibility of dimensionality reduction. On the positive side, operating in a lower dimensional space is easier; negatively, the fidelity of the resultant model will be limited by the quality of the lower dimensional embedding. Moreover, the distribution of the data need not be unified, requiring multiple manifolds to capture its structure.

The manifold assumption is particularly important for semi-supervised learning. It allows us to learn the structure of manifolds from unlabelled data alone, thereby gaining some insights into the learning problem, and to incorporate what we have learnt from the unlabelled data into the process of discriminating between the labelled examples.

1.5.2.3 Semi-supervised learning methods

We have already seen that in semi-supervised learning, strong assumptions need to be made to use unlabelled data effectively. So, in solving any one problem, it is critical to choose a method best suited to the structure of the data. In other words, an uniformed choice of semi-supervised method is risky, and a poor choice may result in worse prediction performance than learning the labelled data alone (Zhu and Goldberg, 2009).

In the following section, we will introduce several methods that are commonly used to tackle semi-supervised learning problems; and outline how these methods make different assumptions to link the marginal distribution and the conditional probability distribution to improve predictive performance. An inclusive, but not exhaustive, list of semi-supervised learning methods is self-training, co-training, probabilistic generative models, graph-based models, and Transductive Support Vector Machines (TSVMs) (Zhu and Goldberg, 2009, Zhu et al., 2003).

1.5.2.3.1 Self-training

In the middle 1990s, a method that treats learning algorithms as evaluation function searches for a good subset of relevant feature selection was proposed as a wrapper approach (John et al., 1994, Kohavi and John, 1997). Self-training as a typical instance of the wrapper approach is recognised as an incremental algorithm

(Nigam and Ghani, 2000). It is probably the simplest semi-supervised learning technique, which initially appeared around 1960s and 1970s but was firstly published in 1995 for solving word sense disambiguation within different contexts (Yarowsky, 1995). The basic idea of self-training is to prepare for a set of data samples with labels and another set of data samples without labels, and to allow the model to teach itself with its own predictions. The initial classifier is trained with the whole set of labelled samples; then it is used to classify the unlabelled data samples. A subset of unlabelled data samples that are learnt with most confident weights together with their assigned labels are then transferred from the original unlabelled dataset to the labelled dataset. By repeating this process of self-teaching or bootstrapping, the classifier can be trained iteratively until all the unlabelled samples are classified or discarded as uncertain.

The advantages of self-training are straightforward and intuitive. As the choice of learner is completely open, self-training can be either inductive or transduction learning (Zhu and Goldberg, 2009). Moreover, it is flexible enough to adopt either a simple or complex classifier, as the task demands. In common with wrapper methods, the disadvantages are the high cost of computation. This is because the learning algorithm has to be conducted iteratively for each featured subset, and is hard to analyse in general (Jelonek and Stefanowski, 1997). To optimise the performance of the self-training process, it is suggested that the unlabelled samples should be randomly selected at a relatively small proportion of the dataset, which means self-training usually requires multiple rounds of operation (Culp and Michailidis, 2008).

As a method of semi-supervised learning, self-training makes strong assumptions. It is possible to improve classification where the learner is optimal and selection metrics are good, but it is also possible to degrade prediction by polluting the original labelled dataset. We can imagine that an initial incorrect prediction may be reinforced by itself during the re-training over multiple iterations, resulting in worse performance. As the selection metrics determine the criteria on which weakly labelled or unlabelled examples can be confidently refined and combined with labelled dataset for training and re-training, the selection metrics in practice are very crucial for the success of self-training (Rosenberg et al., 2005).

1.5.2.3.2 Co-training

Co-training was initially proposed as a means of boosting the learning performance achievable with a small set of labelled data samples by combining with a relatively large set of unlabelled data examples, applied to the task of classifying web pages (Blum and Mitchell, 1998, Mitchell, 2004). In common with self-training, co-training aims to augment the labelled data by learning the unlabelled data iteratively. In addition, co-training and self-training are both wrapper methods (Zhu and Goldberg, 2009), which means co-training is also a computationally intensive process but flexible enough to fit a wide range of learning tasks. Co-training can apply any appropriate learning algorithms to produce classifiers as long as a confidence score can be assigned to the prediction on which a decision if the unlabelled data example can be confidently labelled or not can rest.

In contrast with self-training, co-training requires two distinct "views" of the data on which predictions with two separate classifiers are made over multiple iterations. The views are created by splinting the data into two feature sets naturally, or randomly as an alternative. Co-training is a method that relies on strong assumptions. Its success requires the two views to be good enough in two respects: firstly, the characteristics of the data can be partitioned into two sets of different and complementary features, and each set of features encompassed in a view must be sufficient for good used for classification. Second, both views must be conditionally independent in relation to the target class (Zhu, 2006, Zhu and Goldberg, 2009, Prakash and Nithya, 2014). Conditional independence can be stated as below:

$$P\,(x^{(1)}\,/\,y, x^{(2)}) = P\,(x^{(1)}\,/\,y);$$

$$P\,(x^{(2)}\,/\,y, x^{(1)}) = P\,(x^{(2)}\,/\,y).$$

From this statement, we can see knowing the true label (y) and either of the view (x $^{(1)}$ or x $^{(2)}$) have no effects on observing the other view (x $^{(2)}$ or x $^{(1)}$). In particular, conditional independence assumption is critical for co-training to learn unlabelled data objects (Zhu and Goldberg, 2009, Zhu, 2006).

Co-training begins by applying the original labelled data set to two separate classifiers that learn from each view. Then, each classifier is used to predict the unlabelled data and refine the most confident instances to enrich the training data set for the other classifier. In other words, for each iteration, the newly labelled data examples from unlabelled data set predicted by one classifier will be used to further enhance the learning of other classifier and vice versa. Iteratively, the process is repeated until all the unlabelled data examples are exhausted or no more unlabelled data examples can be assigned a label confidently.

1.5.2.3.3 Generative models

Generative models are possibly the first learning method proposed for semi-supervised learning (Zhu, 2006, Prakash and Nithya, 2014). As generative models assume that all the data are generated by the same model, for a given training data set, generative models understand the data by learning to recreate it.

In machine learning, a generative model is a statistical model of the joint probability distribution of a set of random variables: an observable variable (x), and a target variable (y), where the joint probability distribution is the probability distribution for the simultaneous behaviours of x and y that falls in within a specified range. Specifically, the joint probability distribution that actually describes the relationships and interactions between the variables can be formulated as a cumulative distribution function, a joint probability density function, or a joint probability mass function. The fundamental assumption of generative models can be stated as: $p(x, y) = p(y)\, p(x \mid y)$ (Zhu, 2006). In generative models, $p(x \mid y)$ is an identifiable mixture distribution, which can be in a variety of forms. The concept behind mixture models is critical for semi-supervised or unsupervised learning that incorporates unlabelled data into the models. This is because unlabelled data represents the distribution and of all instances, mixed together. In order to have the instances classified, it is ideal to decompose the mixed entity into individual classes by identifying the instance distribution from each class (Zhu and Goldberg, 2009). Frequently used mixture models include mixture of Gaussian distribution (GMM) which is suitable for image classification (Permuter et al., 2006, Kim and Kang, 2007); mixture of multinomial distributions (Naïve Bayes) which is commonly used for text categorisation (McCallum and Nigam, 1998, Baker and McCallum, 1998, Kibriya et al., 2004); and hidden Markov models (HMM) which is more often applied to speech recognition (Rabiner, 1989, Bahl et al., 1986, Huang et al., 1990). As our tasks are in relation to learning images, we mainly focus on GMM to introduce how generative models contribute to semi-supervised learning in this section.

For a training dataset (X), the problem we are interested is to predict the label (y) by giving a data example (x ∈ X). A generative model is a probabilistic approach that applies Bayes rule to maximise the conditional probability p (y | x) to assign a label to the give data example; and meantime to minimise the expectation loss. The definition of posterior probability and conditional probability (Zhu and Goldberg, 2009) can be stated as follows:

$$p\,(y\,/\,x) = \frac{p\,(x \mid y)\,p\,(y)}{\sum y^{\prime}\ p\,(x\,|y^{\prime}\,)\,p\,(y^{\prime}\,)};$$

$$p\,(y\,/\,x) \in [0,\,1];\ \Sigma_y\,p\,(y\,/\,x) = 1.$$

To demonstrate a GMM, we assume the training dataset is drawn from a multivariate Gaussian distribution. Thus, by estimating the mean of feature vector among the instances (μ_y) and the covariance matrix (Σ_y), the probability density function (Zhu and Goldberg, 2009) of class conditional distribution can be stated as below:

$$p\,(x\,/\,y) = N\,(x;\,\mu_y,\,\Sigma_y) = \frac{1}{(2\pi)^{D/2}\,|\Sigma y|^{1/2}}\ exp\left(-\frac{1}{2}(x-\ \mu_y)^T\,\Sigma_y^{-1}(x-\ \mu_y)\right).$$

From the equations above we can see the prediction can be worked out with the conditional probability (p (x | y)) and prior probability (p (y = C_i)). Specifically, the conditional probability of classes is determined by the mean of feature vector (μ) and covariance matrix (Σ) as model parameters. Thereby, the set of related parameters is usually defined as initialisation of a mixture model: θ = [p (y = C_i), μ, Σ], accordingly, the conditional probability and prior probability can be notated as p (x | y, θ) and p (y = Ci, θ), respectively. The process of training is to optimise θ, which involves another term: maximum likelihood estimation (MLE). Briefly, MLE is commonly used as a criterion while optimizing a problem by maximizing the logarithm likelihood (Zhu and Goldberg, 2009). Intuitively, MLE is the parameter that makes log (p (D | θ)) largest for a given training dataset (D). In semi-supervised learning, as the likelihood is contributed by both labelled and unlabelled instances, the MLE can be combined as bellow:

$$log\, p\ (D\,/\ \ \theta\ ) = log\ (\prod_{i=1}^{l} p\ (x_i,\ y_i \mid \theta)\ \prod_{i=l+1}^{l+u} p\ (x_i \mid \theta)\ )$$

$$= \sum_{i=1}^{l} log\, p\ (y_i \mid \theta)\ p\ (x_i \mid y_i,\ \theta) + \sum_{i=l+1}^{l+u} log\, p\ (x_i \mid \theta),$$

where the training data (D) is composed by labelled data (x_i) and unlabelled data (x_u); p (x_i | θ) refers as the probability of an unlabelled data sample generated by any of the classes.

MLE in semi-supervised learning is not as straightforward as the procedure in supervised learning, but a local optimum can be achieved by the expectation-maximisation (EM) algorithm (Zhu, 2006, Zhu and Goldberg, 2009, Baum et al., 1970, Dempster et al., 1977). Based on the procedure for prediction introduced above, a general EM algorithm makes use of an initialised parameter (θ) whose maximised likelihood is estimated from the original labelled data to assign provisional labels (“soft labels”) to the unlabelled data; then the original parameter (θ) combined with newly labelled data is used to further improve the maximised likelihood. The two steps are referred as expectation and maximisation which is an iterative procedure until p (D | θ) converges. For a training data (D) including labelled data and labels (x_l, y_l), and unlabelled data (x_u), the initial parameter is θ and the MLE is log (p (x_l | θ)). The derivation of the labels (y_u) for unlabelled data can be done by repeating the expectation and maximisation steps iteratively as follows:

$$Initialisation{:}\ D = (x_l, y_l, x_u);$$

$$Expectation\ step{:}\ q(y_u) = p\ (y_u\,/\,D,\ \ \theta\ );$$

$$Maximisation\ step{:}\, \theta'\ = \underset{\theta}{argmax} \sum_{y_u} q\ (y_u)\ log\, p\ (D, y_u \mid \theta).$$

Ultimately, EM maximises p (xl, yl, xu | θ), which is similar to self-training. In each iteration, the EM algorithm takes the current classifier (θ) to label the unlabelled data and augments the labelled data by refining the most confident instances of unlabelled data so as to improve the classifier (θ) for the next iteration. The EM algorithm is highlighted in this section because it is widely recognised as a standard method for optimisation in the context of generative models (Zhu and Goldberg, 2009). But, in machine learning, there are a wide range of optimisation methods, such as variational approximation (Jaakkola, 2001), combinatorial optimisation (Chapelle et al., 2008) and continuous optimisation (Hadley et al., 1990), which are beyond the scope of this thesis.

Generative models offer a mature probabilistic framework for semi-supervised learning as long as the assumptions made by the mixture models are correct or close to the truth. Nonetheless, for every plus there is a minus. It is not always easy to identify or even verify the applicability and correctness of a model. Consequently, prediction fidelity may be hurt by considering the unlabelled data with incorrect models (Zhu, 2006, Castelli and Cover, 1996). The problem can be especially acute where only a small proportion of training data labelled. To minimise the risk of model error, for a specific task, an intuitive way is to select the model that fits the task best with the aid of domain knowledge. Another way is introducing a weight parameter (w $\in [0, 1]$) to weaken the role of unlabelled data during training process (Corduneanu, 2002).

Even if the underlying model is right, performance could still be inferior owing to the difficulty of avoiding purely local optima with EM (Dayan and Hinton, 1997). A local optimum might present the optimal solution within a focal space. The problem can be alleviated by selecting different initialisation parameters (θ) and maximizing the likelihood with the EM algorithm multiple times. That way, the parameter (θ) that leads to the best likelihood is more likely to correspond to the global optimum. An alternative remedy is to select better initialisation (θ) informed by prior domain knowledge.

1.5.2.3.4 Discriminative models

For classification, besides the generative model introduced in the last section, discriminative modelling is the other major approach. In contrast with generative models, discriminative algorithms discard the generation assumptions to categorise instances by estimating the posterior probability (p (y | x)) directly. It has been claimed the straightforward approach adopted by discriminative models theoretically and empirically delivers superior performance in classification tasks, in particular, when the training data is large in proportion to the number of features (Ng and Jordan, 2002, Vapnik, 1998). However, generative models can make use of the joint probability (p (x, y)) to create synthetic instances similar to those in the labelled data, revealing complex relationships between the independent and dependent variables. So, in some circumstances, generative models can perform better than discriminative models for classification and regression tasks (Bernardo et al., 2007).

Generative and discriminative approaches differ from each other, but they are complementary. In the following section, we will introduce two methods based on discriminative models: graph-based methods and Transductive Support Vector Machines (TSVMs).

1.5.2.3.4.1 Graph-based methods

In semi-supervised learning, the idea of graph-based methods is to construct an affinity graph from the given training data, including both labelled and unlabelled data, and assume smoothness of the labels over the graph to capture the similarity between instances simply. Graph-based methods are non-parametric, discriminative, as well as transductive (Zhu, 2006, Zhu and Goldberg, 2009). In a constructed graph, any vertex that represents an instance either labelled or unlabelled will be connected to another vertex by an undirected edge. The length of the edge between two instances indexes their similarity. In graph-based algorithms, a weight parameter (w_{ij}) is usually introduced to use the edges to assign labels to unlabelled instances. Specifically, the weight parameter increases as the Euclidean distance between two instances decreases. Two instances thus tend to be in the same sub-group if the weight of the edge connecting them is large. The Gaussian kernel function (Zhu and Goldberg, 2009) shows the relationships between weight and Euclidean distance when defining similarity thus:

$$wij = exp\left(- \frac{\|x_i - x_j\|^2}{2\sigma^2}\right),$$

where σ is a bandwidth parameter to control the amplitude of weight changes.

Graph-based methods assume any two instances connected by an edge tend to be of the same label. Figure 1.11 presents a simple example of applying graph-based methods to a two-class classification semi-supervised learning problem. The training process begins with fixing the labelled instances, then the vertices that represent unlabelled instances are connected to the labelled instances with sparse edges, assigning similar vertices to the same class. Over training iterations, newly classified vertices are used to measure the similarity between instances that establish further connections. Each newly classified, (originally unlabelled) vertex is thus used as a stepping stone to propagate and proliferate the labels until all the unlabelled instances are assigned. In short, graph-based methods exploit the similarity between neighbors in a graph, employing a process of propagation from a neighbor to a neighbor's neighbors. For example, in Figure 1.11, x_u is assigned the negative class owing to its graph neighborhood even though it is closer to a positive instance in terms of Euclidean distance.

In the following section, a few commonly used graph-based algorithms for semi-supervised learning will be briefly introduced including Mincut, harmonic function and manifold regularisation (Zhu, 2006, Zhu and Goldberg, 2009). Two aspects must be optimised in training these graph-based algorithms: one is the loss function; the other is the smoothness through the whole graph. In the context of this thesis, we do not focus on the mathematical details behind these algorithms in depth; but rather survey the underlying concepts and assumptions to help stimulate ideas on modelling the relationships between brain lesions and cognition and behaviour.

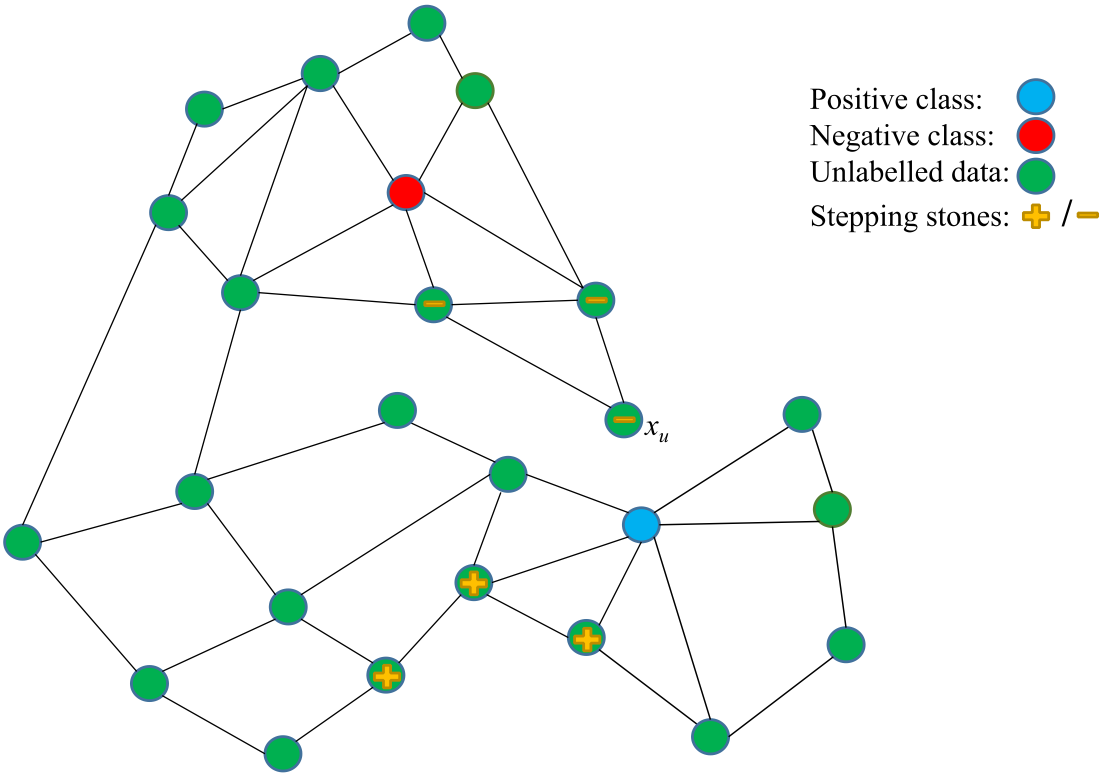


Figure 1.11 Basic concept of graph-based methods in semi-supervised learning

In essence, graph-based method assumes any two instances connected by an edge tend to be the same label. In this figure, the vertex of x_u is a typical instance that is positioned much closer to a positive instance but labelled as negative due to graphical connection.

1.5.2.3.4.1.1 Mincut

Mincut is a graph-based algorithm for semi-supervised learning, where a graph is split into sub-groups by removing minimal relationships. To be more specific, for a set of data that include instances with are unlabelled and labelled with a binary value, Mincut aims to identify the most effective edges that can “block”, if removed, the routes connecting any positive instance to any negative instance. The optimisation seeks to minimise the cuts required to separate positive and negative instances, measured by the weights between connected instances. By splitting the graph, the unlabelled instances are then assigned the same labels as the originally labelled instances in the corresponding sub-graph.

Mathematically, Mincut can be presented as a loss function associated with a regulariser to minimise the risk of overfitting (Zhu and Goldberg, 2009) as below:

$$\min_{f:\,f(x)\in(-1,1)} \infty \cdot \sum_{i=1}^{l}\left(y_i - f(x_i)\right)^2 + \sum_{i,j=1}^{l+u} w_{ij} \left(f(x_i) - f(x_j)\right)^2 .$$

In the equation above, the first term is a loss function for the labelled vertices in the graph whose value is zero if prediction is the same as original labels; otherwise, the value will be infinity. The second term is risk regularisation by minimizing cut size for the unlabelled vertices in the graph. For any pair of vertices, on one side, the risk is theoretically zero if they are not connected with an edge (dissimilar) as the weight (w_{ij}) between them is zero; on the other, there is no risk or misclassification if f(x_i) equals to f(x_j). Pairs of vertices that are confidently predicted to be different classes (small w_{ij}) are then considered cut.

1.5.2.3.4.1.2 Harmonic function

Harmonic Function is another graph-based semi-supervised learning algorithm, which assigns a real value to each unlabelled vertex in the graph by taking all its neighbors' values and calculating their weighted average. Optimisation can be presented in a form similar to Mincut algorithms; but Harmonic function returns a real value rather than the discrete values of Mincut algorithms (Zhu and Goldberg, 2009):

$$\min_{f:\,f(x)\in\mathbb{R}} \infty \cdot \sum_{i=1}^{l}\left(y_i - f(x_i)\right)^2 + \sum_{i,j=1}^{l+u} w_{ij} \left(f(x_i) - f(x_j)\right)^2 .$$

Training involves a process of label propagation, which is an iterative procedure. At the initial iteration, it ensures f(x_i) equals y_i for the labelled vertices and assign an arbitrary value to each unlabelled vertex; then f(x_i) returned on each unlabelled vertex will be iteratively improved with its neighbors' weighted average in the form as below) until the procedure has converged.

$$f(x_i) \leftarrow \frac{\sum_{j=1}^{l+u} w_{ij} f(x_j)}{\sum_{j=1}^{l+u} w_{ij}}$$

As harmonic function satisfies Laplace equation, the graph-based undirected Laplacian matrix can be worked out for a closed form solution (Zhu et al., 2003). An adjacency is defined by edge weights; and a degree matrix is a diagonal matrix, in which, for each vertex, the element is filled with the sum of its neighbor's weights.

$$\textit{Adjacency matrix: } A_{ij} = w_{ij},$$

$$\textit{Degree matrix: } D_{ii} = \textstyle\sum_{i,j=1}^{l+u} w_{ij} \ ,$$

$$\textit{Laplacian matrix: } L = D - A,$$

$$F = (f(x_1), \ldots, f(x_{l+u}))^T.$$

The risk regulariser can then be presented as:

$$\textstyle\sum_{i,j=1}^{l+u} w_{ij} \left(f(x_i) - f(x_j)\right)^2 = 2\, F^T \cdot L \cdot F \ .$$

The labels of unlabelled vertices can be assigned by setting a threshold at zero; thus, the signs of the returned values are used as indicators of positive or negative labels.

1.5.2.3.4.1.3 Manifold regularisation

Both Mincut and harmonic function learn a function to fit labelled and unlabelled vertices that are only involved in the given graph, so their learning processes are intuitive transduction. By constrast, manifold regularisation is an inductive learning algorithm that fits the data globally rather than with the given training data alone. This relies on the manifold assumption, taking a subset from the whole feature space to determine the regularisation norm; it also relies on the fundamental assumption of smoothness. In a manifold regularisation algorithm, similarly to harmonic function, Laplacian norm is the common choice to measure the smoothness of the function. In addition, to help generalise the function beyond the training data, a second regularisation term (Zhu and Goldberg, 2009) based on an integral function to involve the gradient on the manifold, is introduced:

$$\| f \|^2 = \int_{x \in \mathbb{R}} \| \nabla_{\mathbb{R}} f(x) \|^2 \ dx \, .$$

The loss for manifold regularisation is stated as follows:

$$\min_{f:\, f(x) \in \mathbb{R}} \infty \cdot \sum_{i=1}^{l} \big(y_i - f(x_i)\big)^2 + \lambda_1 \| f \|^2 + \lambda_2 F^T \cdot L \cdot F,$$

where λ_1 and λ_2 are used to balance the two regularisation terms.

Manifold regularisation is particularly useful when the number of test instances is large. Moreover, the loss function can accommodate label noise, which helps to improve training accuracy when the training data is imperfectly labelled.

To sum up, graph-based methods are supported by clear mathematical framework and provide strong performance if the learning task fits the graph well. But, conversely, performance will be poor if the these prerequisites are not satisfied.

1.5.2.3.4.2 Transductive support vector machines (TSVMs)

Simply put, TSVMs can be treated as an extension of SVMs. Based on the SVMs we have introduced in the section of supervised learning algorithms, the intuition of TSVMs is to learn a training data including both labelled and unlabelled instances under the cluster assumption (Chapelle and Zien, 2005); then confirm a classifier that ensures the correct classification on labelled instances while maximizing the margin based on all (labelled and unlabelled) instances. S^{3}VMs are commonly referred to as Transductive Support Vector Machines (TSVMs) in of the literature. This is because the original theory of SVMs applied to semi-supervised learning proposes to improve performance by learning an additional set of unlabelled instances via transduction (Bennett and Demiriz, 1999, Joachims, 1999a, Joachims, 1999b). Although SVMs' learning process can be either induction or transduction (Tong and Koller, 2001), the learned function is actually generalised to any unseen test instances. So, some researchers assert that S^{3}VMs is the more appropriate name for SVM applications in semi-supervised learning (Zhu, 2006, Zhu and Goldberg, 2009). In this thesis, S^{3}VMs and TSVMs are treated as the same to avoid confusion.

Figure 1.12 illustrates how the margin between classes changes when considering additional unlabelled instances within a traditionally inductive SVM compared with a semi-supervised Transductive Support Vector Machine. TSVMs maximise the margin by considering both labelled and unlabelled instances, the decision boundary being guided by the unlabelled instances away from the high-density regions so as to minimise the bounds of generalisation error. In SVMs, the discriminators learn the decision boundary explicitly by modeling the posterior probability (p (y | x)) from the labelled instances alone. In TSVMs, the only available estimates from unlabelled instances is p (x). Owing to the potentially strong impact on the final prediction of prior assumptions derived from unlabelled instances, it is essential to establish the connection between p (y | x) and p(x), or identify the shared parameters between them, which is effectively to balance the impact of prior assumptions and judge how unlabelled instances can aid the training procedure (Zhu, 2006, Seeger, 2000).

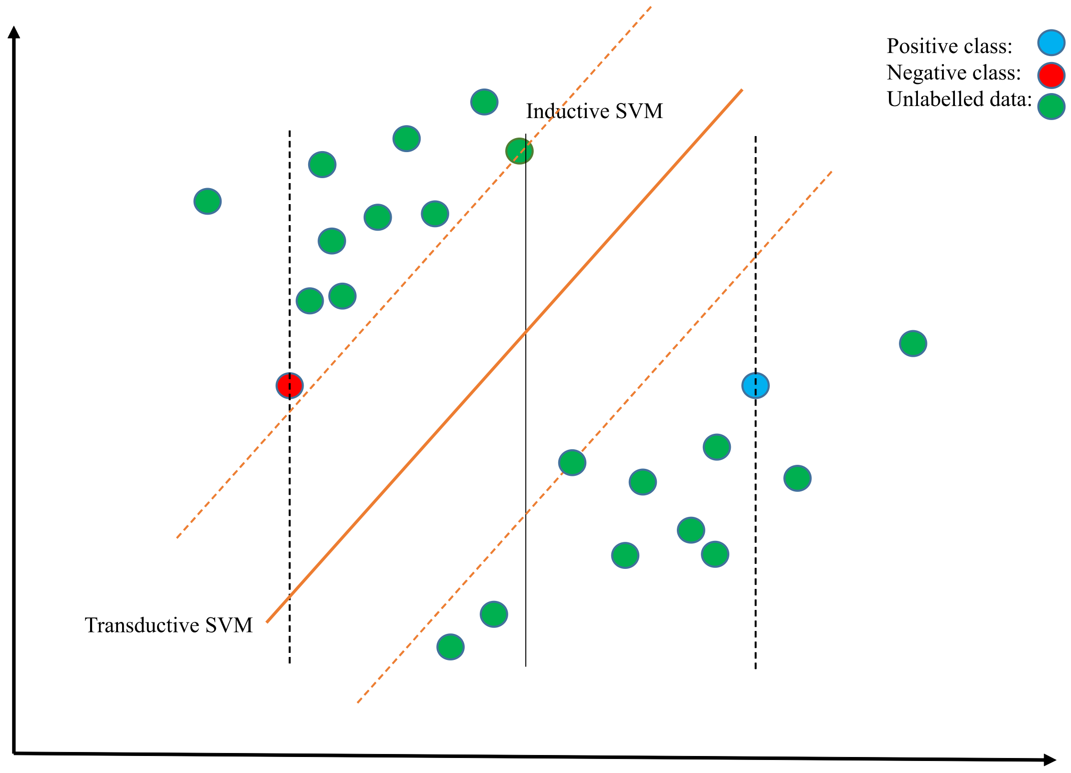


Figure 1.12 The comparison between TSVMs and inductive SVMs

This figure shows how the decision boundary moves from an inductive SVM to a TSVM. In SVM, the decision boundary is based on the posterior probability explicitly modelled by labelled data instances. In TSVM, associated with the posterior probability, prior assumptions are taken into account to aid the optimisation procedure.

Mathematically, we recall the standard SVM hinge loss on labelled instances is as follows:

$$c\left(x, y, f(x)\right) = max(1 - y(w^T \cdot x + b), 0)\, , f(x) = (w^T \cdot x + b),$$

where y denotes the raw labels. For the unlabelled instances, we use the prediction ($\hat{y} = \mathrm{sign}(f(x))$) on instances instead of the raw labels. Thus, the loss function named as hat loss (Zhu and Goldberg, 2009) can be stated as below:

$$c\left(x, \hat{y}, f(x)\right) = max(1 - \hat{y}\,(w^T \cdot x + b), 0) = max(1 - sign(w^T \cdot x + b)(w^T \cdot x + b), 0)$$
$$= max\,(1 - |w^T \cdot x + b|, 0)$$

Now, the function does not rely on the pre-known labels, but is fully determined by the prediction (f(x)). Combining the labelled and unlabelled instances, and taking the minimum regularisation risk on hinge loss and a regulariser for hat loss into account, the learning function can be formulated as below:

$$\min_{w,b} \sum_{i=1}^{l} max(1 - y_i(w^T x_i + b), 0) + \lambda_1 \parallel w \parallel^2 + \lambda_2 \sum_{j=l+1}^{l+u} max\big(1 - \left|w^T x_j + b\right|, 0\big)\,;$$

$$subject\ to\quad \frac{1}{u} \sum_{j=l+1}^{l+u} w^T x_j + b = \gamma\,,$$

where λ_2 is a parameter provided empirically to control the influence of involvement with unlabelled instances. γ is used to estimate the fraction of positive labelled instances among the unlabelled data set, which can be finely tuned via cross validation. In practice, the formulation above is often constrained with the similar label proportion between labelled and unlabelled instances; otherwise, the prediction tends to be imbalanced to classify all the unlabelled instances to only one side (Zhu and Goldberg, 2009, Collobert et al., 2006, Chapelle and Zien, 2005).

TSVMs were initially applied to solving transductive problems with a mix of labelled and unlabelled instances, relying on the methods of overall risk minimisation (ORM); but this method requires more labelled than unlabelled samples to obtain sufficient information for estimating classification (Vapnik, 2006, Vapnik, 1998). In the implementation proposed by Bennet and Demiriz, ten independent datasets were adopted to evaluate the effects of unlabelled instances during training. Three folds were observed to improve generalisation significantly by incorporating up to seventy unlabelled instances. No significant difference in terms of performance was detected in the other datasets (Bennett and Demiriz, 1999). This study successfully demonstrated the benefits of extra unlabelled data during training; but it left two unanswered questions: one is how to derive prior knowledge of the connection between labelled and unlabelled instances; the other is how to incorporate more unlabelled instances in the context of great computational complexity. S^3VM-IQP (Demiriz and Bennett, 2001) and SVM-Light (Joachims, 1999a) are a couple of similar implementations that seek to solve these problems by beginning with an optimised inductive classifier, and then applying penalty terms to optimise the switch from unlabelled instances to labelled. Iteratively, the margin is re-measured in response to new labelled instances, until all unlabelled instances are optimally labelled. The slight difference between the two implementations are the number of unlabelled instances switched in each iteration and the heuristic for avoiding sub-optima. Theoretically, a global minimum is the best solution; but, Demiriz and Bennett suggested that it is worth exploreing solutions based on local minima, which may be sufficient to improve the generalisation (Demiriz and Bennett, 2001). In a laterempirical study, (Joachims, 2006) showed that compared with inductive SVMs, performanceincreased in proportion to the the number of unlabelled instances in a text classification task. Critically, he showed high-dimensional learning problems typically fit TSVMs well; but not all problems are suitable for TSVMs, dependent on the structure of hypothesis space and the difficulty of optimisation.

We can see optimisation is a critical concern for TSVMs. The optimisation of the hinge loss in standard SVMs is convex, which provides a powerful regularisation. But, by adding a set of unlabelled instances, the optimisation in TSVMs becomes a non-convex and mixed-integer problem with multiple local maxima (Chapelle and Zien, 2005), which results in extreme high cost of computation. This is why early implementations of TSVMs struggled with large numbers of unlabelled instances (Collobert et al., 2006, Joachims, 1999b, Bennett and Demiriz, 1999). The major effort in TSVM development has therefore been to extend the established power standard SVMs to semi-supervised learning robustly and practicably. A variety of studies have focused on improving the efficiency of optimisation and alleviating the burden of computation. For example, semi-definite programming, a method based on convex relaxation to reformulate non-convex TSVMs to enable convex optimisation, can reduce computational complexity from exponential to polynomial level (Xu et al., 2008, Bie and Cristianini, 2004). Although it is still difficult to handle large data sets, it is competent at dealing with data of moderate size, which is a step forward compared with early TSVM implementations. ConCave Convex Procedure (CCCP) is another widely implemented method based on concave-convex programming. Collobert et al.'s study applied CCCP in TSVMs by taking advantages of scalability over convexity from objective non-convex function to deal with relatively large (up to 40,000 unlabelled examples) training data. Moreover, steady improvement is shown in the experiments compared with standard SVMs (Collobert et al., 2006). A more recent study that is also based on CCCP proposed a robust and fast TSVM (RTSVM) (Cevikalp and Franc, 2017). In this study, instead of the conventional hinge loss, ramp loss was applied on the labelled examples; and for the unlabelled examples, the non-convex objective function was decomposed into parts as either convex or concave, then CCCP was achieved by solving each part with stochastic gradient. The experiments showed the implementation was capable of handling linear problems with a large-scale training data though it was not markedly superior to the other implementations while dealing with nonlinear problems. Sindhwani and Keerthi proposed a set of TSVM classifiers dedicated to handle large-scale, sparse datasets with high-dimensionality (Sindhwani and Keerthi, 2006, Sindhwani and Keerthi, 2007), which is particularly suited to the context of research on human brain imaging, including focal brain injury. For example, a set of brain images with focal damage are may be spatially registered to a common space for the purpose of group comparison. At any specific resolution, each voxel is naturally treated as an independent variable, yielding a high-dimensional feature space within which anatomy—both normal and pathological—will vary widely. In their study, a multiple-switch TSVM based on Modified Infinite Newton algorithm (Keerthi and DeCoste, 2005) was implemented in the software package SVMLin to provide higher efficiency and scalability, mitigating the effects of searching for multiple local minima during optimisation. Remarkably, experiments on six different test datasets demonstrated the performance of multi-switch TSVM to be superior to both standard SVM and SVM^{light}. Moreover, across the test datasets, it was faster by at least two orders of magnitude compared with SVM^{light} (Sindhwani and Keerthi, 2006, Sindhwani and Keerthi, 2007).

Besides the initial Branch-and-Bound method of global optimisation implemented in the original TSVM (Joachims, 1999b) and the more recent optimisation methods introduced above, a few other implementations are also worth considering, such as difference convex programming (Wang and Shen, 2007), deterministic annealing (Chapelle et al., 2008), and entropy minimisation (Grandvalet and Bengio, 2005).

1.5.3 Multi-label algorithms

In the last two sections, our focus has been on methods of solving classic pattern recognition problems with either supervised or semi-supervised learning models. In such problems, the classes are mutually exclusive, exhibiting no overlap. In our research context, however, we must deal not only with single-label problems, but with multi-label problems too, where a multiplicity of non-exclusive class labels may be attached to each data instance. The development of multi-label classification algorithms has been powerfully driven by problems in the field of medical diagnosis and bioinformatics (Boutell et al., 2004, Tsoumakas and Katakis, 2007, Elisseeff and Weston, 2002, Vens et al., 2008, Sorower, 2010). For example, a patient with focal brain damage may suffer from multiple deficits at the same time, or may need to receive multiple treatments or interventions simultaneously. From a clinical perspective, it may be critically important to identify which one or more treatments are effective or co-effective in any one patient. In other words, it may not be sufficient to model the deficit or recovery of a patient by each material factor separately. In this section, we will introduce multi-label classification for the purposes of predicting multiple targets so as to reveal the relations and interactions between treatments.

We begin by clarifying three terms of art here: multi-label classification, multi-output learning and multi-class classification. Multi-label classification denotes a scenario where a set of target labels may be assigned to each instance simultaneously, for there is no mutual exclusion among the prediction features in each instance. Multi-label classification has strong relations with multi-output learning. In multi-output learning, the supervised learning problem has a response variable with more than one dimension, whether discrete or continuous. Multi-label problems denote a subset of a multi-output problems when the response variables are discrete. Note these two are sharply distinct from multi-class classification, also known as multinomial classification, where to each input instance one of three or more classes is exclusively assigned. Multi-label classification can be understood as a generalised extension of multi-class classification, without the constraint of a single label; instead, each instance may be assigned more than one labels from the three or more classes.

For a formal definition of multi-label classification, we can suppose a d-dimensional feature space, denoted as $\chi = \mathbb{R}^d$, where the set of labels can be denoted as $\mathcal{Y} = [y_1 \ldots y_n]$. The learning procedure for a multi-label problem is to generalise a function h: $\chi \rightarrow \mathcal{Y}$ from a training data D = (X_i, Y_i), where i = 1 … l, $X \in \chi$ and $Y \in \mathcal{Y}$, to predict any unseen given instance. The outputs of a multi-label function can be either binary, or more commonly real values. The real-valued classifier may involve an extra function for threshold calibration to determine a confident label boundary (Barutcuoglu et al., 2006, Zhang and Zhou, 2014).

With regards to the character of multi-label classification, a few key properties require consideration before the approach is adopted in the solution of any given problem. Label cardinality (LCard(D) = $\frac{1}{l}\sum_{i=1}^{l} Y_i$) yields the average number of labels for each instance in a given dataset, indicating the degree of multilabelledness in a dataset. Accordingly, label cardinality can be normalised as label density: LDen(D) = $\frac{1}{|\mathcal{Y}|}$ LCard(D). A further two properties are label diversity (LDiv), indicating the unique list of label sets in the training data, and proportion of label diversity (PLDiv = $\frac{1}{|D|}$LDiv), which is normalised by the number of instances to identify the contribution of each label set.

1.5.3.1 The key concerns in multi-label classification

Compared with classic binary classification tasks, the challenge of multi-label classification is dealing with the overwhelming complexity of possible label outputs. Label diversity obviously grows exponentially with increasing number of labels. For a training data with n distinct labels ($\mathcal{Y} = [y_1 \ldots y_n]$), the full permutation for all possible distinct label sets is 2^n. The effective exploitation of label correlations or dependencies is essential to alleviate this complexity (Zhang and Zhou, 2014). The approach to exploiting the correlations between labels can be generally categorised into first-order strategy, second-order strategy, and high-order strategy (Zhang and Zhang, 2010, Zhang and Zhou, 2014). The first-order strategy decomposes a multi-label classification task into a set of binary classification tasks regardless of the correlations among labels. This straightforward approach provides simplicity and efficiency, and is widely adopted in problem transformation methods (Boutell et al., 2004, Tsoumakas and Katakis, 2007, Zhang and Zhou, 2014). Ignoring label correlations, however, may result in a suboptimal solution. The second-order strategy involves a pairwise solution with regards to the relations and interactions between each pair of labels. The high-order strategy comprehensively considers the relation between each label and all others. This strategy models a full spectrum of label correlations, providing theoretically optimal fidelity, but relies heavily on computation and makes the model less flexible (Zhang and Zhang, 2010, Zhang and Zhou, 2014). All these strategies are widely applied across multi-label classification tasks, which will be detailed in the next section.

1.5.3.2 The overview of methods for multi-label classification

General speaking, multi-label classification methods can be mainly categorised as problem transformation methods and algorithm adaptation methods (Tsoumakas and Katakis, 2007, Sorower, 2010, Madjarov et al., 2012b, Zhang and Zhou, 2014). The former decomposes multi-label classifications problem into multiple simpler single-label classification problems. In other words, its core philosophy is to tailor the data to the algorithm. By contrast, the latter extend single label algorithms to make them fit multi-label learning problems, tailoring the algorithm to the data instead.

1.5.3.2.1 The problem transformation methods

As mentioned above, problem transformation methods change the form of multi-label data so that it can be learnt by well-established, single-label algorithms. Three sub-categories of transformation are recognised: transforming to binary classification problems, transforming to label ranking problems, and transforming to multi-class classification problems.

Early strategies in the literature are very simple. For example, one method is to choose one from the multiple labels for each multi-label instance, and discard the others either manually or randomly. Another way is to discard the multi-label instances from the data and merely retain the single-label ones (Tsoumakas and Katakis, 2007, Boutell et al., 2004). Both methods readily transform a multi-label classification problem to a classic single-label learning problem; but at the cost of discarding a great deal of information. In the following sections, concerned with the order strategies proposed by Zhang and Zhou (Zhang and Zhang, 2010, Zhang and Zhou, 2014), we introduce a few problem transformation methods in more detail to show how these methods with different strategies and generalizability fit a broad spectrum multi-label classification tasks.

1.5.3.2.1.1 Transforming to binary classification

Binary relevance (Boutell et al., 2004, Zhang and Zhou, 2005, Tsoumakas and Katakis, 2007, Zhang and Zhou, 2014) is a first-order approach that decomposes each label (l) from a multiple set of labels into a pair of binary labels ([l, -l]). In this way, the original multi-label learning problem is transformed to be a set of independent two-class classification learning problems. During the training procedure, each instance is trained by the set of binary models in turn and treated as relevant if the output is positive, otherwise irrelevant. This process is also named cross-training (Boutell et al., 2004). For any unseen instance, predictions are combined with the status of relevance from the set of binary classifiers. If the outputs from the set of classifier are all irrelevant, multi-label testing criteria can be applied to decide the outputs. For example, P-Criterion treats this kind of situation as 'unknown'. T-Criterion selects the least negative outputs as predictions (Boutell et al., 2004). As a common first-order approach, binary relevance is straightforward to implement in parallel and build up a set of classifiers independently. The disadvantage is that the correlations between labels are not considered, and the decomposition may lead to marked imbalance for some labels. For this reason, the number of labels in the training data and its label density are worth checking before adopting this approach.

In an attempt to deal with the ignorance of label correlations in binary relevance-based methods, the classifier chains-based method (Read et al., 2009, Read et al., 2011) employs a high-order strategy to transform a multi-label classification task into a chain of binary classification tasks. To be similar to binary relevance methods, a multi-label classification problem is decomposed into a set of binary classification problems, in which each binary classification tasks corresponds with one label. In the classifier chains model, however, a binary label relevance predicted by a classifier will be appended as an extra dimension in the feature space of all input instances, contributing to the prediction for the next classifier. In other words, among the set of decomposed binary classification models, each classifier is learnt in a feature space augmented by adding the prediction of all the preceding classifiers, which forms a chain of classifiers that take account for the label correlations as a whole. To avoid the effect of the chain's order on prediction accuracy, iterative ensembles of classifier chains have been proposed (Read et al., 2009, Read et al., 2011). At each iteration, a set of randomly sampled training instances with a chain ordering allocated by a random permutation function are prepared for learning a classifier. The final predictions create a confidence vector by averaging across the iterations of classifiers, thereafter applying a strategic threshold function to dichotomise the relevance and irrelevance. An alternative approach is presented by Bayes-optimal probabilistic classifier chains (Dembczynski et al., 2010) where the classifier chains are formulated within probability theory. Here instead of binary relevance within ensembles of classifier chains, the feature space is appended by conditional probability. Compared with binary relevance methods, classifier chains-based methods preserve label correlations though they are hard to implement in parallel (Zhang et al., 2014). This increases computational complexity linearly rather than exponentially, which is acceptable.

1.5.3.2.1.2 Transforming to label ranking

Calibrated label ranking (Fürnkranz et al., 2008) applies a second-order strategy to address a multi-label classification problem with the techniques based on pairwise comparison, transforming the original problem into a label ranking problem.

For a set of labels ($\mathcal{Y} = [\mathcal{y}_1, \ldots \mathcal{y}_n]$) in a multi-label training data, the permutation of pairwise comparison can generate n · (n-1) / 2 binary classifiers in total. For each pairwise comparison, the training instances have opposite relevance on the pair of labels (e.g. [$\mathcal{y}_j$, $\mathcal{y}_k$]) included to generate a binary classifier. Specifically, a pairwise comparison can be stated as follows (Zhang and Zhou, 2014):

$$\mathrm{D_{jk=}} \left(\mathrm{x_i},\ \psi\left(\mathrm{Y_i, y_j, y_k}\right)\right), 1 \leq \mathrm{i}\ \leq \mathrm{m},$$

$$\text{where}\ \ \psi\left(\mathrm{Y_i, y_j, y_k}\right) = \begin{cases} +1, & \text{if } \phi\left(\mathrm{Y_i, y_j}\right) = +1 \text{ and } \phi(\mathrm{Y_i, y_k}) = -1; \\ -1, & \text{if } \phi\left(\mathrm{Y_i, y_j}\right) = -1 \text{ and } \phi(\mathrm{Y_i, y_k}) = +1. \end{cases}$$

Then, a learning algorithm can be applied on the binary training subset (D_{jk}) to yield a classifier ($f_{jk}(x)$). The total n · (n-1) / 2 classifiers generated by the procedure above iteratively can yield a vector of votes for any unseen instance. Subsequently, by incorporating a virtual label ($\mathcal{y}_V$) (Zhang and Zhou, 2014), the ranked labels included in the voting vector can be further dichotomised into relevance and irrelevance by a threshold function.

The pairwise comparison adopted in calibrated label ranking is helpful to reduce the effects on class imbalance; but the computational complexity is in quadratic scale though it can be alleviated in some extent by pruning approaches (Mencía et al., 2010, Madjarov et al., 2010, Madjarov et al., 2012a).

1.5.3.2.1.3 Transforming to multi-class classification

The first method of transforming a multi-label classification problem to a multi-class classification problem is label powerset (Boutell et al., 2004, Tsoumakas and Katakis, 2007), in which each subset of multiple labels ('labelset') is converted to a single label. Consider, for example, a patient suffering from acute ischemic stroke complicated by hemorrhage. Ischemic acute stroke and hemorrhage are here two labels which can be combined to generate a single-label – 'ischemic acute stroke ^ hemorrhage'. This method transforms a multi-label learning problem to a multi-class problem with no information loss; but it necessarily increases the number of classes significantly. Furthermore, the sparsity and imbalance of some classes may make prediction very difficult. Additionally, this method only involves the 'labelsets' observed in the training data but not the full permutation of possibilities, some of which may be found in any unseen test data (Tsoumakas et al., 2011).

To cope with the limitations of label powerset mentioned above, random k-labelsets (Tsoumakas and Vlahavas, 2007) have been proposed. The concept here is to split the original set of labels into multiple subsets of labels chosen randomly. For each subset of labels, the method of label powerset is applied to achieve the problem transformation, generating a multi-class classifier. Here k is the parameter that determines the size of subset of labels; 'disjoint' and 'overlapping' are two strategies to break the labelset into a number of subsets (in size of k) of labels. Both strategies have demonstrated improvement in empirical testing. For data with a large number of labels, the strategy of overlapping is more effective in correcting uncorrelated errors (Tsoumakas et al., 2011). In any unseen test instance, the outputs predicted by the set of classifiers determine that a particular label is treated as relevant if the number of votes in relation to this label is more than half of total number of related classes.

Random k-labelsets is an approach that takes into account the label corrections, flexibly controlled by the value of k within a high-order strategy.

1.5.3.2.2 Algorithm adaptation methods

In contrast with algorithm independent problem transformation methods, algorithm adaptation methods rely on a specific learning algorithm, designed to fit multi-label data directly. There is a wide range of proposed approaches, employing decision trees (Vens et al., 2008, Clare and King, 2001), k-nearest neighbors (Zhang and Zhou, 2005, Zhang and Zhou, 2007), support vector machines (Xu, 2011, Elisseeff and Weston, 2002), boosting algorithms (Wu et al., 2010, Schapire and Singer, 2000), entropy calculation (Ghamrawi and McCallum, 2005), Bayesian (Zhang et al., 2009, McCallum, 1999), and neural network (Zhang, 2009, Wei et al., 2014).

Our interest here is in connecting clinical outcome parameters with clinical imaging data, which—as we have argued—will tend to be a high-dimensional problem. In this section, we will introduce two algorithm adaptation methods based on the extensions of decision trees – Multi-label Decision Tree (Clare and King, 2001) – and support vector machines – Ranking Support Vector Machine (Elisseeff and Weston, 2002). Described in terms of order strategies (Zhang and Zhang, 2010), decision trees adapted for multi-label learning problems constitute a first-order approach, which is straightforward and relatively low in computational complexity in the context of a high-dimensional feature space. The SVM is a second-order approach, optimised for non-linear problems.

1.5.3.2.2.1 Multi-label decision tree

To deal with a multi-label classification problem within the classic decision tree algorithm (Quinlan, 2014), for a given training data (D (x_i, Y_i)), the measurement of uncertainty by entropy can be extended as follows (Clare and King, 2001):

$$\mathrm{entropy(D)} = -\sum_{i=1}^{N}(p_i \cdot \log p_i + q_i \cdot \log q_i)\,,$$

where, N is the number of classes in D (x_i, Y_i). p_i is the fraction of instances in D belonging to class i; $q_i = 1 - p_i$ is the fraction being of not class i ($1 \leqslant i \leqslant N$).

Then, by introducing a splitting value (ϑ), the information gain can be achieved in the formulation below (Zhang and Zhou, 2014):

$$\text{information gain}(\mathrm{D}, \mathrm{A}, \vartheta) = \text{entropy}(\mathrm{D}) - \frac{|\mathrm{D}^{+}|}{|\mathrm{D}|} \cdot \text{entropy}(\mathrm{D}^{+}) - \frac{|\mathrm{D}^{-}|}{|\mathrm{D}|} \cdot \text{entropy}(\mathrm{D}^{-}),$$

where A is an attribute in the feature space. D^+ is subject to $x_{iA} \geqslant \vartheta$; D^- is subject to $x_{iA} \leqslant \vartheta$. From the initial root node (D), the attribute and its according splitting value are identified by multi-label decision tree to maximise the information gain so as to generate leaves. The generated two leaves (D^+ and D^-) are treated as root nodes respectively, repeating the process described above. Multi-label decision tree training is a recursive procedure, converged by meeting an empirically specified stopping criterion.

To reduce computational complexity, from the formulation of entropy above we can see multi-label decision trees present the problem as one label versus the rest. In other words, the approach assumes the set of labels are independent of each other. Hence, this is a first-order strategy, which does not take into account label correlations. Potential improvements proposed by other studies mainly focus on two aspects: one is pruning strategy (Clare and King, 2001, Zhang and Zhou, 2014); the other is to employ ensemble methods, shown in Kocev et, al.'s study to provide better predictive performance in both single-objective and multi-objective decision trees (Kocev et al., 2007).

1.5.3.2.2.2 Multi-label support vector machines

To address multi-label classification problems with a kernel method was firstly proposed by Elisseeff and Weston (Elisseeff and Weston, 2002). Here an SVM based ranking system learning model was adapted to retain a large margin with minimum cost. This method, also named a ranking support vector machine (Rank-SVM), employs a second-order strategy (Zhang and Zhang, 2010) to take into account pairwise label correlations. It naturally has the properties of SVMs, a loss function – Ranking Loss (Schapire and Singer, 2000), defined as the fraction of incorrect ordering between pairwise comparisons on average – is used to optimise a set of linear classifiers, and nonlinear problems are addressed via kernel tricks (Elisseeff and Weston, 2002). Specifically, for a multi-label training set (D = (x_i, Y_i), where 1 ⩽ i ⩽ m; $Y_i \in \mathcal{Y}$, where $\mathcal{Y}$ = ($y_1 \ldots y_q$)), each individual label is learnt to generate a binary classifier. As described in the preceding section on SVMs, a set of independent classifiers (Elisseeff and Weston, 2002) for the multi-label training set D can be stated as a sign function in the form of an inner product:

$$f_j(x) = \text{sign}(< w_1, x > +b_1), \text{where } 1 \ \leq j \ \leq q.$$

In light of the need for pairwise comparison, the Ranking Loss (Schapire and Singer, 2000) is stated as below:

$$RL(f, x, Y) = \frac{1}{|Y||\overline{Y}|} \left|(y_k, y_l) \ \in Y \times \overline{Y}, \text{subject to } f_k(x) \geq \ f_l(x)\right|.$$

To solve a multi-label classification problem with the kernel method ranking system, it is ideal to retain a large margin while minimizing the Ranking Loss. For each instance in the training set: (x_i, Y_i), between the label pair in relation to relevant-irrelevant, the margin is the shortest signed Euclidean (L^2) distance to the hyperplanes, which can be expressed as follows:

$$\min_{(x_i, Y_i) \in D} \ \min_{(y_k, y_l) \in Y_i \times \overline{Y}_i} \frac{< w_k - w_l, x_i > +b_k - b_l}{\parallel w_k - w_l \parallel} \ ,$$

where y_k is within the labels sets (Y_i) of x_i; whereas y_l belongs to the complementary label sets ($\overline{Y_i}$). In particular, a positive margin will be reasonably returned if the relevant-irrelevant label pair is well ranked. Thus, based on the formulation above, to obtain a maximum margin we proceed as follows:

$$\max_{f_i(x)} \ \min_{(x_i, Y_i) \in D} \ \min_{(y_k, y_l) \in Y_i \times \overline{Y}_i} \frac{1}{\parallel w_k - w_l \parallel^2}, \text{ subject to: } < w_k - w_l, x_i > +b_k - b_l \geq 1 \ ,$$

which is equivalent to:

$$\min_{f_i(x)} \ \max_{1 \leq k \leq l \leq q} ||w_k - w_l||^2 \, , subject\ to\colon \ < w_k - w_l, x_i > +b_k - b_l \geq 1, (y_k, y_l) \in Y_i \times \overline{Yi}.$$

The problem can be further simplified to use the sum operator to approximate the max operator as follows (Elisseeff and Weston, 2002):

$$\min_{f_i(x)} \sum_{k=1}^{q} ||w_l||^2 \text{, } subject\ to\text{: } < w_k - w_l, x_i > +b_k - b_l \geq 1, (y_k, y_l) \in Y_i \times \overline{Y_i}.$$

Lastly, instead of the hinge loss adopted in the binary SVMs, the adapted Rank-SVM incorporates the Ranking loss as the slack variable to complete the equation:

$$\min_{\mathrm{f_i(x),\ \xi\ _{ikl}}} \sum_{\mathrm{k=1}}^{\mathrm{q}} ||\mathrm{w_k}||^2 + \mathrm{C} \sum_{\mathrm{i=1}}^{\mathrm{m}} \frac{1}{|\mathrm{Y_i}||\overline{\mathrm{Y_i}}|} \sum_{(\mathrm{y_k,y_l}) \in \mathrm{Y_i} \times \overline{\mathrm{Y_i}}} \xi\ _{\mathrm{ikl}}\ ,$$

$$\text{subject to: } < \mathrm{w_k} - \mathrm{w_l}, \mathrm{x_i} > +\mathrm{b_k} - \mathrm{b_l} \geq 1 - \ \xi\ _{\mathrm{ikl,}}\ \ \xi\ _{\mathrm{ikl}} \geq 0, 1 \leq \mathrm{i} \leq \mathrm{m}.$$

Rank-SVM is based on well-established and robust SVM machinery. A major advantage of Rank-SVMs is proper accounting for pairwise label correlations, a characteristic critical for any SVM-based algorithm here. This is because SVM is fundamentally a discriminative method that assumes the set of classes involved in learning to be disjoint, whereas the classes in multi-label data are definitionally correlated and overlapping. Another prominent advantage of Rank-SVM is the ability to handle non-linear classification problems with the aid of kernel methods. Nonetheless, in common with any ranking system, Rank-SVM can order each label pair to be relevant or irrelevant, but it does not output the label sets directly (Elisseeff and Weston, 2002, Tsoumakas and Katakis, 2007, Zhang and Zhou, 2014). To determine the relevance of a label, a threshold strategy needs to be applied. In contrast with conventional and isolated threshold functions, such as least squares, Calibrated Rank-SVM (Jiang et al., 2008) introduced a virtual label providing a natural zero point during the stage of label ranking, thereby bridging label ordering and label ranking to optimally identify the interactive effects of threshold.

Besides Rank-SVM, a range of SVM variants have been proposed during past few years, enriching the arsenal of tools available for tackling multi-label classification with kernel methods. For example, Wang et al. presented a Simplified Constraints Rank-SVM (SCRank-SVM) (Wang et al., 2014) which removes the bias term b to alleviate optimisation constraints, thereby improving the navigability of the solution space. Instead of a ranking loss, Tsochantaridis et al. adopted hamming loss to create a generalised multi-class classification framework that can cope with complex outputs (Tsochantaridis et al., 2004). The technique of multiple kernel learning is helpful to deal with the difficulty of kernel selection during multi-label classification (Ji et al., 2009, Kloft et al., 2009).

1.5.3.3 Evaluation metrics

Since each instance in a multi-label learning problem is definitionally associated with multiple labels simultaneously, the conventional evaluation metrics used in the evaluation of classic binary or single-label classification problems, such as, accuracy, precision, recall, F-score, receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), and so on, must be extended.

Generally, evaluation measures can be applied either on a per example basis or a per label basis. Example-based metrics evaluate the difference between real and predicted label sets on each individual test example; then calculate the average over the whole test set (Godbole and Sarawagi, 2004, Tsoumakas and Vlahavas, 2007, Zhang and Zhou, 2014). By contrast, label-based metrics evaluate the performance of all test examples on each individual label; then calculate the micro- and macro-average over all labels (Sorower, 2010, Tsoumakas and Vlahavas, 2007, Zhang and Zhou, 2014).

1.5.3.3.1 Example-based metrics

For a multi-label test data ($S = (x_i, Y_i)$), where $1 \leqslant i \leqslant n$; $Y_i \in \mathcal{Y}$, where $\mathcal{Y} = (y_1 \ldots y_q)$), based on the learnt classifier ($f(\cdot)$), the major example-based metrics are as follows:

$$\mathrm{HammingLoss}\ (f, S) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{q}\ |f(x_i)\Delta Y_i|,$$

where Δ stands for the symmetric difference between real and predicted label sets, which is equivalent to the operation of exclusive disjunction of in Boolean logic. In other words, hamming loss is referred as the fraction of incorrect labelling (misclassification) over the whole label set.

$$\mathrm{SubsetAccuracy} = \frac{1}{n}\sum_{i=1}^{n} I\,(f(x_i) = Y_i),$$

where I equals to 1 if the label is correctly predicted, or I equals to 0. Subset accuracy is a strict metric that indicates the percentage of exactly correct match with the real label set.

Both hamming loss score and subset accuracy measure the similarity between the predicted and real label set. In addition, the standard information retrieval measures are used for evaluation as well (Godbole and Sarawagi, 2004):

$$\mathrm{Accuracy}(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{|f(x_i) \cap Y_i|}{|f(x_i) \cup Y_i|}\ ;$$

$$\mathrm{Precision}(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{|f(x_i) \cap Y_i|}{|f(x_i)|}\ ;$$

$$\mathrm{Recall}(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{|f(x_i) \cap Y_i|}{|Y_i|}\ ;$$

$$F_\beta = \frac{\left(1 + \beta^2\right)\cdot \mathrm{Precision}(f)\cdot \mathrm{Recall}(f)}{\beta^2\cdot \mathrm{Precision}(f) + \mathrm{Recall}(f)}.$$

β is a balancing factor. When β=1, F_1 refers to the harmonic mean of precision and recall.

As aforementioned, when a ranking based system is selected as a multi-label classification approach, the learning function returns real rather than logical values. For this specific situation, one-error is a straightforward metric to indicate training error; coverage and average precision are the measures based on information retrieval to evaluation the performance of label rankings (Schapire and Singer, 2000):

$$\mathrm{one-error} = \frac{1}{n}\sum_{i=1}^{n}\left[\left[f(x_i) \notin Y_i\right]\right];$$

$$\mathrm{coverage}(f) = \frac{1}{n}\sum_{i=1}^{n}\max_{y\in Y_i}\mathrm{rank}_f(x_i, y) - 1;$$

$$\mathrm{averagePrecision}(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|Y_i|}\sum_{y\in Y_i}\frac{\left|y' \;\middle|\; \mathrm{rank}_f\left(x_i, y'\right) \leq \mathrm{rank}_f(x_i, y), y' \in Y_i\right|}{\mathrm{rank}_f(x_i, y)}.$$

1.5.3.3.2 Label-based metrics

Label-based metrics decompose the multi-label classification evaluation process into a set of individual label evaluations. Similarly to the measures for the classic single-label classifications, four basic quantities: true positive (TP), false negative (FN), true negative (TN) and false positive (FP) can be used to characterise performance. Specifically, for a multi-label test data (S = (x_i, Y_i), where 1 ⩽ i ⩽ n; Y_i ∈ $\mathcal{Y}$, where $\mathcal{Y}$ = (y_1 … y_q)), based on the learnt classifier (f(·)),the evaluation metrics on a specific label y can be expressed as follows:

$$\mathrm{TP}_y = |y \in Y_i \wedge y \in f(x_i), 1 \leq i \leq n|;$$

$$\mathrm{FP}_y = |y \notin Y_i \wedge y \in f(x_i), 1 \leq i \leq n|;$$

$$\mathrm{TN}_y = |y \notin Y_i \wedge y \notin f(x_i), 1 \leq i \leq n|;$$

$$\mathrm{FN}_y = |y \in Y_i \wedge y \notin f(x_i), 1 \leq i \leq n|.$$

Based on the evaluation on each binary label M (TP_y, FP_y, TN_y, FN_y), all labels can be evaluated with the macro- and micro-averaged version:

$$\mathrm{Macro-averaging} = \frac{1}{q}\sum_{j=1}^{q} M\left(\mathrm{TP}_y, \mathrm{FP}_y, \mathrm{TN}_y, \mathrm{FN}_y\right);$$

$$\mathrm{Micro} - \mathrm{averaging} = \mathrm{M}\left(\sum_{j=1}^{q} \mathrm{TP_y}, \sum_{j=1}^{q} \mathrm{FP_y}, \sum_{j=1}^{q} \mathrm{TN_y}, \sum_{j=1}^{q} \mathrm{FN_y}\right).$$

It can be seen that the accuracy of macro-averaging and micro-averaging is the same; while precision and recall are different. In addition, macro- / micro-accuracy and hamming loss is 1. If a ranking based system selected, label-based measures can use AUC instead (Zhang and Zhou, 2014):

$$\mathrm{Macro} - \mathrm{averaged} - \mathrm{AUC} = \frac{1}{q} \sum_{j=1}^{q} \frac{\left|f(x,y_j) \geq f\left(x' \ ,y_j\right),\left(x,x' \ \right) \in Z_j \times \overline{Z_j}\right|}{|Z_j||\overline{Z_j}|},$$

where Z_j is the set including all the test examples associated with label y_j; $\overline{Z_j}$ is the set excluding the examples with label y_j.

$$\mathrm{Micro} - \mathrm{averaged} - \mathrm{AUC} = \frac{|f(x,y) \geq f\left(x' \ \geq y' \ \right),(x,y) \in S^{+},(x' \ ,y' \ ) \in S^{-}|}{|S^{+}||S^{-}|},$$

where S+ and S- are referred as a pair of sets: instance-relevant labels and instance-irrelevant labels.

1.5.4 Accuracy estimation and assessment of model generalizability

In section 1.4.3, we introduced a set of machine learning based inferential models dedicated to probing the relationships between lesion architecture and brain functional anatomy. A crucial concern is how to evaluate these models, and select the optimal amongst all the candidates.

Applied to human brain mapping, the ultimate goal of a trained classifier is to model the true relationship across the entire population. In this way, the full diversity of complex patterns integrating the spatial correlations between brain structure and function can be identified. Nonetheless, we are compelled by practicability to rely on samples of the population from which we must generalise to the whole. When using a sub set of the population to build an inferential model, there are two major inter-related concerns: one is to prepare sufficient samples for approximating the true population; the other is to apply appropriate techniques for estimating how well the trained classifiers model the true population. Assuming we have succeeded in maximizing the volume of available data, our focus is to optimise its use, and the evaluation of the models it is used to build.

When a trained model is shown to predict the outcome of interest from the input feature space without any errors, the question arises whether it has successfully modelled the true relationships across the population or merely memorised the training data, i.e. overfitted. As the true structure of the underlying anatomical-functional relationships is unknown, we cannot assess spatial inference objectively. Instead, we need techniques that enable the sample data to be chosen to approximate the true population, and allow us to detect errors of generalisation so as to facilitate the process of model optimisation. In practice, a learning problem, in particular one that involves complex patterns such as the lesion architecture of the human brain, usually suffers from the difficulty of considerably more input features than the number of samples, a scenario inherently prone to incur overfitting. This is because the trained models may not be able to capture sufficient information to absorb the diversity across the population. In this situation, the trained models can be parameterised to perform well in the given training dataset; but deliver poor performance on a separate test dataset. Consequently, the problem of overfitting limits the generalizability of a learning model and reduces its ultimate utility.

Underfitting, fitting, and overfitting scenarios are visualised simultaneously in figure 1.13 to illustrate the effects implications for generalizability. The straight line in orange colour depicts a underfitting classification function that separates the red and blue markers but has them substantially mixed on either side of the line. Poor performance on a given training set here does not generalise to the true population, for a better fit can be found. When this model is tested with an independent set of data drawn from the true population, the performance will be most likely to be similarly poor unless the training set shown in the figure is a completely unrepresentative. By contrast, the curvy green line presents a classifier in the overfitting scenario, which weaves along the red or blue dots to separate the groups with exquisite precision. It is reasonable to suspect the generalizability of a classifier that is tightly customised for a give training set. When tested against a new set of data, the performance will not to be as good as the performance gained from the training set unless the sub sample of true population given for training is completely presentative to the entire population. Hence, overfitting reduces generalizability.

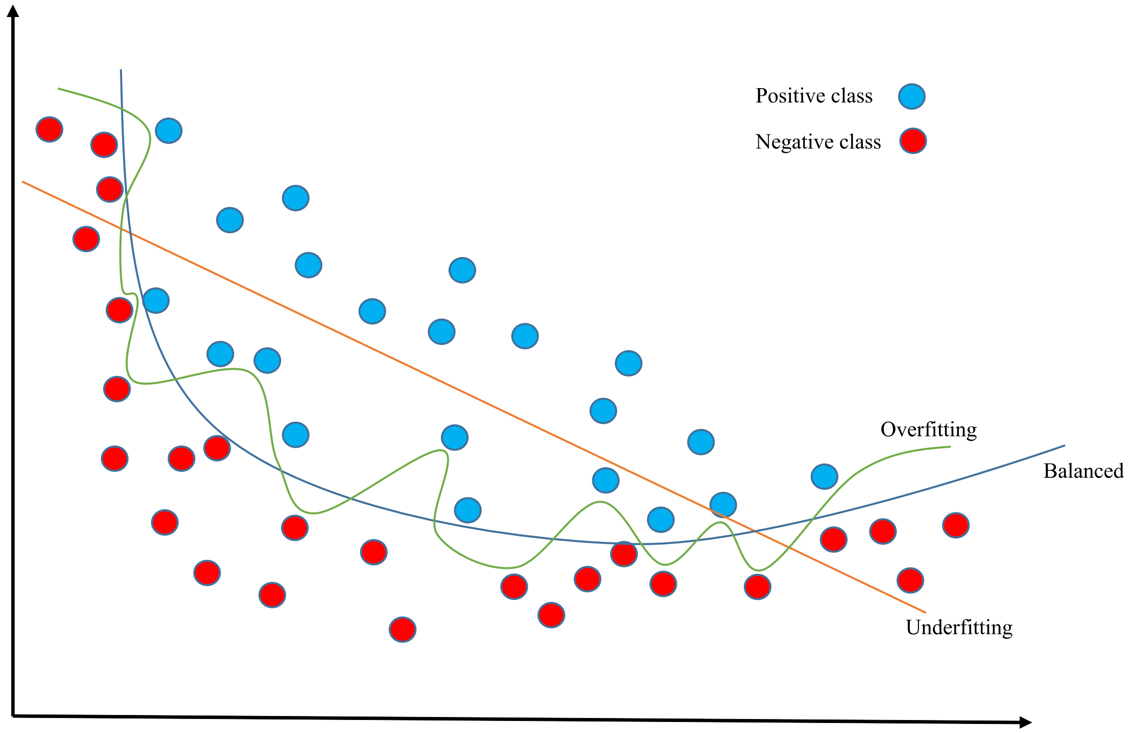


Figure 1.13 underfitting, balanced and overfitting

Two sets of markers in blue and red representing a positive and negative class, respectively, are drawn in the figure to illustrate a classifier overfitting (green), underfitting (orange) and correctly fitting (blue). An overfitted boundary fits each datum in either group very tightly. By contrast, an underfitted boundary is less good at separating the classes. As a result, there are quite a few instances misclassified in either membership. Between the two extremes, a balanced solution presents a reasonable curve with optimised generalizability.

A successfully generalizable solution for a learning problem should combine an optimised learning model with methods to avoid overfitting. On one hand, we need accuracy estimation so as to approach the optimal model; on the other, we need a means of evaluating the generalizability of a selected model. In the following section, we will introduce two most common methods: cross-validation and bootstrap; and review how these methods can be applied for accuracy estimation, model selection and the assessment of generalizability.

1.5.4.1 Cross-validation

Cross-validation is a process for deriving estimates of how the prediction of a trained model can be generalised to an independent dataset to which the model is naive. In essence, in cross-validation the inferential model is leant and validated through a series of training and test sets derived from different divisions of the entire dataset in a rotational manner. In this rotation independent procedure, each model is trained and tested explicitly by a completely separate training and test set to minimise overfitting. Moreover, each sample is rotated as test data at least once to avoid selection bias. A range of strategic variations provided by cross-validation will be introduced in details as follows, which are designated to be sensitive to an inferential model distorted by overfitting or selection bias; meantime, to use the available data in a most desirable way.

1.5.4.1.1 Exhaustive cross-validation

For a given training set of N samples, the method of exhaustive cross-validation trains and tests on the full set of possible permutations of the data, divided into training and validation sets. A typical method is leave-p-out cross-validation (Kohavi, 1995), which fixes the number of validation set as p, and uses the remaining as training set. It is an iterative procedure, ensuring the training and validation are rotated in all possible combinations. The number of iterations is the binomial coefficient: $\mathrm{C}_{\mathrm{p}}^{\mathrm{N}}$. Computational complexity makes the process infeasible where p is greater than one and N is moderately large (e.g. C_{10}^{500}).

For this reason, more commonly used method of exhaustive cross-validation is to fix p at 1, which is referred as leave-one-out cross-validation (Kohavi, 1995). Compared with leave-p-out cross-validation, leave-one-out cross-validation repeats the process N times to ensure each sample involved in the training set is used for model validation once. The process is similar to the resampling technique of jackknife (Miller, 1974). To be clear, the jackknife is a linear approximation of the bootstrap with the aim of estimating the variance and bias of a statistic, which excludes a sample each time and computes the statistic from the remaining set;

whereas, leave-one-out computes the statistic on the one left-out sample each time to estimate accuracy.

1.5.4.1.2 Non-exhaustive cross-validation

The methods of non-exhaustive cross-validation are approximations to leave-p-out cross-validation, evaluating only a representative sample of possible train/test splits.

A simple method, called the holdout method, is to randomly split the entire dataset into two parts for training and test, respectively. Typically, a large part (e.g. 70% or 80% of the whole set) is used for training a model, and unseen remainder is used for evaluating performance. It is a single run method, with no necessity for either repetition or rotation, so it is really a type of validation rather than cross-validation (Arlot and Celisse, 2010, Kohavi, 1995).

The most common variation of this is k-fold cross-validation (Kohavi, 1995, Zhang, 1993) where the data is randomly partitioned into k subsets of equal size, and each subset is used for validation group while the other k-1 subsets are used for training. The number of iterations through the whole process is k, meaning each subset is used in evaluation once. Such k-fold cross-validation can be treated as an approximation of leave-p-out cross-validation, where $\mathrm{p} = \frac{\mathrm{N}}{\mathrm{k}}$ are satisfied. For example, if k equals N, k-fold cross-validation and leave-one-out cross-validation will collapse to the same method. Compared with the holdout method, k-fold cross-validation involves each sample into either training or testing at least once and still retains low computational complexity. A wide range of studies recommend 10-fold cross-validation as an optimally balanced choice in terms of stratification and efficiency (Braga-Neto and Dougherty, 2004, Kohavi, 1995, Kohavi, 1996, Zhang, 1993), though k is definitely not a fixed parameter and can be flexibly adjusted to suit different scenarios. In this thesis, we will largely apply binary classification model to reveal the correlations between spatial brain lesion architecture and functions, this stratified method selects each fold for validation with approximately equal mean response value, which ensure the proportion of either sub-group in each fold to be roughly the same. At the end, a single estimate can be averaged from the k validations.

1.5.4.1.3 Cross-validation methods and model selection

A general conclusion about the optimal cross-validation method for model selection is difficult to draw owing to its dependence on the detailed properties of the underlying learning framework. Nonetheless, three major aspects – bias, variance and computational complexity –can be used as basic criteria (Arlot and Celisse, 2010).

Cross-validation methods essentially estimate prediction errors. Bias and variance are two types of fundamental errors usually presented as a trade-off to reflect the accuracy and consistency of an inferential model. As aforementioned, a large-scale dataset is crucial for our proposed research. The closer to the true population, the better performance will be delivered by an inferential model. When the sample size is fixed, cross-validation estimates the performance with a pessimistic bias as it estimates a trained model with a subset of the original training set. To some extent, cross-validation tends to over-estimate the error compared with the error estimated by the original sample set. Leave-one-out cross-validation will be less biased, or approximately unbiased, as there is merely one sample different from the original training set in each repetition through the entire procedure. Hence, leave-one-out cross-validation is preferable to the case of the continuous error functions, in other words, the functions with a large signal-to-noise ratio to avoid sub optima asymptotically (Breiman, 1996, Arlot and Celisse, 2010, Mah, 2014b).

On the other aspect, an optimal model should ideally have the lowest variance possible. During a cross-validation procedure, the variance usually decreases when the number of samples for test increases. Although leave-one-out cross-validation is approximately unbiased, it tends to have a high variance. Compared with leave-p-out or k-fold cross-validation, we can imagine leave-one-out cross-validation will be more sensitive to any possible noise or artifacts included in the original sample set, so the performance may vary substantially with each fold. Yet, both leave-p-out and k-fold cross-validation could suffer from the variance as well if the number of original training samples is small. For example, if 10-fold cross-validation is applied for estimating models trained by a small sample set, each split removed from the original set will affect the training performance significantly, resulting in high variance. Hence, it is highly framework dependent to decide between exhaustive and non-exhaustive cross-validation methods (Arlot and Celisse, 2010), though, for a given proper sample set, a general conclusion is leave-p-out or k-fold cross-validation tends to be more balanced in regards with bias and variance. In addition, where the error function is discontinuous, in other words, the goal of the error function is to identify the number of misclassifications, such as a binary classification problem, leave-p-out or k-fold cross-validation is more effective at reducing variance (Breiman, 1996, Arlot and Celisse, 2010, Mah, 2014b). In contrast with most cross-validation methods employing multiple runs, the holdout method performs the validation with a single run, which is preferable to the residual method but needs to be adopted with caution. Evaluation with the holdout method is strongly dependent on how the division of training and test set are made and may end up with high variance.

The fundamental concerns about bias and variance aside, computational complexity is a key factor to be considered. For the binary classification problems that are the focus of this thesis, the preferable methods are leave-p-out and k-fold cross-validation whose computational complexity are C_{p}^{N} and k, respectively. For a relatively large training set, it is apparent leave-p-out cross-validation is infeasible in term of computational complexity though its estimations could be averaged by all the possible splits of training and test division so as to minimise the bias and variance. As a compromise, the computational complexity of k-fold cross-validation is feasible and reasonable. Besides choosing a k value empirically, some studies also suggested to choose a small k (e.g. 2 or 5), but use multiple ways to do the k splits, which was helpful to optimise the bias and variance of k-fold cross-validation. In this way, the computational complexity could increase linearly as $n \cdot k$, which is still feasible (Mah, 2014b, Breiman and Spector, 1992).

1.5.4.2 The bootstrap method

Before the variety of cross-validation methods were sufficiently extended, leave-one-out cross-validation was the dominant method for providing approximately unbiased estimates, at the cost of high variance. The bootstrap method was proposed to estimate prediction errors as a smoothed version of leave-one-out cross-validation, aiming to reduce variance and quantify the uncertainty of a model (Efron, 1983, Efron and Gong, 1983, Efron, 1986, Efron and Tibshirani, 1997).

Bootstrapping performs an iterative process that estimates the statistics of the population by resampling a dataset with replacement. As a resampling technique, it usually involves intensive computation. The computational complexity is dependent on the size of sample and the number of repeats. Specifically, in creating a bootstrap procedure, the preliminaries are to decide the size of bootstrap sample (n) based on the original sample set (n ⩽ N), and the iterations of this procedure. For each iteration, resampling is followed by randomly choosing one sample each time from the given training sample set and recording its value; then replacing the chosen sample back into the original training set, ensuring the next time resampling is still based on the entire training set. Resampling continues until n samples are chosen, forming a bootstrap sample set for constructing a learning model. In any bootstrap sample, some samples may appear multiple time and some may not be involved at all. The samples absent from the bootstrap sample set are named “out-of-bag” samples, can be used for model estimates. The iteration will be repeated for pre-defined times to calculate the final estimates. Sufficient iterations, such as hundreds or thousands of times, are usually required to produce meaningful statistics.

A critical prerequisite for the validity of the non-parametric bootstrap method is the assumption that the sample distribution is identical with that of the true population. Resampling a bootstrap sample from the original training set can then be validly portrayed as a process of randomly taking the original training set from the population. Another key principle is that the samples are treated as completely independent from each other. That is why the bootstrapping requires replacement. Based to the adoption of bootstrapping summarised by Adèr and Mah (Adèr, 2008, Mah, 2014b) and associated with our research context, the spatial architecture of human brain lesions is extremely hard to link with the resultant functional deficits.

Moreover, the sample size we are dealing with in our studies (N = 1333) is arguably insufficient given the variability of the. Here k-fold cross-validation is preferable and performs at intermediate computational complexity. In this thesis, we will therefore use k-fold cross-validation for the purpose of accuracy estimation and model selection. Nonetheless, we can potentially combine cross-validation with the bootstrap method to implement a more comprehensive estimate strategy with the continuously increasing scale of available data.

1.6 Summary

For more than a century and a half, there has been awareness of some of the basic organisational principles underlying the functional architecture in human brain (Broca, 1861, Wernicke, 1874). Nonetheless, how this functional anatomy can be comprehensively deciphered remains a puzzle.

A variety of tools and techniques have developed in the last few decades to tackle this challenge. For example, TMS, first demonstrated by Barker, et al. (Barker et al., 1985) is a widely used investigative tool to probe the brain, especially, in the field of neurophysiology and neuropsychology (Sack and Linden, 2003, Walsh and Cowey, 2000). However, it cannot penetrate deep enough to influence regions beneath the cortex, restricting the range of insight it could conceivably disclose. fMRI (Ogawa et al., 1990a) has revolutionised functional mapping of the human brain, but despite its popularity, leaves the absence of observed activity uninterpretable, greatly limiting the conclusions that can be licitly drawn from it. Moreover, any indirect measurement of neuronal activity can be influenced by a mix of factors that may not be captured by experimental manipulations. Most importantly, the claim that a specific region is necessary for a particular function cannot be sustained by any purely correlative technique.

Lesion mapping - historically the first method of brain mapping – has theoretically greater power because it seeks to determine the necessity of a putative brain region for a given function or deficit. Our knowledge on functional anatomy relies heavily on studies of patients with focal brain lesions. Though lesion-deficit studies have great inferential potential, their power is limited by a crucial vulnerability. First, their sample sizes are typically small; second, the inferential approach in widespread use is blind to consistent errors arising from interactive high-dimensional patterns of focal damage. Hence, understanding this extremely complex distribution, inevitably with the aid of large-scale data, is essential to further progress in the field.

In the following chapters of this thesis, a series of studies on therapeutic inference and lesion prediction will be conducted to examine the benefits of a relatively novel approach – high-dimensional inference on focally damaged human brain with the aid of machine learning techniques. The largest available (at the time of writing) brain lesion dataset will be parameterised into a high-dimensional structure to model the correlation between focal damage and functional deficits. In Chapter 2, a series of experiments are conducted to establish a methodological foundation by comparing the limits of inferential power achieved from a conventionally low-dimensional statistical model and a machine learning driven high-dimensional model. In Chapter 3, a semi-supervised machine learning technique will be employed to show how taking advantage of all available clinical data—whether labelled or not—can deliver optimal prediction. In Chapter 4, a study of therapeutic inference on focally damaged brain will demonstrate how a large-scale high-dimensional model with advanced and robust inferential power can help to detect the consistent errors so as to improve the sensitivity of clinical trials. In Chapter 5, a prescriptive inference study will be designed, in which multiple interventions are simulated to identify the effectiveness of the treatments as well as their interactions. This will be accomplished by a multi-label machine learning technique.

Chapter 2 The limits of behavioural outcome prediction following focal brain injury

2.1 Introduction

Efforts to model the relationship between focal brain damage and behavioural outcomes have a long history, traceable to over one and a half centuries ago when the correlation between ventroposterior frontal lobe damage and speech production was first demonstrated by Broca (Broca, 1861), and later the correlation between superior temporal gyrus damage and language comprehension was reported by Wernicke (Wernicke, 1874). Since then, a complex functionally specified architecture of human brain has been revealed by relating lesions to deficits. In the contemporary era, although the epoch-making introduction of non-invasive techniques has expanded the options for conducting functional localisation studies, lesion-deficit mapping still retains its value as arguably the most powerful method, uniquely capable of establishing the necessity of a given substrate. Though also disruptive, TMS (Barker et al., 1985) is technically restricted to superficial cortex and practically limited in the number of possible concurrent stimulations: spatial resolution and coverage are therefore low even if temporal resolution is in the order of milliseconds. Functional MRI (Ogawa and Lee, 1990, Ogawa et al., 1990b, Ogawa et al., 1990a, Ogawa et al., 1992), though widely used over the past three decades, is merely correlative and hence weaker in its inferential power. By contrast, lesion-deficit mapping can theoretically sustain a strong claim to a particular region being required for a specific function, for it is on the loss of the function that the inference rests. With the advent of CT and structural MRI, lesion-deficit mapping no longer relies on post-mortem examination; furthermore its excellent spatial resolution and temporal correspondence invite us to return to this traditional brain mapping method (Fox, 2018, Mah et al., 2014b, Rorden and Karnath, 2004, Rorden et al., 2009).

The fidelity of lesion-deficit mapping relies on two distinct aspects – the properties of the source data and the validity of the inferential models applied to it. The quality of imaging data is no longer a limitation, though its volume likely always will be. But whereas the properties of data can be judged empirically, the other crucial aspect – whether or not an inferential model is coherent – requires conceptual analysis. Moreover, fatal biases in a model could be concealed under the cloak of "statistical significance", causing systematic, consistent errors that cannot be detected empirically because there is here no ground truth (Xu et al., 2017a). Inevitably, dealing with the complex distributed architecture of the human brain, is bound to require inferential models of commensurate complexity. One typical example is visuospatial neglect, shown across a range of studies to be associated with damage to either inferior frontal lobe (Heilman and Valenstein, 1972, Husain and Kennard, 1996, Vallar, 2001) or posterior parietal lobe (Halligan et al., 2003, Karnath et al., 2001b). At the same time, other studies have shown *recovery* of spatial neglect following sequential strokes in *both* parietal and frontal lobes (Daffner et al., 1990, Vuilleumier et al., 1996). In these studies, the functional outcome is non-monotonically dependent on multiple brain components, indicating a nonlinear relationship between function and anatomy, and supporting the operation of an underlying distributed functional network in the human brain. Thus, one of the crucial criteria of a valid inferential model here is the capacity to deal with multiple variables, and moreover, their interactions.

To tackle the interactions between brain components, it is critical to clarify the determinants of the intrinsic dimensionality of human brain. However, a second anatomical dimensionality – lesion architecture – is widely neglected which together with the functional architecture determines lesion-deficit relations (Mah et al., 2015, Mah et al., 2014b, Mah et al., 2014a, Xu et al., 2017a). For a specific pathological process, whenever a voxel is affected, it is generally inevitable that other voxels will be "collaterally" affected. This is induced by the nature of the pathology, and need bear no relation to the functional architecture. If these voxels are "parasitically" associated with the function merely because of covariance with voxels on which the function is truly dependent, the spatial inference will be biased. Neglecting the complex distributed lesion architecture will invalidate any models that seeks to relate lesion and deficit.

The conventional approach to relating lesions to deficits is the template overlay method, from which voxel-wise lesion symptom mapping (VLSM) has evolved. By superimposing a set of lesions to a template with discrete anatomical landmarks, the overlay method extracts the overlap to compare with a control group, assuming the commonly affected brain regions are associated with the deficit (Frey et al., 1987, Robertson et al., 1988, Robertson and Lamb, 1991). VLSM is based on the same concept, but performs the comparisons voxel-by-voxel so as to generate a voxel-based measure of confidence (Bates et al., 2003). Across both methods, the comparisons are conducted by a statistical test which quantifies the contribution of each part or point of human brain independently – hence, it is a univariate test, applied *en masse*.

The intrinsic anatomical relationships underlying the natural pattern of pathological damage are potential sources of systematic biases which complicate the architecture of natural human brain lesions. The core defect of univariate-based models is to neglect complex spatial correlations that may distort the inferred map. Such errors cannot be identified by any mass univariate models, for they are driven by dependencies mass univariate models definitionally ignore. In other words, formally significant statistics obtained from a univariate model may not reflect the true anatomical picture. Regardless of the spatial resolution, including region-wise or voxel-wise parcellation, the errors can neither be detected nor corrected, for the interwoven relationships from the critical perspective of anatomy are completely isolated. Instead, massive univariate tests merely entrench the statistical biases over redundant repetition.

It is tempting to try to extend the VLSM approach by escalating the anatomical dimensionality to some degree while preserving its mass-univariate essence. A common way is to include lesion size as a covariate or a nuisance regressor in the model (Karnath et al., 2004, Karnath and Smith, 2014, Schwartz et al., 2012, Kümmerer et al., 2013). Another method makes use of a random Gaussian fields to describe the relations between each voxel and its neighbours, in which the value of each voxel is adjusted by the estimate of Gaussian smoothness. (Kimberg et al., 2007). The former method simply adds lesion volume as an additional variable in the voxel-wise model. The latter method applies a Gaussian estimate to weight each voxel. Although the spatial dependence between each voxel and its adjacency is taken into account, the set of models are still established by each single voxel independently.

Unfortunately, the problem here cannot be remedied unless the entire lesion is treated as a whole. This is because the anatomy of the lesion is driven by a pathological process which is neither perfectly random nor perfectly biased. This partial bias will vary from one pathological process to another, and will cause a pattern of anatomical interactions specific to each pathology. When the relations between lesion and behaviour are modelled, it is critical to taken in account the covariance between voxels. Instead of establishing the models voxel-by-voxel, a fundamental prerequisite for a valid model here will be to involve an irreducible multiplicity of voxels that describe sufficient dimensions of the lesioned brain – hence, a high-dimensional model (Mah et al., 2014b, Mah et al., 2014a, Xu et al., 2017a). The contrast of between a mass-univariate prediction approach against a high-dimensional multivariate approach is illustrated in Table 2.1.

It has previously been shown that high-dimensional modelling can remedy the spatial biases identified in mass-univariate prediction models in the context of spatial inference (Mah et al., 2014b). In this chapter, we use an analogous approach to model outcome prediction rather than spatial inference.

	Dimensionality of functional anatomy	**Dimensionality of lesion anatomy**
Mass-univariate prediction	Low sensitivity, regardless of data size	With spatial biases, regardless of data size
High-dimensional multivariate prediction	High sensitivity, given sufficient data	Without spatial biases, given sufficient data

Table 2.1 The necessity of high-dimensional model in lesion-deficit mapping (Xu et al., 2017a)

Correlation-based machine learning techniques are often required to deal with the interactions between a multiplicity of variables in high-dimensional models. In this study, the high-dimensional data will be modelled by SVM, which attempts to separate two classes labelled by a binary outcome with a hyperplane as described in the introduction. In particular, SVM is a typical machine learning technique that is capable to train sparse datasets in high-dimensionality and address the dimensionality mismatch between the number of training features and cases (Joachims, 1998, Saur et al., 2010). In the following chapter, based on aforementioned demonstration in terms of concept and methodology, a series of inferential models will be established to illustrate the relationship between the biases derived from neglecting anatomical dimensionality of lesion architecture and prediction performance. To be more specific, two series of simulations are modelled to assess the performance between a mass-univariate technique and multivariate based SVM. The outcome in first series of lesion-behaviour predictions is determined by a synthetic ground truth, where if a certain proportion of each lesion falls within each Brodmann area the patient is labelled as “affected”. In the other series, the ground truth outcome is based on a full combinatorial of Brodmann pairs, where a positive label is allocated where each lesion drops within *either* of each pair of Brodmann areas. The critical question is whether or not the systematic biases can be revealed and corrected; and to what extent the predictive power in relation to lesion-behaviour can be improved with the use of multivariate techniques.

2.2 Methods

2.2.1 Patients

A set of 1333 patients admitted to University College London Hospital (UCLH) Hyper-acute Stroke Unit (HASU) over the past decade were selected. All patients were clinically diagnosed with acute ischemic stroke, and received MRI, including diffusion-weighted imaging (DWI), in which at least one acute ischemic lesion could be identified. A threshold for minimal DWI-quantified lesion volume was set at 50 milliliters which covered the 90th centile of our entire clinical population in order to reduce the disproportionate impact from very large lesions. An adjusted cohort including 1172 patients was thus produced.

The age distribution from the set of 1172 patients was in range of 18 to 97 years old with a mean of 63.89 and a standard deviation of 15.91. The gender ratio (Male) was 0.561 (Appendix B. Figure 7.1). Ethnicity for 952 (81.2%) from the full set of patients was disclosed in the clinical record which is illustrated in Appendix B. Figure 7.2. The distribution of age and sex, and the constitution of ethnicity reflected the catchment of clinical population in UCLH naturally.

2.2.2 Imaging

2.2.2.1 MR data acquisition

All acquisitions performed on the scanners manufactured by General Electric (GE) (Discovery MR 450, Genesis Signa and Signa Excite), Philips (Achieva, Ingenia and Intera), or Siemens (Avanto, Biograph mMR, Espree, Skyra, Symphony, Symphony Tim, Trio Tim and Verio) with field strength of either 1.5 or 3 Tesla in a single session for each patient were collected via a picture archiving and communication system (PACS) from UCLH (the specific range of scanners and related parameters are listed in appendix A). The originated voxel sizes were different in some extent between the scanners, and so were the according spatial resolutions. Typically, for 1.5 Tesla GE medical system, the voxel size was sampled as 1mm x 1mm x 6.5mm (Columns = 256, Rows = 256). For 1.5 Tesla Philips

medical system, the typical voxel size was 1.2mm x 1.2mm x 6mm (Columns = 192, Rows = 192); the voxel size was 1mm x 1mm x 6mm (Columns = 256, Rows = 256) performed on 3 Tesla Philips scanners. For Siemens, the voxels sizes and spatial resolution were 1.8mm x 1.8mm x 6.5mm (Columns = 128, Rows = 128) and 1.2mm x 1.2mm x 6.5mm (Columns = 192, Rows = 192) on 1.5 Tesla and 3 Tesla scanners, respectively. The routine of clinical practice with regular equipment upgrades and replacement is the main reason for the diversity of scanners over the period of data collection. All brain scans were performed for the purpose of clinical routine obeying clinical protocols.

The specific sequences extracted from each MRI session were axially required echo planar DWI with b-values of 0 s/mm^2 and 1000 s/mm^2 which is used for lesion segmentation (Mah et al., 2014c) so as to implement the series of simulations designed for this study. DWI is able to indicate the restricted ability of Brownian motion for extracellular water protons where cytotoxic edema causes imbalance. So, DWI is sensitive for detecting ischemic stroke and widely used for locating acute ischemic lesions (Warach et al., 1995, Löuvbld et al., 1997, Lövblad et al., 1998). In the clinical application, a DWI sequence usually includes images with b values of 0 s/mm^2 (a b0 image) and 1000 s/mm^2 (a b1000 image). The former type image shows stronger contrast on normal tissues but relatively insensitive to ischaemia. Conversely, an image with high diffusion weighting is used to sense water molecules movement. So, it is sensitive to ischaemia but poor to the contrast of normal tissues. Both types of images are complementary for brain registration and lesion segmentation, which rely on the contrast between normal tissue types and the discrimination between normal tissues and lesions, respectively.

2.2.2.2 Image pre-processing

A processing pipeline was implemented within MATLAB (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.) associated with a configured SPM12 toolbox (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/) to ensure the images to be processed systematically.

For the purpose of scientific imaging analysis, typically, the raw clinical MR images needed to be firstly converted from the complicated format of Digital Imaging and Communications in Medicine (DICOM) images to Neuroimaging Informatics Technology Initiative (NIfTI) format. Subsequently, in order to provide accurate and robust performance on image registration and segmentation in the later stage, we employed a pre-processing pipeline with the following steps. We empirically clamped the signal between 0.1% and 99.9% of the cumulative distribution estimated with a kernel density method (Botev et al., 2010), removing unusually low and high signals in each NIfTI file. Then an oracle-based 3D discrete cosine transform (ODCT3D) (Manjón et al., 2012) denoising method was applied for further noise reduction. In the next step, image registration was conducted based on each pair of DWI files. In order to optimise the alignment between the b0 and b1000 images in each pair, we used SPM12's standard co-registration function to have the b0 image rigidly co-registered to the b1000 image. Then, the normalisation / segmentation routine integrated in SPM12 (Ashburner and Friston, 2005) was applied to the b0 image to calculate a deformation field which described the optimal non-linear transformation of the b0 image into Montreal Neurological Institute (MNI) stereotactic space. Next, the deformation field derived from the b0 image was applied to the b1000 image to transform it into MNI space using the same routine. The parameters involved in the whole process of image registration were set at defaults by SPM12. Finally, both b0 and b1000 images in each pair were resampled to 2mm^3 isotropic with 6th degree b-spline interpolation (Mah et al., 2014b, Mah et al., 2014c) and manually checked against the SPM template to confirm the satisfactory of registration.

2.2.2.3 Lesion segmentation

A previously validated method based on the anomaly metric *zeta* (Mah et al., 2014c) was applied to segment lesions from the 1172 normalised b1000 images in a voxel-wise manner.

A set of preliminaries conducted to optimise the performance of lesion segmentation were performed as follows. First, 492 DWI images derived from patients suspected of acute stroke but found to have normal imaging were selected

to form a reference normal image set. No acute lesion was visible here in any patient, and none was reported by the corresponding radiological record. Every scan was pre-processed using the exact same pipeline and resliced at the same resolution of 2mm isotropic voxels, which spatially presented as a 91 x 109 x 91 matrix. Then, by combing the tissue maps provided by SPM12 including white matter tissue, grey matter tissue and cerebrospinal fluid (CSF), an inclusive brain mask was created. So as to remove image artefacts, we set an empirical threshold to exclude areas in the frontal and temporal poles commonly prone to artefacts on DWI sequences, and adjusted the inclusive brain mask accordingly. Subsequent lesion segmentation was performed within the confines of this adjusted mask. Next, the signal distribution was normalised for both lesion and reference datasets. Such signal normalisation sought to reduce instrumental and other incidental signal heterogeneity, so that the comparison between individual images could be maximally sensitive. The procedure was performed as follows. A binary white matter mask was created from SPM12's white matter tissue probability map by thresholding at a value of >0.9. A robust kernel density estimate method (Botev et al., 2010) was then applied to the voxels falling within both customised maps in each b1000 image to obtain the peak white matter distribution which was then subtracted from each corresponding b1000 image to normalise the signal distribution. Guided by information in the radiological report, signal normalisation on unilateral and bilateral lesions was performed differently. For unilateral lesions—the vast majority—only the unaffected hemisphere was used to estimate the peak. For bilateral lesions, since ischemic lesions tended to be small in these circumstances, the kernel density estimate was applied to the whole brain. In both cases, the signal normalisation was equally effective in adjusting the differences in the signal distribution across all scans performed by a variety of scanners and related technical and practical variance.

Lesion segmentation was performed in a voxel-wise manner. We calculated the zeta anomaly metric for each single voxel in each image independently against the reference set. To be specific, the voxel-wise zeta value in the test image is derived from the mean distance to the k nearest neighbors drawn from the anatomically homologous voxels in the unlesioned reference set, normalised by the mean distance between the k neighbors themselves (Mah et al., 2014c). The measure was the Euclidean distance in signal space. The only manipulable

parameter here is the number of nearest neighbors – k – commonly chosen with reference to the size of the reference dataset. In this study, we set k at 23 based on the calculation as below:

$$k = Ceiling\left(\sqrt{492}\right).$$

Zeta is a continuous metric. An adaptive threshold was therefore determined for each image. In brief, volume connectivity computed by *bwlabeln*, an integrated function of MATLAB (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.) was applied to segment each lesion map into connected clusters. Then, each cluster was estimated under a generalised extreme value (GEV) distribution to derive mean and variance values which were used to decide threshold adaptively as described in Mah et al, 2014.

2.2.3 Dimensionality reduction

The predictive performance of a model is often dependent on the ratio of the number of instances to the number of material features. It is therefore helpful to reduce the dimensionality of individual features, while retaining their diversity. In human neuroanatomy, brain asymmetry between two hemispheres has been observed in two distinct aspects – neuroanatomical differences, and functional and behavioural lateralisation (Hugdahl, 2005, Toga and Thompson, 2003), but these differences are minor and do not generally extend to the vascular tree (Wright et al., 2013). Vascular lesions rarely cross the midline of brain and are generally literalized (Nachev et al., 2008). It is therefore reasonable to collapse our stack of vascular lesions onto one hemisphere to reduce dimensionality for following predictive models.

2.2.4 Simulations

Our focus is the impact of the dimensionality of the distributed lesion architecture on predictive modelling, a commonly neglected aspect. In order to visualise the dimensionality of lesions, t stochastic neighbour embedding (t-SNE)

(Maaten and Hinton, 2008) was applied to cluster the set of 1172 lesions labelled by their volumes. In Figure 2.1, the resultant distribution was neither clearly separated nor wholly random, which indicated a single or a few concrete features were definitely insufficient to describe the relationships between lesions. Instead, this implied complex, inter-connected features must be considered as a whole to reveal hidden relationships and achieve optimal predictive performance. The impact on predictive power of neglecting lesion dimensionality within mass-univariate models, and its remedy by high-dimensional models, will now be explored with a series of lesion-deficit synthetic models for clinical prediction rather than lesion-deficit mapping.

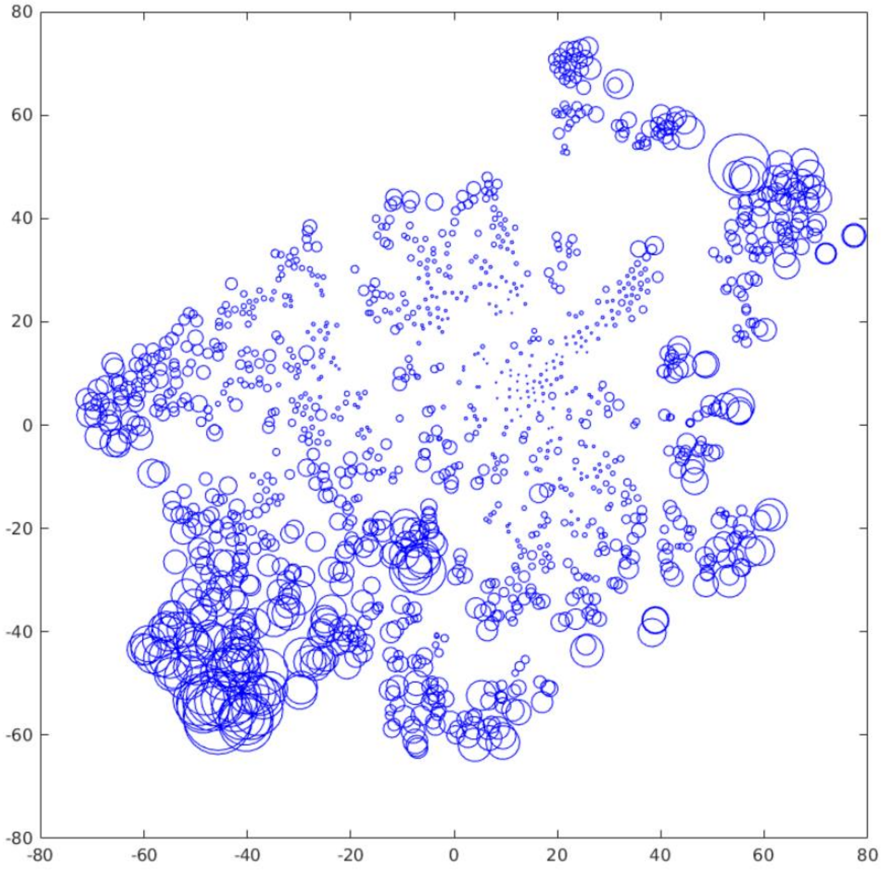


Figure 2.1 Lesion clustering by lesion volume

For the stack of 1172 patients, two-dimensional embedding of the lesion anatomy labelled by the volume of each lesion was generated by t-SNE (Maaten and Hinton, 2008). The clustering was distributed in the way neither simple nor wholly random, which necessitated the use of multivariate methods to achieve optimal predictive performance and minimise bias.

2.2.4.1 Simulation one: lesion-deficit prediction depending on single Brodmann region performed by a mass-univariate technique (Fisher's exact test) against a multivariate technique (SVM)

2.2.4.1.1 Data preparation

The set of 1172 segmented lesion maps were binarised to denote each voxel in the brain as being hit or not for the purpose of subsequent simulations. All lesions were collapsed onto the right hemisphere, then a threshold was applied to ensure every lesion voxel was collectively presented at least four times. The lesion distribution is shown in Figure 2.2. Thus, a data matrix involving 1172 patients with 66770 features at the voxel size of 2mm was created.

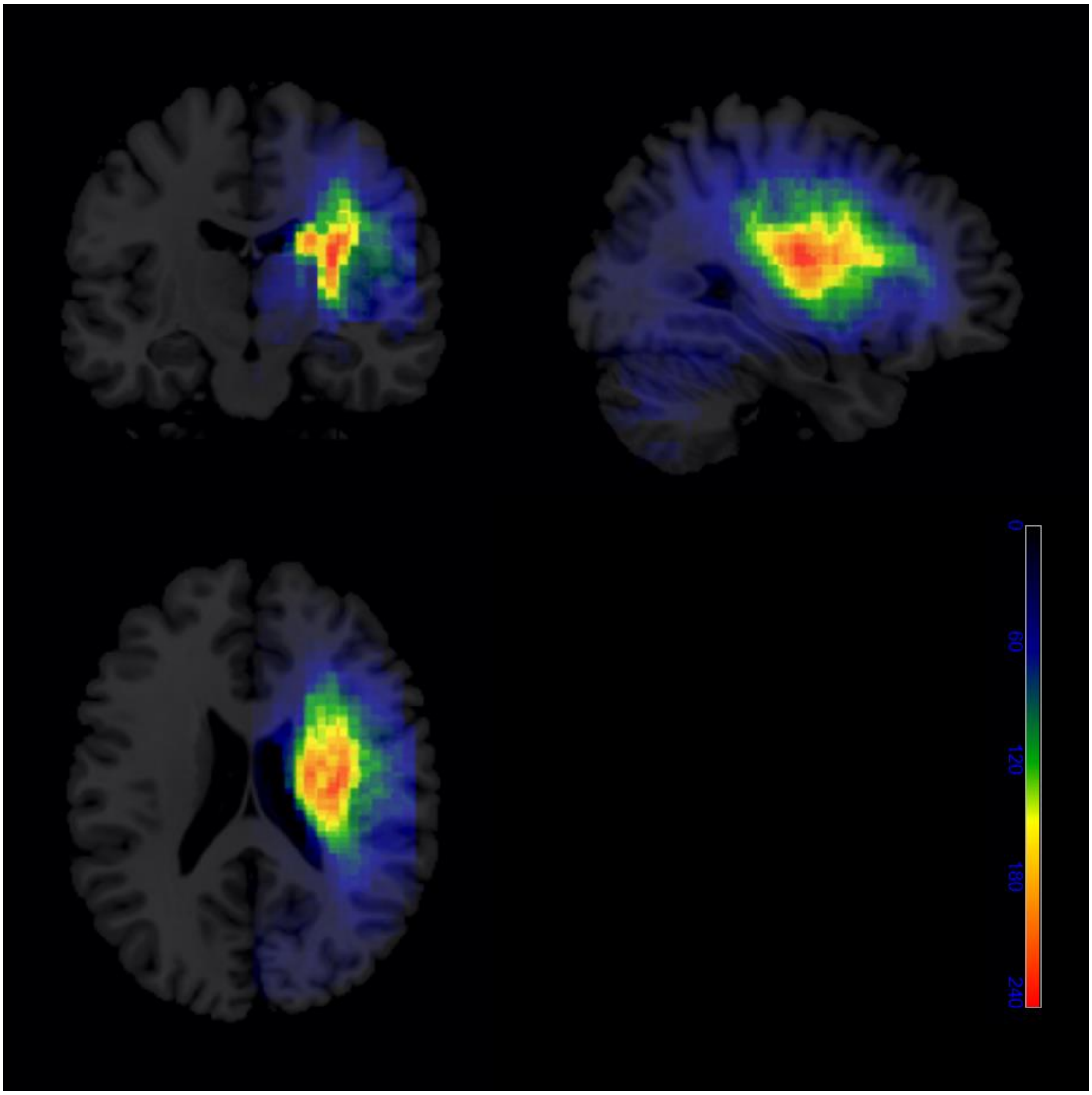


Figure 2.2 Lesion distribution

1172 lesions collapsed on the right hemisphere and were summed in a template in voxel size of 2mm. The colour spectrum demonstrated the density of lesion voxels distribution.

2.2.4.1.2 Lesion-deficit models for clinical prediction

In this series of simulations, we assumed a binary behavioural outcome was sensitive to neuroanatomy. In order to simulate this, each image from the stack of 1172 binary lesion masks was labelled as “affected” or “unaffected” dependent on whether or not at least 20% fell within a specific single Brodmann area. 10% noise was introduced to naturalise the rate for hypothetical deficit of interest. In other words, there was a 90% possibility that 20% or more involvement of a specific Brodmann area by a lesion would lead to the affection of behavioural outcome (deficit).

The set of 1172 lesion masks covered a total of 39 Brodmann areas. Within each area, the lesion-behaviour correlation was independently demonstrated by a pair of simulations modelled by a mass-univariate (Fisher’s exact test) and a multivariate (SVM) technique, respectively.

2.2.4.1.3 Mass-univariate (Fisher’s exact test) analysis

Based on the processed data matrix, the behavioural outcomes of 1172 patients were labelled as “affected” or “unaffected”; and each of the 66770 binarised features were marked as “presented” or “absent”. Thus, a 2 x 2 confusion table was constructed for the following analysis. The entire lesion masks were split up into two parts, where 70% (911 lesions) were randomly selected for training purpose in order to identify a set of discriminating voxels; and the remaining 30% (261 lesions) were for the use of test. The mass-univariate analysis was performed using Fisher’s Exact Test which yielded a p value for each voxel to represent a voxel-wise likelihood indicating how an injured voxel could be associated with the hypothetical functional deficit. In this way, a probability map of the brain was created to illustrate a voxel-wise significance - the lower a voxel’s p value was, the more significantly it correlated with the symptom. As the statistical test performed on each voxel was independent and simultaneous, Bonferroni correction was applied to correct for multiple comparisons, where the alpha value (α) was set at 0.01 (Bland and Altman, 1995, Bonferroni, 1935). Fisher's method (Fisher, 2006) was then employed to derive an image-level simulated outcome prediction from the set of significant voxels.

A total of 39 models as described above were established and evaluated to cover any single Brodmann areas. In each model, 240 different randomisations of the data for training and test were performed iteratively.

2.2.4.1.4 Multivariate (SVM) analysis

1172 labelled cases ("affected" vs "unaffected") were modelled as a two-class classification problem using LibSVM (Chang and Lin, 2011). In contrast to the mass-univariate analysis in which each voxel contributed to the clinical lesion-deficit prediction independently, 66770 dimensions involved in each lesion were modelled simultaneously in multivariate analysis. Moreover, the interactions between voxels representing the architectural interactions and correlations intrinsically derived from the complex lesion architecture were also taken into account.

In order to optimise the predictive performance, a 10-fold cross-validation was used to search for the optimal cost parameter (C) in the radial basis function (RBF) kernel. 31 different C values in data type of integer were assessed as below:

$$\mathrm{C} = 2^{\mathrm{m}}, \text{where m is integer in the range between} - 15 \text{ and } 15.$$

The cost parameter that yielded the best classification performance from the 10-fold cross-validation was applied to the training dataset to generate the optimal predictive model. The classifier created by the optimised model was voxel-wise in which each voxel was weighted in terms of polarity and magnitude representing the membership of either class, and the significance of the specific voxel in relation to the behavioural deficit, respectively. Subsequently, this weighted classifier was used for prediction on the test dataset.

A total of 39 models as described above were established and evaluated to cover any single Brodmann areas. In each model, 240 different randomisations of the data for training and test were performed iteratively.

2.2.4.1.5 Comparison of behavioural outcome prediction with mass-univariate and multivariate analysis

Both methods derived a voxel-wise probability brain map from a dataset that was then used to predict the outcome status of an independent set of test data. In the mass-univariate analysis, Fisher's exact test generated a voxel-wise p value for each independent hypothesis test. Then, this p-value brain map was multiplied with each test brain to calculate the prediction on each voxel. The voxel-wise prediction was combined for the whole brain using the chi-square distribution-based Fisher's method. In the multivariate analysis, the whole set of voxels involved in the brain mask was modelled by SVM RBF kernel simultaneously to generate a weighted probability brain map. Similarly, this brain map was used as a classifier to calculate the overlap between each test lesion mask so as to decide the membership by summing up the overlapped weights.

The decision values derived from both methods were used to calculate the related evaluation metrics including accuracy, sensitivity and specificity. The details are as follows:

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{the total number of cases}}\ ;$$

$$\begin{aligned}\text{sensitivity} &= \frac{\text{number of true positves}}{\text{number of true positives} + \text{number of false negatives}} \\ &= \frac{\text{number of true positives}}{\text{the total number of "affected" cases}}\ ;\end{aligned}$$

$$\begin{aligned}\text{specificity} &= \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}} \\ &= \frac{\text{number of true negatives}}{\text{the total number of "unaffected" cases}}\,.\end{aligned}$$

In addition, the comparison between mass-univariate and multivariate analysis was also interpreted and visualised using a receiver operating characteristic (ROC) curve and the area under the curve (AUC) (Hanley and McNeil, 1982, Pauker and Kassirer, 1975, Swets, 1973, Thornbury et al., 1975). ROC curve plotted with true positive rate (sensitivity) on y-axis against false positive rate (1 - specificity) on x-axis illustrates the capacity to distinguish two classes at all classification thresholds. AUC stands for the area under a ROC curve, which aggregates the performance of classification at all thresholds. So, the higher AUC is, the stronger separability of the classification model is. The confidence interval at the level of 95% was applied to ROC curves for both methods to demonstrate the discrimination reliably.

2.2.4.2 Simulation two: lesion-deficit prediction depending on dual Brodmann regions performed by a mass-univariate technique (Fisher's exact test) against a multivariate technique (SVM)

2.2.4.2.1 Data preparation

The exact same stack of lesion masks as described in simulation one was used to perform this series of simulation. The 1172 binarised lesion maps collapsed onto the right hemisphere formed a date matrix featured 66770 dimensions (voxel size: 2mm). All features presented at least four times among the total 1172 lesion maps.

2.2.4.2.2 Lesion-deficit model for clinical prediction

In this set of simulations, we extended the single Brodmann area based lesion-deficit models to dual Brodmann areas in order to demonstrate the discrimination between mass-univariate and multivariate analysis in the context of distributed functional anatomical dependence. Due to the complex neural network underlying human brain, any given function or behavioural deficit tends to dependent on multiple loci (Mah et al., 2014b, Power et al., 2011, Fox, 2018). To establish such models, each lesion mask was labelled as “affected” or “unaffected” decided by whether or not its 20% or more fell within either of specific dual Brodmann regions. Minimum 20% volume of the lesion could solely fall within either of the specific dual Brodmann areas or cover both Brodmann areas. 10% noise rate was applied to present the naturalisation of hypothetical deficit of interest. In this way, the established models illustrate a 90% possibility that at least 20% involvement of a specific dual Brodmann area overlapped by a lesion would result in the behavioural deficit.

We used a Brodmann brain map including forty-one areas. The number of fully combinatorial dual areas is 820. In particular, the region combined by Brodmann area 28 and 36 was not covered by any of the 1172 lesion masks, so a total of 819 predictive models based on dual Brodmann areas were established to perform mass-univariate (Fisher's Exact Test) and multivariate (SVM) analysis, respectively.

2.2.4.2.3 Mass-univariate (Fisher's exact test) analysis

As before, a 2 x 2 confusion table was constructed by labelling each lesion mask as either "affected" or "unaffected" and marking each dimension as either "present" or "absent" dependent on its binary value. Then, the mass-univariate analysis was performed using Fisher's Exact Test by modelling each dimension independently to identify the correlation between its presenting flags and the hypothetical behavioural deficit. The stack of 1172 lesion masks was randomly split into two portions (70% vs 30%) for the purpose of training and test, respectively. During the training procedure, a voxel-wise p value was produced to describe the likelihood of how a "presented" voxel was associated with the affection of behavioural outcome. The lower the p value was, the high impact the voxel played. To adjust the p value derived from multiple independent significance tests, we applied Bonferroni correction (Bland and Altman, 1995, Bonferroni, 1935) at alpha level of 0.01. Then, the corrected voxel-wise p values formed a probability map is used to evaluate the model performance with the remaining 30% lesion masks. During the test process, associated with the Fisher's method (Fisher, 2006), each lesion mask was multiplied by the probability map to sum up as a probability score used for indicating the membership of class.

A total of 819 models as described above were established and evaluated to cover any combination of two Brodmann areas. In each model, 240 different randomisations of the data for training and test were performed iteratively. With the concern of time efficiency, all models were processed on a distributed 24-nodes server machine in parallel to optimise the feasibility.

2.2.4.2.4 Multivariate (SVM) analysis

In multivariate analysis, LibSVM (Chang and Lin, 2011) was used to model the set of 1172 labelled lesion masks as a binary classification problem. A total of 66770 dimensions in each lesion map were modelled simultaneously. Moreover, the interactions between dimensions were also taken in account by the correlation-based technique. By contrast with the mass-univariate analysis, multivariate analysis was conducted to illustrate the complexity of intrinsic lesion architecture and reveal mechanisms underlying the damage. To match the mass-univariate analysis, 70% of the lesion masks were used for training and the others were used to test the prediction performance. In order to optimise the training process for the RBF kernel, a 10-fold cross-validation was performed to assess the optimal cost parameter in the range of -15 and 15 to the power of 2 ($\mathrm{C} = 2^{\mathrm{m}}$*,* where m is integer between -15 and 15). Thus, the optimal predictive model established with optimised hyper-parameter was applied to the set of training lesion maps to generate a classifier, in which the polarity and magnitude of voxel-wise weights represented to the membership of either class and the extent of confidence with the classification. This classifier was subsequently used to predict the memberships from an independent set of lesion masks.

A total of 819 models as described above were established and evaluated to cover any combination of two Brodmann areas. In each model, 240 different randomisations of the data for training and test were performed iteratively. With the concern of time efficiency, all models were processed on a distributed 24-nodes server machine in parallel to optimised the feasibility.

2.2.4.2.5 Comparison of behavioural outcome prediction between mass-univariate and multivariate analysis

By training on a subset of the stack of 1172 lesion masks, both models generated a probability map that could be used to classify the remaining test lesion masks. In the mass-univariate models, the classifier was comprised of p values from independent significance tests; in multivariate analysis, the classifier was produced by taking into account all the dimensions together as well as their inter-connections with a correlation-based Gaussian kernel machine. As mentioned in simulation one, the basic evaluation metrics to compare with the performance of two types predict models were the level of accuracy and the according sensitivity and specificity. In addition, AUC-ROC curves were plotted to interpret and visualise the discrimination between both analyses. The confidence interval at the level of 95% was applied to ROC curves to prove reliability.

2.2.5 Software and hardware

All simulated prediction models were implemented in the environment of MATLAB (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.).

In simulation one, a 12-core Inter® Xeon® CPU E5-2620 2.00GHz processor with 64GB RAM and 9TB 7200 RPM SATA hard drive was used to conduct both Fisher’s Exact Tests and SVM models under a 64 bit Linux operation system (Ubuntu version 15.04).

In simulation two, as the models needed to cover the full combination of any two Brodmann regions, enormous computation was essentially required. A Supermicro® MicroCloud with 24 modular nodes was used to perform mass-univariate and multivariate analysis in parallel, which could be theoretically 12 times faster comparing to a standard sole desktop. In each modular node, there is an Inter® Xeon® E3-1200 processor with 4GB RAM and 2TB SATA hard drive. The operation system across the 24 modular nodes was Windows® 10 professional 64 bit.

2.3 Results

2.3.1 Simulation one: lesion-deficit prediction dependent on single Brodmann regions performed by a mass-univariate technique (Fisher’s exact test) versus a multivariate technique (SVM)

The stack of 1172 lesion masks was labelled with a hypothetical functional deficit dependent on a specified proportion of each lesion overlapping with a specific Brodmann area. Thus, based on all of Brodmann areas covered by the set of lesion masks, a total of 39 lesion-deficit prediction models were established iteratively. Each model was performed by a mass-univariate (Fisher’s Exact Test) and a multivariate (SVM) technique respectively in order to illustrate how the complexity of lesion architecture and the inter-connection between dimensions will contribute to the lesion-deficit correlation for clinical prediction. The discrimination of separability was interpreted with an AUC-ROC curve.

2.3.1.1 Comparison of Fisher’s exact test against SVM

The set of 1172 lesion maps were randomly split into 70% and 30% to train a predictive model and test its performance for 240 iterations. In each iteration, a Fisher’s Exact Test was used in the mass-univariate analysis to generate a model for each single Brodmann area; whilst multivariate analysis was implemented by SVM. Thereby, a total of 9360 (39 Brodmann areas for 240 iterations) mass-univariate analysis models matched with 9360 models for multivariate analysis to compare the performance of clinical lesion-behavioural prediction. Both approaches took less one minute to generate a model and the whole series of simulation took approximately five calendar days.

In the case of mass-univariate training procedure, voxel-wise p values derived from Fisher’s Exact Test indicated the likelihood of how a lesioned voxel is independently associated with the hypothetical deficit. By multiplying each test lesion map with the set of p values and then using Fisher’s method, the outcome predictions were generated. We applied Bonferroni correction (alpha level at 0.01) for multiple comparisons (Bonferroni, 1935, Bland and Altman, 1995) to ensure the voxels retained with reliable influence and significance above the defined threshold. In the multivariate analysis, a correlated weighting map derived from SVM training process was applied to each test lesion mask for prediction outcome. Preliminarily, the optimal model for each single Brodmann area was confirmed with a search of the cost parameter (C), which assessed a total of 31 different values with a 10-fold cross-validation.

For vast majority (32 of 39, 82.1%) of the models dependent on a single Brodmann region, the prediction performed by a high-dimensional multivariate approach (SVM) was superior to a low-dimensional mass-univariate technique (Fisher’s Exact Test). Figure 2.3 (a) shows the contrast of AUC value in each predictive model between Fisher’s Exact Test and LIBSVM; and Figure 2.3 (b) shows the extent of their differences. Among the superior models, the average of relative difference in term of AUC value between two approaches was 0.012. For the minority (7 of 32, 17.9%) of the models who performed better in mass-univariate analysis against multivariate analysis, we found the average number of lesion masks involved into those Brodmann areas was 14.57; whereas the average number of lesion masks within the other 32 Brodmann areas remarkably increased to 49.0 (Table 2.2). This reflected the high-dimensional multivariate approach could provide greater performance, given sufficient data size. In other words, multivariate analysis investigated the lesion mechanism more thoroughly than the mass-univariate analysis.

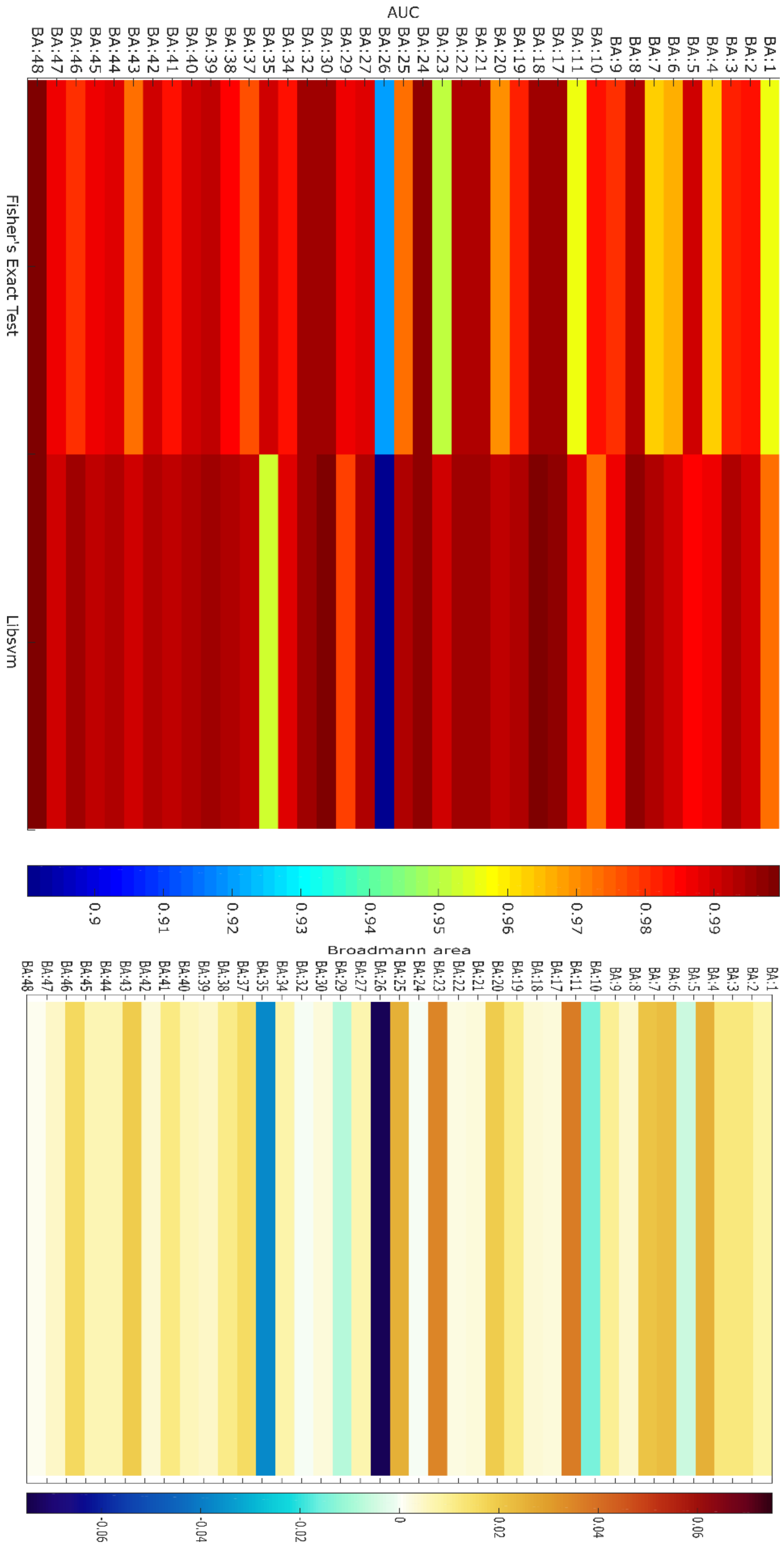
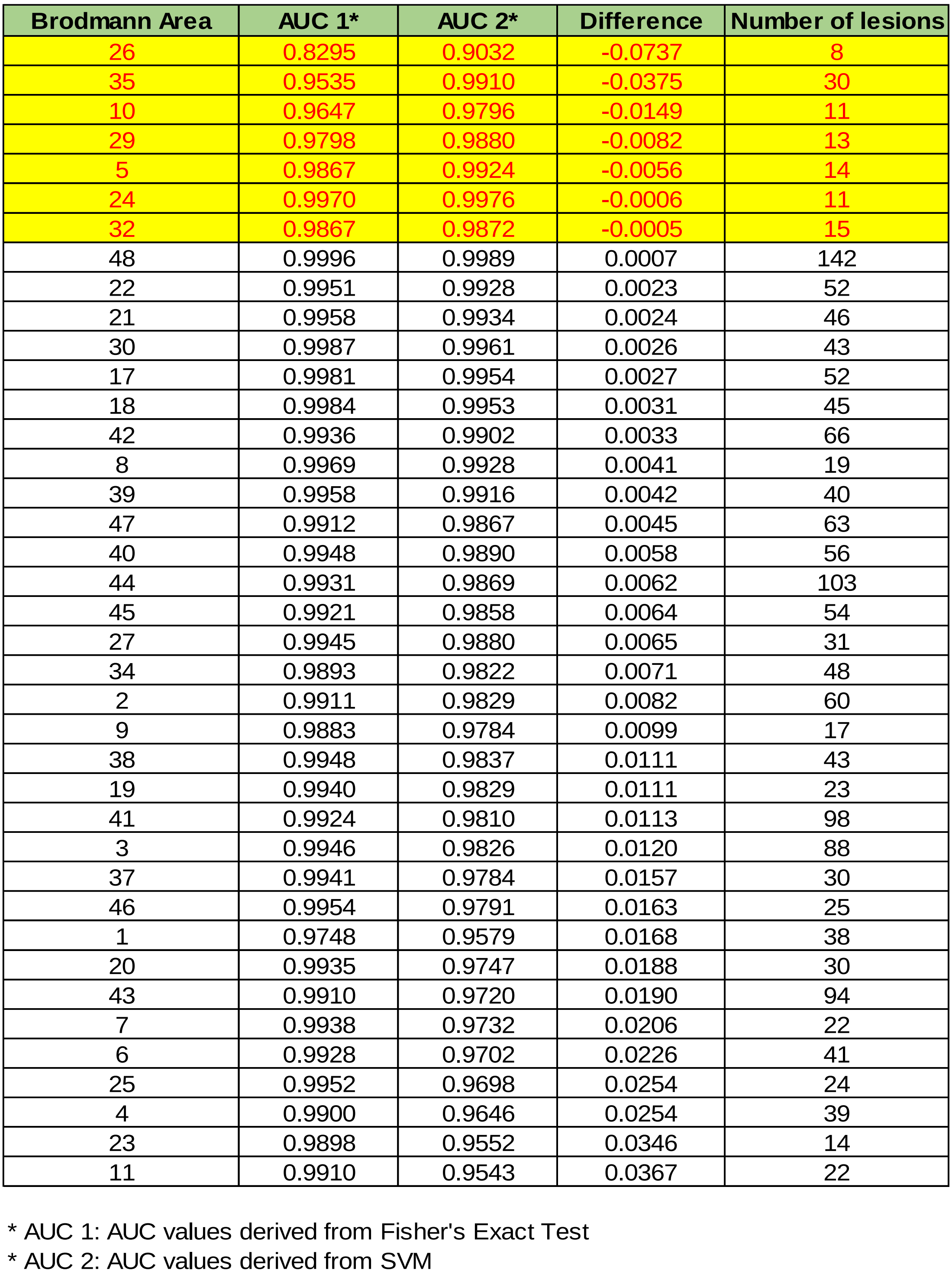


Figure 2.3 The comparison of AUC between Fisher's exact test and SVM

The discriminations of predictive performance between mass-univariate and multivariate analysis, where majority of the case illustrated the superior predictive power achieved from multivariate analysis.



Brodmann Area	AUC 1*	AUC 2*	Difference	Number of lesions
26	0.8295	0.9032	-0.0737	8
35	0.9535	0.9910	-0.0375	30
10	0.9647	0.9796	-0.0149	11
29	0.9798	0.9880	-0.0082	13
5	0.9867	0.9924	-0.0056	14
24	0.9970	0.9976	-0.0006	11
32	0.9867	0.9872	-0.0005	15
48	0.9996	0.9989	0.0007	142
22	0.9951	0.9928	0.0023	52
21	0.9958	0.9934	0.0024	46
30	0.9987	0.9961	0.0026	43
17	0.9981	0.9954	0.0027	52
18	0.9984	0.9953	0.0031	45
42	0.9936	0.9902	0.0033	66
8	0.9969	0.9928	0.0041	19
39	0.9958	0.9916	0.0042	40
47	0.9912	0.9867	0.0045	63
40	0.9948	0.9890	0.0058	56
44	0.9931	0.9869	0.0062	103
45	0.9921	0.9858	0.0064	54
27	0.9945	0.9880	0.0065	31
34	0.9893	0.9822	0.0071	48
2	0.9911	0.9829	0.0082	60
9	0.9883	0.9784	0.0099	17
38	0.9948	0.9837	0.0111	43
19	0.9940	0.9829	0.0111	23
41	0.9924	0.9810	0.0113	98
3	0.9946	0.9826	0.0120	88
37	0.9941	0.9784	0.0157	30
46	0.9954	0.9791	0.0163	25
1	0.9748	0.9579	0.0168	38
20	0.9935	0.9747	0.0188	30
43	0.9910	0.9720	0.0190	94
7	0.9938	0.9732	0.0206	22
6	0.9928	0.9702	0.0226	41
25	0.9952	0.9698	0.0254	24
4	0.9900	0.9646	0.0254	39
23	0.9898	0.9552	0.0346	14
11	0.9910	0.9543	0.0367	22

* AUC 1: AUC values derived from Fisher's Exact Test
* AUC 2: AUC values derived from SVM

Table 2.2 The detailed performance between Fisher's exact test and SVM

Overall, the absolute AUC values from both approaches are high because the series of simulations were modelled in idealised conditions. The critical point that matters here is their relative difference, which indicated the benefits from the fidelity of lesion anatomical dimensionality. The average AUC value derived from SVM was

0.988 (95%CI: 0.986 – 0.990); whereas the mean AUC score derived from Fisher's Exact Test was 0.981 (95%CI: 0.978 – 0.983) (Figure 2.4). Compared with the multivariate case, mass-variate analysis was less sensitive and less reliant on data size.

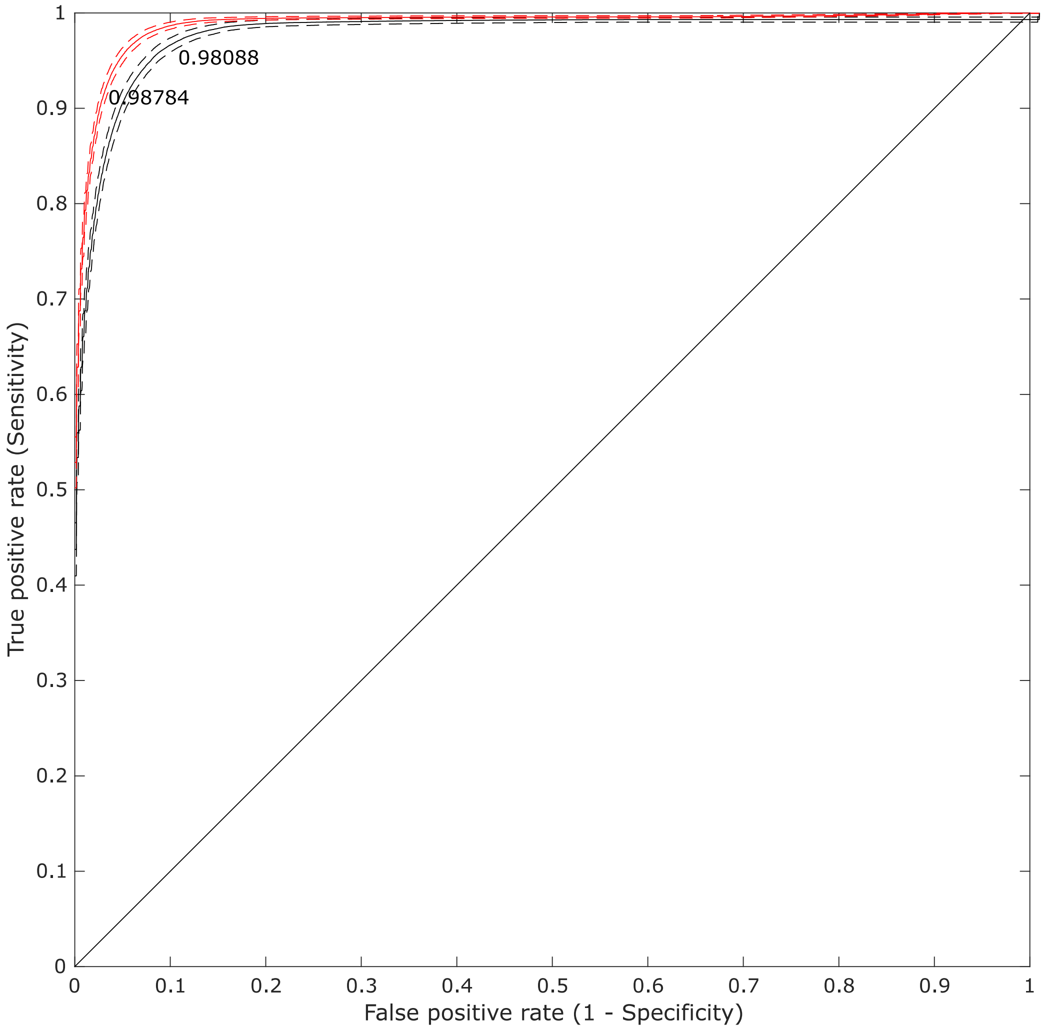


Figure 2.4 Comparison of ROC curve between Fisher's exact test and SVM

The ROC curves derived from both low- and high-dimensional models demonstrated the averaged predictive performance across 39 single Brodmann areas. A total of 240 randomisations that split the lesion data for training and test purpose was iteratively conducted to enhance the performance stability and reliability.

2.3.2 Simulation two: lesion-deficit prediction dependent on dual Brodmann regions performed by a mass-univariate technique (Fisher’s exact test) versus a multivariate technique (SVM)

The stack of 1172 lesion masks was labelled with a hypothetically functional deficit dependent on a certain proportion of each lesion overlapping with any combination of two Brodmann regions. Thus, except for the combination of Brodmann 28 and 36 which was not covered any of the lesion mask, a total of 819 lesion-deficit prediction models were established independently. Each model was performed by a mass-univariate (Fisher’s Exact Test) and a multivariate (SVM) technique respectively in order to illustrate how the complexity of lesion architecture and the inter-connection between dimensions will contribute to the lesion-deficit correlation for clinical prediction. The discrimination of separability was interpreted with a AUC-ROC curve.

2.3.2.1 Comparison of Fisher’s exact test against SVM

For correspondence with the series of simulation one, the whole set of 1172 lesion mask was randomly split into two subsets (70% and 30%): one was for training a classifier and the other subset independent from training was for the purpose of performance testing. The number of randomisations was 240. Iteratively, a Fisher’s Exact test was used in the mass-univariate analysis to generate a model for each combinatorial two Brodmann areas; whilst multivariate analysis was implemented by SVM. Thereby, a total of 196,560 (819 combinatorial dual Brodmann areas for 240 randomisations) mass-univariate analysis models matched with 196,560 models for multivariate analysis to compare the performance of clinical lesion-behavioural prediction. Although both approaches took less one minute to generate a model, in regards with the enormous total number of models, we used a Supermicro MicroCloud machine including 24 modular nodes which could conduct 12 models simultaneously in parallel. The entire process of evaluation took approximately ten calendar days.

In the case of mass-univariate training procedure, voxel-wise p values derived from Fisher's Exact test indicated the likelihood of how a lesioned voxel is independently associated with the hypothetical deficit. By multiplying each test lesion map with the set of p values and then using Fisher's method, the outcome predictions were generated. We applied Bonferroni correction (alpha level at 0.01) for multiple comparisons (Bonferroni, 1935, Bland and Altman, 1995) to ensure the voxels retained with reliable influence and significance above the defined threshold. In the multivariate analysis, a correlated weighting map derived from SVM training process was applied to each test lesion mask for prediction outcome. Preliminarily, the optimal model for each combinatorial dual Brodmann areas was confirmed with a search of the cost parameter (C), which assessed a total of 31 different values with a 10-fold cross-validation.

For vast majority (782 of 819, 95.5%) of the models depending on dual Brodmann regions, the prediction performed by a high-dimensional multivariate approach (SVM) was superior to a low-dimensional mass-variate technique (Fisher's Exact Test). The schema ball in Figure 2.5 shows the contrast of AUC value in each predictive model between Fisher's Exact Test and LIBSVM; and the brightness of the lines connecting any two Brodmann areas shows the extent of their differences. Among the models with advanced performance from SVM, the relative difference of AUC value in average was 0.027. Overall, the average AUC value derived from SVM was 0.987 (95%CI: 0.985 – 0.989); whereas the mean AUC score derived from Fisher's Exact Test was 0.963 (95%CI: 0.958 – 0.967) (Figure 2.6).

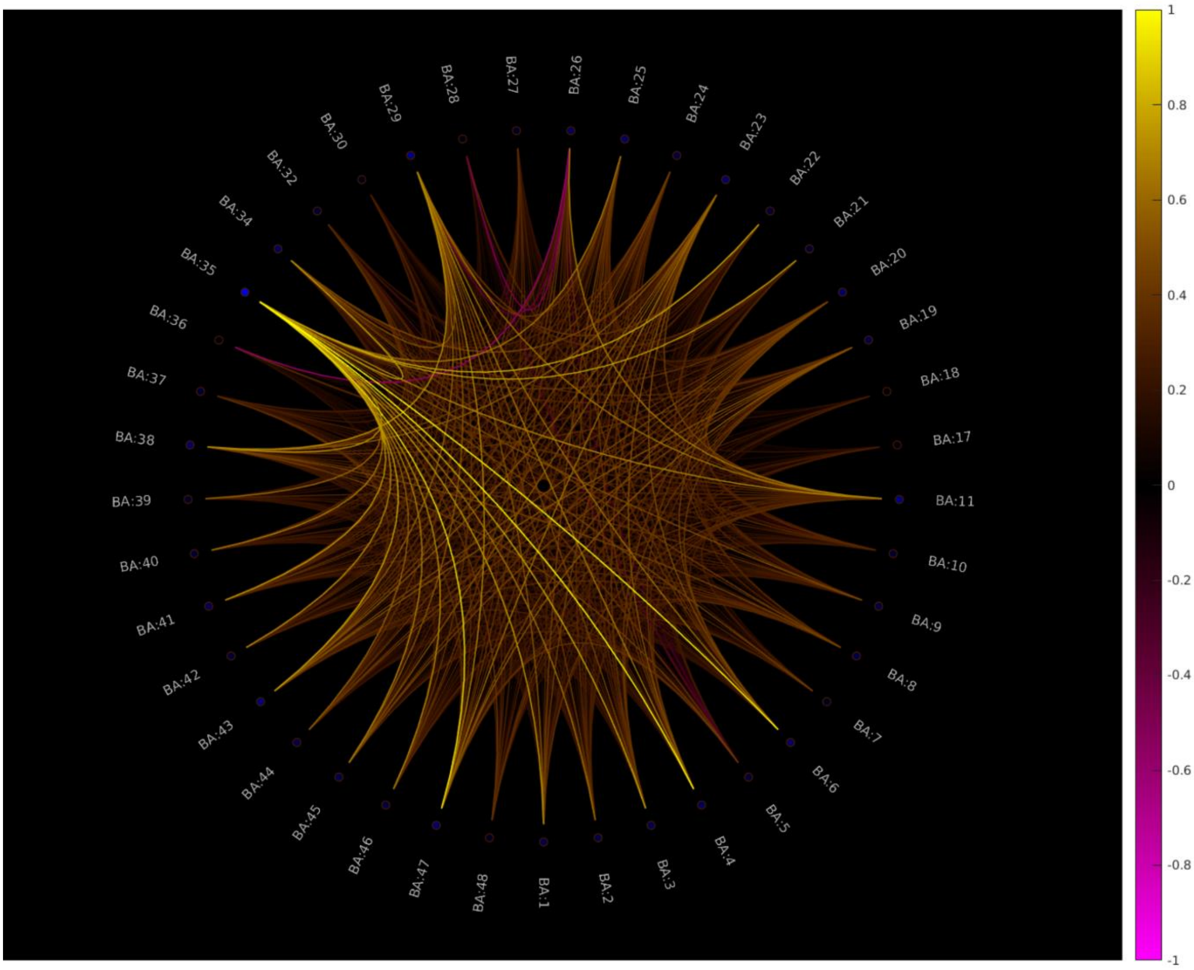


Figure 2.5 Differences between Fisher's exact test and LIBSVM across 819 combinatorial two Brodmann regions

The scheme ball illustrated the discriminations of predictive performance between mass-univariate and multivariate analysis, where majority of the case showed superiority achieved from multivariate analysis. Each line was connected with two Brodmann areas representing a combinatorial Brodmann pair. The brightness of the connected lines was referred as the extent of differences. As there was a total of 819 comparisons, in case there was any extreme case that might present relatively much larger differences than the other cases, the scale of differences marked in colour bar is the value of logarithm.

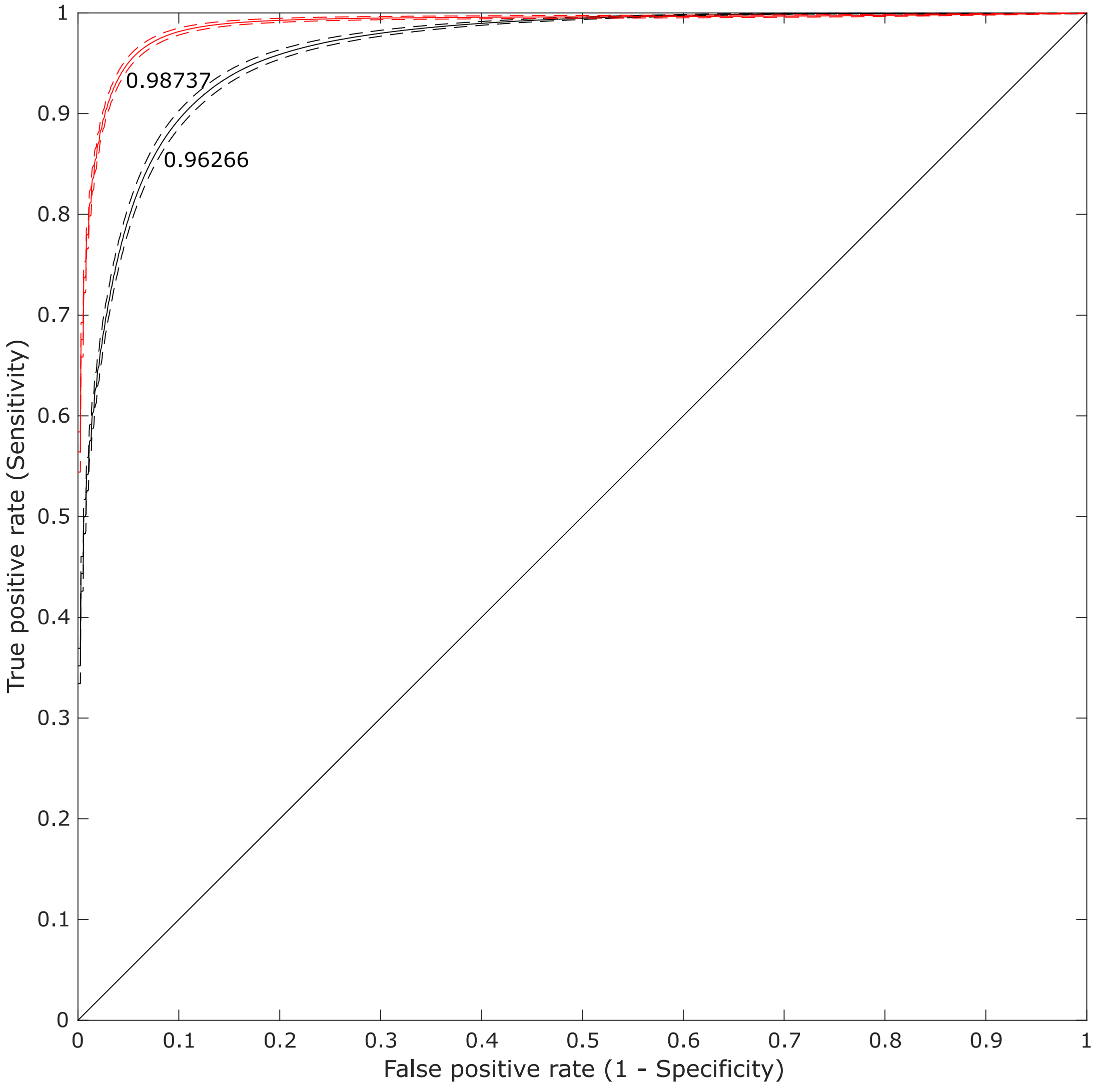


Figure 2.6 Comparison of ROC curve between Fisher's exact test and SVM

The ROC curves derived from both low- and high-dimensional models demonstrated the averaged predictive performance across 819 combinatorial Brodmann pairs. As a more complex model comparing to the case of single region, the extent of discriminations between two analysis methods was further enlarged. A total of 240 randomisations that split the lesion data for training and test purpose was iteratively conducted to enhance the performance stability and reliability.

2.4 Discussion

2.4.1 Simulation one: lesion-deficit prediction dependent on single Brodmann regions performed by a mass-univariate technique (Fisher's exact test) versus a multivariate technique (SVM)

In this series of simulations, a hypothetical deficit was determined by a single locus model dependent on 20% or more of a lesion mask overlapping with a specific Brodmann area. In order to illustrate the limits of lesion-behavioural prediction, every segmented region in Brodmann brain parcellation covered by our lesion dataset was modelled to compare the performance between mass-univariate against multivariate analysis. The result reveals the high-dimensional multivariate analysis that takes into account the complex lesion architecture and its inter-woven relations is overall superior in terms of prediction sensitivity compared with the conventional low-dimensional approach that relies on mass-univariate analysis.

The assumption behind the hypothetical behavioural outcome modelled here is sensitivity of the focal brain injury to the underlying neuroanatomy, reflected in the different lesion patterns in the affected group compared with the other. Thus, the simulations were modelled as a binary classification problem to discriminate two groups of lesion masks. The multivariate approach showed significant, but not huge superiority (95%CI: 0.986 – 0.990 vs. 95%CI: 0.978 – 0.983), because we idealised the conditions and simplified model complexity for prediction. In the mass-univariate test, how the set of identified discriminating voxels independently coincided with the causal relations of lesion-deficit could be reflected by how effectively the cluster of voxels indicated the membership of a lesion mask. With regards to the differential vulnerability induced by the vasculature (Sosa et al., 2014), the strong lesion bias contaminating the mass-univariate analysis tended to displace the identified cluster of discriminating voxels away from the truly critical locus. Instead, a set of parasitic voxels from "collateral" damage tended to be included in the prediction, leading to misclassification (Mah et al., 2014a, Mah et al., 2014b, Xu et al., 2017a). This illustrates the fundamental drawback of mass-univariate analysis, here extended to outcome prediction. Note the matter is not the magnitude of the difference between the two approaches but its very presence: these are idealised conditions.

Although voxel-wise lesion-symptom mapping was a step forward compared with the original region-of-interest based methods, such as template overlay (Frey et al., 1987) and overlap subtraction (Rorden and Karnath, 2004), there have been increasing criticisms on the sensitivity and validity of the mass-univariate approaches to lesion-deficit prediction in the recent years. The relationships between lesions and deficits are constitutionally non-linear and multivariate. The fundamental weakness of mass-univariate based VLSM is its inability to cope with complex inter-related patterns, which forces more and more studies turned to either region-wise or voxel-wise multivariate methods (Chechlacz et al., 2018, Husain and Nachev, 2007, Karnath and Smith, 2014, Lessov-Schlaggar et al., 2016, Mah et al., 2014b, Mah et al., 2014a, Norman et al., 2006, Pustina et al., 2018, Smith and Nichols, 2018, Smith et al., 2013, Xu et al., 2017a, Yourganov et al., 2016, Zhang et al., 2014, Zhao et al., 2017). As aforementioned, for lesion-behavioural prediction, two crucial aspects contribute to the complexity – the lesion and the functional architecture. With regards with the region-wise, or ROIs based multivariate analysis, the underlying network of human brain function is complex and distributed, which cannot be precisely segmented by any atlas reliably. Although the inter-region correlations can be taken into account, the region-wise, or ROIs based multivariate analysis is conducted in an ambiguous and coarse spatial granularity, in which spatial bias and reduction of predictive sensitivity may still be incurred by the potential uncertainty derived from brain segmentations.

Prior to any methodological consideration, the essence of the lesion architecture is intrinsically high dimensionality where the inter-voxel correlations may be critically related to the lesion-behavioural mapping. This high-dimensional problem inevitably requires a multivariate solution. The higher the resolution, the better the chance of identifying true lesion-deficit relations, though this necessarily incurs data and computational costs. Our results are consistent with previous studies that employed voxel-wise multivariate analysis which showed advanced performance in multivariate analysis compared with the conventional voxel-wise mass-univariate approach (Mah et al., 2014b, Zhang et al., 2014, Zhao et al., 2017). Going beyond these studies, we did not constrain the prediction to particular brain area or a fixed functional deficit; instead, we conducted the analysis on each Brodmann single area with a hypothetical deficit. Methodologically, we illustrated the discrimination between two approaches. Conceptually, we generalised the model to potentially fit clinical lesion-deficit prediction globally.

2.4.2 Simulation two: lesion-deficit prediction dependent on dual Brodmann regions performed by a mass-univariate technique (Fisher's exact test) versus a multivariate technique (SVM)

To provide greater fidelity to the likely underlying patters of functional-anatomical dependence, a second series of simulation based on two Brodmann areas were conducted. A total of 819 two loci models were established, covering any two Brodmann area combinations. Analogously to the first simulation, we also employed a hypothetical deficit dependent on at least 20% of a lesion mask falling within the specific pair of Brodmann areas. Thus, a two-class classification problem was modelled with mass-univariate and multivariate analysis, respectively. The overall performance derived from multivariate analysis was significantly superior to mass-univariate analysis. Compared with the contrast in simulation one, the extent of difference between two approaches was markedly magnified.

In the single locus model, prediction was artificially simplified to dependence on a relatively small, single region, and so the validity of mass-univariate analysis was affected by spatial bias to a limited extent; whereas a more complicated two loci scenario exposed the limitation of the mass-univariate analysis further. From the results, we can see the average performance of multivariate analysis in both series of simulation remained more or less the same (0.988 vs. 0.987). Whereas in the case of mass-variate analysis, there was an apparent difference between one-regional and two-regional models (0.981 vs. 0.963). This reflected the significant effects on the complexity of the underlying prediction models. Mass-univariate analysis was much less sensitive to model complexity. It is reasonable to expect multivariate analysis will present overwhelming advantages further if we were to simulate the hypothetical deficit in relation to three or more brain regions simultaneously.

As aforementioned, by contrast with most other studies (Pustina et al., 2018, Smith et al., 2013, Zhang et al., 2014, Zhao et al., 2017), we used a hypothetical deficit in both series simulations rather than a specific cognitive impairment or assessment. This enabled us to examine how any behavioural outcome that is sensitive to neuroanatomy will differ across different lesion patterns and modelling methods. The results in both series of simulations supported our intuition, that high-dimensional multivariate analysis, by explicitly modelling spatial interactions, could overcome the systemic bias incurred by mass-univariate analysis. Moreover, the essence of a large-scale dataset is not for improving marginal accuracy in mass-univariate models but for matching the dimensionality of intrinsically complex lesion architecture.

2.5 Conclusion

Lesion-deficit mapping plays an enormously important role in both system neuroscience and clinical behaviour prediction. The interactions between the complex lesion distribution and the underlying functional network are critical to relating neuroanatomy to function in the human brain. Our results showed the advanced fidelity of lesion-deficit prediction was achieved by a high-dimensional multivariate approach which explicitly modelled the spatial structure of natural focal lesions. The performance of high-dimensional modelling reflects its individuating power, which indicates the clinical applicability of lesion-deficit prediction.

The contrast between a high-dimensional multivariate approach against a low-dimensional mass-univariate approach was illustrated in two sets of simulations where affection of outcome followed a single locus and multi-locus distributed network, respectively. The multivariate approach achieved significant superiority to mass-univariate approach and remained more or less the same level of performance in both sets of simulations; whereas, the performance of mass-univariate analysis dropped remarkably in multi-locus models indicating the vital weakness to deal with complexity. With regards with the full complexity of lesion architecture, a high-dimensional multivariate framework associated with correlation based machine learning techniques is inevitable. Moreover, the dimensionality of lesion architecture is highly relied on a large-scale dataset. The more lesions are involved into modelling, the better characterisation of lesion distribution can be conducted; meanwhile, less dimensionality mismatch and extent of over-fitting happens. Any easy alternative with neglect of lesion dimensionality will lower the limits of lesion-deficit prediction.

It is crucial to realise that the complexities of the architecture of lesions is comparable to that of the macroscopic functional organisation of the brain. Crucially, to characterise lesion architecture with high-dimensional modelling is the prerequisite to capture the relationships between lesion and outcome not only for clinical prediction but functional mapping.

Chapter 3 The benefits of unlabelled data in lesion-deficit prediction

3.1 Introduction

The history of lesion-deficit studies can be traced back to the mid of 19th century, as we have already seen. The advent of non-invasive imaging techniques, such as computed tomography (CT) (Hounsfield, 1973) and Magnetic Resonance Imaging (MRI) (Lauterbur, 1973), has allowed brain lesions to be visualised, assessed, and quantified *in vivo*. Thus, a series of remarkable lesion-deficit associations have been made during the past few decades including emotion (Adolphs et al., 1994, Adolphs et al., 1995), decision-making (Bechara et al., 1994, Bechara et al., 2000), motor skills (Maldonado et al., 2008), intelligence (Barbey et al., 2012, Gläscher et al., 2010) and visuospatial processing (Kravitz et al., 2011).

Lesion-deficit mapping infers that a specific region is necessary for a function from its functional loss or impairment when the region is damaged. This is why lesion methods still retain their distinct value compared with other techniques, such as Transcranial Magnetic Stimulation (TMS) (Barker et al., 1985) and functional Magnetic Resonance Imaging (fMRI) (Ogawa and Lee, 1990, Ogawa et al., 1990b, Ogawa et al., 1990a, Ogawa et al., 1992), and compel us to return to this historic method. Yet, the conventional approaches widely applied to analyse lesion mapping models, particularly, voxel-based lesion symptom mapping (VLSM) have involved growing criticisms in the recent years (Mah, 2014a, Nachev et al., 2008, Xu et al., 2017a). The critical concerns, as we have seen, are derived from the intrinsic complexity of human brain damage which is derived from the anatomical relationships underlying the natural patterns of pathological damage. We have seen (Chapter 2) that the correct inferential model here must be able to handle multiple variables simultaneously so as to model anatomically inter-woven functional correlations and interactions; and meantime, the validity of an inferential model also requires to take into account sufficient variables to match the dimensionality of lesion architecture. Both aspects are indispensable – a high-dimensional multivariate analysis.

The dimensionality of lesion anatomy establishes the principal need for high-dimensional multivariate modelling for lesion-deficit prediction, which not only reflects the spatial validity of the model, but also achieves superior predictive fidelity, as we have seen. One previous study (Mah et al., 2014b) illustrated the high-dimensional models associated with multivariate analysis could remedy the systematically spatial biases identified in mass-univariate prediction models, which fundamentally proved the validity of inferential models for lesion-deficit prediction. Subsequently, another study (Xu et al., 2017a) demonstrated how a validated model would contribute to the fidelity of lesion prediction. In the study, Xu et al. established models for clinical lesion-deficit prediction based two typical pairs of Brodmann areas (areas 39 & 44; and 37 & 38), which indicated superior predictive sensitivity and specificity achieved by a high-dimensional multivariate models against the conventional mass-univariate model. In Chapter 2, we extended the previous study to model the prediction covering any single Brodmann area and fully combinatorial two regions, respectively, demonstrating the limits of clinical lesion-deficit prediction using a high-dimensional multivariate model compared with conventional mass-univariate models. The results showed the overall superiority of multivariate models that took into account the lesion dimensionality simultaneously and sufficiently.

A high-dimensional multivariate model is irreplaceable while modelling lesion-behavioural prediction; but we have to be aware of its dependence on sample size and computational power. Nowadays, with the rapid development of hardware, the need for intensive computation tends to be less of a problem. Nevertheless, a sufficient dataset with complete labelling of deficits as well as imaging is still a major challenge. Available resource and time will always limit data scale here, especially where data representative with respect to age, gender, ethnic, pathology and socioeconomic diversity is preferred. However, one of the key characteristics of clinical data is the difficulty of assuring completeness of each record where the data is not material to direct clinical care (Lee and Yoon, 2017). Large scale clinical datasets take a long time to assemble and cannot be easily tailored to research purposes. Specifically, for lesion studies, it is common that merely a proportion of the available imaging will be linked to behavioural outcomes, leaving many unlabelled scans. A learning technique that can take advantage of both labelled and unlabelled data to learn from relatively fewer labelled instances and complement the whole dataset with an extra set of unlabelled instances, will make more efficient use of available data. Theoretically and practically, semi-supervised learning methods are of great interest because the unlabelled data can be treated as additional optimisation variables, together with labelled data to reduce the variability of trained classifiers so as to stabilise and improve the performance (Peikari et al., 2018, Sindhwani and Keerthi, 2006, Zhu, 2006, Zhu and Goldberg, 2009, Chapelle et al., 2008).

In this chapter, we extend the previous chapter to employ a transductive semi-supervised SVM (Sindhwani and Keerthi, 2006, Sindhwani and Keerthi, 2007) in comparison with an ordinary SVM to investigate how the size of a labelled sample influences the fidelity of clinical lesion-deficit prediction; and the extent to which prediction models can benefit from a set of extra unlabelled samples. We adopt SVM for it is capable to training on sparse datasets with high-dimensionality and address the dimensionality mismatch between the number of training features and cases (Joachims, 1998, Saur et al., 2010). Semi-supervised SVMs can be understood as advanced modifications of regular SVMs, which maintain the original characteristics but extend the capacity to learn from unlabelled data points.

Specifically, two series of simulations are performed to compare the predictive performance of a purely supervised learning model and a semi-supervised model at different proportions of available labelled data. On the one hand, we would like to quantify how predictive performance could be improved with increasing the number of labelled lesion maps. On the other hand, for each fixed number of labelled lesions, we wish to quantify the benefit of adding an extra set of unlabelled lesions within a semi-supervised framework. For comparison with the two series of simulations in the previous chapter, a hypothetical behavioural deficit is introduced, based on a certain proportion of each lesion falling within either a specific Brodmann area or any combination of two Brodmann areas. Constrained with the volume of the available data and time, we randomly pick up five single-region models and ten dual-region models for evaluation.

3.2 Method

3.2.1 Patients

A set of 1333 patients admitted to University College London Hospital (UCLH) Hyper-acute Stroke Unit (HASU) over the past decade were selected. All patients were clinically diagnosed with acute ischemic stroke, and received MRI, including diffusion-weighted imaging (DWI), in which at least one acute ischemic lesion could be identified. A threshold for minimal DWI-quantified lesion volume was set at 50 milliliters which covered the 90th centile of our entire clinical population in order to reduce the disproportionate impact from very large lesions. An adjusted cohort including 1172 patients was thus produced.

The age distribution from the set of 1172 patients was in range of 18 to 97 years old with a mean of 63.89 and a standard deviation of 15.91. The gender ratio (Male) was 0.561 (Appendix B. Figure 7.1). Ethnicity for 952 (81.2%) from the full set of patients was disclosed in the clinical record which is illustrated in Appendix B. Figure 7.2. The distribution of age and sex, and the constitution of ethnicity reflected the catchment of clinical population in UCLH naturally.

3.2.2 Imaging

3.2.2.1 MR data acquisition

All acquisitions performed on the scanners manufactured by General Electric (GE) (Discovery MR 450, Genesis Signa and Signa Excite), Philips (Achieva, Ingenia and Intera), or Siemens (Avanto, Biograph mMR, Espree, Skyra, Symphony, Symphony Tim, Trio Tim and Verio) with field strength of either 1.5 or 3 Tesla in a single session for each patient were collected via a picture archiving and communication system (PACS) from UCLH (the specific range of scanners and related parameters are listed in appendix A). The originated voxel sizes were different in some extent between the scanners, and so were the according spatial resolutions. Typically, for 1.5 Tesla GE medical system, the voxel size was sampled as 1mm x 1mm x 6.5mm (Columns = 256, Rows = 256). For 1.5 Tesla Philips

medical system, the typical voxel size was 1.2mm x 1.2mm x 6mm (Columns = 192, Rows = 192); the voxel size was 1mm x 1mm x 6mm (Columns = 256, Rows = 256) performed on 3 Tesla Philips scanners. For Siemens, the voxels sizes and spatial resolution were 1.8mm x 1.8mm x 6.5mm (Columns = 128, Rows = 128) and 1.2mm x 1.2mm x 6.5mm (Columns = 192, Rows = 192) on 1.5 Tesla and 3 Tesla scanners, respectively. The routine of clinical practice with regular equipment upgrades and replacement is the main reason for the diversity of scanners over the period of data collection. All brain scans were performed for the purpose of clinical routine obeying clinical protocols.

The specific sequences extracted from each MRI session were axially acquired echo planar DWI with b-values of 0 s/mm^2 and 1000 s/mm^2 which is used for lesion segmentation (Mah et al., 2014c) so as to implement the series of simulations designed for this study. DWI is able to indicate the restricted ability of Brownian motion for extracellular water protons where cytotoxic edema causes imbalance. So, DWI is sensitive for detecting ischemic stroke and widely used for locating acute ischemic lesions (Warach et al., 1995, Löuvbld et al., 1997, Lövblad et al., 1998). In the clinical application, a DWI sequence usually includes images with b values of 0 s/mm^2 (a b0 image) and 1000 s/mm^2 (a b1000 image). The former type image shows stronger contrast on normal tissues but relatively insensitive to ischaemia. Conversely, an image with high diffusion weighting is used to sense water molecules movement. So, it is sensitive to ischaemia but poor to the contrast of normal tissues. Both types of images are complementary for brain registration and lesion segmentation, which rely on the contrast between normal tissue types and the discrimination between normal tissues and lesions, respectively.

3.2.3.2 Image pre-processing

A processing pipeline was implemented within MATLAB (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.) associated with a configured SPM12 toolbox (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/) to ensure the images to be processed systematically.

For the purpose of scientific imaging analysis, typically, the raw clinical MR images needed to be firstly converted from the complicated format of Digital Imaging and Communications in Medicine (DICOM) images to Neuroimaging Informatics Technology Initiative (NIfTI) format. Subsequently, in order to provide accurate and robust performance on image registration and segmentation in the later stage, we employed a pre-processing pipeline with the following steps. We empirically clamped the signal between 0.1% and 99.9% of the cumulative distribution estimated with a kernel density method (Botev et al., 2010), removing unusually low and high signals in each NIfTI file. Then an oracle-based 3D discrete cosine transform (ODCT3D) (Manjón et al., 2012) denoising method was applied for further noise reduction. In the next step, image registration was conducted based on each pair of DWI files. In order to optimise the alignment between the b0 and b1000 images in each pair, we used SPM12's standard co-registration function to have the b0 image rigidly co-registered to the b1000 image. Then, the normalisation / segmentation routine integrated in SPM12 (Ashburner and Friston, 2005) was applied to the b0 image to calculate a deformation field which described the optimal non-linear transformation of the b0 image into Montreal Neurological Institute (MNI) stereotactic space. Next, the deformation field derived from the b0 image was applied to the b1000 image to transform it into MNI space using the same routine. The parameters involved in the whole process of image registration were set at defaults by SPM12. Finally, both b0 and b1000 images in each pair were resampled to 2mm^3 isotropic with 6th degree b-spline interpolation (Mah et al., 2014b, Mah et al., 2014c) and manually checked against the SPM template to confirm the satisfactory of registration.

3.2.2.3 Lesion segmentation

A previously validated method based on the anomaly metric *zeta* (Mah et al., 2014c) was applied to segment lesions from the 1172 normalised b1000 images in a voxel-wise manner.

A set of preliminaries conducted to optimise the performance of lesion segmentation were performed as follows. First, 492 DWI images derived from patients suspected of acute stroke but found to have normal imaging were selected

to form a reference normal image set. No acute lesion was visible here in any patient, and none was reported by the corresponding radiological record. Every scan was pre-processed using the exact same pipeline and resliced at the same resolution of 2mm isotropic voxels, which spatially presented as a 91 x 109 x 91 matrix. Then, by combing the tissue maps provided by SPM12 including white matter tissue, grey matter tissue and cerebrospinal fluid (CSF), an inclusive brain mask was created. So as to remove image artefacts, we set an empirical threshold to exclude areas in the frontal and temporal poles commonly prone to artefacts on DWI sequences, and adjusted the inclusive brain mask accordingly. Subsequent lesion segmentation was performed within the confines of this adjusted mask. Next, the signal distribution was normalised for both lesion and reference datasets. Such signal normalisation sought to reduce instrumental and other incidental signal heterogeneity, so that the comparison between individual images could be maximally sensitive. The procedure was performed as follows. A binary white matter mask was created from SPM12's white matter tissue probability map by thresholding at a value of >0.9. A robust kernel density estimate method (Botev et al., 2010) was then applied to the voxels falling within both customised maps in each b1000 image to obtain the peak white matter distribution which was then subtracted from each corresponding b1000 image to normalise the signal distribution. Guided by information in the radiological report, signal normalisation on unilateral and bilateral lesions was performed differently. For unilateral lesions—the vast majority—only the unaffected hemisphere was used to estimate the peak. For bilateral lesions, since ischemic lesions tended to be small in these circumstances, the kernel density estimate was applied to the whole brain. In both cases, the signal normalisation was equally effective in adjusting the differences in the signal distribution across all scans performed by a variety of scanners and related technical and practical variance.

Lesion segmentation was performed in a voxel-wise manner. We calculated the zeta anomaly metric for each single voxel in each image independently against the reference set. To be specific, the voxel-wise zeta value in the test image is derived from the mean distance to the k nearest neighbors drawn from the anatomically homologous voxels in the unlesioned reference set, normalised by the mean distance between the k neighbors themselves (Mah et al., 2014c). The measure was the Euclidean distance in signal space. The only manipulable

parameter here is the number of nearest neighbors – k – commonly chosen with reference to the size of the reference dataset. In this study, we set k at 23 based on the calculation as below:

$$k = Ceiling\left(\sqrt{492}\right).$$

Zeta is a continuous metric. An adaptive threshold was therefore determined for each image. In brief, volume connectivity computed by *bwlabeln*, an integrated function of MATLAB (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.) was applied to segment each lesion map into connected clusters. Then, each cluster was estimated under a generalised extreme value (GEV) distribution to derive mean and variance values which were used to decide threshold adaptively as described in Mah et al, 2014.

3.2.3 Dimensionality reduction

The predictive performance of a model is often dependent on the ratio of the number of instances to the number of material features. It is therefore helpful to reduce the dimensionality of individual features, while retaining their diversity. In human neuroanatomy, brain asymmetry between two hemispheres has been observed in two distinct aspects – neuroanatomical differences, and functional and behavioural lateralisation (Hugdahl, 2005, Toga and Thompson, 2003), but these differences are minor and do not generally extend to the vascular tree (Wright et al., 2013). Vascular lesions rarely cross the midline of brain and are generally literalised (Nachev et al., 2008). It is therefore reasonable to collapse our stack of vascular lesions onto one hemisphere to reduce dimensionality for following predictive models.

3.2.4 Technique

In the present chapter, we employ SVMlin (Sindhwani and Keerthi, 2006, Sindhwani and Keerthi, 2007) for supervised and semi-supervised modelling. TSVM is a natural extension of the original SVM to model binary classification problems in a semi-supervised manner. In brief, TSVM was initially proposed to drive the optimised hyperplane with labelled data points alone towards regions with lower data density by appending an extra set of data points without labels. Further details about the theory of transductive learning has been mentioned in Chapter 1 and can also been found in previous studies (Bennett and Demiriz, 1999, Fung and Mangasarian, 2001, Joachims, 1999b). As a variation of TSVM, SVMlin can provide superior scalability and a multiple switch heuristic to not only optimise but also speed up the training process. Moreover, it is capable of coping with large scale datasets exhibiting sparsity which fits the properties of our lesion set well.

3.2.5 Simulations

The previous studies (Mah et al., 2014b, Xu et al., 2017a) and previous chapter have proven the modelled dimensionality of the lesion architecture plays a vital role for both functional localisation and deficit prediction. In this chapter, we quantify the benefit of adding unlabelled data to a lesion-deficit prediction problem within the framework of semi-supervised learning.

3.2.5.1 Simulation one: single Brodmann area based lesion-deficit prediction performed by a supervised learning technique (SVM) versus a semi-supervised learning technique (transductive SVM)

3.2.5.1.1 Data preparation

The set of 1172 segmented lesion maps were binarised to denote each voxel in the brain as being hit or not hit for the purpose of subsequent simulations. All lesions were collapsed onto the right hemisphere, then a threshold was applied to ensure every lesion voxel was collectively presented at least four times. A data matrix involving 1172 patients with 2640 features at the voxel size of 6mm was created for modelling analysis. The lesion distribution (Figure 3.1) indicated the voxel-wise hit frequency followed a Gaussian distribution which potentially supported us to adopt a transductive SVM (TSVM) to identify the benefits from extra data points during the optimisation procedure.

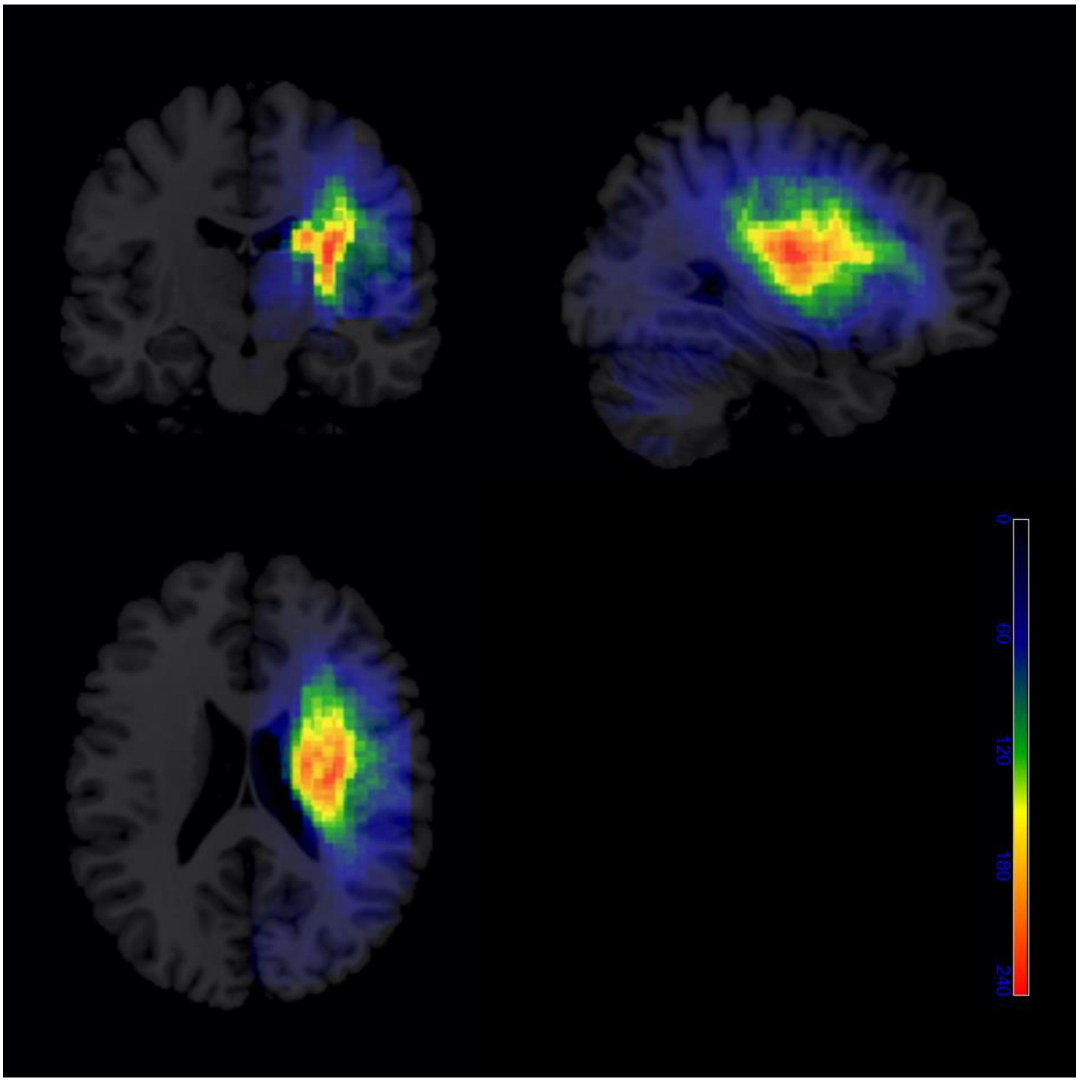


Figure 3.1 Lesion distribution

1172 lesions collapsed on the right hemisphere and were summed in a template in voxel size of 2mm. The colour spectrum demonstrated the density of lesion voxels distribution.

3.2.5.1.2 Lesion-deficit models for clinical prediction

As in the previous chapter, in this series of simulations, we posited a binary behavioural outcome that was sensitive to the underlying neuroanatomy. The lesion masks were divided into two groups, differing in their lesion patterns according to their intersection with each Brodmann area over separate experiments. The ground truth for labelling ("affected" or "unaffected") depended on whether or not at least 20% of a lesion fell within a specific single Brodmann area. 10% noise was introduced to make the hypothetical deficit more realistic. In other words, there was a 90% possibility that 20% or more involvement of a specific Brodmann area by a lesion would lead to a deficit.

In order to demonstrate the effect of unlabelled dataset while correlating the hypothetical functional deficit and lesion maps, the way to establish the prediction models were considered with two aspects. On one hand, a set of proportions were pre-defined to divide the entire lesion dataset into labelled and unlabelled sample sets, respectively. In this way, we expected to examine how the proportion of unlabelled examples in a dataset would contribute to the prediction. On the other hand, for each above proportion, we implemented supervised learning models with the labelled instances alone; and semi-supervised learning models with both labelled and unlabelled instances. This is to discriminate learning methods.

The set of 1172 lesion masks covered a total of 39 Brodmann areas. For each proportion of labelled lesion in each area, the lesion-behaviour correlation was independently modelled by a supervised (SVM) and a semi-supervised (TSVM) technique, respectively.

3.2.5.1.3 Supervised learning models (SVM)

SVMlin (Sindhwani and Keerthi, 2006, Sindhwani and Keerthi, 2007) was used to solve the two-class classification problem within a supervised learning framework. In order to quantify how the number of labelled data instances can contribute to prediction performance, 30%, 45%, 60%, 75% and 100% of the images used for training were labelled. In order to optimise predictive performance, a 10-fold cross-validation was used to search for the optimal regularisation parameter (lambda) across 9 levels of lambda values in range from 10^{-3} to 10^{5} (The step was 10 times.). The regularisation parameter that yielded the best classification performance from the cross-validation process was applied to the training dataset to generate the optimal predictive models. The classifier created by the optimised model was voxel-wise in which each voxel was weighted in terms of polarity and magnitude representing to the membership of either class and the significance of the specific voxel in relation to the behavioural deficit, respectively. Subsequently, this weighted classifier was used for prediction on the test dataset.

As mentioned above, the classifiers were trained with a set of different proportions randomly extracted from the whole lesion set, which required a relatively balanced label distribution for any specific Brodmann area. Meantime, the time expense was also considered. Thus, 5 Brodmann areas from a total of 39 areas were randomly selected to model the lesion-deficit prediction at 5 different portions of samples. In each Brodmann area, each proportion of samples were randomly split into 70% and 30% for training and test, separately. The number of randomisations for the split of training of test was 120 times. Hence, a total of 3000 (5 Brodmann areas x 5 portions x 120 iterations) high-dimensional multivariate analysis models were established for evaluating the prediction performance with a supervised learning method. The whole process took approximately 15 hours.

3.2.5.1.4 Semi-supervised learning models (TSVM)

SVMlin (Sindhwani and Keerthi, 2006, Sindhwani and Keerthi, 2007) can model a two-classification problem within a semi-supervised framework (TSVM) by appending the training data with an extra set of instances labelled as neutral (0). In addition, the maximum number of switches can be set empirically during the optimisation process. For direct comparison with supervised models, in the series of semi-supervised models, we still set five proportions of labelled data across 30%, 45%, 60%, 75% and 100%, leaving the remaining lesion maps as unlabelled data. We can now quantify the impact of adding unlabelled data, as a function of its relative proportion.

The optimisation procedure was also accomplished by 10-fold cross-validation, where the optimal regularisation parameter (lambda) was searched in the range as follows:

$L = 10^m$, where m is integer in the range between -3 and 5.

The maximum number of switches was derived from the proportion of unlabelled data points. Specifically, the optimisation was across a set of proportions between 0.5 and 1, with a step of 0.1. The optimal parameters were then used to generate the final classifiers.

The same 5 areas randomly chosen for the supervised learning models were used to predict the lesion-deficit correlation at 5 different proportions of unlabelled samples. Train/test splits were 70% and 30%, as before. The number of randomisations for each split was 120 times. Hence, a total of 3000 (5 Brodmann areas x 5 proportions x 120 iterations) semi-supervised learning models were established for evaluating the prediction performance. The whole process costed approximately 30 hours.

3.2.5.1.5 Performance comparison between supervised learning and semi-supervised learning approaches

By learning a set of purely labelled data and a mix of labelled and unlabelled data respectively, both methods generated a stack of discriminating voxels in the form of a probability brain map that can subsequently be used as a classifier to predict the hypothetical deficit.

In order to evaluate the predictive abilities of each classifier, the predicted values on the test dataset were used to calculate the related evaluation metrics including accuracy, sensitivity, specificity and balanced accuracy. The details are as follows:

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{the total number of cases}}\ ;$$

$$\begin{aligned}\text{sensitivity} &= \frac{\text{number of true positves}}{\text{number of true positives} + \text{number of false negatives}} \\ &= \frac{\text{number of true positives}}{\text{the total number of "affected" cases}}\ ;\end{aligned}$$

$$\begin{aligned}\text{specificity} &= \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}} \\ &= \frac{\text{number of true negatives}}{\text{the total number of "unaffected" cases}}\ ;\end{aligned}$$

$$\text{balanced accuray} = \frac{\text{sensitivity} + \text{specificity}}{2}.$$

As aforementioned, the set of simulated lesion-deficit prediction models began with a relatively small proportion of labelled data instances randomly extracted from the stack of 1172 lesions; then increased the proportions incrementally in order to identify the benefits from the size of labelled training dataset and an additional set of unlabelled data points. As a result, our simulations were inevitably to some extent constrained by imbalances in the label distribution. We therefore employed balanced accuracy (BAC) as one of the evaluation metrics, which not only allowed for the deviation of meaningful confidence intervals but avoided the over optimistic estimates derived from a biased classifier owing to label imbalance (Brodersen et al., 2010).

In addition, the judgement of discrimination between supervised and semi-supervised analysis was also interpreted and visualised using a receiver operating characteristic (ROC) curve and the area under curve (AUC) (Hanley and McNeil, 1982, Pauker and Kassirer, 1975, Swets, 1973, Thornbury et al., 1975). ROC curve plotted with true positive rate (sensitivity) on y-axis against false positive rate (1 - specificity) on x-axis illustrates the capacity to distinguish two classes at all classification thresholds. AUC stands for the area under a ROC curve, which aggregates the performance of classification at all thresholds. So, the higher AUC is, the stronger separability of the classification model is. The confidence interval at the level of 95% was applied to ROC curves for both methods to demonstrate the discrimination reliably. In this series of simulation, the AUC scores and ROC plotting were associate with a performance metrics software – PERF (http://osmot.cs.cornell.edu/kddcup/software.html).

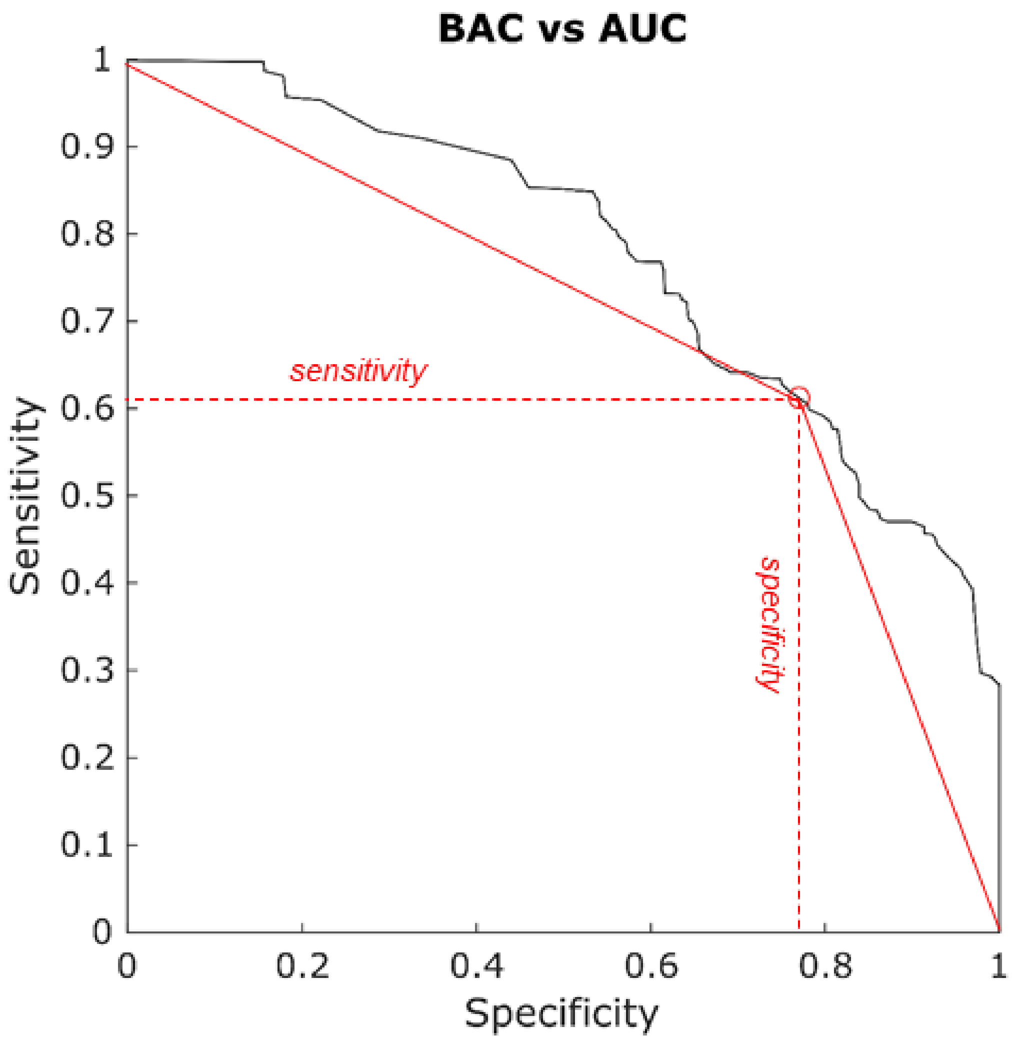


Figure 3.2 Balanced accuracy and AUC

BAC is special occasion of AUC. It is averaged by sensitivity and specificity by thresholding the prediction value at zero.

To be specific, for both supervised and semi-supervised models, evaluations were conducted to examine if a larger proportion of labelled data points could produce better predictive performance. Furthermore, for each specific proportion of labelled data instances, the evaluations mentioned above were performed to identify whether or not the semi-supervised models were superior to the supervised models overall by learning an additional set of data instances without a label.

3.2.5.2 Simulation two: dual Brodmann area based lesion-deficit prediction performed by a supervised learning technique (SVM) versus a semi-supervised learning technique (transductive SVM)

3.2.5.2.1 Data preparation

The same stack of lesion masks as described in simulation one was used to perform this series of simulations. In total, 1172 binarised lesion maps collapsed onto the right hemisphere, forming a data matrix with 2640 dimensions. All features were positive in at least four instances amongst the total 1172 lesion maps.

3.2.5.2.2 Lesion-deficit model for clinical prediction

In this series of simulations, we extended the single Brodmann area based lesion-deficit models to dual Brodmann areas in order to demonstrate the discrimination between supervised and semi-supervised analysis while dealing with a more complex problem.

Realistically, any given function or behavioural deficit is likely to be dependent on multiple loci (Mah et al., 2014b, Power et al., 2011, Fox, 2018). Thus, in the series of simulations, each lesion mask was labelled as “affected” or “unaffected” dependent on whether or not 20% or more of it fell within a specific pair of Brodmann regions. A minimum 20% volume of the lesion could solely fall within either of the specific dual Brodmann areas or cover both Brodmann areas. 10% noise rate was applied to render the hypothetical deficit of interest more realistic. Thus, the established models illustrate a 90% possibility that at least 20% involvement of a specific dual Brodmann areas by a lesion would result in the behavioural deficit.

We used a Brodmann brain map including forty-one areas. The number of fully combinatorial dual areas is 820. Since Brodmann areas 28 and 36 were not covered by any of the 1172 lesion masks, a total of 819 combinatorial regions were used for modelling supervised and semi-supervised analysis, respectively.

3.2.5.2.3 Supervised learning models (SVM)

To relate the lesion masks and the hypothetically defined behavioural deficit, SVMlin (Sindhwani and Keerthi, 2006, Sindhwani and Keerthi, 2007) was used to classify the lesion patterns in a supervised learning manner. We set five different proportions of labelled data ranging from 30%, 45%, 60%, 75% to 100%. Each set of extracted lesions was used to learn a classifier separately. During optimisation process, a 10-fold cross-validation was used to tune regularisation parameter (lambda) in the range of -3 to 5 to the power of 10 ($\mathrm{C} = 10^{\mathrm{m}}$, where m is integer between -3 and 5). The optimal regularisation parameter was then used for training process. The trained classifier created by the optimised model was voxel-wise, weighting in terms of polarity and magnitude the membership of either class and the significance of the specific voxel in relation to the behavioural deficit, respectively. This weighted classifier was subsequently used for prediction on the test dataset.

As outlined above, the labelled proportion started at a low fraction—30%—complicating the task of achieving reasonable class balance across each pair of Brodmann areas. Here 10 pairs of Brodmann areas from a total of 819 combinations were randomly selected. In each two-region model, each proportion of samples was randomly split into 70% and 30% for training and test, separately. The number of randomisations for the split of training of test was 120 times. Hence, a total of 6000 (10 combinatorial two-region x 5 portions x 120 iterations) models were run. The whole process costed approximately 30 hours.

3.2.5.2.4 Semi-supervised learning models (TSVM)

By mixing labelled and unlabelled data together, SVMlin (Sindhwani and Keerthi, 2006, Sindhwani and Keerthi, 2007) takes into account both types of data to optimise the hyperplane, performing classification in a semi-supervised manner. For consistency with the previous experiments, in this series of semi-supervised models, five proportions ranging from 30%, 45%, 60%, 75% to 100% were evaluated, treating the remainder of the data as unlabelled. On one hand, it was available to identify if a larger proportion of labelled samples included in a mixed dataset would improve training performance.

As before, the optimisation procedure was also accomplished by 10-fold cross-validation, where the optimal regularisation parameter (lambda) was searched in the range as follows:

L = 10^m, where m is integer in the range between -3 and 5.

The maximum number of switches was across a set of proportions between 0.5 and 1, with a step of 0.1. The optimal parameters were then used to generate the final classifiers.

The same 10 randomly chosen pairs of Brodmann areas were used, with the same random split into 70% and 30% for training and test, respectively. The number of randomisations for the split of training of test was 120. Hence, a total of 6000 (10 combinatorial two Brodmann areas x 5 portions x 120 iterations) semi-supervised learning models were used to evaluate performance. The whole process took approximately 60 hours of computational time.

3.2.5.2.5 Performance comparison between supervised learning and semi-supervised learning approaches

For each two-region model, by learning a subset from the whole training dataset at five pre-defined proportions, both supervised and semi-supervised analysis generated probability brain maps used for classifying an independent test dataset.

As before, accuracy, sensitivity, specificity, and balanced accuracy were used to evaluate performance. Balanced accuracy was used to permit the derivation of meaningful confidence intervals and avoid over optimistic estimates derived from a biased classifier in the context of class imbalance (Brodersen et al., 2010). In addition, AUC-ROC curves were plotted to interpret and visualise the discrimination between both analyses at a set of varying thresholds (http://osmot.cs.cornell.edu/kddcup/software.html). Confidence intervals were set at 95%.

3.2.6 Software and hardware

All simulated prediction models were implemented in the environment of MATLAB (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.).

For both sets of simulations, a 12-core Inter® Xeon® CPU E5-2620 2.00GHz processor with 64GB RAM and 9TB 7200 RPM SATA hard drive was used to conduct both SVM and TSVM models under a 64 bit Linux operation system (Ubuntu version 15.04).

3.3 Results

3.3.1 Simulation one: single Brodmann region based lesion-deficit prediction performed by a supervised approach (SVM) versus a semi-supervised approach (TSVM)

The stack of 1172 lesion masks were labelled with a hypothetical functional deficit dependent on a certain proportion of each lesion overlapping with a specified Brodmann region. Thus, except for Brodmann areas 28 and 36, which were not covered any of the lesion mask, a total of 39 single Brodmann regions were prepared for building up lesion-deficit prediction models. In this series of simulations, the prediction was performed by a supervised learning method (SVM) and a semi-supervised learning method (TSVM), respectively. Within each learning approach, we would like to quantify how the number of labelled data instances included in the training procedure would contribute to classification fidelity. We also wish to quantify the benefit from an additional set of unlabelled data across the two approaches.

3.3.1.1 Performance comparison between supervised learning and semi-supervised learning models

Lesion-deficit prediction based on five randomly sampled single Brodmann areas was performed with five subsets randomly extracted from the set of 1172 lesion maps. Within each subset, 120 randomisations were split into 70% and 30% for training and testing, respectively. Hence, the results from a total of 3000 supervised learning models and 3000 semi-supervised learning models were evaluated.

For each chosen area, the balance of “affected” vs “unaffected” depended on the lesion distribution, requiring the use of balanced accuracy as the key metric of fidelity. The results are shown in Figure 3.3. Overall, the semi-supervised models were superior to the supervised models at the 95% confidence interval (Table 3.1). The largest difference (0.712±0.015 vs. 0.752±0.015) between the two approaches occurred within the smallest proportion (30%), and diminished with increasing proportions, emphasizing the importance of unlabelled data were the labelled number is relatively small.

In the purely supervised learning models, performance rose with training data size, as might be naturally expected. In the semi-supervised learning models, the picture was more complex, exhibiting a plateau over the intermediate proportions.

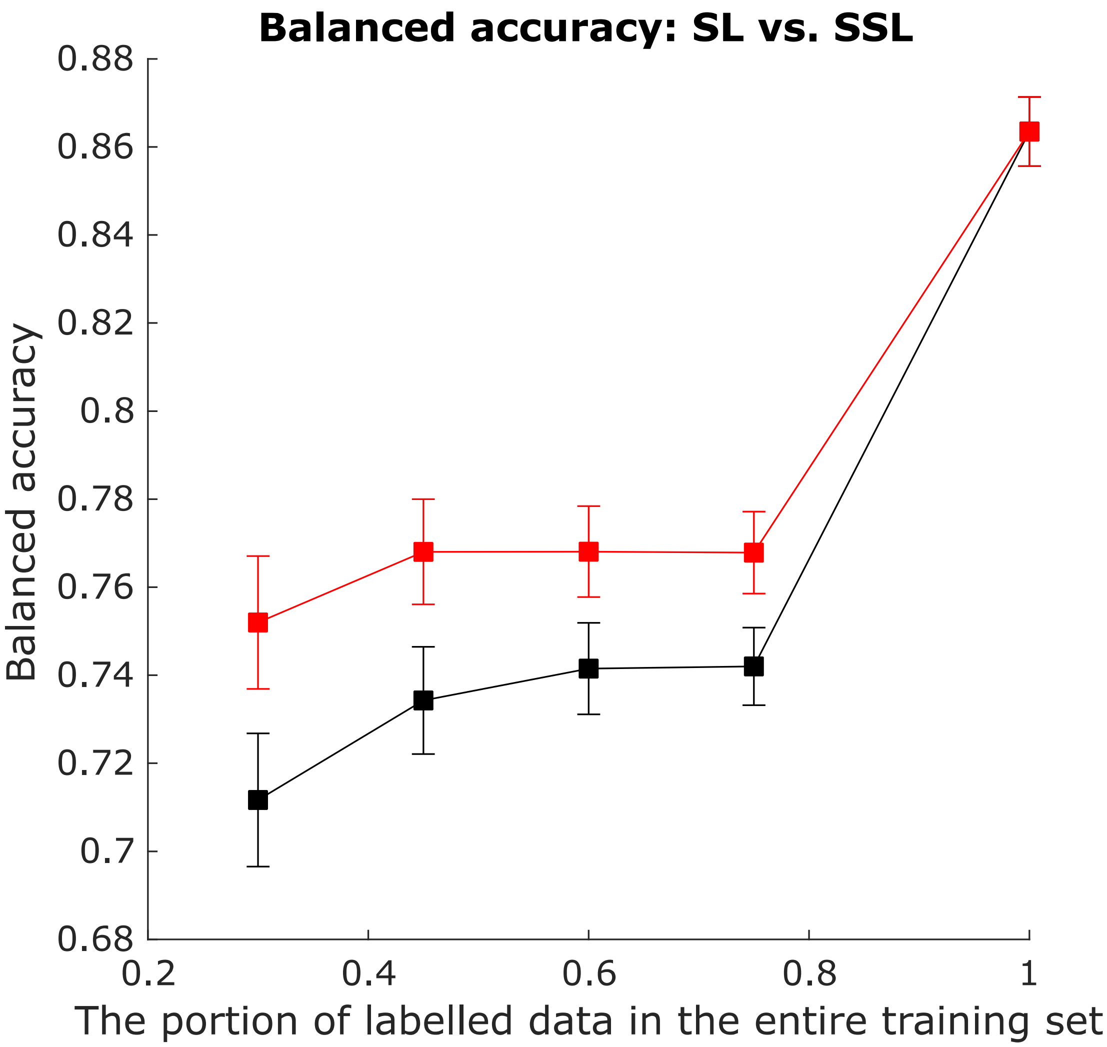


Figure 3.3 Balanced accuracy

Balanced accuracy was used for estimating the generalizability between supervised and semi-supervised learning methods. The estimates were derived from five single Brodmann regions sampled from the total 39 regions at five pre-defined portions of subsets. For each region of each proportion of subset, 120 randomisations were iteratively conducted to train and test with difference set of lesion data. The predictive performance achieved by the supervised learning method is in black, the semi-supervised in red. Error bars show 95% confidence intervals.

	30%	45%	60%	75%	100%
SL: balanced accuracy	0.712±0.015	0.734±0.012	0.741±0.010	0.742±0.009	0.863±0.008
SSL: balanced accuracy	0.752±0.015	0.768±0.012	0.768±0.010	0.768±0.009	0.863±0.008

Table 3.1 The details of balanced accuracy

Given the class imbalance, balanced accuracy values were helpful here in illustrating the contrast between supervised and semi-supervised analysis methods, as well as the contribution of labelled and unlabelled data points in each learning approach. To extend the interpretation, the ROC curve and related AUC scores were also computed for supervised and semi-supervised simulations at each proportion. Rather than choosing the prediction threshold at zero in the metric of balanced accuracy, ROC curves were computed with a set of varying thresholds to visualise the impact of threshold choice on classification fidelity.

Overall, the results of ROC curves (Figure 3.5 to 3.9) and related AUC scores (Figure 3.4) showed the semi-supervised models to be superior to the supervised models. Compared with the evaluation with balanced accuracy, two slight differences were identified in the measurement of AUC scores. First, the performance of the semi-supervised models at the proportions of 30%, 45% and 60% was not distinguished from the corresponding supervised models at the 95% confidence interval, but a clear distinction was found at 75%. This may have reflected the instability of prediction performance at small scales of labelled data in the presence of class imbalance. Performance variability fell with increasing training data size, as indicated by descending standard error rates across the five pre-defined proportions, where the error rates dropped from 0.018 to 0.002 in supervised models and the range of error rates in semi-supervised models was between 0.017 and 0.002 (Table 3.2). For the three intermediate ranges of proportions (45%, 60% and 75%), the measurement of balanced accuracy was estimated to be the same, while the AUC scores were subtly improved (0.893, 0.897 and 0.902). This is consistent with the expected pattern.

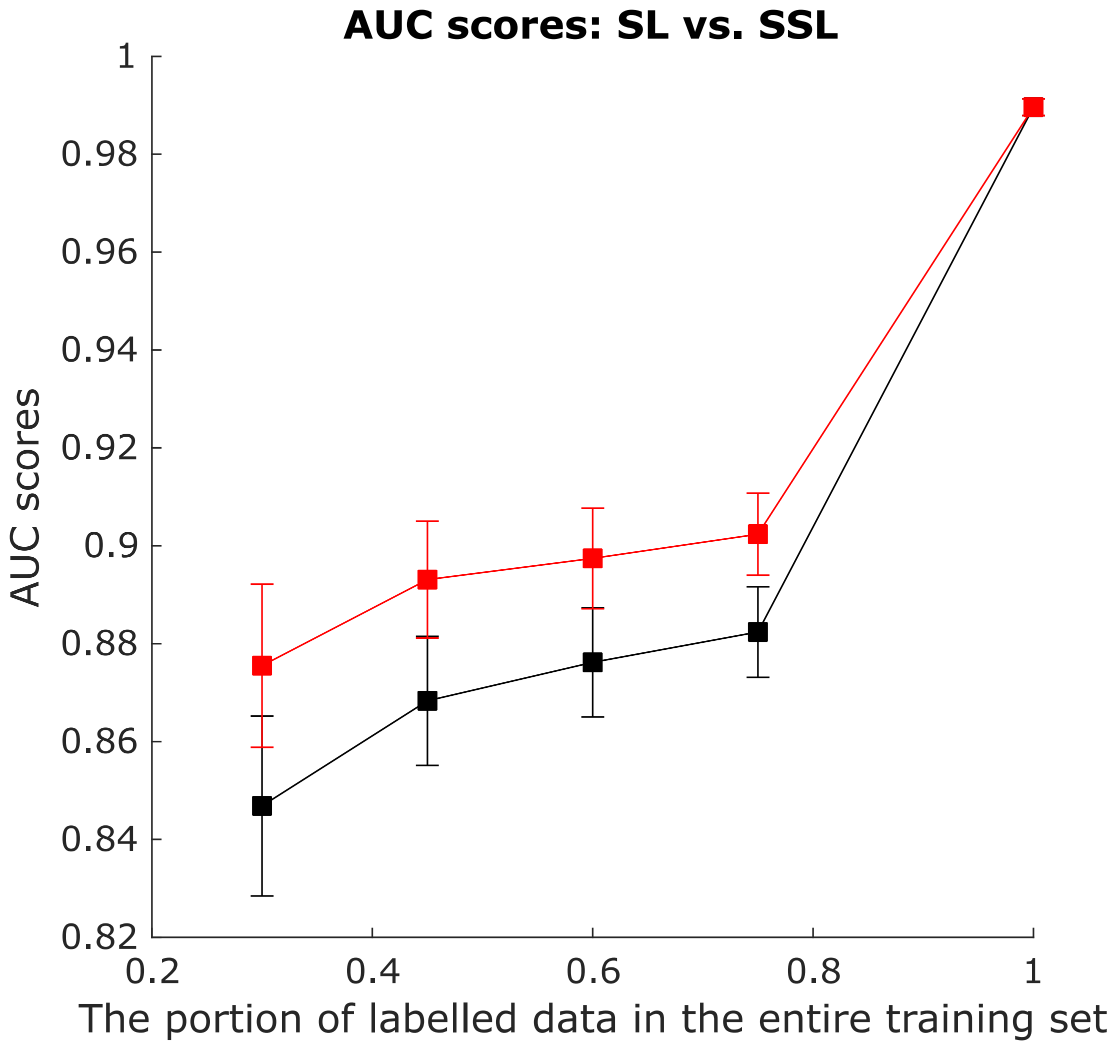


Figure 3.4 AUC scores

AUC scores were calculated to estimate the generalizability between supervised and semi-supervised learning methods with an extensive range of prediction value. The estimates were derived from five single Brodmann regions sampled from the total 39 regions at five pre-defined portions of subsets. For each region of each proportion of subset, 120 randomisations were iteratively conducted to train and test with difference set of lesion data. The predictive performance achieved by supervised learning method is in black, semi-supervised in red. Due to limitation of our lesion data, the improvement was not consistently significant across all the thresholds in particular that the proportion of subset was relative small, where the distribution of labelled samples tended to be seriously imbalanced.

	30%	**45%**	**60%**	**75%**	**100%**
SL: AUC scores	0.847±0.018	0.868±0.013	0.876±0.011	0.882±0.009	0.990±0.002
SSL: AUC scores	0.875±0.017	0.893±0.012	0.897±0.010	0.902±0.008	0.990±0.002

Table 3.2 AUC scores

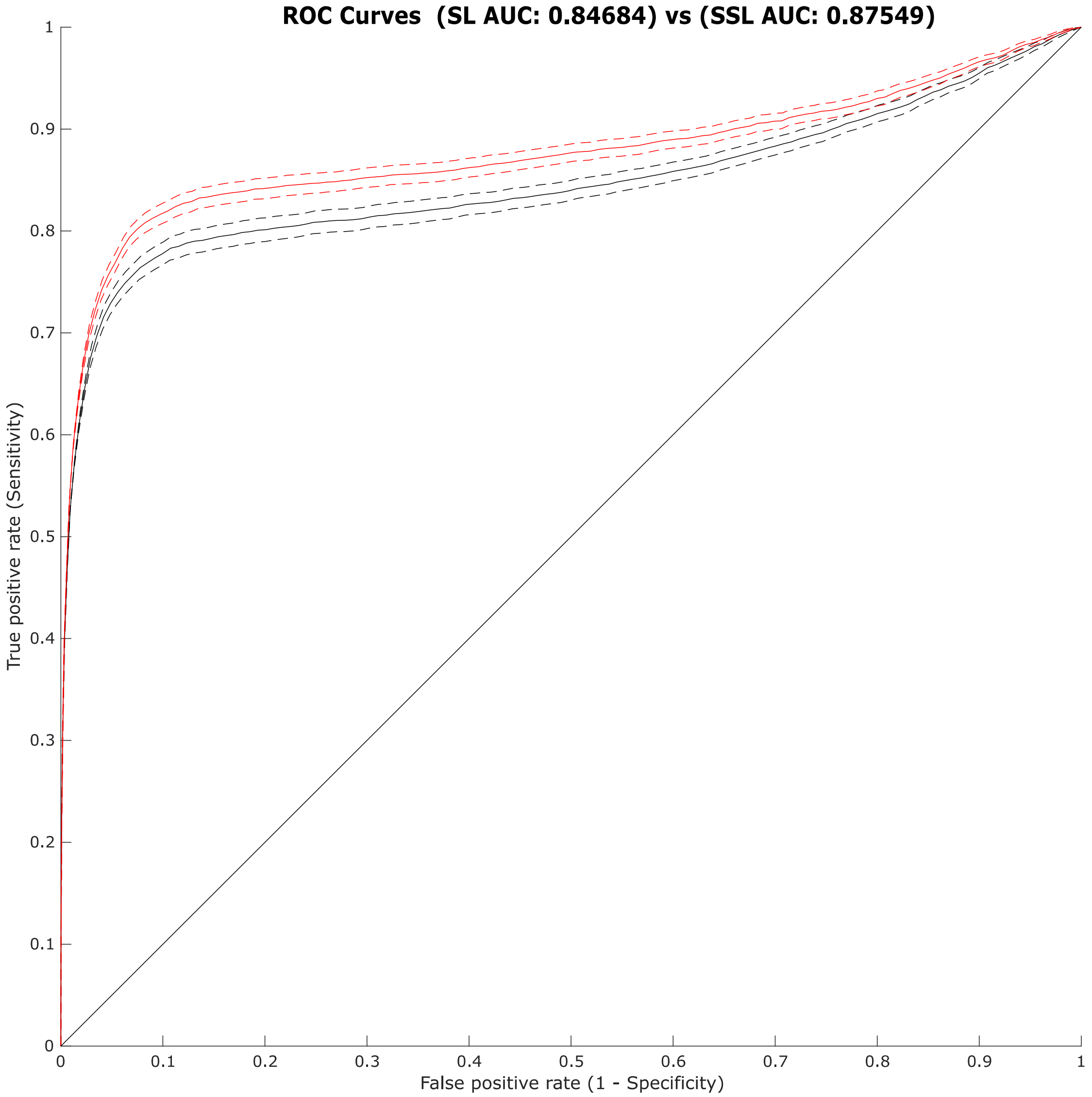


Figure 3.5 ROC curve (subset portion: 30%)

Refer to the combined description affiliated under Figure 3.8

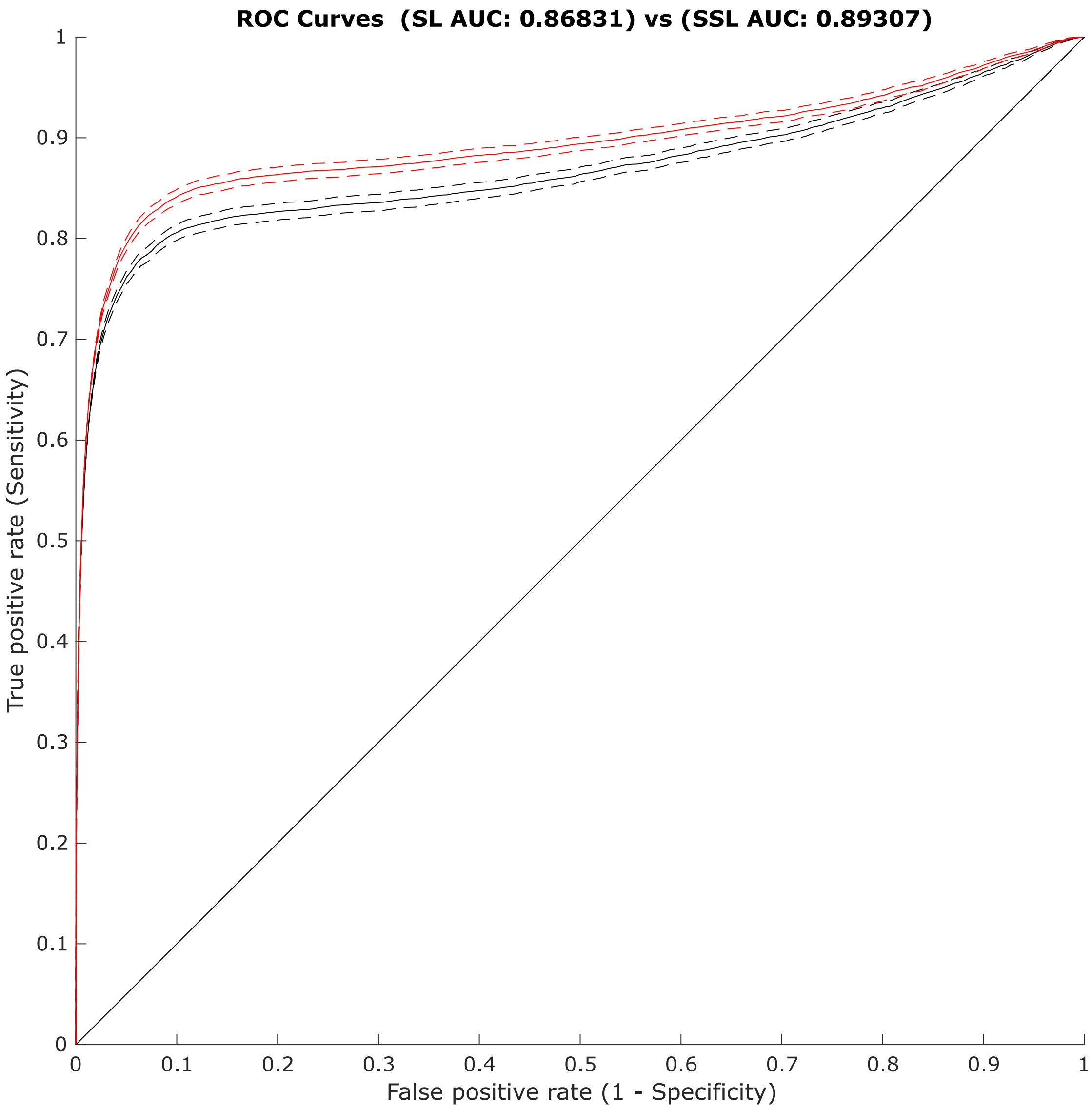


Figure 3.6 ROC curve (subset portion: 45%)

Refer to the combined description affiliated under Figure 3.8

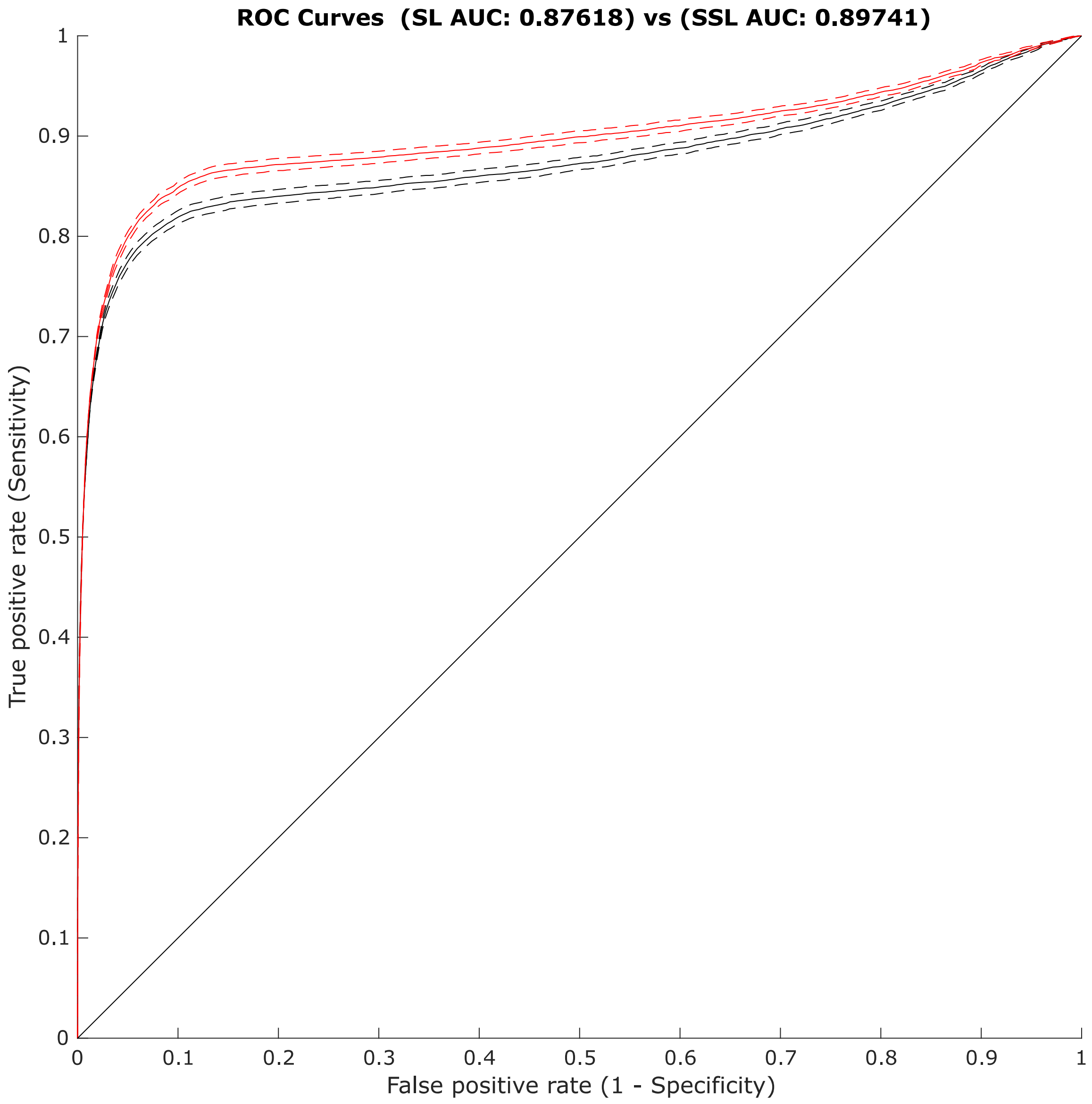


Figure 3.7 ROC curve (subset portion: 60%)

Refer to the combined description affiliated under Figure 3.8

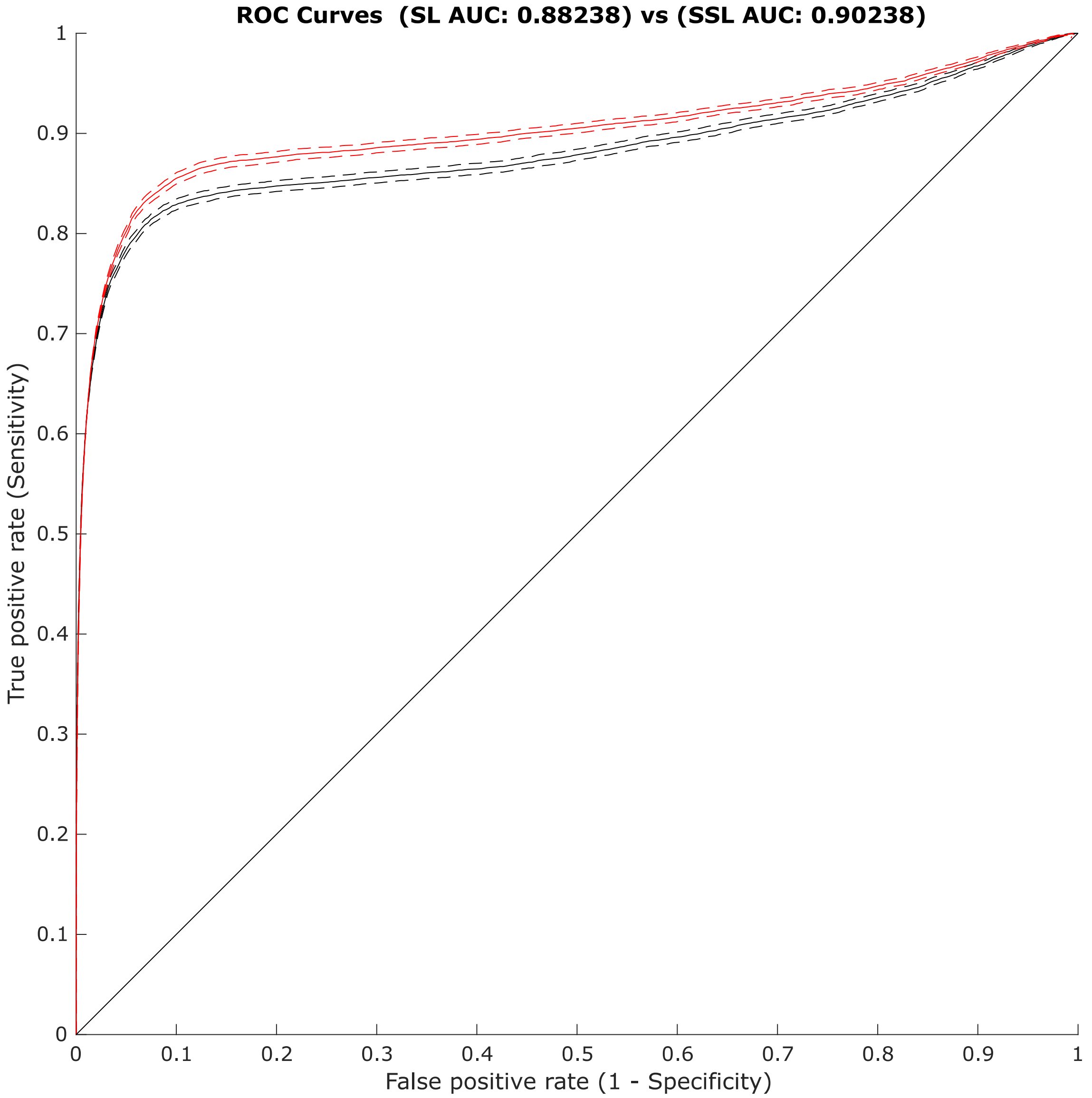


Figure 3.8 ROC curve (subset portion: 75%)

In the above four figures (3.7 – 3.10), a set of portions (30%, 45%, 60% and 75%) of the whole lesion data using as labelled data was randomly sampled for 120 iterations to discriminate the predictive power between a supervised and semi-supervised learning methods, respectively. The ROC curves for both learning methods were therefore plotted in the same figure for illustrating the contrast. The black solid line and relevant dotted lines referred to ROC curve and corresponding 95% confidence interval achieved in supervised analysis. Whilst the significantly superior performance achieved in semi-supervised analysis was shown in red lines. The more labelled data samples involved, the higher performance was observed via ROC curves.

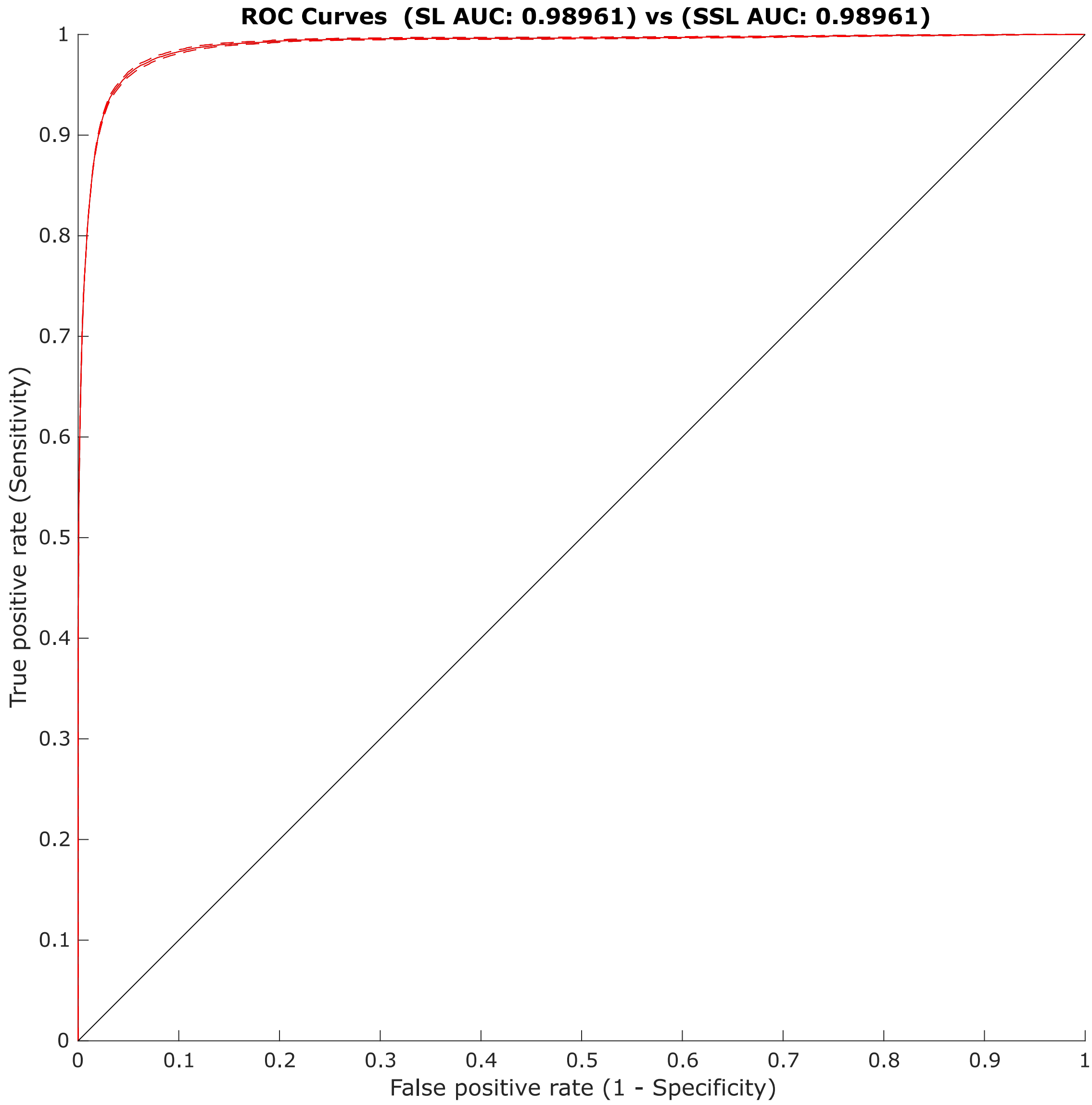


Figure 3.9 ROC curve (the whole lesion set)

The whole lesion data using as labelled data was randomly sampled for 120 iterations to examine the predictive power between a supervised and semi-supervised learning methods, respectively. The ROC curves for both learning methods were therefore plotted in the same figure for illustrating the contrast. The black solid line and relevant dotted lines referred to ROC curve and corresponding 95% confidence interval achieved in supervised analysis. Whilst the exact same performance achieved in semi-supervised analysis was shown in red lines. Two curves were completely overlapped. Comparing the performance achieved by a subset, the predictive power was boosted.

Overall, the series of one-region based simulations showed the prediction performance was proportional to the size of labelled data points while learning a classifier. Moreover, for any labelled training dataset, an additional unlabelled dataset would probably lift the performance further in some extent. Nonetheless, as the inherent properties of our lesion maps decided the imbalanced label distribution, the stability and reliability of prediction performance could be enhanced if the size of dataset could be expanded further.

3.3.2 Simulation two: dual Brodmann regions based lesion-deficit prediction performed by a supervised approach (SVM) versus a semi-supervised approach (TSVM)

The stack of 1172 lesion masks were labelled with a hypothetical functional deficit based on overlap with either of two specified Brodmann regions. Thus, except for the combination of Brodmann 28 and 36 which was not covered any of the lesion mask, a total of 819 combinatorial pairs of Brodmann regions were prepared. As in the last series of simulations, a supervised learning method (SVM) and a semi-supervised learning method (TSVM) were compared, with exactly the same objective, now evaluated in the context of a more complex ground truth.

3.3.2.1 Performance comparison between supervised learning and semi-supervised learning models

Ten combinatorial pairs of Brodmann areas was randomly sampled to establish lesion-deficit prediction. In each set of two-region models, prediction was performed with five proportions (30%, 45%, 60%, 75% and 100%) extracted from the set of 1172 lesion maps randomly. In each subset, 120 randomisations, split into 70% and 30% for training and testing were performed. Hence, the results from a total of 6000 supervised learning matched with according 6000 semi-supervised learning models were available for analysis.

Balanced accuracy was again used owing to the presence of class imbalance. Figure 3.10 shows the results. Overall, the semi-supervised models tended to be superior, but not within the 95% confidence intervals. As summarised in Table 3.3, the largest difference in terms of balanced accuracy (0.807±0.013 vs. 0.824±0.013) between two approaches occurred at the smallest proportion (30% labelled data and 70% unlabelled data sampled from the full 1172 lesion masks). This emphasised the contribution of an additional unlabelled data set on top of a relatively small set of labelled data points. As before, the semi-supervised advantage fell with increasing proportions of labels. When the instances used for learning a classifier were all labelled, the performance achieved obtained from both approaches converged as expected. The results were aligned with expectations; but the reliability of contrast between two approaches did not show statistical significance.

Within each method, predictive fidelity grew with increasing training data size. Owing to the biased vascular brain lesion distribution, the ratios between two classes in this series of simulations was approximately 1:7 ("affected" / "unaffected") on average, which means there were approximately a total of 150 lesions labelled as "affected" from the full 1172 lesion maps. From Figure 3.12 associated with the details in Table 3.3, we can see the trends of improvement in both methods were more or less the same. When the labelled training dataset increased from 30% to 45% of the 1172 lesion set, the ascendance was in a relatively large amplitude; meantime, the standard error rates in both methods dropped from 0.013 to 0.010. Among three middle range proportions (45%, 60%, 75%), the performance achieved by both methods improved mildly, which also further stabilised by slightly dropped standard error rates. At last, the value rose dramatically when all the labelled data points contributed to training the classifiers with a standard error rate at 0.002. This indicated classifiers might not be sufficiently learnt by only 30% labelled data points. Then, the classifiers were trained fairly between 45% and 75% labelled data points involvement.

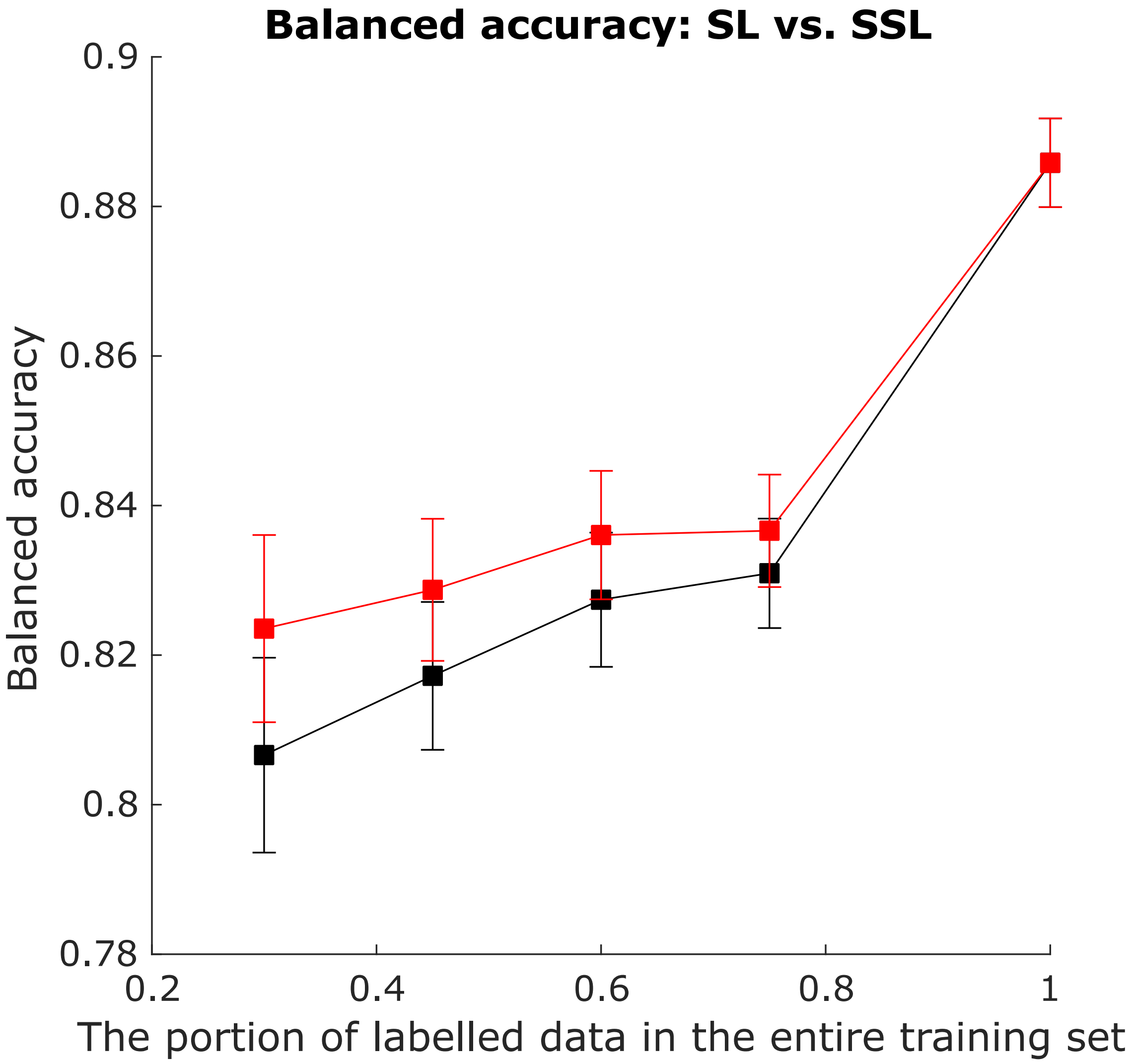


Figure 3.10 balanced accuracy

Balanced accuracy was used for estimating the generalizability between supervised and semi-supervised learning methods. The estimates were derived from ten combinatorial Brodmann pairs sampled from the total 819 pairs at five pre-defined portions of subsets. For each region of each proportion of subset, 120 randomisations were iteratively conducted to train and test with difference set of lesion data. The predictive performance achieved by supervised learning method is in black, semi-supervised in red. Superior performance was observed in semi-supervised analysis, but not significantly so.

	30%	45%	60%	75%	100%
SL: balanced accuracy	0.807±0.013	0.817±0.010	0.827±0.009	0.831±0.007	0.886±0.006
SSL: balanced accuracy	0.824±0.013	0.829±0.009	0.836±0.009	0.837±0.007	0.886±0.006

Table 3.3 balanced accuracy

The results of ROC curves (Figure 3.12 to 3.16) and related AUC scores (Figure 3.11) show a similar picture. The semi-supervised models were superior overall, but not within the 95% confidence interval. Moreover, the differences were less marked than for balanced accuracy (a detailed comparison is shown in Table 3.4). Compared with balanced accuracy, the ROC curves and according AUC scores were estimated by a variety of thresholds to illustrate the contrast between supervised and semi-supervised learning models comprehensively.

Within each method, performance improved with more labelled data as expected.

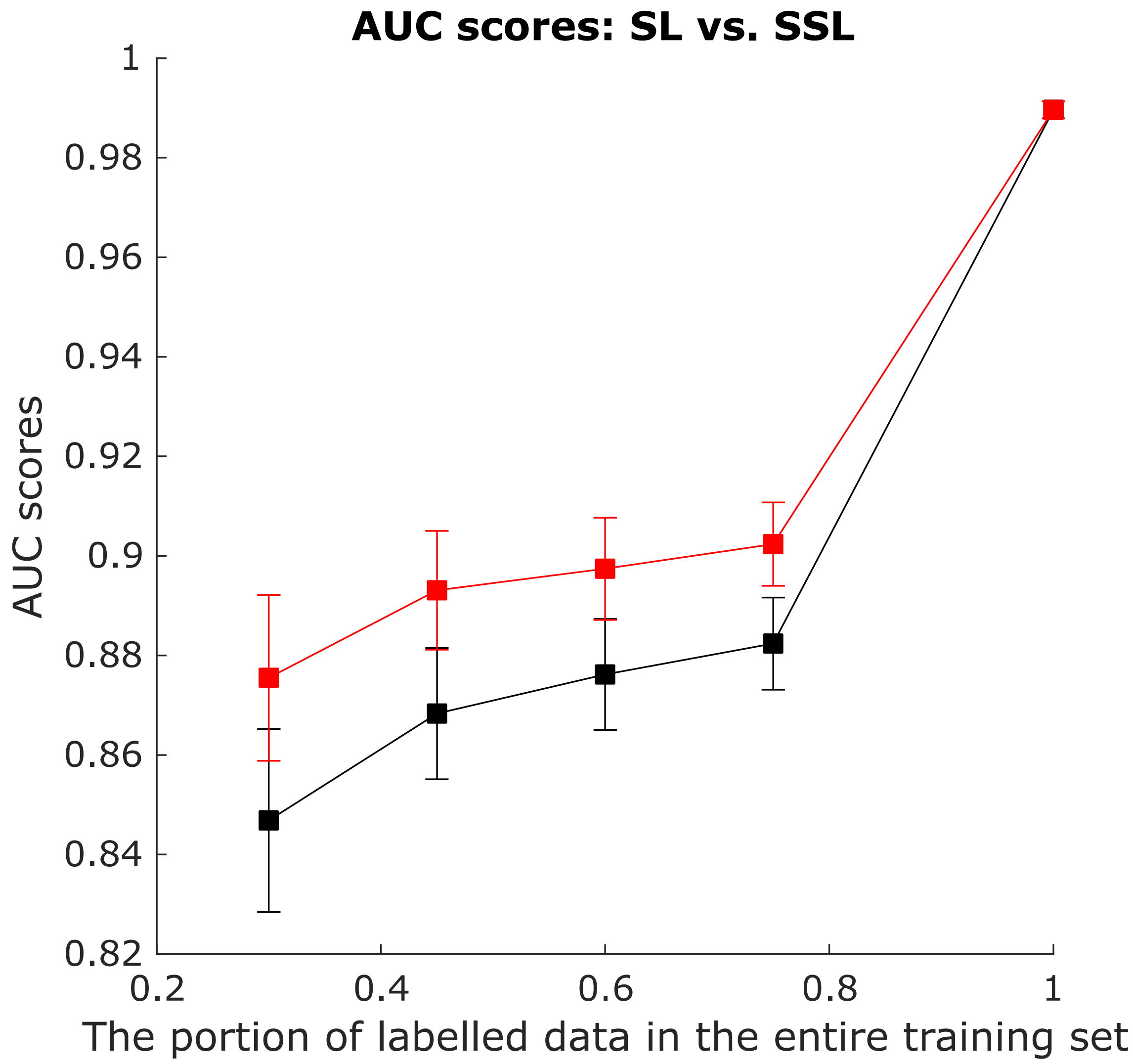


Figure 3.11 AUC scores

AUC scores were calculated to estimate the generalizability between supervised and semi-supervised learning methods with an extensive range of prediction values. The estimates were derived from ten combinatorial Brodmann regions sampled from the total 819 regions at five pre-defined portions of subsets. For each region of each proportion of subset, 120 randomisations were iteratively conducted to train and test with difference set of lesion data. The predictive performance achieved by supervised learning method is in black, semi-supervised in red. As a more complex model, the limitation of lesion data was further emphasised. The improvement was not significant across all the thresholds.

	30%	**45%**	**60%**	**75%**	**100%**
SL: balanced accuracy	0.914±0.011	0.928±0.008	0.936±0.007	0.937±0.006	0.985±0.002
SSL: balanced accuracy	0.922±0.010	0.931±0.007	0.938±0.006	0.939±0.005	0.985±0.002

Table 3.4 balanced accuracy

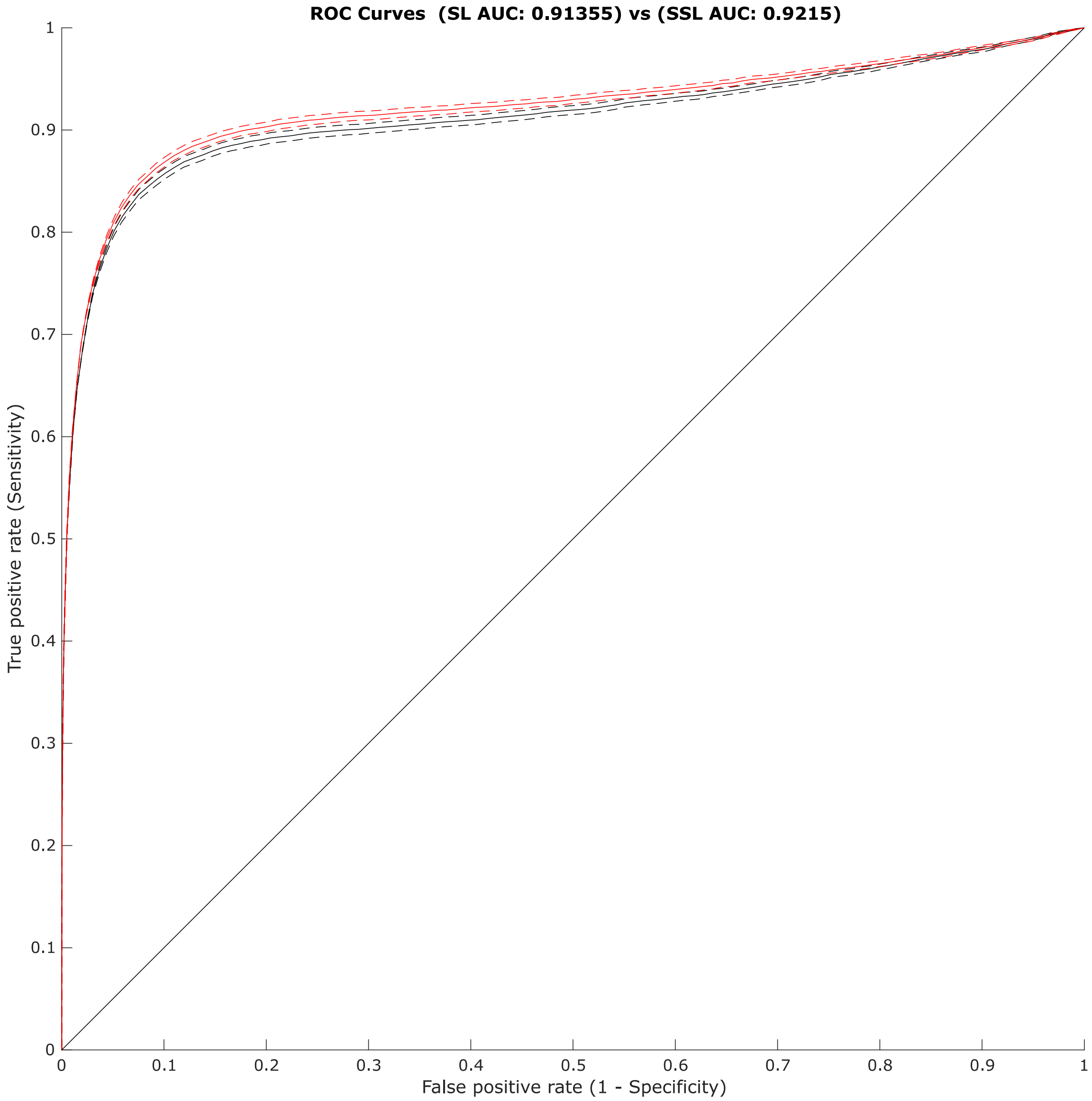


Figure 3.12 ROC curve (subset portion: 30%)

Refer to the combined description affiliated under Figure 3.15

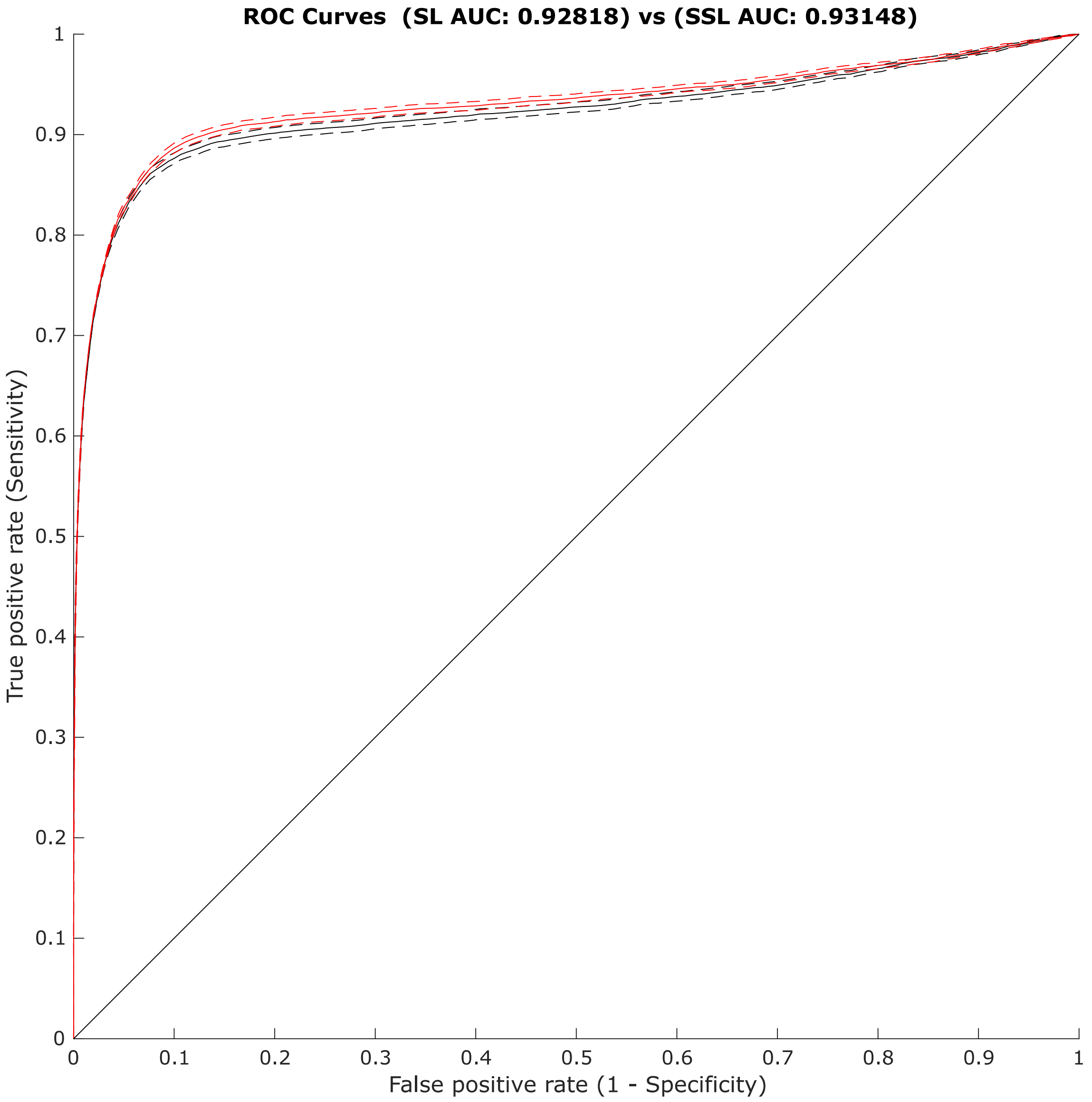


Figure 3.13 ROC curve (subset portion: 45%)

Refer to the combined description affiliated under Figure 3.15

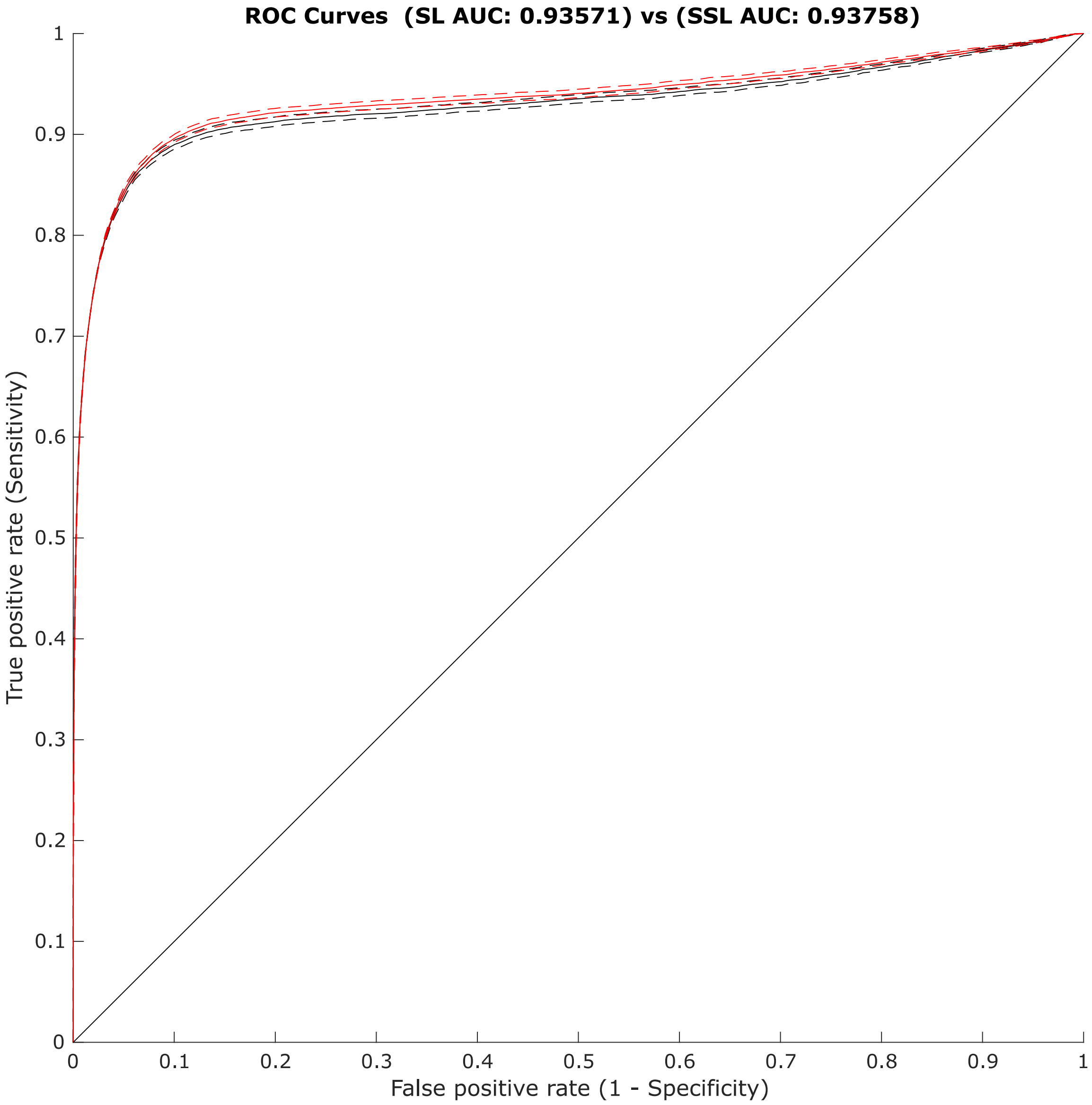


Figure 3.14 ROC curve (subset portion: 60%)

Refer to the combined description affiliated under Figure 3.15

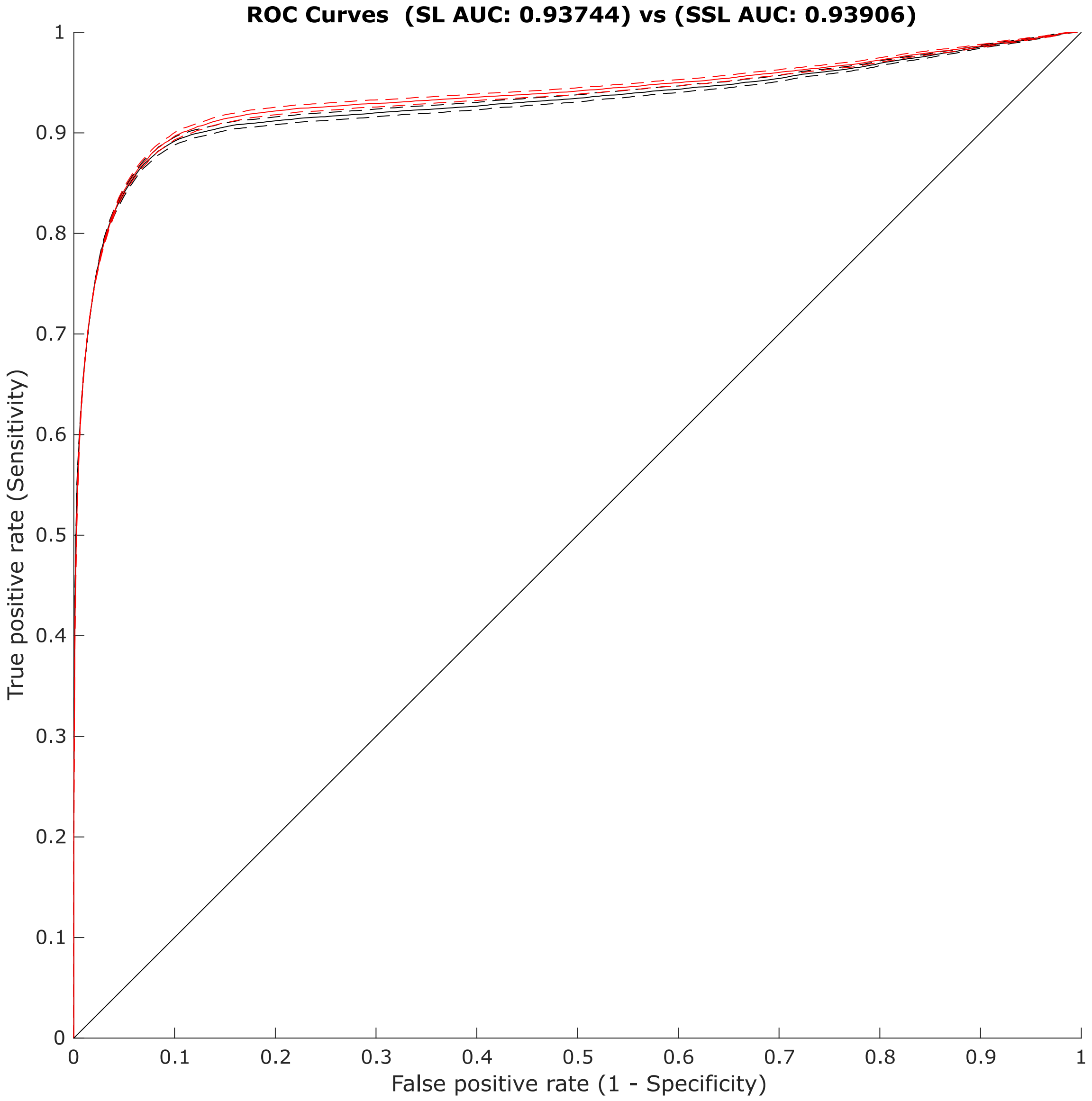


Figure 3.15 ROC curve (subset portion: 75%)

In the above four figures (3.14 – 3.17), a set of portions (30%, 45%, 60% and 75%) of the whole lesion data using as labelled data was randomly sampled for 120 iterations to discriminate the predictive power between a supervised and semi-supervised learning methods, respectively. The ROC curves for both learning methods were therefore plotted in the same figure for illustrating the contrast. The black solid line and relevant dotted lines referred to ROC curve and corresponding 95% confidence interval achieved in supervised analysis. Whilst the performance achieved in semi-supervised analysis was shown in red lines. There is no significance between two analysis methods observed.

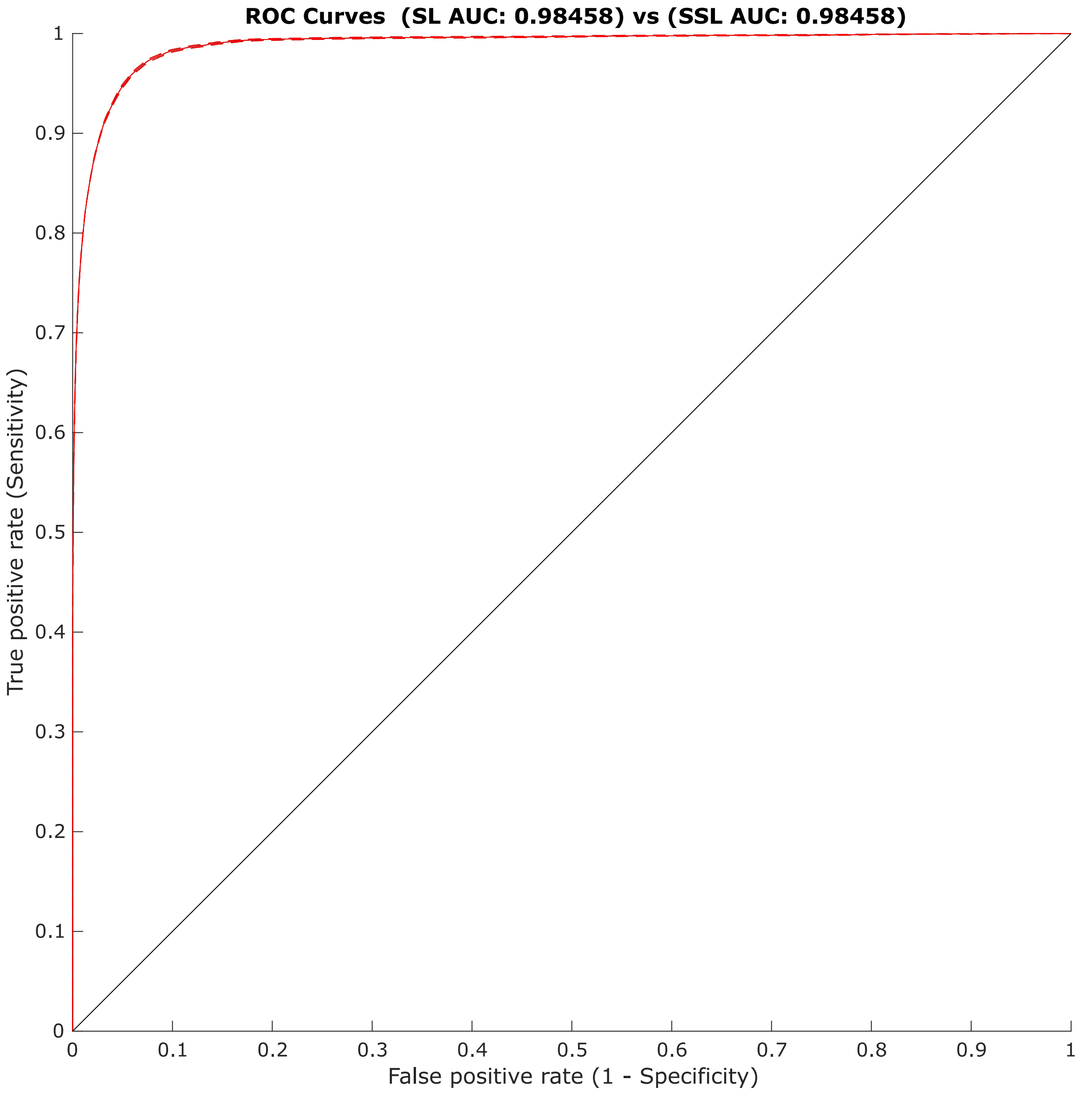


Figure 3.16 ROC curve (the whole lesion set)

The whole lesion data using as labelled data was randomly sampled for 120 iterations to examine the predictive power between a supervised and semi-supervised learning methods, respectively. The ROC curves for both learning methods were therefore plotted in the same figure for illustrating the contrast. The black solid line and relevant dotted lines referred to ROC curve and corresponding 95% confidence interval achieved in supervised analysis. Whilst the exact same performance achieved in semi-supervised analysis was shown in red lines. Two curves were completely overlapped. Comparing the performance achieved by a subset, the predictive power was boosted.

In agreement with the series of simulations based on single Brodmann regions, the two-region based simulations showed enhanced performance with larger proportions of labelled data, but the benefit of semi-supervised learning was less clear.

3.4 Discussion

3.4.1 Simulation one: lesion-deficit prediction depending on single Brodmann region performed by a supervised learning method (SVM) versus a semi-supervised learning method (TSVM)

In the previous chapter, we showed that high-dimensional multivariate analysis (SVM) that takes into account the complex lesion architecture and its inter-woven relations is overall superior in terms of prediction fidelity compared with the conventional low-dimensional approach reliant on mass-univariate analysis. This series of simulations replicated the analysis, now employing supervised (SVM) and semi-supervised (TSVM) high-dimensional methods, and exploring the benefit of incorporating unlabelled data into the model. As aforementioned, the nature of clinical data makes large, completely labelled datasets difficult to acquire (Lee and Yoon, 2017). The use of additional unlabelled data may contribute to the optimisation of the learning process further, and to reduce the variability of a learnt classifier, though it could also degrade performance if the unlabelled dataset is constitutively dissimilar (Peikari et al., 2018, Sindhwani and Keerthi, 2006, Zhu, 2006, Zhu and Goldberg, 2009, Chapelle et al., 2008).

Our analysis explores the impact of a trade-off between the volume of labelled and unlabelled data. Here the most illuminating difference can reasonably be assumed to be located at the point of the greatest disparity in the proportions of labelled and unlabelled data. Within each individual learning method, the results obtained from five pre-defined proportions of data showed the prediction performance was proportional to the size of training dataset. Comparison between the two methods showed the semi-supervised learning models to be superior, though the magnitude of the advantage could be potentially enhanced with the addition of more data, and were sensitive to the complexity of the hypothesised underlying functional anatomy.

The analysis of balanced accuracy suggested a clear benefit for semi-supervised learning; AUC-ROC less so, both in terms of accuracy and its variance. It should be noted that imbalance—and relatively low numbers of lesions falling within each test region—will have reduced our ability to quantify the benefit definitively, even with a dataset of unprecedented size.

Overall, given the intrinsically high-dimensional lesion architecture (Mah et al., 2015, Mah et al., 2014b, Xu et al., 2017a, Xu et al., 2017b), semi-supervised learning is a complementary strategy of potential use in lesion-deficit prediction. This is because an extra set of unlabelled data instance not only increases the number of cases involved into a learning procedure; but enhances the naturalisation of feature distribution (Klöppel et al., 2012). Moreover, for a predictive model parameterised with spatial complexity, it is reasonable to believe the more unlabelled data is mixed with labelled data, the more natural a description of the general population is likely to be achieved. Hence, predictive performance is improved.

3.4.2 Simulation two: lesion-deficit prediction depending on a combinatorial pair of Brodmann regions performed by a supervised learning method (SVM) against a semi-supervised learning method (TSVM)

The single locus simulations fundamentally revealed the contribution of an additional unlabelled dataset for modelling lesion-deficit prediction, though the correlations between focal damage and functional deficit were idealised within a single Brodmann region. In order to identify if an extra set of unlabelled data could improve predictive performance in a more complex scenario, two-region based simulations were conducted. As before, a hypothetical deficit was determined by a dual loci model dependent on 20% or more of a lesion mask overlapping with a combination of Brodmann areas. Then, lesion-deficit prediction was modelled by a supervised (SVM) and semi-supervised (TSVM) learning methods, respectively. The trade-off between labelled and unlabelled samples was still followed by five pre-defined proportions. The critical point of comparison was assumed to be located at the point of the greatest difference between labelled and unlabelled data.

The results fulfilled expectations. For both balanced accuracy and AUC-ROC, the maximum contrast between the two learning approaches was 0.17 (balanced accuracy: 0.824 vs. 0.807) and 0.008 (AUC: 0.922 vs. 0.914) respectively when 30% of all lesions were extracted randomly using as labelled samples; whilst, the contrast reduced to 0.006 (0.837 vs. 0.831) and 0.002 (0.939 vs. 0.937). Note that the performance obtained in these simulations was better overall that in the one area simulations. For example, in one-region simulations, the balanced accuracies were in the range from 0.712 to 0.863 in supervised learning models and from 0.752 to 0.863 in semi-supervised learning models. Whereas, in two-region simulations, the value ranges were between 0.807 and 0.886 in supervised models; and they were between 0.824 and 0.886 in semi-supervised models. Similar comparisons were also observed in the estimates of AUC. In regard to the design of both series simulations, the imbalance of class distribution in first series of simulation was more severe than it was in the two-region based simulation. Specifically, the positive and negative class ratio was roughly 1:12 in one-region simulations; whereas, the ratio increased to approximately 1:7. In other words, the balance of label distribution in the two-region simulations was much better. This may explain why a more complex predictive model that based on two Brodmann areas performed better than a model that simplified lesion-deficit correlation within a single Brodmann area – the properties of dataset played a vital role during lesion prediction. Nevertheless, in contrast with one-region simulations, where the performance of semi-supervised learning models was significantly distinguished from the corresponding supervised learning models, in the two-region simulations, semi-supervised models showed superiority, but not within the 95% confidence interval. This may be interpreted as implying the current dataset was less than sufficient to distinguish two learning approaches engaged with more complex predictive models.

Overall, two set of simulations were complementary to each other. The simpler predictive models based on a single Brodmann area assured the value of unlabelled dataset. While coping with more complex predictive models, the other series of simulations showed the possibilities implicitly; meantime, encouraged us to expand existing dataset.

3.5 Conclusion

The intrinsic dimensionality of complex lesion architecture compels us to employ a high-dimensional multivariate analysis while modelling lesion-deficit prediction. Nonetheless, the power of high-dimensional modelling is highly dependent on a sufficient dataset with complete labelling which is still a major challenge, particularly, in clinical domain. In order to alleviate the constraints, we involved a mix of data samples with both complete and incomplete labelling to optimise the performance of high-dimensional modelling for lesion-deficit prediction. In light of the results obtained from two series of simulations, predictive performance was clearly proportional to the sample size with complete labelling. Accordingly, an additional set of unlabelled samples could improve the performance further in some extent, though the significance of discrimination was varied upon the complexity of simulated models.

This study was consistent with the intuition that the addition of extra unlabelled samples together with labelled samples will improve the naturalisation of lesion distribution, and reduce variability. Theoretically, the sample lesion distribution could be better approximated to the general population with the increase of either labelled or unlabelled samples. In practice, this study demonstrated an effective alternative when the completely labelled data instances are insufficient, or an additional dataset with incomplete labelling are available.

Chapter 4 Therapeutic inference in the focally damaged human brain

4.1 Introduction

To make a therapeutic inference in the focally damaged human brain is to establish a causal relation between a therapeutic intervention and its behavioural outcome in the context of focal brain injury. We have seen that on this problem are likely to impinge biological factors reflecting two distinct complex dimensionalities intrinsically to the human brain in health and disease: the distributed lesion architecture (Mah et al., 2014b, Xu et al., 2017a, Xu et al., 2017b) and the distributed functional network (Glasser et al., 2016a, Sporns, 2013, Sporns, 2011, Sporns et al., 2005).

In the preceding chapters, we have examined the impact on lesion-deficit prediction of neglecting the dimensionality of lesions through the use of modelling methods that ignore the high-dimensional spatial correlations present in lesion data. Here we focus on a related but distinct problem: the identification of a therapeutic effect within an interventional study. In the conventional interventional studies, a low-dimensional model relying on a few crude anatomical factors such as the volume of damage or crudely discretised brain regions are almost invariably employed. Thus, the variation arising from variability of individual outcomes tends to be ignored owing to insufficient parameterisation. As a result, the effect of an intervention embedded in a casual field of multiple contributory factors can only be weakly isolated.

In essence, conventional interventional trials are group studies, where a cohort of patients is assumed to be homogeneous and individual differences are treated as noise, approximated no better than by the mean of the group. Consequently, the loss of individual information may lead to poor sensitivity. Given the complexity of the human brain, it is reasonable to assume the factors on which the effect of the intervention depends and from which it must be isolated should be extracted from as many features of the patient as are material to the outcome. Indeed, the inadequate dimensionality of the models in common use may explain the striking failure rate of translational medicine, where significant interventional effects observed in animal trials are difficult to reproduce in humans (Brodie, 1962, Wang and Johnson, 2008, Bracken, 2009, Jucker, 2010). Of course, translation may fail for many reasons including differences in genetics, molecular or cellular mechanisms, immunity and so on. But a failure caused by unaddressed complexity in both the functional and lesion architecture (Xu et al., 2017b, Xu and Pan, 2013, Mak et al., 2014, Perel et al., 2007) is at least addressable through the use of the right modelling method. Though the distributed functional network underlying any one function remains largely unknown, a critical prerequisite for establishing a causal relation between a therapeutic intervention and its outcome ought to be the sufficient parameterisation of the complexity of the lesion architecture as reflected in its complex spatial distribution. We have seen this can be done with the right high-dimensional multivariate model.

Now to establish a high-dimensional model of therapeutic inference we need sufficiently parameterised lesion patterns associated with a behavioural outcome. In this chapter we use gaze deviation, as a behavioural measure that is directly quantifiable from brain imaging alone, removing the noise and uncertainty arising from clinical labelling. Patients with acute ischemic stroke in whom critical oculomotor or attentional neural circuits are disrupted often exhibit a gaze palsy: a shift in the preferred direction of gaze at rest (Ramat et al., 2006). Since our population of stroke patients has been imaged twice—on admission with CT, and typically 24 to 72 hours later with MRI—we are able to quantify gaze at two time points, yielding a vector of change over time. Thus, an outcome ground truth can be set up by relating each lesion map to the behavioural parameter of a change in gaze. We can implement such high-dimensional modelling with support vector machines, generating a classifier predicting which lesion patterns tend to cause persistent gaze deviation rather than spontaneous recovery. Simulated interventions with varying effect sizes can then be evaluated to derive a continuous “therapeutic function” which describes the ability to identify the effect of an intervention with the aid of low or high-dimensional multivariate models. This is the approach pursued in this chapter.

4.2 Methods

4.2.1 Patients

A set of 1333 patients admitted to University College London Hospital (UCLH) Hyper-acute Stroke Unit (HASU) over the past decade were selected. All patients were clinically diagnosed with acute ischemic stroke, and received MRI, including diffusion-weighted imaging (DWI), in which at least one acute ischemic lesion could be identified. A threshold for minimal DWI-quantified lesion volume was set at 50 milliliters which covered the 90th centile of our entire clinical population in order to reduce the disproportionate impact from very large lesions. An adjusted cohort including 1172 patients was thus produced.

The age distribution from the set of 1172 patients was in range of 18 to 97 years old with a mean of 63.89 and a standard deviation of 15.91. The gender ratio (Male) was 0.561 (Appendix B. Figure 7.1). Ethnicity for 952 (81.2%) from the full set of patients was disclosed in the clinical record which is illustrated in Appendix B. Figure 7.2. The distribution of age and sex, and the constitution of ethnicity reflected the catchment of clinical population in UCLH naturally.

4.2.2 Imaging

All patients underwent both CT within a few hours of the first onset of acute ischemic stroke and MR imaging sessions typically in the following 24 to 72 hours.

4.2.2.1 CT data acquisition

All CT imaging sessions acquired by either a Somatom Definition or a Somatom Sensation scanner manufactured by Siemens (Erlangen, Germany) with a typical resolution for each uncontrasted head image of 0.48 mm x 0.48mm x 5mm (Columns = 512, Rows = 512), and retrieved via a picture archiving and communication system (PACS) from UCLH. All scans were obtained within clinical protocols in the course of routine clinical care. The accompanying radiological reports were checked for precise diagnosis, which helped us eliminate alternative diagnoses, such as primary intracerebral haemorrhage.

4.2.2.2 CT data analysis

4.2.2.2.1 Imaging pre-processing

A dedicated pipeline for processing CT head images was implemented within MATLAB (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.). First, the raw clinical CT head images were converted from Digital Imaging and Communications in Medicine (DICOM) images to Neuroimaging Informatics Technology Initiative (NIfTI) format for the purpose of scientific imaging analysis. Then, by rigidly co-registering to the standard SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/) tissue probability map based on normalised mutual information with adjustment from a Procrustes analysis weighted by the white and gray matter compartments, each CT image was rigidly transformed into MNI stereotactic space (Jha et al., 2016, Xu et al., 2017b).

4.2.2.2.2 Semi-automatic gaze segmentation

The degree of head yaw was estimated by the co-registering procedure. Then, in every three-dimensional co-registered head scan, each intraocular lens, typically best seen in the axial slice that passed closest to the lens's centroid, was segmented. This was performed by a trained operator (Parashkev Nachev) in a semi-manual manner with the aid of MIPAV's level-set VOI tool (https://mipav.cit.nih.gov/) on its corresponding (usually the same or immediately neighboring) slice thereafter (Figure 4.1). This method was able to segment both natural or prosthetic lenses. In cases where only one lens was visible, the other eye was ignored. Next, a MATLAB integrated function (*poly2mask*) (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.) was applied to generate a binary mask based on the voxel coordinates of each segmented lens boundary, which derived the centroid of each mask and according orientation derived with another MATLAB function (*regionprops*)

(MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.). Finally, in combination with the estimate of head yaw derived from the procedure of image co-registration, the calculated centroid coordinates and its orientation automatically determined the side of the corresponding eye and it is gaze direction relative to the axis of the body.

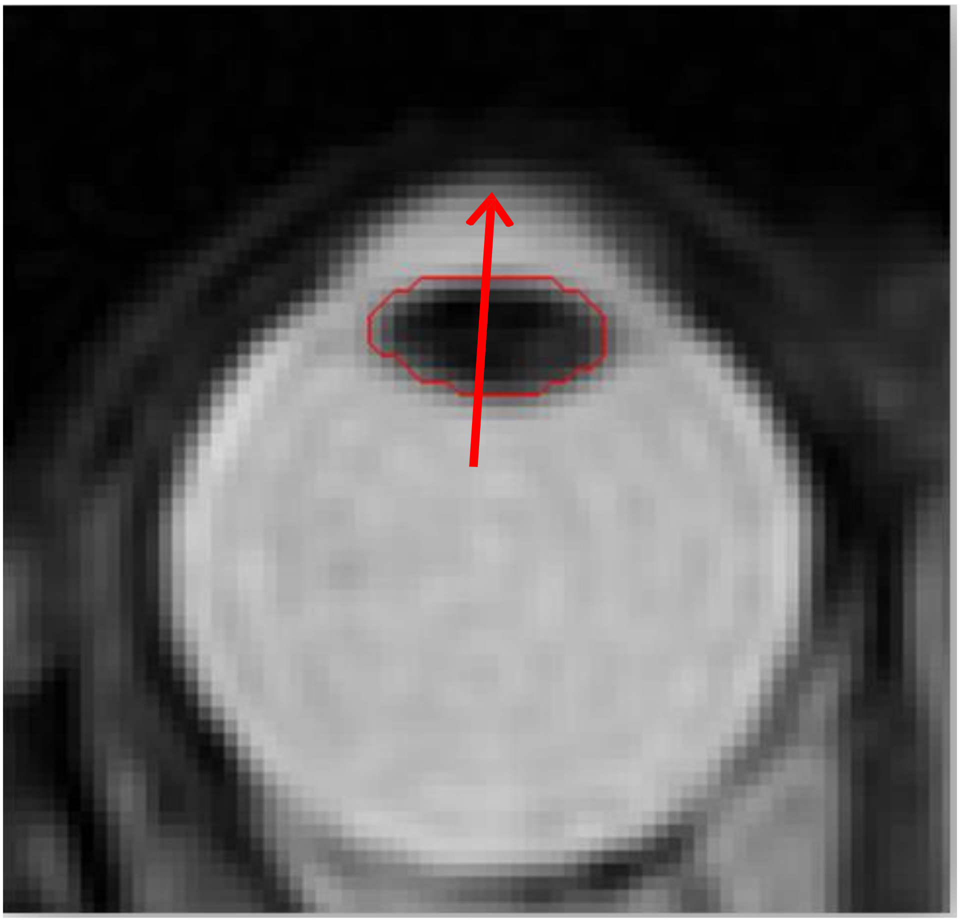


Figure 4.1 Gaze segmentation

In each brain sequence, the slice that best demonstrated the intraocular lens was manually identified by an experienced neurologist. Then, the yaw angle was automatically calculated by the related function integrated in MATLAB.

4.2.2.3 MR data acquisition

All acquisitions performed on the scanners manufactured by General Electric (GE) (Discovery MR 450, Genesis Signa and Signa Excite), Philips (Achieva, Ingenia and Intera), or Siemens (Avanto, Biograph mMR, Espree, Skyra, Symphony, Symphony Tim, Trio Tim and Verio) with field strength of either 1.5 or 3 Tesla in a single session for each patient were collected via a picture archiving and communication system (PACS) from UCLH (the specific range of scanners and related parameters are listed in appendix A). The originated voxel sizes were different in some extent between the scanners, and so were the according spatial resolutions. Typically, for 1.5 Tesla GE medical system, the voxel size was sampled as 1mm x 1mm x 6.5mm (Columns = 256, Rows = 256). For 1.5 Tesla Philips medical system, the typical voxel size was 1.2mm x 1.2mm x 6mm (Columns = 192, Rows = 192); the voxel size was 1mm x 1mm x 6mm (Columns = 256, Rows = 256) performed on 3 Tesla Philips scanners. For Siemens, the voxels sizes and spatial resolution were 1.8mm x 1.8mm x 6.5mm (Columns = 128, Rows = 128) and 1.2mm x 1.2mm x 6.5mm (Columns = 192, Rows = 192) on 1.5 Tesla and 3 Tesla scanners, respectively. The routine of clinical practice with regular equipment upgrades and replacement is the main reason for the diversity of scanners over the period of data collection. All brain scans were performed for the purpose of clinical routine obeying clinical protocols.

The specific sequences extracted from each MRI session were axially required echo planar DWI with b-values of 0 s/mm^2 and 1000 s/mm^2 which is used for lesion segmentation (Mah et al., 2014c) so as to implement the series of simulations designed for this study. DWI is able to indicate the restricted ability of Brownian motion for extracellular water protons where cytotoxic edema causes imbalance. So, DWI is sensitive for detecting ischemic stroke and widely used for locating acute ischemic lesions (Warach et al., 1995, Löuvbld et al., 1997, Lövblad et al., 1998). In the clinical application, a DWI sequence usually includes images with b values of 0 s/mm^2 (a b0 image) and 1000 s/mm^2 (a b1000 image). The former type image shows stronger contrast on normal tissues but relatively insensitive to ischaemia. Conversely, an image with high diffusion weighting is used to sense water molecules movement. So, it is sensitive to ischaemia but poor to the contrast of normal tissues. Both types of images are complementary for brain registration and lesion segmentation, which rely on the contrast between normal tissue types and the discrimination between normal tissues and lesions, respectively.

4.2.2.4 MR data analysis

4.2.2.4.1 DWI pre-processing

A processing pipeline was implemented within MATLAB (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.) associated with a configured SPM12 toolbox (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/) to ensure the images to be processed systematically.

For the purpose of scientific imaging analysis, typically, the raw clinical MR images needed to be firstly converted from the complicated format of Digital Imaging and Communications in Medicine (DICOM) images to Neuroimaging Informatics Technology Initiative (NIfTI) format. Subsequently, in order to provide accurate and robust performance on image registration and segmentation in the later stage, we employed a pre-processing pipeline with the following steps. We empirically clamped the signal between 0.1% and 99.9% of the cumulative distribution estimated with a kernel density method (Botev et al., 2010), removing unusually low and high signals in each NIfTI file. Then an oracle-based 3D discrete cosine transform (ODCT3D) (Manjón et al., 2012) denoising method was applied for further noise reduction. In the next step, image registration was conducted based on each pair of DWI files. In order to optimise the alignment between the b0 and b1000 images in each pair, we used SPM12's standard co-registration function to have the b0 image rigidly co-registered to the b1000 image. Then, the normalisation / segmentation routine integrated in SPM12 (Ashburner and Friston, 2005) was applied to the b0 image to calculate a deformation field which described the optimal non-linear transformation of the b0 image into Montreal Neurological Institute (MNI) stereotactic space. Next, the deformation field derived from the b0 image was applied to the b1000 image to transform it into MNI space using the same routine. The parameters involved in the whole process of image registration were set at defaults by SPM12. Finally, both b0 and b1000 images in each pair were resampled to 2mm^3 isotropic with 6th degree b-spline interpolation (Mah et al., 2014b, Mah et al., 2014c) and manually checked against the SPM template to confirm the satisfactory of registration.

4.2.2.4.2 T2 pre-processing and semi-automatic gaze segmentation

Similar to the pre-processing procedure of CT head images, MR T2 sequences were also rigidly co-registered to the tissue probability map provided by SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/) in standard MNI space based on the normalised mutual information which was adjusted with a white and gray matter compartments weighted Procrustes analysis (Jha et al., 2016, Xu et al., 2017b).

As for the CT images, the degree of head yaw estimated during the co-registering procedure was first recorded. Then, in every co-registered T2 scan, the axial slice that best showed each intraocular lens was manually identified by a trained operator (Parashkev Nachev), and semi-manually segmented with the aid of MIPAV's level-set VOI tool (https://mipav.cit.nih.gov/) on its corresponding slice (usually the slice that passed the closest to each lens's centroid or its immediately neighboring) (Figure 4.1). This method is compatible with segmenting both natural and prosthetic lenses. In the minority of patients where only one lens is visible, we segmented the one lens and treated the other as absent. Subsequently, MATLAB's *poly2mask* function (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.) was applied to create a binary mask based on the voxel coordinates of each lean boundary. Thereafter, with aids of another MATLAB's function: *regionprops* (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.), the centroid of the binary mask and its orientation was calculated, which allowed for determining the side of the corresponding eye and the direction of gaze for each eye in the axial plane relative to the head automatically. Finally, in combination with the degree of head yaw estimated from the image registration procedure, the direction of gaze relative to the axis of the body was confirmed.

4.2.2.5 Lesion segmentation

A previously validated method based on the anomaly metric *zeta* (Mah et al., 2014c) was applied to segment lesions from the 1172 normalised b1000 images in a voxel-wise manner.

A set of preliminaries conducted to optimise the performance of lesion segmentation were performed as follows. First, 492 DWI images derived from patients suspected of acute stroke but found to have normal imaging were selected to form a reference normal image set. No acute lesion was visible here in any patient, and none was reported by the corresponding radiological record. Every scan was pre-processed using the exact same pipeline and resliced at the same resolution of 2mm isotropic voxels, which spatially presented as a 91 x 109 x 91 matrix. Then, by combing the tissue maps provided by SPM12 including white matter tissue, grey matter tissue and cerebrospinal fluid (CSF), an inclusive brain mask was created. So as to remove image artefacts, we set an empirical threshold to exclude areas in the frontal and temporal poles commonly prone to artefacts on DWI sequences, and adjusted the inclusive brain mask accordingly. Subsequent lesion segmentation was performed within the confines of this adjusted mask. Next, the signal distribution was normalised for both lesion and reference datasets. Such signal normalisation sought to reduce instrumental and other incidental signal heterogeneity, so that the comparison between individual images could be maximally sensitive. The procedure was performed as follows. A binary white matter mask was created from SPM12’s white matter tissue probability map by thresholding at a value of >0.9. A robust kernel density estimate method (Botev et al., 2010) was then applied to the voxels falling within both customised maps in each b1000 image to obtain the peak white matter distribution which was then subtracted from each corresponding b1000 image to normalise the signal distribution. Guided by information in the radiological report, signal normalisation on unilateral and bilateral lesions was performed differently. For unilateral lesions—the vast majority—only the unaffected hemisphere was used to estimate the peak. For bilateral lesions, since ischemic lesions tended to be small in these circumstances, the kernel density estimate was applied to the whole brain. In both cases, the signal normalisation was equally effective in adjusting the differences in the signal distribution across all scans performed by a variety of scanners and related technical and practical variance.

Lesion segmentation was performed in a voxel-wise manner. We calculated the zeta anomaly metric for each single voxel in each image independently against the reference set. To be specific, the voxel-wise zeta value in the test image is derived from the mean distance to the k nearest neighbors drawn from the anatomically homologous voxels in the unlesioned reference set, normalised by the mean distance between the k neighbors themselves (Mah et al., 2014c). The measure was the Euclidean distance in signal space. The only manipulable parameter here is the number of nearest neighbors – k – commonly chosen with reference to the size of the reference dataset. In this study, we set k at 23 based on the calculation as below:

$$k = Ceiling\left(\sqrt{492}\right).$$

Zeta is a continuous metric. An adaptive threshold was therefore determined for each image. In brief, volume connectivity computed by *bwlabeln*, an integrated function of MATLAB (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.) was applied to segment each lesion map into connected clusters. Then, each cluster was estimated under a generalised extreme value (GEV) distribution to derive mean and variance values which were used to decide threshold adaptively as described in Mah et al, 2014.

4.2.2.6 Behaviour: gaze deviation

The preferred direction of gaze was employed to describe patients' neurological state in the present study. As a crucial component of the neurological examination, the direction of gaze reflects the disruption of critical motor and attentional neural circuits distributed across multiple brain regions (Corbetta et al., 1998), which has been widely used as a standard parameterisation of clinical deficits in stroke, such as NIH Stroke Scale (NIHSS) (Group, 1997, Muir et al., 1996).

In practice, the way to quantify gaze deviation has not been standardised though it is an important indicator clinically. The methods we proposed above is to segment the lens of each eye on a specific slice of each head image (both CT and MRI), which allowed for objective and contemporaneous measures of preferred gaze direction in the axial plane of the head. The procedure of scanning has no strong lateralised features, hence, is free from external attentional bias. Prior to gaze segmentation, both estimates of head-centric and body-centric yaw could be derived from image co-registration procedure. The two estimates were highly correlated as head is generally supported in a midline-aligned position during scan sessions. We opted to head-centric value for the purpose of gaze segmentation. Thus, for the subsequent modelling, the gaze of each patient was parameterised at two time points: the acquisition time of the CT, and the time of the MRI in 24-72 hours typically (a minority of patients were scanned by MRI later, up to 10 days.).

Patients with acute ischemic stroke are naturally immediately treated after CT scanning. Treatments that altered the anatomical properties of the lesion, such as thrombolysis, were always carried out before the MRI scanning from which the lesion map was segmented. Hence, the lesion pattern presented the neurological state at the second time point, which reflected all the changes during the period of both spontaneous and treatment induced intervention.

To demonstrate the correlation between the orientation of gaze deviation and crude parameterisations of lesion masks, a generalised linear model was fitted to relate the gaze angle distributed in seven bins to the laterality of the corresponding lesions in each bin indexed by the mean proportion of lesion voxels found in the right hemisphere. As expected, gaze deviation was sensitively modulated by the lesion laterality at the time of onset (Figure 4.2 A). At the later time of the MRI, however, there was no apparent relation between gaze deviation and the laterality of the brain lesion (Figure 4.3). This was consistent with the idea that recovery might be related to more detailed features of the underlying neuroanatomy not easily captured by any simple parameterisation.

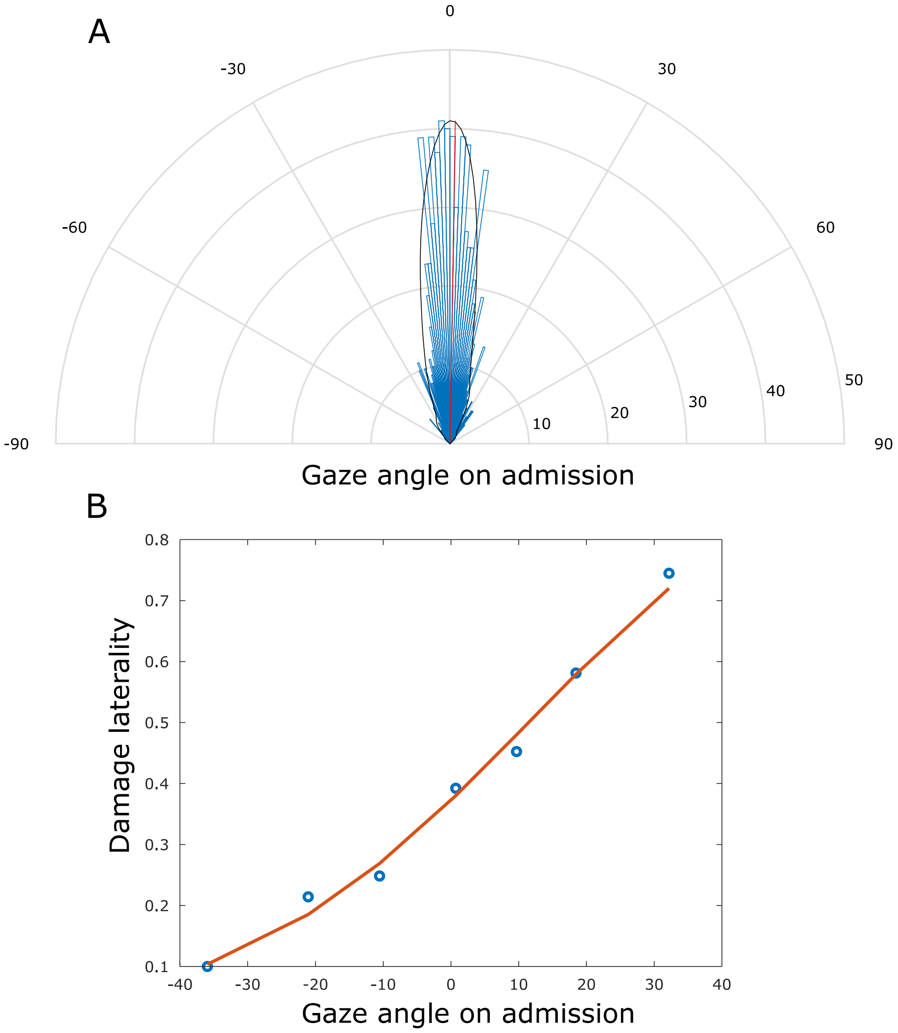


Figure 4.2 Distribution of patient gaze on admission (CT)

The gaze on admission was derived from a semi-automated segmentation algorithm based on the intraocular lenses on CT scans. Part A demonstrated the distribution of gaze deviation which was visualised by polar plot of the histogram (in blue) and related kernel density estimate (in black). Circular mean (in red) of patient gaze was 0.93 degrees to the midline. Part B demonstrated the relations between gaze direction on admission and the laterality of brain lesion. Seven bins of deviated gaze angle were empirically pre-defined. For each bin, the averaged ratio of lesion volume dropped within right hemisphere was plotted in blue circle. The relation across gaze was plotted with a general linear model fitted by maximum likelihood (in red). The plots showed gaze direction was strongly correlated with the laterality of brain damage. Hereby, the variation of gaze shown in Part would not be noise.

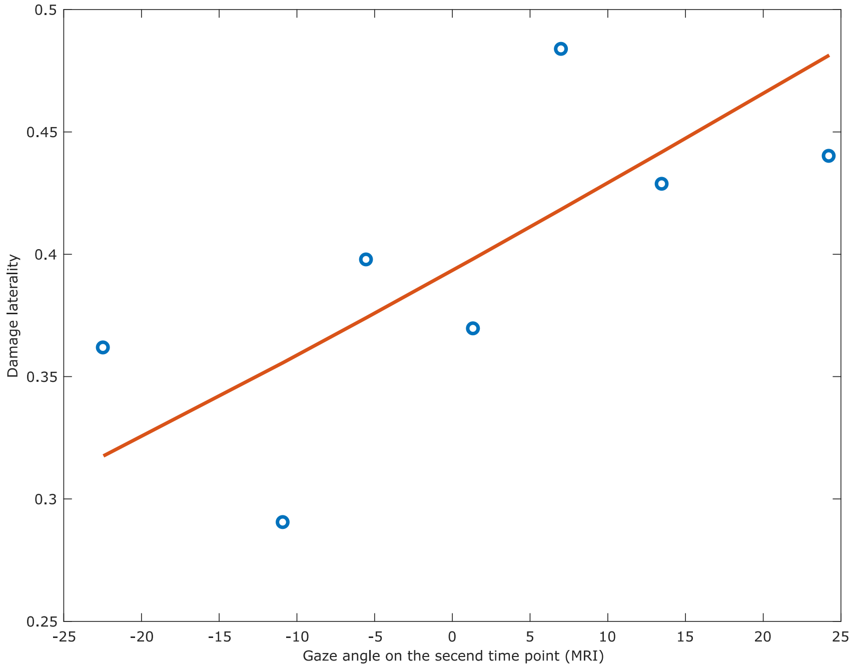


Figure 4.3 Distribution of patient gaze at the time point of the MRI

The gaze on the second time point was derived from the same semi-automated segmentation algorithm based on the intraocular lenses on MR scans. Seven bins of deviated gaze angle were empirically pre-defined. For each bin, the averaged ratio of lesion volume dropped within right hemisphere was plotted in blue circle. The relation across gaze was plotted with a general linear model fitted by maximum likelihood (in red). The relations between gaze direction typically 24-72 hours after admission and the laterality of brain lesion showed much less dependence, as the mechanism of neurological recovery is more complex.

To model therapeutic inference, the patients exhibiting a gaze deviation to the left of 12 degrees or more at the time point of CT, returning to within 3 degrees of the midline by the time of the MRI were labelled as ‘recovered’, and those whose deviation remained at least 6 degrees to the left were labelled as ‘persistent (unrecovered)’. Subjectively and empirically, we set the thresholds above to divide the whole lesion maps into two illustrative categories. Naturally, there is no hard boundaries for either abnormality or normality. We chose deviation to the left side because left-hemispheric ischemic strokes tend to be more frequent and appear to result in a worse outcome comparing to their counterparts induced in right hemisphere (Hedna et al., 2013, Yamamoto et al., 2014). In addition, one sided model was chosen to minimise the complexity of the underlying neural dependents and make the contrast with low-dimensional approaches more conservative.

4.2.3 Therapeutic inference

We established two series of therapeutic simulations modelled by low- and high-dimensional approaches respectively to identify and quantify the relative inferential sensitivities to detecting the effect of each intervention.

Specifically, interventions were simulated with varying effect sizes, which allowed us to estimate the ability of detecting an effect across a range of sizes. We opted to use simulated rather than real interventions in order to obtain reliable ground truths against which the comparative fidelity of low and high-dimensional models could be compared. Two broad categories of interventions in focal brain damage were simulated as follows: lesion-reducing and lesion-retaining interventions. The former is generally deployed in the "hyper-acute" period to salvage the threatened brain, such as drug-induced or mechanical thrombolysis in ischemic stroke; the latter is usually deployed at a later stage after the lesion is established, such as drug-assisted or behavioural rehabilitation. For the purpose of this study, a lesion-reducing intervention was simulated to modify the lesion anatomy in its periphery at a variety of pre-defined levels of effect; and a lesion-retaining intervention that did not change the lesion itself was simulated to assist the functional adaptation of the brain during recovery with varying effect on the whole patient.

4.2.3.1 Data preparation

To implement the therapeutic inference models, preliminarily, the stack of segmented 1172 lesion maps was re-oriented to voxel size of 6mm, which finalised 5789 voxels in each lesion mask. Then, by binarizing the lesions, the voxels involved in each lesion were simply referred as features that hit by a focal brain injury or not.

4.2.3.2 Lesion non-altering interventions

To simulate a lesion-retaining intervention, we randomly divided the whole set of patients into two equal parts: one half was marked as treated, and the other half was referred as non-treated. By varying the proportion of effectiveness from 10% to 90%, we randomly chose the proportion of corresponding effect size from the treated patients and simply changed their outcome as "*recovered*". In this way, a continuous function simulating therapies can be estimated to describe the relation between the size of therapeutic effect and models' ability to detecting its presence (Figure 4.4). We evaluated the interval of effect size between 10% and 90% at a step of 10% in order to obtain adequate coverage. A *p* value less than 0.05 achieved by the statistical test using for evaluating the given intervention was set as the criterion for successful detection. A total of 600 iterations of sample randomisations at each effect size were performed to stabilise the estimate and assure the statistical confidence.

4.2.3.3 Lesion altering interventions

Here we assume the behavioural effect of the interventions is dependent on the lesion anatomy, which means the individual outcome may change with change of the lesion itself. To simulate this intervention with controllable amounts of lesion reduction, we applied a variety of effect sizes in range of 10% to 90% stepped by 10% to shrink the lesion at according proportion in a way of morphological surface erosion. The outcome was thereby determined by whether or not the critical functional areas that the individual outcome is dependent on were treated. The therapeutic function and the way of evaluation were the same as above describe in lesion-retaining interventions.

4.2.3.4 Low- and high-dimensional models

For each class of intervention, we established two sets of analysis models differing in the parameterisation of the lesion (low- and high-dimensionality) to identify whether or not the sensitivity to detecting the effectiveness of an intervention could be improved by taking into account the intrinsic complexity of lesion architecture and its inner interactions.

In the case of low-dimensionality, the lesion was only parameterised by its total volume. First, the lesion volume together with the induced outcome by a specific intervention were used as multivariate predictors to infer the ability of detecting each patient to be treated or not. Then, each patient’s age and sex referred as co-factors were appended to extend the predictors to identify whether or not the extra crude parameters could improve the sensitivity. In common with conventional therapeutic studies, all low-dimensional models in this chapter were implemented as standard general linear models (GLMs) which are applicable to a small number of variables and provide estimates with conventional least-square minimisation.

In the high-dimensional case, the parameterisation of lesions was the whole 5789 voxels in each lesion map which were binarised to indicate presence or absence of damage across the entire brain. We set a threshold to restrict all the voxels involved into the models to be hit at least twice among the set of patients. In this way, the very rarely damaged voxels which might be artifacts were removed from parameterisation so as to improve the reliability of modelling. Similar to the low-dimensional models, the whole brain parameterisation together with the outcome induced by a specific intervention or the extra co-factors: age and sex were combined as predictors to identify whether or not a patient was treated.

To cope with the predictors including a large number of variables in particular that were substantially more than the number of cases, we need a different inferential framework rather than GLMs chosen for low-dimensional models because the performance and reliability of conventionally estimated GLMs is seriously degraded by the irrelevant variables in overdetermined models. We thus employed SVMLin – a linear transductive support vector machine (TSVM) (Sindhwani and Keerthi, 2006, Sindhwani and Keerthi, 2007) as the choice of technique for the following reasons. First, the foundations of kernel machines are well established and widely used. Among their diverse implementations, SVM is the best known member. Given the presence of unlabelled data here—scans where gaze information was not available—TSVM is a natural extension of original SVM, well-suited to our specific situation, for modelling this binary classification problem in a semi-supervised approach. The details of the theory of transductive learning has been discussed in the introduction chapter, and can also been found in previous studies (Bennett and

Demiriz, 1999, Fung and Mangasarian, 2001, Joachims, 1999b). As a variation of TSVM, SVMlin can provide superior scalability and a multiple switch heuristic not only to optimise but also to speed up the training process. Moreover, it can cope with large scale data characterised by sparsity, which fits the properties of our lesion set well.

Specifically, a multi-switch classification model involving all within-brain voxels (5789 variables) hit at least twice within the set of lesions was trained by SVMLin to predict the status of change in gaze – recovery or persistence, in which the adjustable regularisation parameters λ and λ', the maximum number of switches in TSVM and the possible class fraction of unlabelled data were optimised by 10-fold cross-validation. A total of 10 randomisations were conducted to divide the stack of 1172 lesion maps into two independent portions: 80% of lesions was used as training dataset, and the other proportion was for the purpose of evaluation. The reliably optimised model that provided best predictive performance was used as a classifier for outcome prediction in each patient.

In order to make a straightforward comparison between high- and low-dimensional models, the optimal SVMLin classifier described above was used to generate a single regressor used for predicting each patient's behavioural outcome which was then added to a GLM as a co-variate. The response variable was exactly the same as treatment of each patient in both high- and low-dimensional GLMs. The predictors in the low-dimensional models included the intervention, lesion volume, age and sex; in the high-dimensional models they were identical other than the extra co-variate derived from the classifier indicating the prediction of untreated outcome in each patient. As aforementioned, we randomly chose half of the entire patient cohort as treated, and the other half as non-treated, which would have no relations to the outcome prediction. Hence, we expect the natural outcome predictor derived from the high-dimensional TSVM classifier would improve the sensitivity of detecting a treatment effect: the better the predictive power, the greater the sensitivity (Figure 4.4).

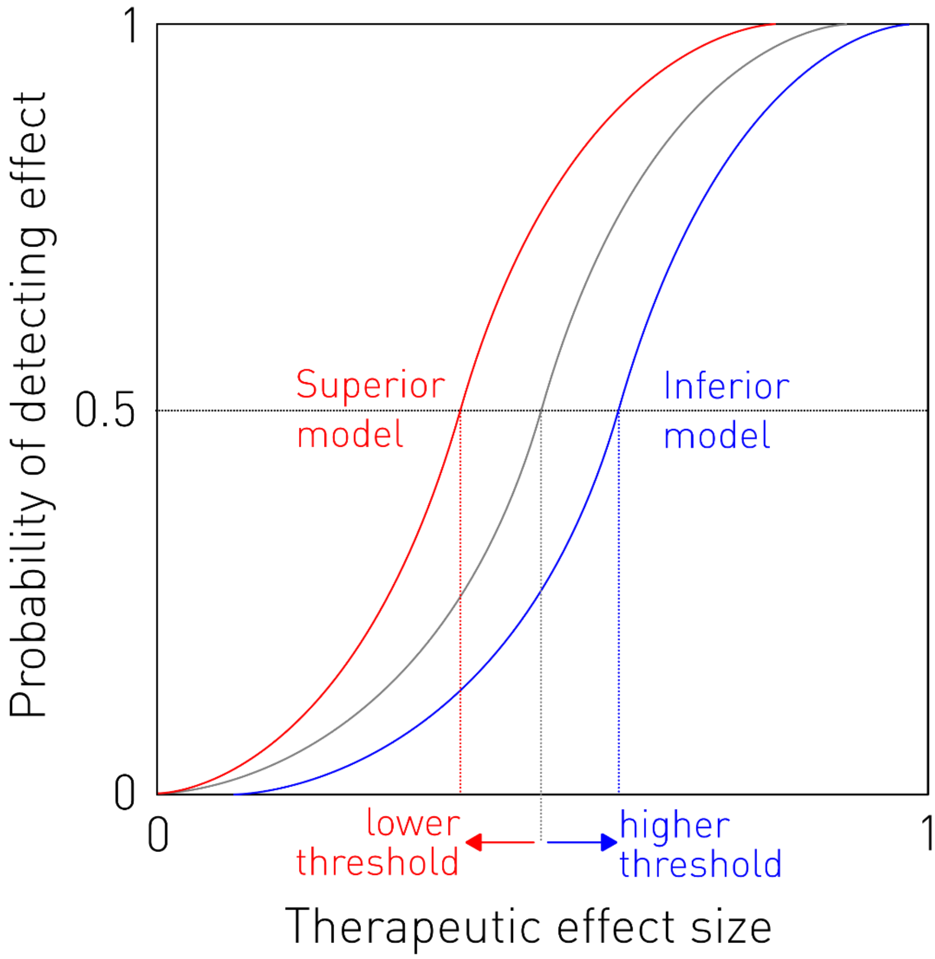


Figure 4.4 Therapeutic function

A continuous monotonic function is used to relate a range of therapeutic effect size (0 to 1, stepped in 10%) and the probability of correctly detecting it. The midpoint indicates the point where half of the trials successfully identify the intervention. Correspondingly, the thresholds of minimum effect size required to identify intervention as successful is used to evaluate the detectability of modelling – a sensitive model will be shifted to the left (in red) and less sensitivity will lead to shifting to the right (in blue).

Based on above description of the lesion-retaining intervention, we modelled 600 randomisations of the data with high- and low-dimensional GLMs which differed only in the inclusion of the predictive factor derived from the SVMLin estimates. For each effect size in range from 10% to 90%, the models which achieved statistical significance (the *p* value associated with the response variable was less than 0.05) were labelled as effectively identifying the treatment. Thus, an estimate of sensitivity and according confidence intervals were available to generate separate therapeutic functions for discriminating the ability of detecting treatment effectiveness between high- and low-dimensional models.

Similarly, the high- and low-dimensional GLMs were performed in the same way in the case of lesion-altering interventions except that the effect size was exerted via lesion shrinkage to a range of percentage from 10% to 90% in all patients who were randomly selected as treated. To be different from the non-lesion altering interventions where the outcomes of treated patients were simply changed as "*recovered*", each treated image shrunk on a specific effect size in lesion-altering interventions was therefore estimated with the SVMLin classifier for an induced outcome. In low-dimensional models, this induced outcome associated with lesion volume and optional co-factors referred as age and sex were used to identify if a patient was treated or not. Whereas, in high-dimensional models, the natural outcome predictive factor derived from the originally untreated lesions was added to the above multivariate predictor to detect the treatment in patients. For each effect size (the proportion of lesion erosion), the treatment was labelled as successfully identified if the statistical significance was achieved ($p < 0.05$). Thus, the estimates of sensitivity and related confidence intervals could be visualised by the therapeutic functions of high- and low-dimensional models, respectively.

4.2.4 Software and hardware

All simulated prediction models were implemented in the environment of MATLAB (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.).

For both classes of simulated interventions, a 12-core Inter® Xeon® CPU E5-2620 2.00GHz processor with 64GB RAM and 9TB 7200 RPM SATA hard drive was used to conduct both TSVM training and GLMs analysis under a 64-bit Linux operation system (Ubuntu version 15.04).

4.3 Results

For the stack of 1172 patients, the distribution of gaze was centred on the midline at both time points of CT and MRI but strongly biased by the lesion, which indicated the dependence of gaze deviation on the neuroanatomy (Figure 4.1 and 4.5).

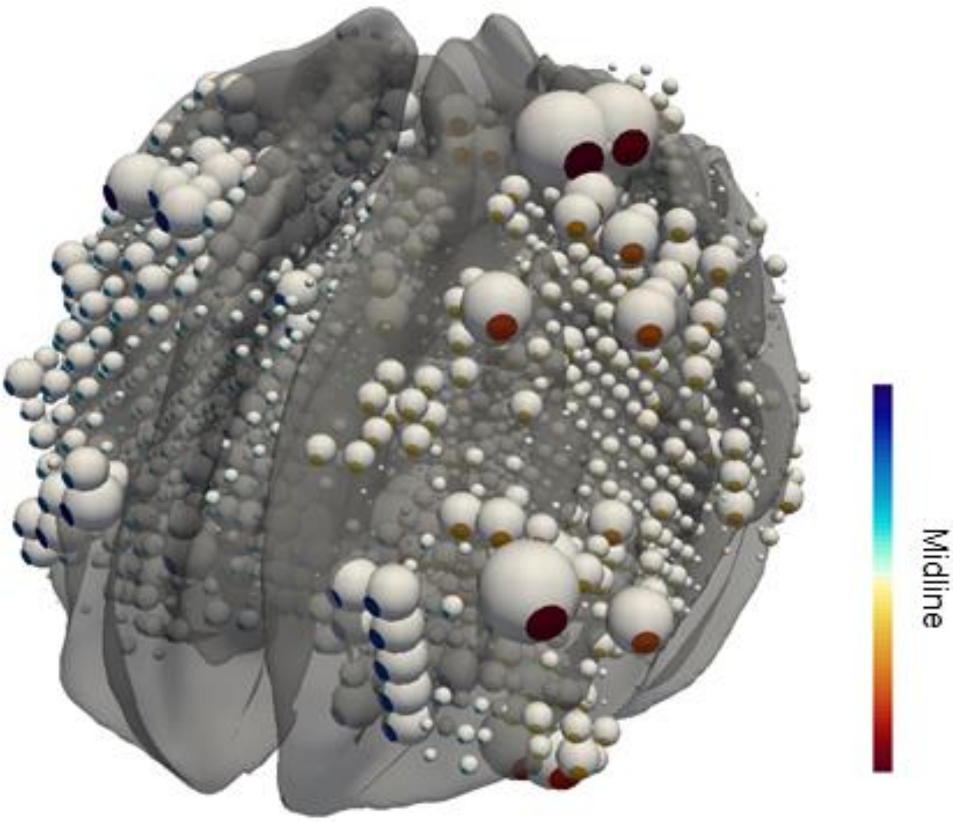


Figure 4.5 Voxel-wise tuning of gaze deviation on admission

Gaze direction is strongly correlated with the anatomical patterns of brain damage.

Based on the standard approach applied in conventional previous studies, we first fitted multivariate linear regression models to the data with low-dimensionality that regressed the outcome to the crude factors of intervention, the volume of the lesion, as well as age and sex. A corresponding set of high-dimensional models that adequately parameterised the lesion within the entire damaged brain rather than merely lesion volume, were fitted to determine the sensitivity of therapeutic inference better lesion parameterisation. To quantify the impact of this theoretical advantage, the high-dimensional models were established with exactly the same predictors and response variable as applied in the standard multivariate linear regression in low-dimensionality except for an additional covariate predictive factor indicating a natural outcome prediction regardless of any intervention which was derived from a classifier learnt by SVMLin (Sindhwani and Keerthi, 2006, Sindhwani and Keerthi, 2007) to relate the high-dimensional anatomical pattern of damage to the gaze deviation. The predictive performance of this classifier on a set of independent test samples of data was 78.33% (se = 1.70%) in sensitivity and 82.78% (se = 0.56%) in specificity, which substantially captured the relations between natural outcome variance and the distributed pattern of damage in a high-dimensionality (Figure 4.6).

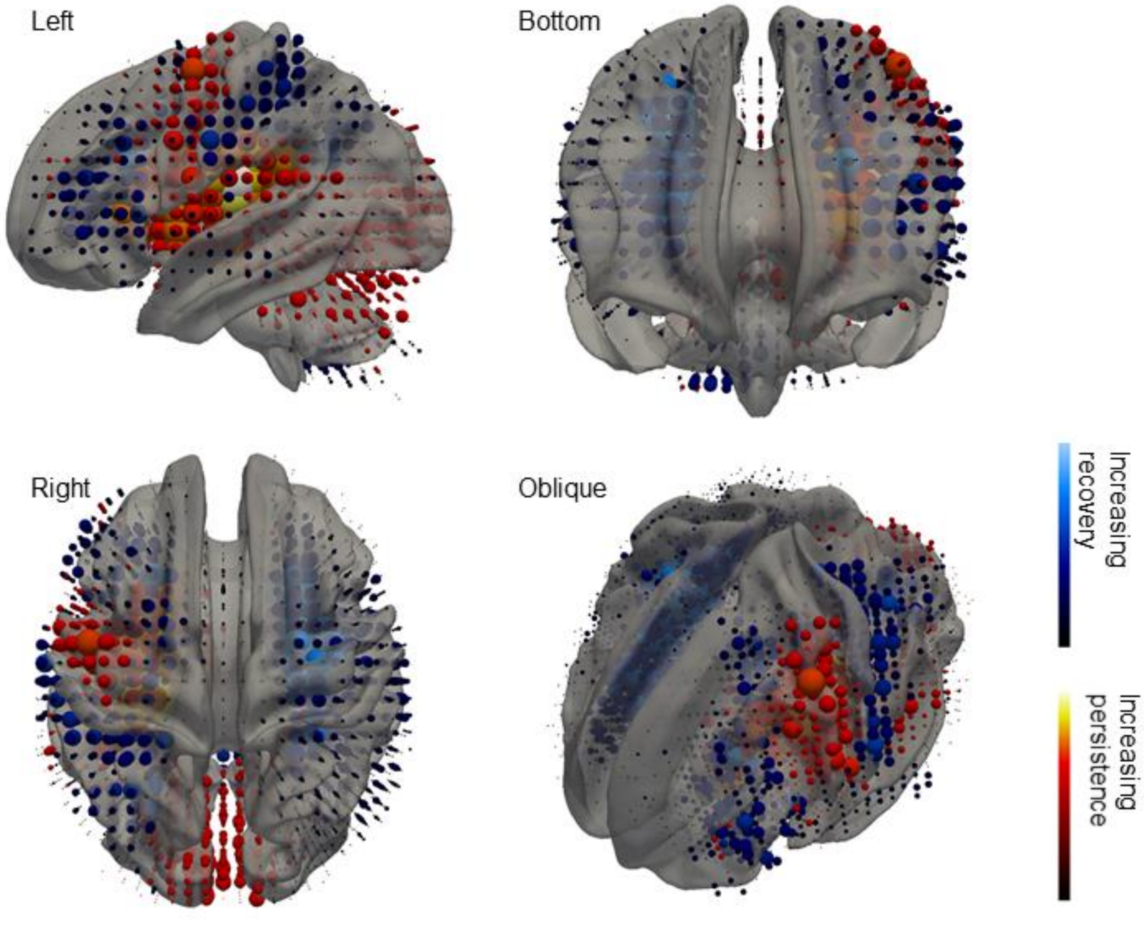


Figure 4.6 High-dimensional classifier derived from SVMLin

The proper classifier derived from the high-dimensional analysis revealed the anatomical patterns to discriminate the voxels that contribute to either neural recovery or persistence.

4.3.1 Non-lesion altering intervention

The above low- and high-dimensional analysis were modelled in each class of interventions. First, the non-lesion altering interventions were used to simulate a process of rehabilitation or physiotherapy after ischemic acute stroke without altering the lesion itself (Johansson, 2000) that hypothetically increased the proportion of patients varying by the pre-defined effect size from 10% to 90% to be recovered from a deviated gaze to normal range between two observed time points: CT on admission and MRI in 24 to 72 hours typically. Specifically, for each effect size, a series of models were evaluated by 600 randomisations where a hypothetical intervention was iteratively effective (to induce the outcome as '*recovered*') on a corresponding proportion (effect size) of the patients who were randomly selected to receive treatment. At the first time point (CT), the gaze deviated leftwards at 12 degrees or more was labelled as abnormal; whilst a normal gaze measured at the second time point of MRI was defined as within 3 degrees of the midline. To evaluate both set of models, the statistical significance for a '*trial*' to be positive or negative was determined at $p < 0.05$.

The therapeutic function produced by the low-dimensional models exhibited a 50% detection rate threshold of 62.90% (95% confidence intervals: 61.50% ~ 64.40%) of those treated responding to the intervention (Figure 4.7 in black), which is remarkably insensitive. Here, analogously to a meta-analysis, the threshold was determined as the midpoint of the therapeutic function (horizontal dotted line) where half of all trials were correctly identified as positive. Another series of low-dimensional analyses that were exactly the same as above but appended the predictors with age and sex as co-factors, were conducted to identify whether or not the extra crude parameters in relation to the patients would improve the sensitivity of therapeutic inference. The therapeutic function produced almost the same threshold at 62.91% (95% confidence intervals: 61.58% ~ 64.40%) (Figure 4.8 in green). Age and sex used as extra variables in low-dimensional multivariate linear regression models were not significantly effective in enhancing sensitivity.

The high-dimensional analyses were conducted in exactly the same way except that the output of a high-dimensional classifier trained by a transductive linear support vector machine was involved as an extra predictive factor. The models enhanced in this way showed substantial improvement of the sensitivity of therapeutic inference, which significantly shifted the threshold leftward to 56.00% (95% confidence intervals: 54.65% ~ 57.35%) (Figure 4.7 in red). As expected, a high-dimensional parameterisations of the complex lesion architecture increased sensitivity to detecting the effectiveness of interventions. Similarly, age and sex used as co-factors were also conducted in the high-dimensional analyses. The threshold corresponded to 56.10% (95% confidence intervals: 54.73% ~ 57.46%) of the patients that received treatment responding to the intervention (Figure 4.8 in blue). As a result, the crude parameters such as age and sex made no significant contribution in high-dimensional models either.

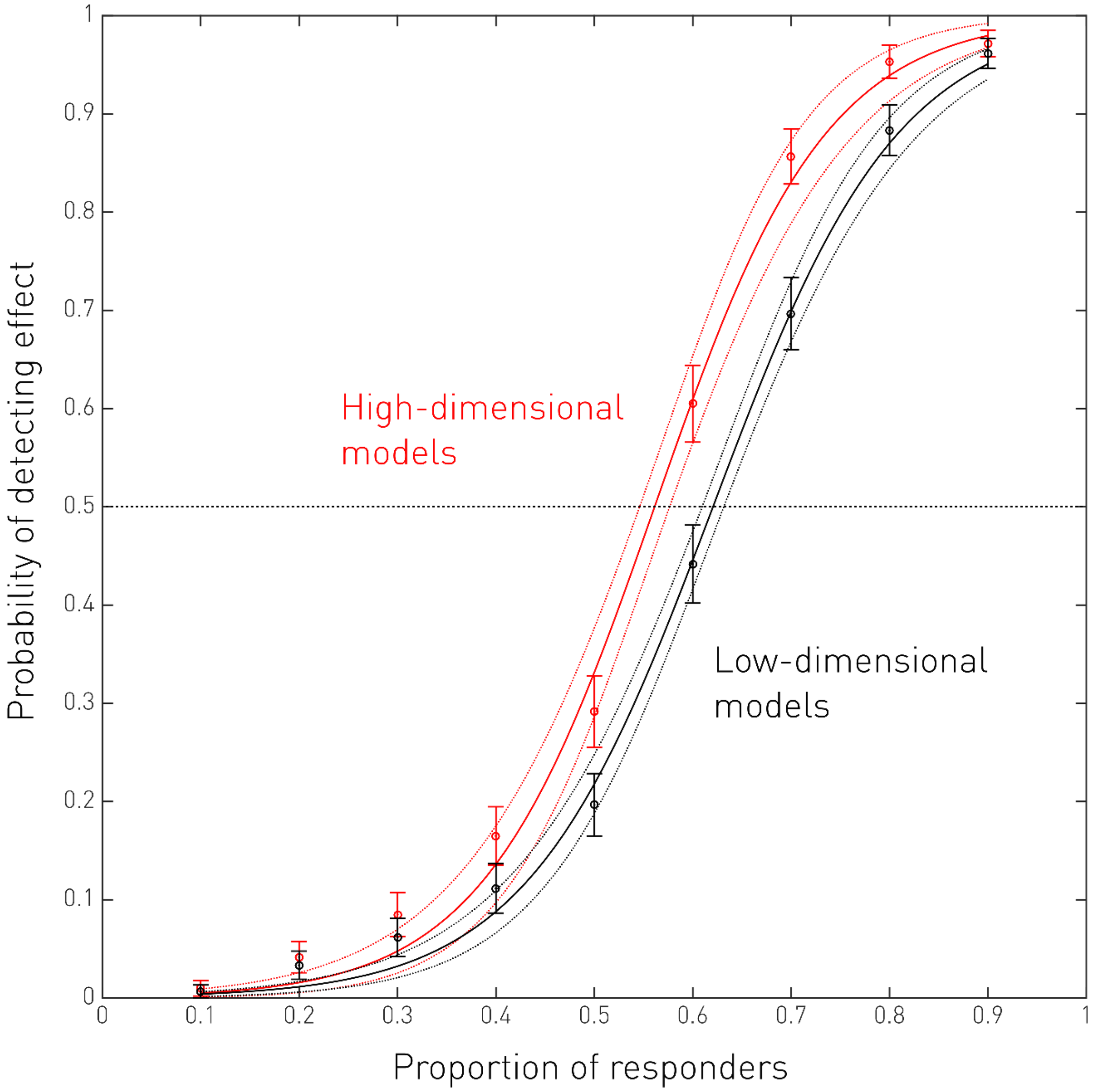


Figure 4.7 Empirical therapeutic inference function of non-lesion altering intervention

A set of effect sizes varying from 0.1 to 0.9 stepped by 0.1 was applied to a hypothetical non-lesion altering intervention modelled in low- and high-dimensionality, respectively. For each ratio of effectiveness, the corresponding proportional patients were successfully treated as recovered. In both types of models, the probability of detecting an intervention as successful was averaged by a total of 600 iterative randomisations. In the low dimensional models (in black), only the factors of intervention and lesion volume were used to construct the linear regression models. The 'trial' would be labelled as positive if the *p*-value for the intervention was smaller than 0.05. The error bars correspond to 95% confidence interval of the means. The mean performance was fitted with a robust spline fit and relevant estimate of 95% confidence intervals were plotted in dotted lines. For the high-dimensional case (in red), the model construction and subsequent analysis were identical other than a high-dimensional classifier correlated gaze and outcome was involved. The high-dimensional analysis method substantially shifted the threshold to the left, which illustrated the superior sensitivity for detecting the effectiveness of a given therapeutic.

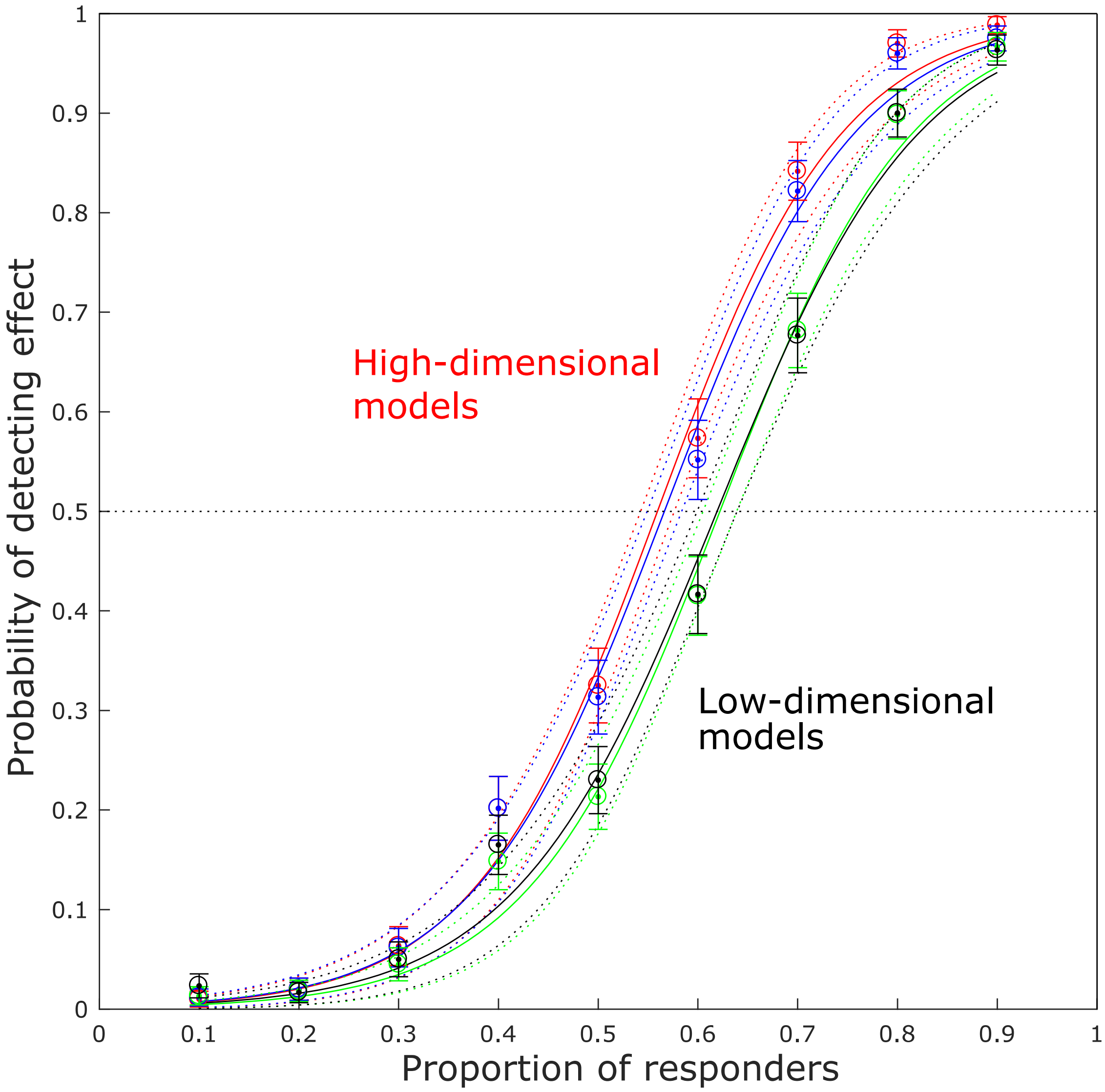


Figure 4.8 Empirical therapeutic inference function of non-lesion altering intervention (with additional co-factors)

The exact same model construction and subsequent analysis were conducted, except that age and sex were appended as co-predictors in both low- and high-dimensional models. On top of plots in Figure 4.7, the performance achieved in low-dimensional models with age and sex did not show significant change (in blue). The same situation was observed in high-dimensional approach as well (in green). This revealed the crude parameters, like age and sex did not present essential contribution to a therapeutic inference.

4.3.2 Lesion-altering intervention

In lesion-retaining interventions, modelling the correlation between the distributed pattern of brain damage and gaze deviation so as to identify the spontaneous recovery regardless of any intervention, the sensitivity of responding to intervention in therapeutic inference was remarkably increased. Based on this, a lesion-altering intervention simulating a process of a thrombolytic therapy (e.g. thrombolysis) (Wardlaw et al., 1997) was also analysed in low- and high-dimensional models, respectively. As aforementioned, the anatomical architecture of lesion was directly linked to the behavioural outcome, we therefore expected the involvement of a high-dimensional classifier using as a predictive factor that learnt the relation between lesion changed by the intervention and outcome would reasonably amplify the discrimination between the analyses modelled in low- and high-dimensionality. During the process of simulated thrombolysis, lesion volume was peripherally reduced by a morphological surface erosion at proportions of 10% to 90% separately. An extreme case could be a large lesion centred in a small critical neural locus which required a great deal of lesion shrinkage to reflect the intervention. At the other extreme may be a small lesion closely located the edge of diffuse neural locus which would produce effect on behavioural outcome with minimal change of lesion volume. Thus, the diversity of behavioural consequences induced by the lesion-altering intervention may be highly reliant on the complex intersection between anatomical lesion architecture and functional anatomy, which could be plausibly captured by a high-dimensional model only. By contrast, a low-dimensional method was insensitive if saved lesion volume merely induced physiological change rather than effect on behavioural outcome.

To quantify the theoretical advantage of high-dimensional models in the scenario of lesion-altering intervention, as above, for each effect size from 10% to 90%, we constructed 600 randomised models in low- and high-dimensionality, respectively. In the low-dimensional method, the standard linear regression models were fitted with induced behavioural outcome and volume of lesion as predictors to predict intervention. Then, age and sex using as co-factors to identify whether or not the extra crude parameters could contribute to the sensitivity of therapeutic inference. The '*trials*' were labelled as positive if *p* value for the intervention was less than *0.05*. To be similar with the criterion used in non-lesion altering intervention, we also use the midpoint of a therapeutic function where half of all trials were identified as positive to determine the threshold. The therapeutic function yielded by this series of analyses illustrated a threshold at an effect size of 78.40% (95% confidence intervals: 75.75% ~ 81.05%) (Figure 4.9 in black), which was a very substantial effect size that identified the intervention to be successful. Appending age and sex as extra predictive factors in the standard linear regression models produced a very similar therapeutic function with the above. The threshold showed an effect size of 78.41% (95% confidence intervals: 75.75% ~ 81.05%) (Figure 4.10 in green).

In the case of high-dimensional models, a series of the same models other than an additional covariate predictive factor derived from a transductive linear support vector machine capturing the relation between damaged brain patterns prior to the interventions and behavioural outcome of gaze were constructed. The threshold of the therapeutic function was remarkably shifted leftward to 55.00% (95% confidence intervals@ 53.10% ~ 56.90%) (Figure 4.9 in red). As expected, the two approaches were differentiated to a greater extent compared with the difference gained in non-lesion altering interventions. Again, we reran exactly the same high-dimensional models with age and sex as extra predictors. There was no significant contribution shown in the results: 55.20% (95% confidence intervals: 53.31% ~ 57.13%) (Figure 4.10 in blue).

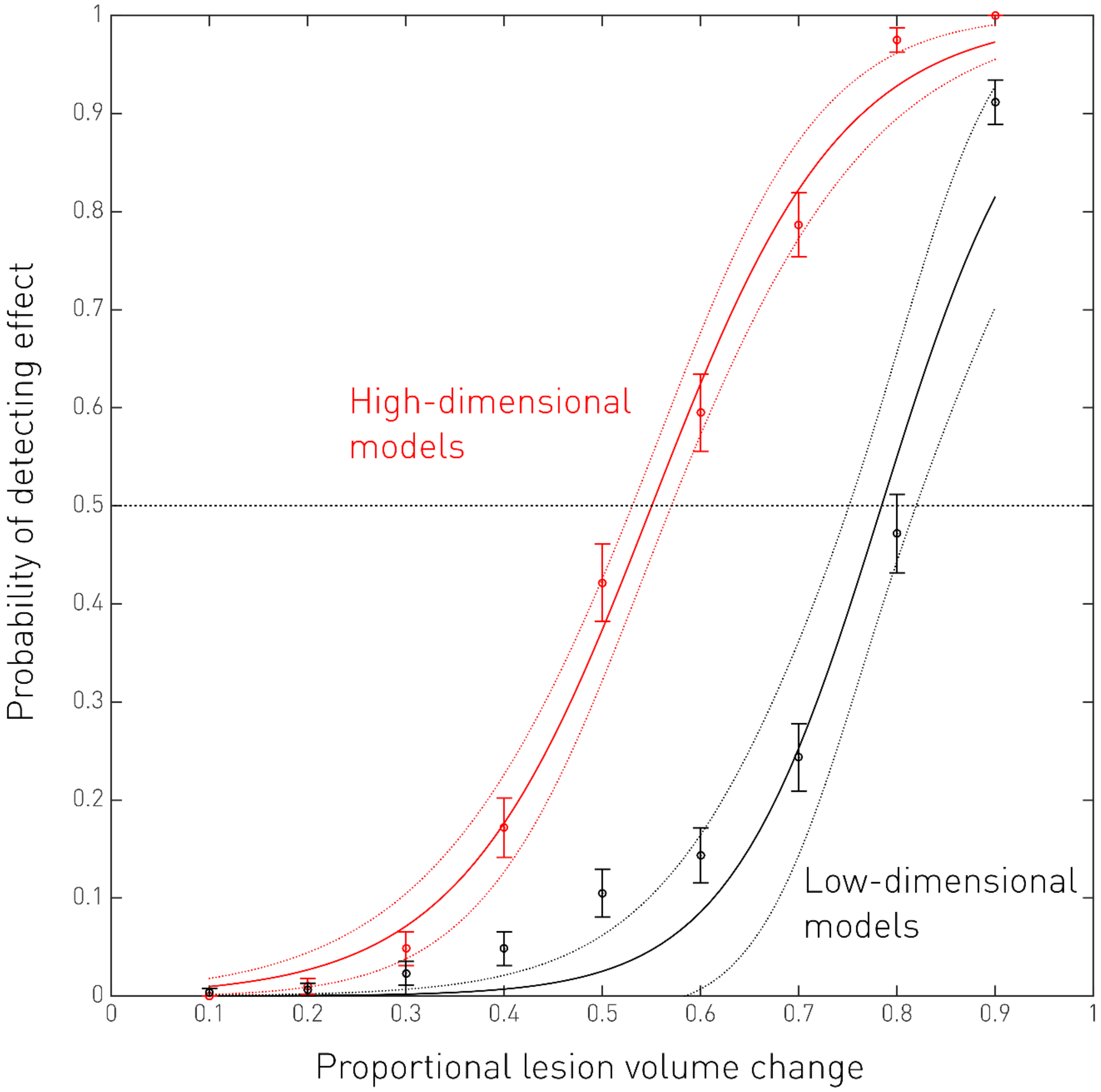


Figure 4.9 Empirical therapeutic inference function of lesion-altering intervention

A set of effect sizes varying from 0.1 to 0.9 stepped by 0.1 was applied to a hypothetical lesion-altering intervention modelled in low- and high-dimensionality, respectively. For each ratio of effectiveness, the corresponding proportional lesion volume was shrunk. In both types of models, the probability of detecting an intervention as successful was averaged by a total of 600 iterative randomisations. In the low dimensional models (in black), only the factors of intervention and lesion volume were used to construct the linear regression models. The 'trial' would be labelled as positive if the *p*-value for the intervention was smaller than 0.05. The error bars correspond to 95% confidence interval of the means. The mean performance was fitted with a robust spline fit and relevant estimate of 95% confidence intervals were plotted in dotted lines. For the high-dimensional case (in red), the model construction and subsequent analysis were identical other than a high-dimensional classifier correlated gaze and outcome was involved. The high-dimensional analysis method substantially shifted the threshold to the left, which illustrated the superior sensitivity for detecting the effectiveness of a given therapeutic.

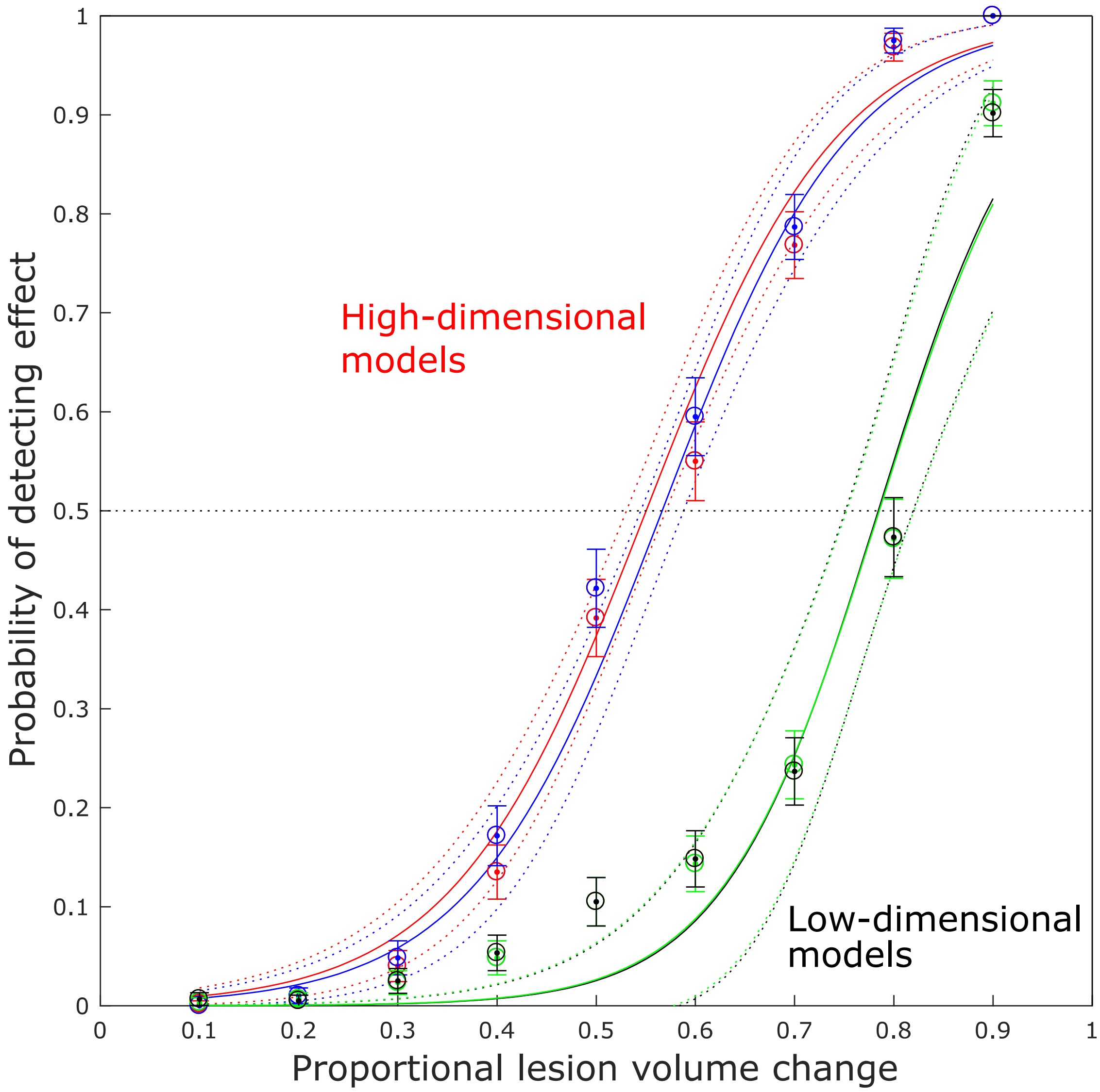


Figure 4.10 Empirical therapeutic inference function of lesion-altering intervention (with additional co-factors)

The exact same model construction and subsequent analysis were conducted, except that age and sex were appended as co-predictors in both low- and high-dimensional models. On top of plots in Figure 4.8, the performance achieved in low-dimensional models with age and sex did not show significant change (in blue). The same situation was observed in high-dimensional approach as well (in green). This revealed the crude parameters, like age and sex did not present essential contribution to a therapeutic inference.

4.4 Discussion

The therapeutic functions produced by analyses of low- and high-dimensionality demonstrated remarkable differences of sensitivity to detecting interventions. This provides empirical support to what is given analytically: where a multiplicity of interacting factors determine the causal relation between brain damage and behavioural outcome, better parameterisation of the relation is bound to improve our ability to identify any one causal factor, such as treatment. Owing to the intrinsic complexity of anatomical lesion architecture (Mah et al., 2015, Mah et al., 2014b, Xu et al., 2017a, Xu et al., 2017b), such collateral and covariant factors are bound to be numerous in their quantity and complex in their interactions.

To benefit from the rapid improvement in the inferential power of machine learning techniques and the computational ability of hardware, a high-dimensional method to sufficiently parameterise human brain is necessary in theory, and is now a feasible option in practice. For a long time, we have been aware that the dimensionality of the human brain is grounded in highly complex functional anatomy that is as yet largely unknown (Sporns et al., 2005, Sporns, 2011, Bullmore and Sporns, 2009). The architecture of lesions is also complex, but potentially modellable with the right high-dimensional methods (Mah et al., 2014b, Xu et al., 2017a). We therefore simulated non-lesion altering and lesion-altering interventions modelled by both low- and high-dimensional methods to identify the impact on inferential sensitivity of modelling the lesion-deficit relation with greater fidelity. We have seen that the impact is substantial.

It is crucial to recognise that we propose a general approach to identify the relation between the distributed pattern of brain damage and behavioural outcome rather than a method for subgroup analysis merely identifying a minority of idiosyncratic cases for whom a high-dimensional method exceptionally works. For every case of therapeutic inference, we claim the outcome is influenced by numerous interacting factors of lesion anatomy. Hence, the whole brain map needs to be parameterised so as to isolate any specific effect predicted by any single factor.

As a general approach intended to illustrate therapeutic inference for a wide range of neurological disorders associated with focal brain injury, we therefore constructed both non-lesion altering and lesion-altering interventions to simulate cardinal forms of intervention. The former simulated an intervention such as rehabilitation, where a treatment facilitates adaptation to a lesion rather than altering it directly. By hypothetically varying the proportions of patients recovering from the deficit, a therapeutic function was produced, describing the relation between the proportion of responders and the probability of detecting a therapeutic effect. Three crude parameters (volume of lesion, age and sex) widely employed in conventional studies were used to fit our low-dimensional models. The threshold for detecting a positive therapeutic effect with the low-dimensional approach was estimated to be very high at 62.9% of patients recovering. By contrast, the high-dimensional method significantly improved the threshold to 56% by adding a covariate predictor capturing individual variability. The improvement was achieved by identifying patients whose gaze deviation recovered regardless of any treatment, a predictive signal no simple parameterisation could conceivably capture.

A lesion-altering intervention corresponds to an even more important therapeutic scenario, where the intervention alters the morphology of the lesion itself, and thus the anatomy of the damage. Given the complex dependence on the underlying substrate revealed by our SVM model of gaze recovery, it is implausible that the interaction between a change in lesion morphology and the resultant deficit would be adequately captured by any simple covariate such as lesion volume. This likely explains the superiority of the high-dimensional models observed here.

Comparing the two therapeutic contexts, the thresholds produced by the high-dimensional method were more or less the same (56.00% and 55.00%), which demonstrated the consistent contribution of the covariate predictive factor learnt by a transductive linear support vector machine. The performance achieved by the low-dimensional method varied dramatically (62.90% and 78.40%), being substantially worse in the lesion-altering case. This indicated the prediction from the multivariate model parameterised with a few crude variables independently was very limited in terms of sensitivity – the more distributed relation underlying the model, the worse performance gained.

Since the brain mechanisms to control gaze are relatively simple, we therefore expect the contrast of estimates between two approaches would be considerably greater with more complex behaviour. Indeed, in future work, our explanatory models must ascend to greater complexity not only to deepen our understanding of brain, but to identify the effectiveness of new treatments, including many hitherto erroneously thought to be ineffective through conventional studies. Accordingly, it is also crucial to continuously grow the scale of modelled data to achieve the right balance of dimensionality and data scale.

4.5 Conclusion

Therapeutic inference in focally damaged human brain concerns the relation between induced structural change in the brain and its behavioural consequences. Conventionally, this causal relation has been established with models that rely on crude descriptions of the brain, and its damage, such as volume of lesion, age, sex, and discretised brain parcellations. As a result, individual variability arising from finely detailed differences in lesion anatomy is treated as noise, assuming each patient can be approximated no better than by the sample mean. We have shown that awareness of the intrinsic dimensionality of lesions, examined across a set of low- and high-dimensional models, has substantial impact on the ability to isolate a therapeutic effect within a hypothetical interventional trial. Conventional low-dimensional models are shown to be remarkably insensitive to therapeutic effects, for both lesion-retaining and lesion-altering interventions. By contrast, the sensitivities obtained from high-dimensional models were dramatically and consistently better, more so for lesion-altering interventions. These results have substantial implications for translational research, providing one remediable explanation for the common failure of interventional studies in humans involving agents shown to be effective in simpler animals. A re-evaluation of such studies with high-dimensional techniques is shown to be warranted.

Chapter 5 The identification of individual susceptibility to treatments with prescriptive inference

5.1 Introduction

We have argued that the complexity of the human brain in the context of focal brain injury arises not only from its functional organisation at multiple levels (Bassett and Gazzaniga, 2011), but also the structure of pathological damage. The fidelity of individual description is then intrinsically dependent on the fidelity of the description of each of these two aspects. The observed functional architecture may be grounded in relatively simple neuronal properties, but is clearly complexity and diverse as a whole (Yagi, 2013, Yoshimura et al., 2005). Equally, at the molecular level, though genetic factors specify individual neuronal trajectories of development, the overall organisation in any one individual cannot be easily derived from them. (Baaré et al., 2001, Peper et al., 2007, Pol et al., 2006). So, it will take some time before this level or organisation is adequately described.

Lesion architectures, by contrast, though very complex, may be easier to learn, and provide a mechanism through which the functional architecture may be further illuminated, as well as facilitating the management of patients with focal brain injury. In order to systematically demonstrate the dimensionality of distributed lesion anatomical patterns and reveal the hidden spatial bias caused by the conventional analysis, by modelling lesion-deficit analysis in low- and high-dimensionality, respectively, a couple of previous studies illustrated the inherent mislocalisation affected by univariate techniques (Mah et al., 2014b) and superior predictive performance obtained from high-dimensional multivariate analysis (Xu et al., 2017a). Crucially, these studies established a foundation which revealed the dimensionality of lesion to be distinct from the functional dimensionality and contribute to understanding the brain mechanisms in parallel.

This is relevant to clinical medicine, where intelligence drawn from populations needs to be “inverted” to the specific individual that is always the focus of clinical interest. A critical question arises: how to make the research findings derived from group analyses fit individuals well? In other words, how to model individuality when our intelligence is drawn from populations, and in particular in relation to *interventions*: the ultimate focus of clinical care. To investigate this broad question, two series of therapeutic inference models with high- and low-dimensionality, were evaluated to discriminate the ability to detecting the effect of a specific intervention that facilitated the change of behavioural outcome in a previous study (Xu et al., 2017b). The results showed the sensitivity of therapeutic inference was dramatically improved in high-dimensional models, where the lesion architecture was modelled in detail commensurate with the complex functional architecture to identify the patient those were recovered spontaneously. This demonstrated the dependence of the causal relation between a therapeutic intervention and its functional outcome on numerous neurological factors as well as their interactions -- all of these are inherently sourced from descriptions of the brain with sufficient dimensionality to capture its individuality. Further to the studies on lesion-deficit prediction (Mah et al., 2014b, Xu et al., 2017a) mentioned above, the models seeking to infer therapeutic effects in focal brain injury further illustrate the importance and necessity of complex parameterisation – a high-dimensional model of focally damaged human brain.

In the studies mentioned above, high-dimensional parameterisation of brain damage played a crucial role in extracting maximum performance from group level analysis. The reliability and sensitivity of any group study is naturally dependent on individual fidelity, but that is not its critical measure. In medical practice, however, the focus is the individual not the group: we wish to know how specific individuals respond to treatment. Although awareness of concepts has existed in clinical research and patient care for a long time – one instance is multi-dimensional immunological matching in transplantation – the adoption of tailored prescription in more complex pathological scenarios is very limited. Conventional medical practice is still reliant on the averaged responses across groups and uses them as the standards to fit everyone. Owing to the wide variance of the individual susceptibilities to a given treatment (Park et al., 2017), this low-dimensional “one-size-fits-all” approach will miss the information integrated in individualities, and consequently fail to provide the necessary precision of prescription. We therefore need a valid model that is able to extract the individual patterns on susceptibility from the group data.

In the following chapter, in order to identify individual susceptibility to treatments, we will combine patients with treatment to construct prescriptive inference with low- and high-dimensionality. We aim to examine whether or not a cohort of patients can be separated into sub-divisions to sensitively discriminate which medical decision, intervention or treatment is safe and efficacious to a specific patient; and which one is not. Methodologically, we expect the high-dimensional models that take into account individual features sufficiently will uncover the hidden sensitivity so as to contribute to the precision in tailoring treatments in practice.

5.2 Method

5.2.1 Patients

A set of 1333 patients admitted to University College London Hospital (UCLH) Hyper-acute Stroke Unit (HASU) over the past decade were selected. All patients were clinically diagnosed with acute ischemic stroke, and received MRI, including diffusion-weighted imaging (DWI), in which at least one acute ischemic lesion could be identified. A threshold for minimal DWI-quantified lesion volume was set at 50 milliliters which covered the 90th centile of our entire clinical population in order to reduce the disproportionate impact from very large lesions. An adjusted cohort including 1172 patients was thus produced.

The age distribution from the set of 1172 patients was in range of 18 to 97 years old with a mean of 63.89 and a standard deviation of 15.91. The gender ratio (Male) was 0.561 (Appendix B. Figure 7.1). Ethnicity for 952 (81.2%) from the full set of patients was disclosed in the clinical record which is illustrated in Appendix B. Figure 7.2. The distribution of age and sex, and the constitution of ethnicity reflected the catchment of clinical population in UCLH naturally.

5.2.2 Imaging

5.2.2.1 MR data acquisition

All acquisitions performed on the scanners manufactured by General Electric (GE) (Discovery MR 450, Genesis Signa and Signa Excite), Philips (Achieva, Ingenia and Intera), or Siemens (Avanto, Biograph mMR, Espree, Skyra, Symphony, Symphony Tim, Trio Tim and Verio) with field strength of either 1.5 or 3 Tesla in a single session for each patient were collected via a picture archiving and communication system (PACS) from UCLH (the specific range of scanners and related parameters are listed in appendix A). The originated voxel sizes were different in some extent between the scanners, and so were the according spatial resolutions. Typically, for 1.5 Tesla GE medical system, the voxel size was sampled as 1mm x 1mm x 6.5mm (Columns = 256, Rows = 256). For 1.5 Tesla Philips

medical system, the typical voxel size was 1.2mm x 1.2mm x 6mm (Columns = 192, Rows = 192); the voxel size was 1mm x 1mm x 6mm (Columns = 256, Rows = 256) performed on 3 Tesla Philips scanners. For Siemens, the voxels sizes and spatial resolution were 1.8mm x 1.8mm x 6.5mm (Columns = 128, Rows = 128) and 1.2mm x 1.2mm x 6.5mm (Columns = 192, Rows = 192) on 1.5 Tesla and 3 Tesla scanners, respectively. The routine of clinical practice with regular equipment upgrades and replacement is the main reason for the diversity of scanners over the period of data collection. All brain scans were performed for the purpose of clinical routine obeying clinical protocols.

The specific sequences extracted from each MRI session were axially required echo planar DWI with b-values of 0 s/mm^2 and 1000 s/mm^2 which is used for lesion segmentation (Mah et al., 2014c) so as to implement the series of simulations designed for this study. DWI is able to indicate the restricted ability of Brownian motion for extracellular water protons where cytotoxic edema causes imbalance. So, DWI is sensitive for detecting ischemic stroke and widely used for locating acute ischemic lesions (Warach et al., 1995, Löuvbld et al., 1997, Lövblad et al., 1998). In the clinical application, a DWI sequence usually includes images with b values of 0 s/mm^2 (a b0 image) and 1000 s/mm^2 (a b1000 image). The former type image shows stronger contrast on normal tissues but relatively insensitive to ischaemia. Conversely, an image with high diffusion weighting is used to sense water molecules movement. So, it is sensitive to ischaemia but poor to the contrast of normal tissues. Both types of images are complementary for brain registration and lesion segmentation, which rely on the contrast between normal tissue types and the discrimination between normal tissues and lesions, respectively.

5.2.5.2 Image pre-processing

A processing pipeline was implemented within MATLAB (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.) associated with a configured SPM12 toolbox (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/) to ensure the images to be processed systematically.

For the purpose of scientific imaging analysis, typically, the raw clinical MR images needed to be firstly converted from the complicated format of Digital Imaging and Communications in Medicine (DICOM) images to Neuroimaging Informatics Technology Initiative (NIfTI) format. Subsequently, in order to provide accurate and robust performance on image registration and segmentation in the later stage, we employed a pre-processing pipeline with the following steps. We empirically clamped the signal between 0.1% and 99.9% of the cumulative distribution estimated with a kernel density method (Botev et al., 2010), removing unusually low and high signals in each NIfTI file. Then an oracle-based 3D discrete cosine transform (ODCT3D) (Manjón et al., 2012) denoising method was applied for further noise reduction. In the next step, image registration was conducted based on each pair of DWI files. In order to optimise the alignment between the b0 and b1000 images in each pair, we used SPM12's standard co-registration function to have the b0 image rigidly co-registered to the b1000 image. Then, the normalisation / segmentation routine integrated in SPM12 (Ashburner and Friston, 2005) was applied to the b0 image to calculate a deformation field which described the optimal non-linear transformation of the b0 image into Montreal Neurological Institute (MNI) stereotactic space. Next, the deformation field derived from the b0 image was applied to the b1000 image to transform it into MNI space using the same routine. The parameters involved in the whole process of image registration were set at defaults by SPM12. Finally, both b0 and b1000 images in each pair were resampled to 2mm^3 isotropic with 6th degree b-spline interpolation (Mah et al., 2014b, Mah et al., 2014c) and manually checked against the SPM template to confirm the satisfactory of registration.

5.2.2.3 Lesion segmentation

A previously validated method based on the anomaly metric *zeta* (Mah et al., 2014c) was applied to segment lesions from the 1172 normalised b1000 images in a voxel-wise manner.

A set of preliminaries conducted to optimise the performance of lesion segmentation were performed as follows. First, 492 DWI images derived from patients suspected of acute stroke but found to have normal imaging were selected

to form a reference normal image set. No acute lesion was visible here in any patient, and none was reported by the corresponding radiological record. Every scan was pre-processed using the exact same pipeline and resliced at the same resolution of 2mm isotropic voxels, which spatially presented as a 91 x 109 x 91 matrix. Then, by combing the tissue maps provided by SPM12 including white matter tissue, grey matter tissue and cerebrospinal fluid (CSF), an inclusive brain mask was created. So as to remove image artefacts, we set an empirical threshold to exclude areas in the frontal and temporal poles commonly prone to artefacts on DWI sequences, and adjusted the inclusive brain mask accordingly. Subsequent lesion segmentation was performed within the confines of this adjusted mask. Next, the signal distribution was normalised for both lesion and reference datasets. Such signal normalisation sought to reduce instrumental and other incidental signal heterogeneity, so that the comparison between individual images could be maximally sensitive. The procedure was performed as follows. A binary white matter mask was created from SPM12's white matter tissue probability map by thresholding at a value of >0.9. A robust kernel density estimate method (Botev et al., 2010) was then applied to the voxels falling within both customised maps in each b1000 image to obtain the peak white matter distribution which was then subtracted from each corresponding b1000 image to normalise the signal distribution. Guided by information in the radiological report, signal normalisation on unilateral and bilateral lesions was performed differently. For unilateral lesions—the vast majority—only the unaffected hemisphere was used to estimate the peak. For bilateral lesions, since ischemic lesions tended to be small in these circumstances, the kernel density estimate was applied to the whole brain. In both cases, the signal normalisation was equally effective in adjusting the differences in the signal distribution across all scans performed by a variety of scanners and related technical and practical variance.

Lesion segmentation was performed in a voxel-wise manner. We calculated the zeta anomaly metric for each single voxel in each image independently against the reference set. To be specific, the voxel-wise zeta value in the test image is derived from the mean distance to the k nearest neighbors drawn from the anatomically homologous voxels in the unlesioned reference set, normalised by the mean distance between the k neighbors themselves (Mah et al., 2014c). The measure was the Euclidean distance in signal space. The only manipulable

parameter here is the number of nearest neighbors – k – commonly chosen with reference to the size of the reference dataset. In this study, we set k at 23 based on the calculation as below:

$$k = Ceiling\left(\sqrt{492}\right).$$

Zeta is a continuous metric. An adaptive threshold was therefore determined for each image. In brief, volume connectivity computed by *bwlabeln*, an integrated function of MATLAB (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.) was applied to segment each lesion map into connected clusters. Then, each cluster was estimated under a generalised extreme value (GEV) distribution to derive mean and variance values which were used to decide threshold adaptively as described in Mah et al, 2014.

5.2.3 Dimensionality reduction

Modelling performance is sensitive to the ratio of number of cases to number of features. A helpful preliminary is therefore to reduce the dimensionality of individual features while retaining the discriminability of each case. In human neuroanatomy, brain asymmetry between two hemispheres has been observed in two distinct aspects – neuroanatomical differences, and functional and behavioural lateralisation (Hugdahl, 2005, Toga and Thompson, 2003), but these differences are minor and do not generally extend to the vascular tree (Wright et al., 2013). Vascular lesions rarely cross the midline of brain and are generally literalised (Nachev et al., 2008). It is therefore reasonable to collapse our stack of vascular lesions onto one hemisphere to reduce dimensionality for following predictive models.

Based on the collapsed lesion maps, a study on classifying stroke lesion anatomy employed non-negative matrix factorisation (NMF) (Lee and Seung, 1999) to further reduce the dimensionality to fifty. As a parts-learning paradigm, NMF is capable to capture the natural hierarchical structure of vascular lesions to yield reasonably compact representations. Although we have other options, such as principal component analysis (PCA), we did not exhaustively estimate the marginal differences between techniques because our focus here is to compare the performance of high- and low-dimensional models rather than identify optimal technique for dimensionality reduction.

The fifty-dimensional representation was further reduced with t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton, 2008) to derive a two-dimensional representation. t-SNE is a robust technique to effectively capture non-linear structure in the data, analogous in its utility to uniform manifold approximation and projection (UMAP) (McInnes et al., 2018).

5.2.4 Synthetic ground truths of lesion-deficit and susceptibility to treatments

To model response to treatment, two sets of ground truths were prepared empirically. The first determined whether or not a lesion caused a behavioural deficit. The second determined whether or not the patient responded to the intervention.

The anatomical patterns of 1333 ischemic lesions were reduced to twenty-one distinct clusters (Figure 5.1) (Bonkhoff et al., 2019), from which two sets of clusters were selected as definitions for each of the two ground truths (Figure 5.2 and Figure 5.3). The details are specified in Table 5.1 and Table 5.2. An SVM model evaluated by 10-fold cross-validation provided (Chang and Lin, 2011) applied to both sets of ground truths showed reliable separability (sensitivity: 96.38%-97.03%; specificity: 88.45%-89.07% and sensitivity: 95.37%-96.24%; specificity: 93.29%-94.13%).

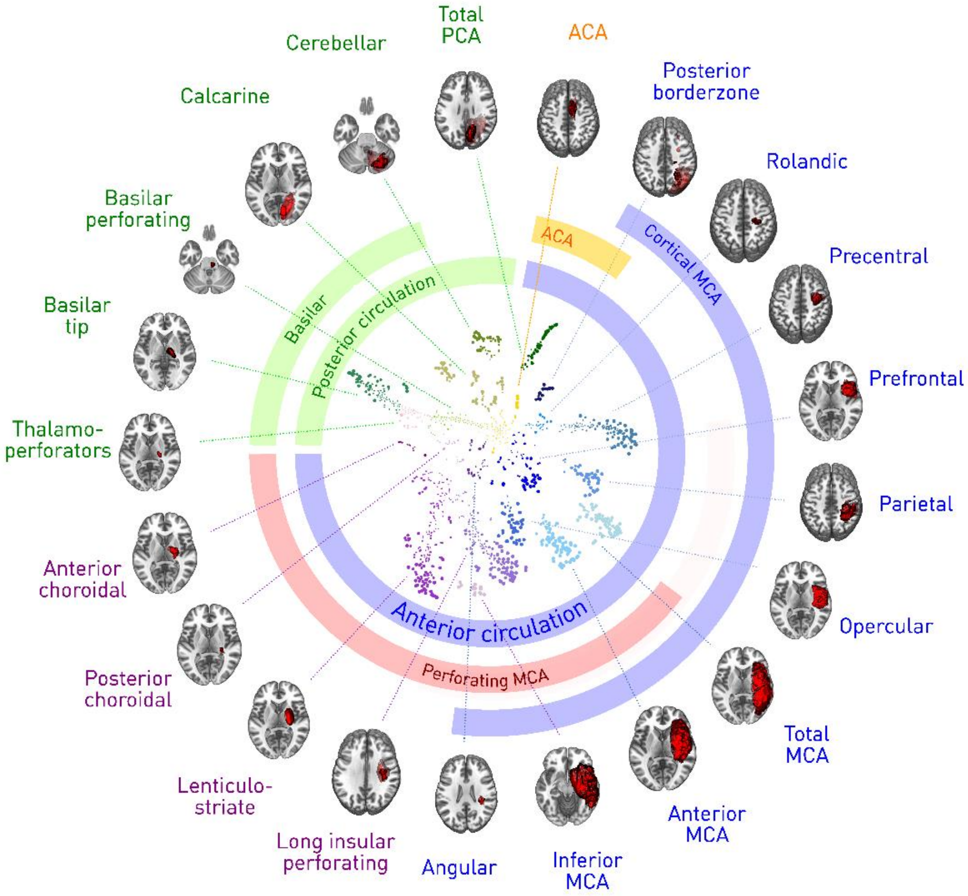


Figure 5.1 Two-dimensional clusters of lesion anatomy (Bonkhoff et al., 2019)

The compact lesion representation derived from t-SNE (Maaten and Hinton, 2008) illustrates a total of 21 stereotyped clusters. Base on this clustering, the ground truths for subsequence analysis would be established empirically.

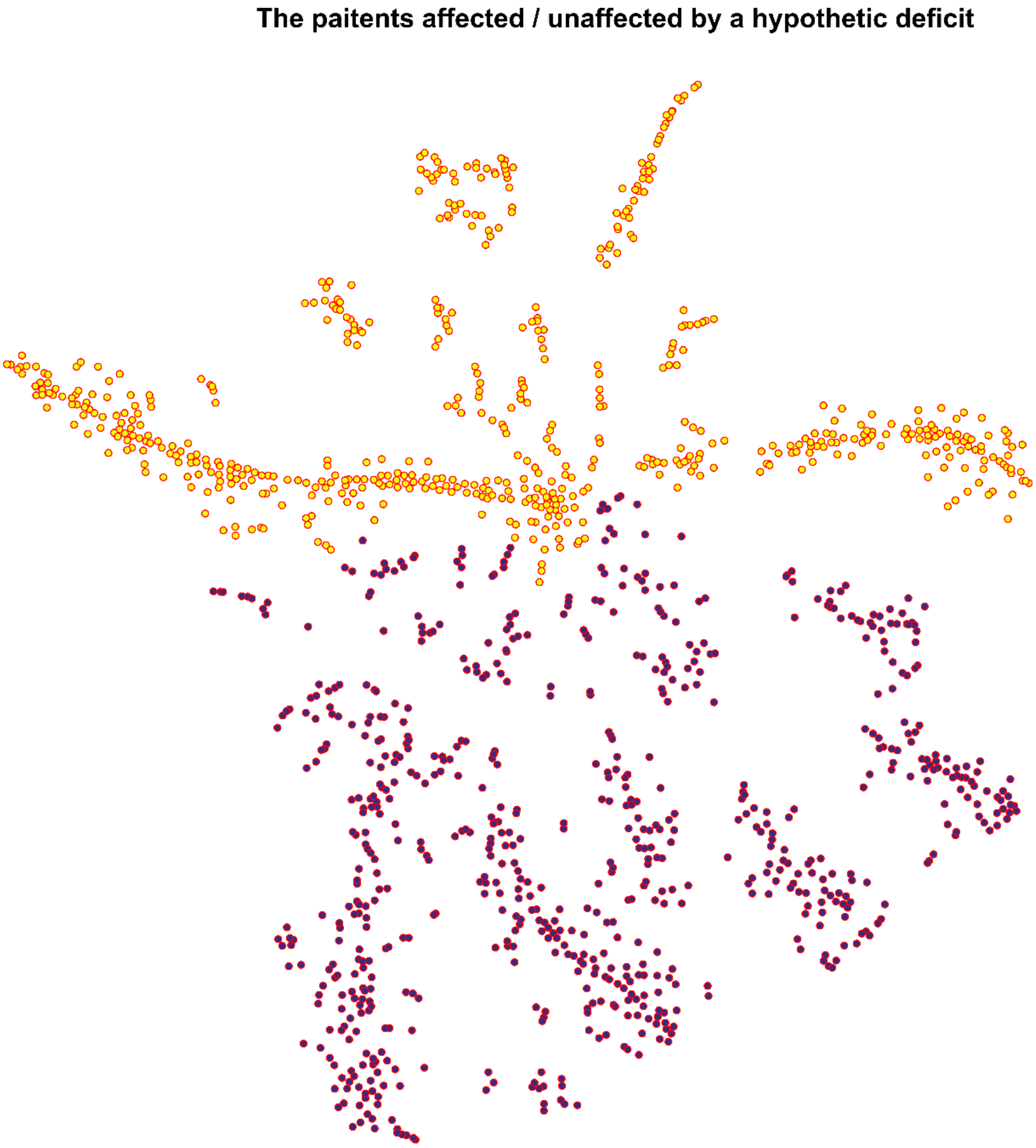


Figure 5.2 Ground truth: the separation of affected and unaffected lesions

Based on the 21 subdivisions, the first ground truth referred as affect (orange dots) versus unaffected (cranberry dots) by a hypothetical functional deficit was empirically sampled.

Affected Clusters	**Unaffected Clusters**
Thalamoperforators	Anterior choroidal
Basilar tip	Posterior choroidal
Basilar perforating	Lenticulostriate
Calcarine	Long insular perforating
Cerebellar	Angular
Total PCA	Inferior MCA
ACA	Anterior MCA
Posterior borderzone	Total MCA
Rolandic	Opercular
Precentral	Parietal
	Prefrontal

Table 5.1 The separation of lesions by the underlying functional deficit

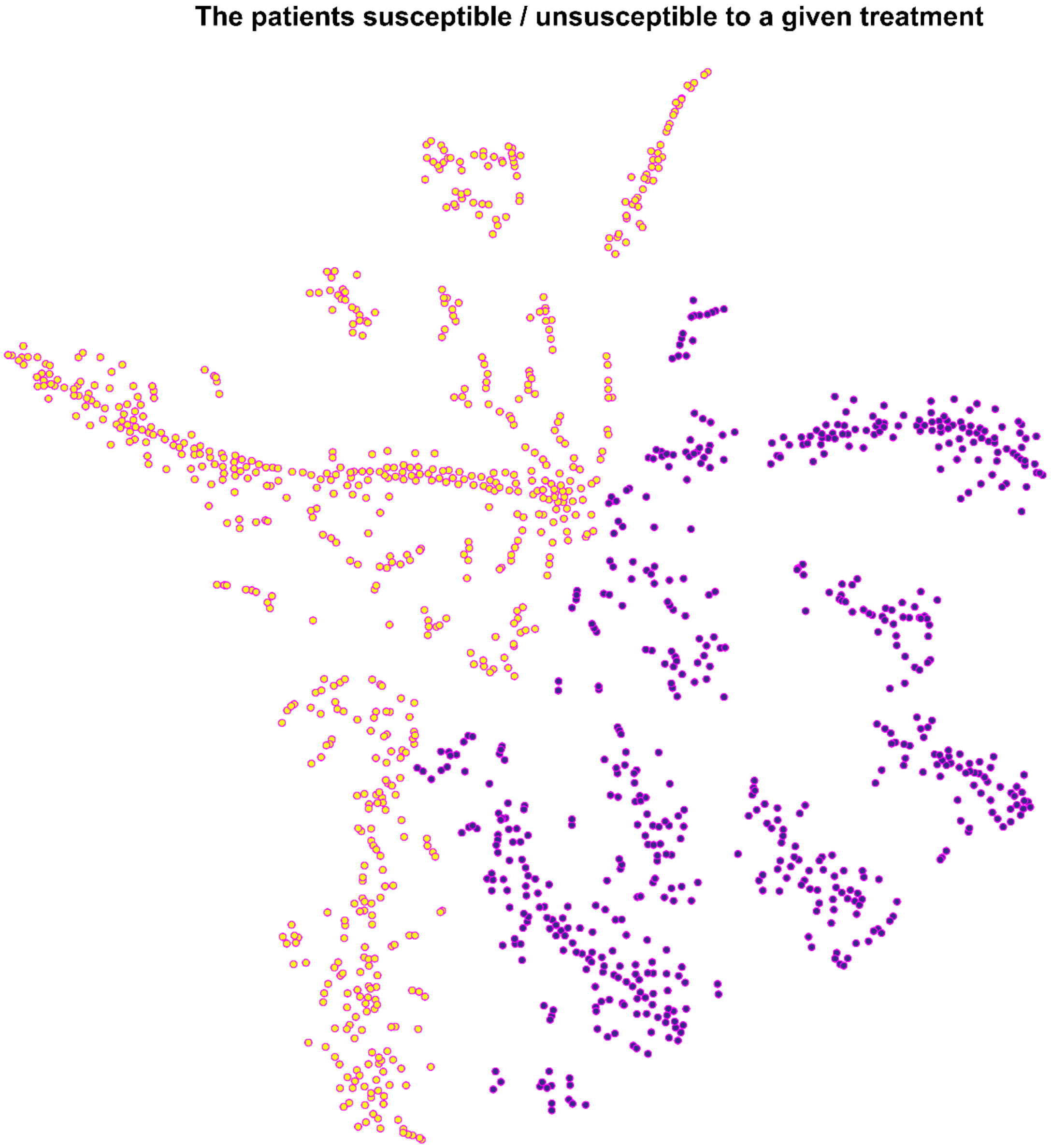


Figure 5.3 Ground truth: the separation of susceptibility

Based on the 21 subdivisions, the first ground truth referred as susceptible (pink dots) versus unsusceptible (purple dots) to a hypothetical treatment was empirically sampled.

Susceptible Clusters	Unsusceptible Clusters
ACA	Posterior borderzone
Total PCA	Rolandic
Cerebellar	Precentral
Calcarine	Prefrontal
Basilar perforating	Parietal
Basilar tip	Opercular
Thalamoperforators	Total MCA
Anterior choroidal	Anterior MCA
Posterior choroidal	Inferior MCA
Lenticulostriate	Long insular perforating
Angular	

Table 5.2 The separation of lesions by the susceptibility to a specific treatment

5.2.5 Prescriptive inference

We have seen that the complex structure of lesion is directed by the underlying pathological process. Evolving from the work presented in previous chapters, in order to further examine the importance of lesion anatomy in its interplay with the functional network, here we evaluated in simulation two series of models, now to address the question of detecting an individual response to treatment. We opted for a hypothetical functional deficit and determination of susceptibility rather than real ground truths because our major focus was to demonstrate the impact of the parameterisation of the lesion, and that can only be isolated if the other components of the causal chain are explicitly known.

Briefly, in both types of models, a simulated treatment varying in efficacy was applied to half of the patients, sampled randomly. In the first series of simulations, patients with the hypothetical deficit and who were susceptible to the treatment were proportionally responsive to the treatment with a positive outcome ("*recovery*") based on the specified interventional efficacy, and those who were not responsive were not affected by the treatment at all. In the other series of simulations, the arrangement was the same, except that those defined as non-responsive were associated with a *worse* outcome if randomised to the treatment arm. In the second series of simulations, we involved counterfactual analysis, on one hand, to enhance the causality derived from the impact evaluation; on the other, to extent the contrast of inferential models in low- and high-dimensionality. In the following sections, two series of inferential models will be descripted in detail.

5.2.5.1 Prescriptive inference with a positive response to a given treatment

In this series of simulations, the ground truth was that patients affected by a hypothetical functional deficit and susceptible to the treatment recovered if treated, with a probability dependent on the set treatment efficacy. Treatment was allocated randomly in equal proportion. For each patient, treatment status was registered as a binary variable and appended to the space of lesion anatomy as an additional feature. A variety of treatment effect sizes was defined from 10% to 90% in steps of 20%, and as well as 100%. For each effect size, the specified proportion of patients who were susceptible recovered from the functional deficit, whereas the others were unaffected. Thus, a series of models for prescriptive inference was simulated to identify the patients with susceptibility to a specific treatment.

5.2.5.2 Prescriptive inference with a polarised response to a given treatment

Here the ground truth was more complex. Whereas susceptible patients behaved as above, those who were labelled as not susceptible would now receive a poor outcome if treated. The idea was to model a two-sided treatment-outcome relationship where treatment is not cost-free, and may cause a worse outcome in patients not suited to it. To establish the models, we also allocated a binary marker to each patient and appended it to according lesion anatomy as an extra feature. In this way, the whole set of patients was randomly divided into two equal sub-groups referred as treated or non-treated. By varying the effect size mentioned in the above simulations, the corresponding proportion of affected susceptible treated patients recovered from the functional deficit, and a corresponding proportion of unaffected non-susceptible treated patients became affected. All the other patients were unaffected by the treatment. Thus, the second series of prescriptive inference was simulated to identify the patients susceptible to a specific treatment with a dual, polarised response.

5.2.5.3 Low- and high-dimensional models

Two sets of lesion representations derived from dimensionality reduction (Bonkhoff et al., 2019) were used for modelling predictive inference with low- and high-dimensionality. Specifically, the 1333 lesion anatomy presented in fifty-dimensional space derived from NMF (Lee and Seung, 1999) was employed to model high-dimensional models, and the two-dimensional feature matrix compacted by t-SNE (Maaten and Hinton, 2008) was used for low-dimensional models. Contrasting two levels of lesion parameterisation, two sets of analysis models were thus used for estimating the sensitivity to identifying the susceptibility of patients to a specific treatment.

In order to reveal the information hidden behind the lesion anatomical patterns, we employed LIBSVM – A Library for Support Vector Machines (Chang and Lin, 2011) to implement the both low- and high-dimensional analysis. We have seen SVM is flexible enough to cope with complex models, and the radial basis function (RBF) kernel integrated in LIBSVM is available to incorporate the interactions between the attributes. In addition, the well maintained LIBSVM (https://www.csie.ntu.edu.tw/~cjlin/libsvm/) provides a wide range of extensions and related functions, such as AUC calculation and ROC plotting, which allows for more convenience and reliability for our purpose of analysis and estimates.

5.2.5.3.1 Model implementation

Following the arrangement described above, 500 randomisations were performed to divide the stack of 1333 patients into two equal proportions assigned to the treated and untreated groups, mimicking a randomised controlled trial. In each randomisation, a set of treatment efficacies varying across 10%, 30%, 50%, 70%, 90% and 100% were sequentially conducted to identify the relation between lesion architecture and susceptibility of patients iteratively. For each effect size, 50 iterations that randomly split the set of patients into two parts (70% and 30%) were performed for the purpose of training a classifier and performance testing. Over each iteration, the selection of patients for training and test was completely independent. During the training procedure, a binary marker indicating whether or not a patient received the treatment was appended to the corresponding lesion feature vector. Thus, treated patients who recovered from the hypothetical functional deficit due to the susceptibility to the treatment could be recorded. The learnt classifier was then applied to each datum in the test dataset twice: first, with the treated feature set to 1, i.e. treated, and second with the treated feature set to 0, i.e. untreated. By quantifying the difference in the predicted outcome under the two different treatment value, the inferred susceptibility of the test patient was thus identified. A total of 150000 (500 x 6 x 50) classifiers and 300000 (500 x 6 x 50 x 2) predictions with low- and high-dimensional models were computed.

The second series of simulations were implemented in exactly the same way, except for the difference in ground truth to reflect a polarised outcome as outlined in the preceding section.

5.2.5.3.2 Model evaluation

The evaluations were focused on two aspects. Within each series of simulations, we were interested in the difference in fidelity between the low- and high-dimensional models. Across both series of simulations, we were interested in the impact the relationship between the treatment and the outcome where susceptibility is polarised into both positive and negative outcomes.

The evaluation metrics to evaluate the classifiers' predictive power included accuracy, sensitivity, specificity. A confidence interval at level of 95% was applied to yield the range of reliability. The details are as follows:

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{the total number of cases}};$$

$$\begin{aligned}\text{sensitivity} &= \frac{\text{number of true positves}}{\text{number of true positives} + \text{number of false negatives}} \\ &= \frac{\text{number of true positives}}{\text{the total number of "affected" cases}};\end{aligned}$$

$$\begin{aligned}\text{specificity} &= \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}} \\ &= \frac{\text{number of true negatives}}{\text{the total number of "unaffected" cases}}.\end{aligned}$$

Then, for each series of simulations, the performance of inferential models through the 500 randomisations was further interpreted with violin plots. As an adaptation of the box plot, a violin plot synergistically combines the box plot and the density distribution of the data to enrich the visualisation (Hintze and Nelson, 1998, Potter et al., 2006). Specifically, in a typical violin plot (Figure 5.x), the data are usually partitioned into quartiles. A box positioned in the centre of the violin indicates the range of upper and lower quartiles, and a line extended through the box shows the limits of data at 95% confidence intervals. The violin plot is shaped by the kernel density estimate (KDE).

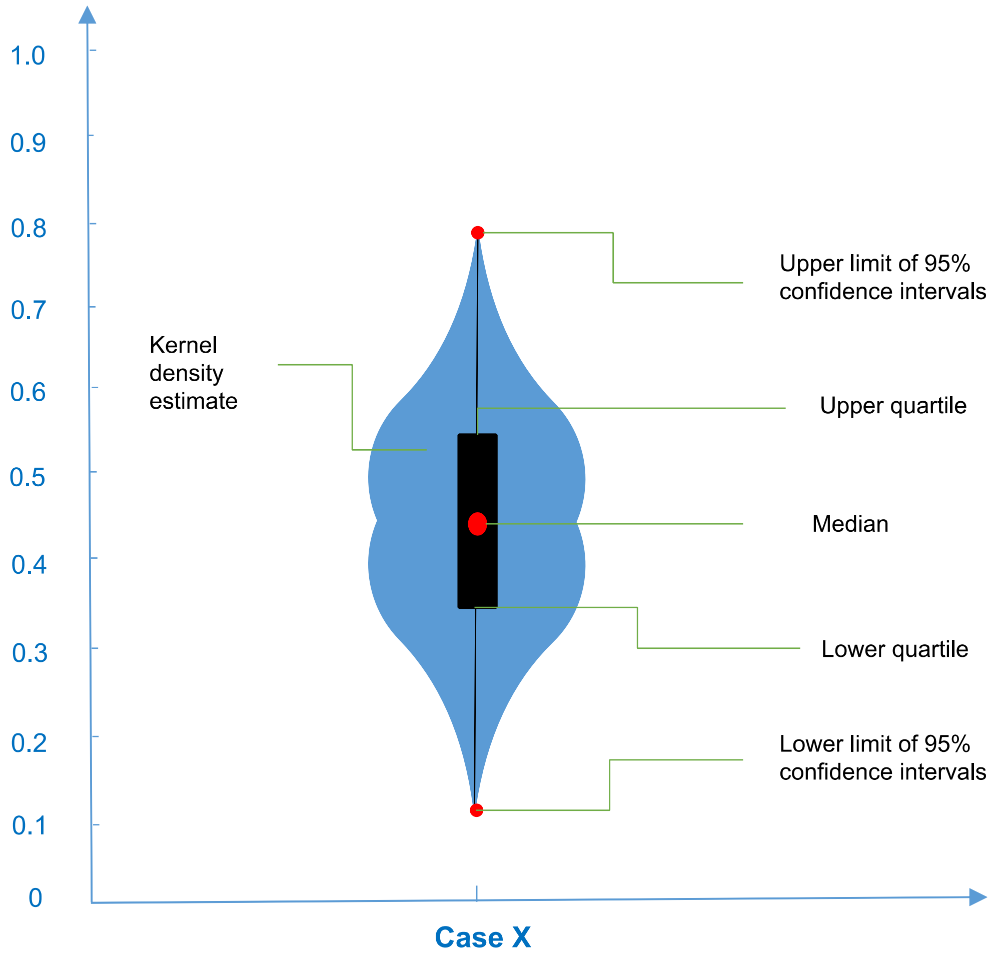


Figure 5.4 The components of violin plot

As a variation of the original boxplot. Violin plot is used to further enrich the visualisation. The central mark (in red) is the median. The bottom and top boundaries of the box (in black) indicate the 1st and 3rd quartiles. The whisker through the box is extended to indicate upper and lower limits of 95% confidence intervals. At last, the violin is shaped by a kernel density estimate.

5.2.6 Software and hardware

All simulated prediction models were implemented in the environment of MATLAB (MATLAB and Statistics Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.).

In both series of simulations, a 12-core Inter® Xeon® CPU E5-2620 2.00GHz processor with 64GB RAM and 9TB 7200 RPM SATA hard drive was used to conduct analysis under 64-bit Linux operation system (Ubuntu version 15.04).

5.3 Results

5.3.1 Prescriptive inference with positive response to a given treatment

The purpose of this series of simulations was two-fold. First, to determine whether empirically observed lesion patterns that could confer individual susceptibility to a specific treatment could be learnt to infer individual treatment susceptibility. Second, if so, to quantify the differential performance of low- and high-dimensional models.

Both low- or high-dimensional models were capable of identifying susceptible patients, with a fidelity that increased with effect size. At each specific effect size, high-dimensional models were markedly superior. The contrast was greater the higher the effect size (Figure 5.5). The mean range of sensitivity of the high-dimensional models across the 500 randomisations was from 11.86% (95% confidence interval: 1.87^10^{-5}) to 77.92% (95% confidence interval: 0.56^10^{-5}); whereas the range of the low-dimensional models was 4.42% (95% confidence interval: 7.43^10^{-5}) to 44.62% (95% confidence interval: 12.81^10^{-5}). This indicates the sensitivity of the prescriptive inference was highly relied on the efficacy of provided treatment, but only given sufficient parameterisation. The mean value across the 500 randomisations and related confidence intervals are listed in table 5.3.

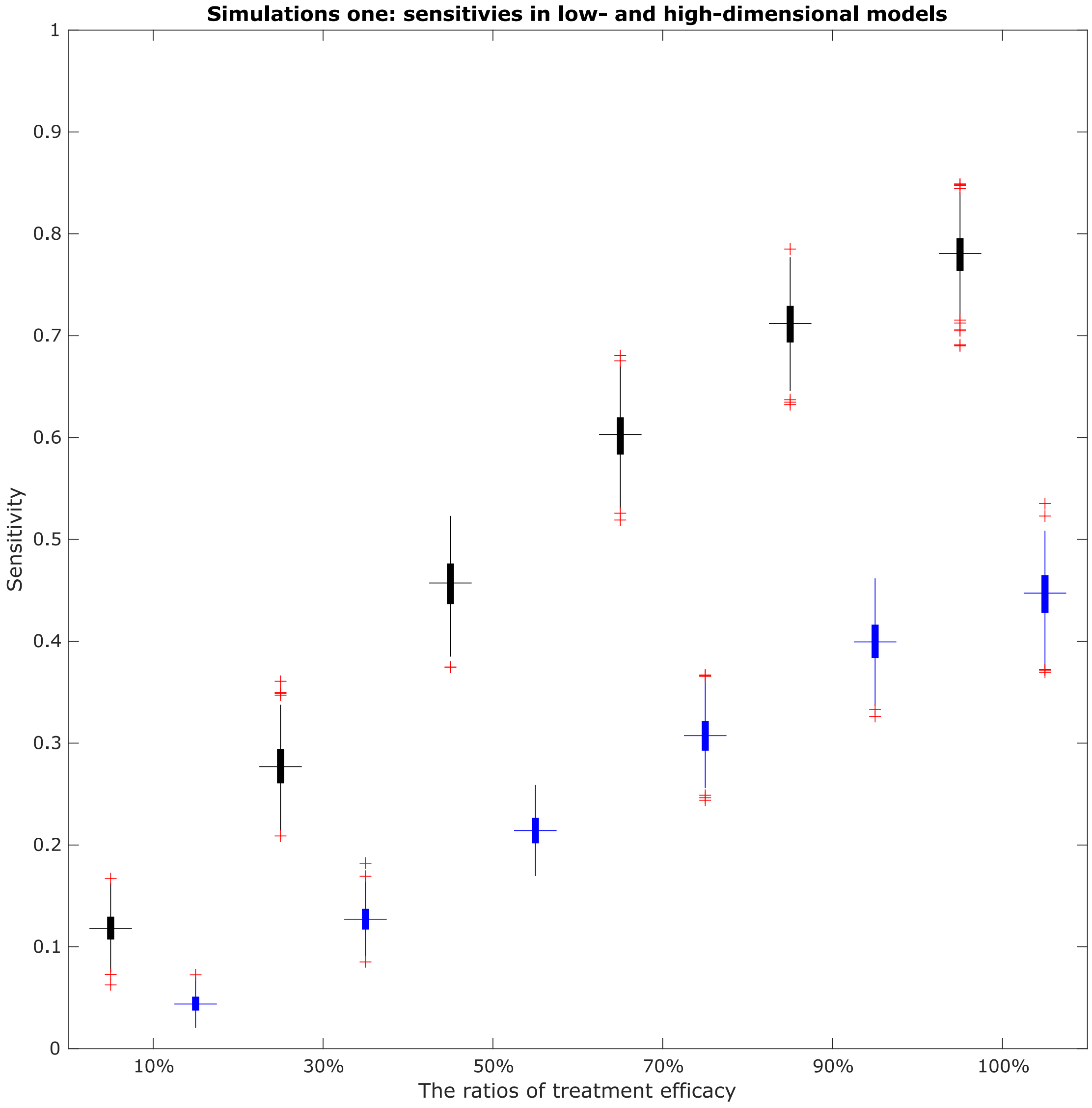


Figure 5.5 Sensitivities in low- and high-dimensional models

For both low- and high-dimensional models, a set of effect sizes was applied to determine the corresponding proportion of treated susceptible patients affected by a hypothetical functional deficit who recovered as a result of the treatment. For each effect size, a total of 500 randomisations were conducted to enhance model reliability. The ability to detect susceptible patients achieved in low- (in blue) and high-dimensional models (in black) were visualised with boxplot. For each box, the central mark refers to the median. The bottom and top edges are 25th and 75th percentiles, respectively. The maximum and minimum values were shown by the extended whiskers through the box. Outliers were plotted individually with red '+'.

Evaluation Efficacy	Mean of sensitivity (high)	Mean of sensitivity (low)	Difference
10%	11.86% (95%CI:1.87^10^{-5})	4.42% (95%CI:7.43^10^{-5})	7.44%
30%	27.72% (95%CI:4.40^10^{-5})	12.73% (95%CI:8.98^10^{-5})	14.99%
50%	45.57% (95%CI:8.67^10^{-5})	21.42% (95%CI:2.21^10^{-5})	24.15%
70%	60.26% (95%CI:30.51^10^{-5})	30.75% (95%CI:9.05^10^{-5})	29.51%
90%	71.17% (95%CI:10.17^10^{-5})	39.91% (95%CI:2.32^10^{-5})	31.26%
100%	77.92% (95%CI:0.56^10^{-5})	44.59% (95%CI:12.81^10^{-5})	33.33%

Table 5.3 Simulation one: sensitivity

5.3.2 Prescriptive inference with a polarised response to a given treatment

Elaborating the series of simulations described above, the models were extended to further examine whether or not the patients susceptible to a specific treatment could be effectively identified and whether or not high-dimensional models consistently performed superiorly in the context of a more complex, polarised response to treatment.

In this series, the pre-defined ground truths allowed for two set of divisions across the stack of 1333 patients referred as ‘affected’ versus ‘unaffected’, and ‘susceptible’ versus ‘non-susceptible’. A set of effect sizes varying from 10% to 100% was applied. In contrast with the first series of simulations, for each effect size, whereas the patients who were given treatment and were susceptible to the treatment recovered, treated patients who were not susceptible to the treatment became affected.

As before, both low- and high-dimensional models were capable of identifying susceptible patients, with a fidelity that increased with effect size. Moreover, high-dimensional models were markedly superior. The difference was closely correlated to the efficacy of treatment (Figure 5.6). For instance, when the effect size was 10%, the contrast was 7.74% (95% confidence interval: 5.57^10^{-5}) versus 6.21% (95% confidence interval: 1.20^10^{-5}); whereas, in the case of 100%, the values were 84.64% (95% confidence interval: 0.92^10^{-5}) versus 50.77% (95% confidence interval: 6.90^10^{-5}). Once again, the results also reflected the importance of treatment efficacy as well as dimensionality. The mean value across the 500 randomisations and related confidence intervals were list in table 5.4.

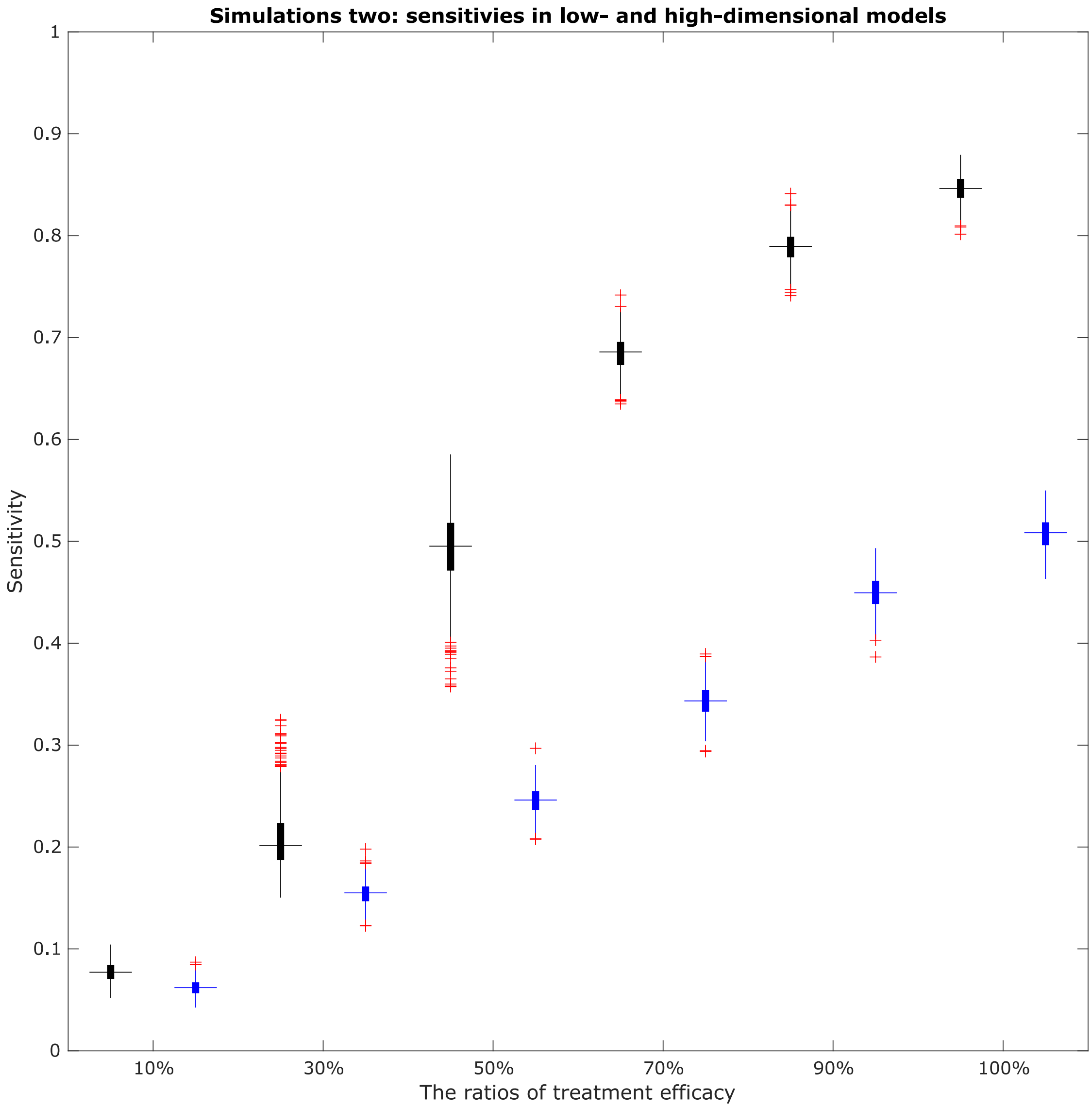


Figure 5.6 Sensitivities in low- and high-dimensional models

For both low- and high-dimensional models, a set of effect sizes was applied to determine the according proportion of patients 1. affected by a hypothetical functional deficit who recovered as a result of the treatment; 2. unaffected by the hypothetical functional deficit who were accidentally attacked as a result of the treatment. For each effect size, a total of 500 randomisations were conducted to enhance model reliability. The ability to detect susceptible patients achieved in low- (in blue) and high-dimensional models (in black) were visualised with boxplot. For each box, the central mark refers to the median. The bottom and top edges are 25th and 75th percentiles, respectively. The maximum and minimum values were shown by the extended whiskers through the box. Outliers were plotted individually with red '+'.

Evaluation / Efficacy	Mean of sensitivity (high)	Mean of sensitivity (low)	Difference
10%	7.74% (95%CI:5.57^10^{-5})	6.21% (95%CI:1.20^10^{-5})	1.53%
30%	20.88% (95%CI:0.48^10^{-5})	15.45% (95%CI:1.11^10^{-5})	5.43%
50%	49.23% (95%CI:11.15^10^{-5})	24.55% (95%CI:2.87^10^{-5})	24.68%
70%	68.48% (95%CI:1.99^10^{-5})	34.31% (95%CI:1.24^10^{-5})	34.17%
90%	78.91% (95%CI:4.62^10^{-5})	44.94% (95%CI:11.43^10^{-5})	33.97%
100%	84.64% (95%CI:0.92^10^{-5})	50.77% (95%CI:6.90^10^{-5})	33.87%

Table 5.4 Simulation two: sensitivity

5.3.3 Comparison between the two series of simulations.

In the first series of simulations, we identified which patients could be benefit from the correct treatment. In the second series of simulations, we quantified fidelity with a polarised treatment. The latter series was more complex as the behavioural outcome changed in either direction. It was reasonable for the inferential models in the latter series of simulations to achieve higher sensitivity because the outcome changed for more patients.

Specifically, to compare the performance between two series of simulations, a set of violin plot varying by efficacious ratios were generated. From the figures (5.9 – 5.14), we can see the functional outcome of untreated patients remained more or less the same – the difference was consistently in the range of 0 to 0.2 whatever the effect size or the dimensionality. Moreover, in the set of plots, the violin shapes were very flat, and their central whiskers were very short. In some extent, this reflected the stability of the inferential models in both series.

The low- and high-dimensional model results for the treated group in the first series of simulations are plotted in the middle of each figure. The difference after treatment in the high-dimensional models was consistently larger than in the low-dimensional models. The higher the effect sizes, the greater the difference between the two types of models. Similarly, the pair of violins on the right in each figure illustrated the difference of behavioural outcome after treatment in the second series of simulations. The larger difference showed not only the patients who were susceptible were correctly identified; but the patients whose outcome changed from ‘unaffected’ to ‘affected’ were also identified. This further proved the superiority of high-dimensional models and enhanced the feasibility of the proposed inferential framework.

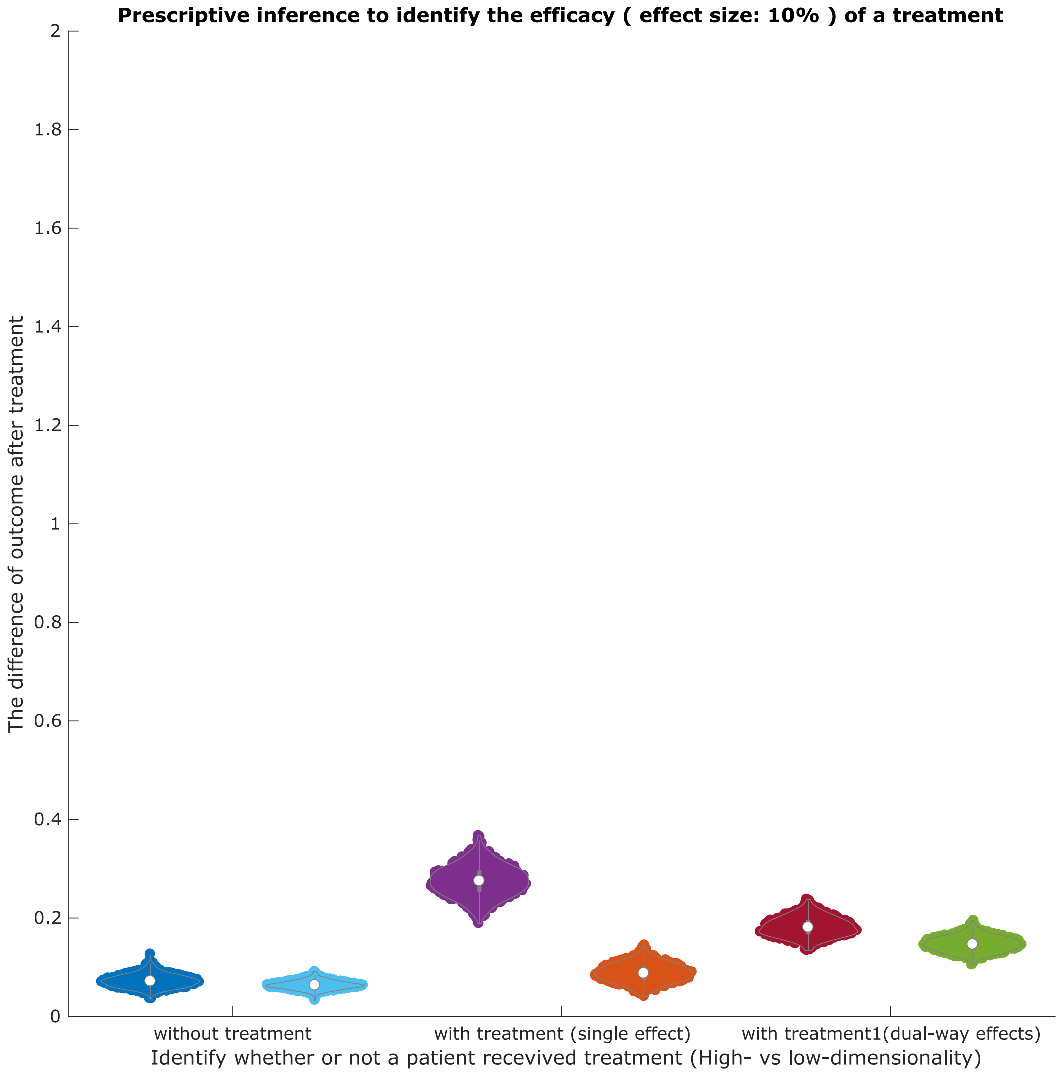


Figure 5.7 Identification of patients with treatment (efficacy ratio: 10%)

By plotting the outcome change before and after treatment, the ability to detect whether or not a patient is treated is thus visualised. As the outcome adopted in this study is binarised, the absolute range of change is 0 to 2. Three pairs of violins are plotted in the same figure to visualise the straightforward contrast. In each pair, left violin represents high-dimensional approach; and right violin represents the low-dimensional approach. For the first pair of violins in blue, the outcome almost barely changed (<0.2) in proportion of patients who were not treated. Fundamentally, this indicated the reliability and stability of the models in both low- and high-dimensionality. For the second pair of violins, the susceptible patients with treatment could be recovered from the hypothetical functional deficit. Where the treatment effect was 10%, a small proportion of patients were affected by the treatment, so the high-dimensional approach (purple) showed slightly larger impact responding to the treatment than the low-dimensional approach (in orange). For the first pair of violins, the susceptible patients with treatment could be recovered from the hypothetical functional deficit. Meanwhile, the unsusceptible patients with treatment could develop a deficit even though they were originally unaffected. Plausibly, the impact of treatment should have been further enlarged, but at the case of 10% treatment efficacy, the relatively small number of treated patients leaded to the instability of inferential models. Comparing to the former scenario, high-dimensional approach (in cranberry) showed slightly small impact and low-dimensional approach (in green) showed marginally larger impact.

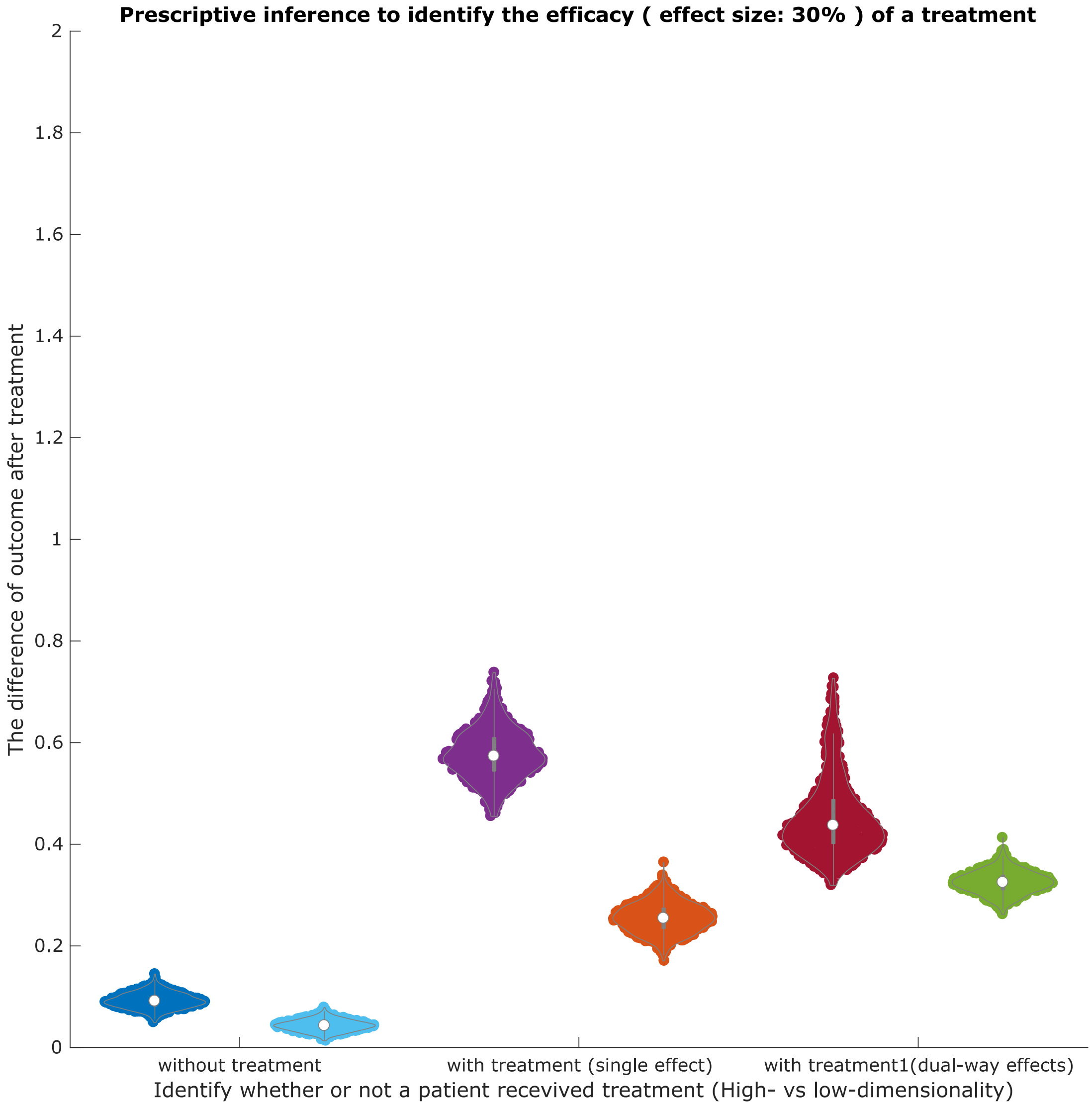


Figure 5.8 Identification of patients with treatment (efficacy ratio: 30%)

For the case of treatment efficacy at 30%, in both scenarios, high-dimensional approach shows large impact responding to the treatment comparing to the low-dimensional approach. Comparing to the case of 10% efficacy, the discrimination between two approaches was extended. To be similar to the case at 10% efficacy, the small proportional involved patients resulted in instability. The latter scenario did not demonstrate superior impact.

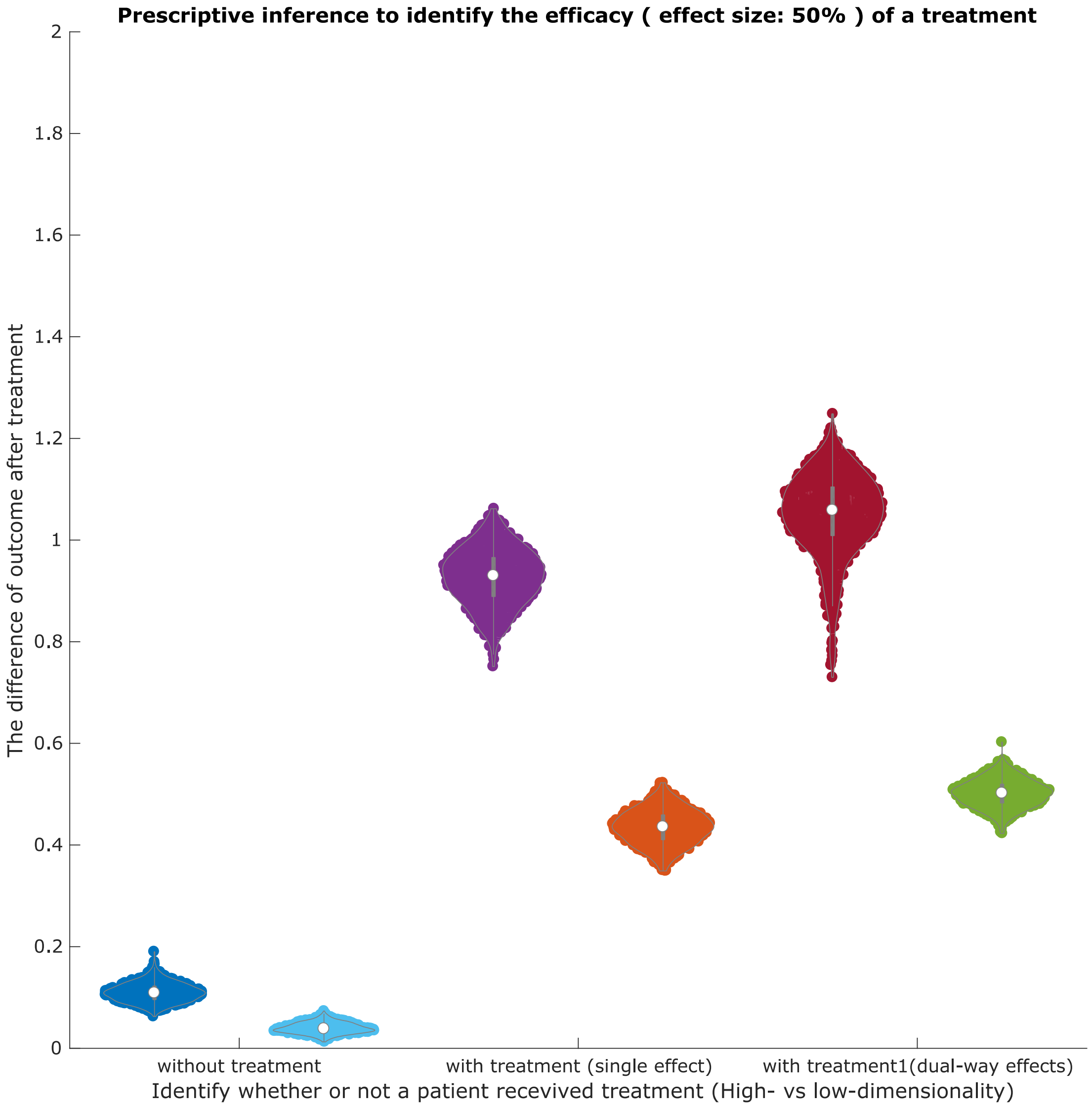


Figure 5.9 Identification of patients with treatment (efficacy ratio: 50%)

For the case of treatment efficacy at 50%, in both scenarios, high-dimensional approach shows large impact responding to the treatment comparing to the low-dimensional approach. Comparing to the case of 30% efficacy, the discrimination between two approaches was extended. Moreover, the latter scenario where the treated patients might change their outcome from dual directions demonstrated further superior impact.

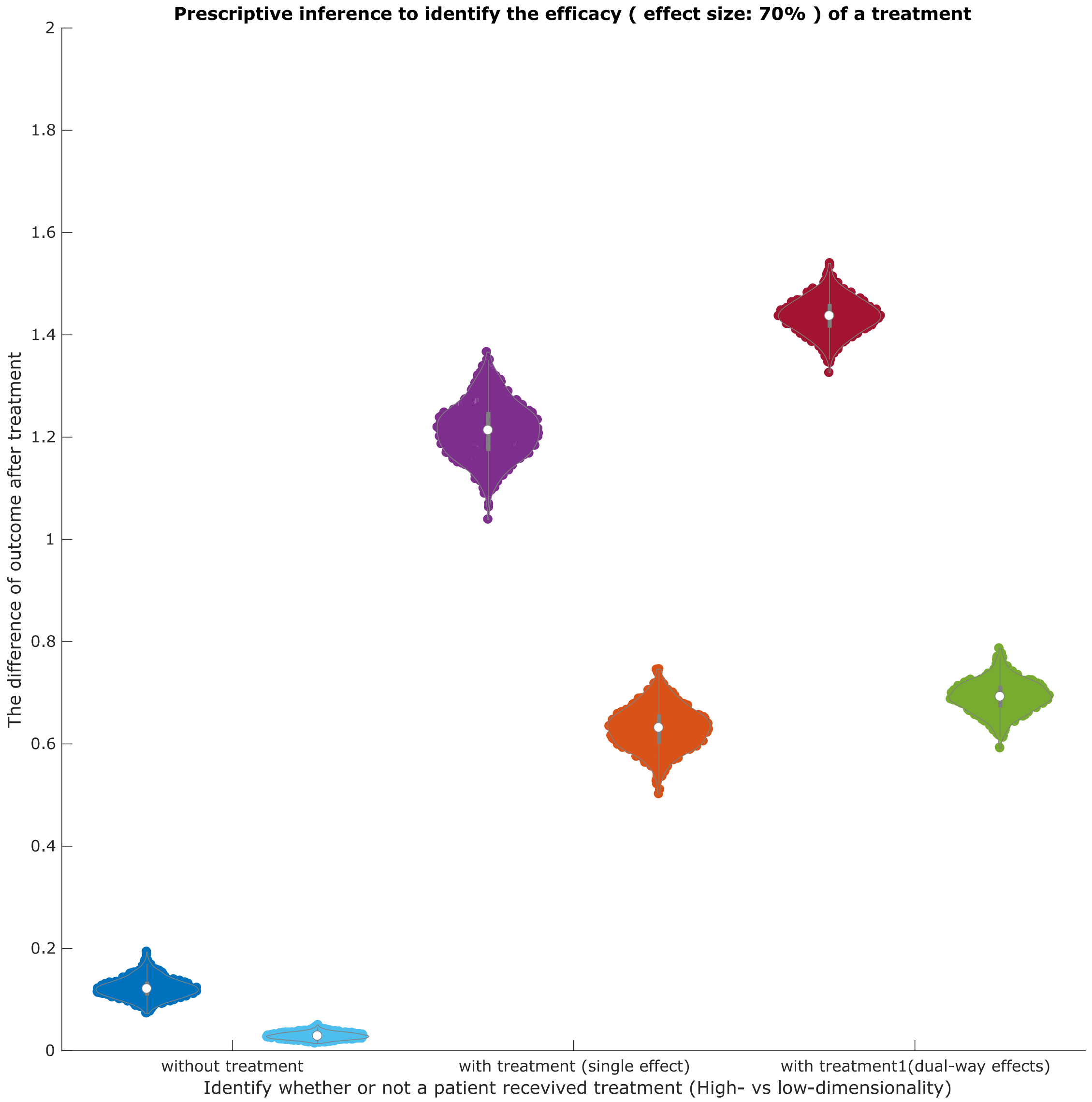


Figure 5.10 Identification of patients with treatment (efficacy ratio: 70%)

For the case of treatment efficacy at 70%, in both scenarios, high-dimensional approach shows large impact responding to the treatment comparing to the low-dimensional approach. Comparing to the case of 50% efficacy, the discrimination between two approaches was further extended. Moreover, the latter scenario where the treated patients might change their outcome from dual directions demonstrated further superior impact.

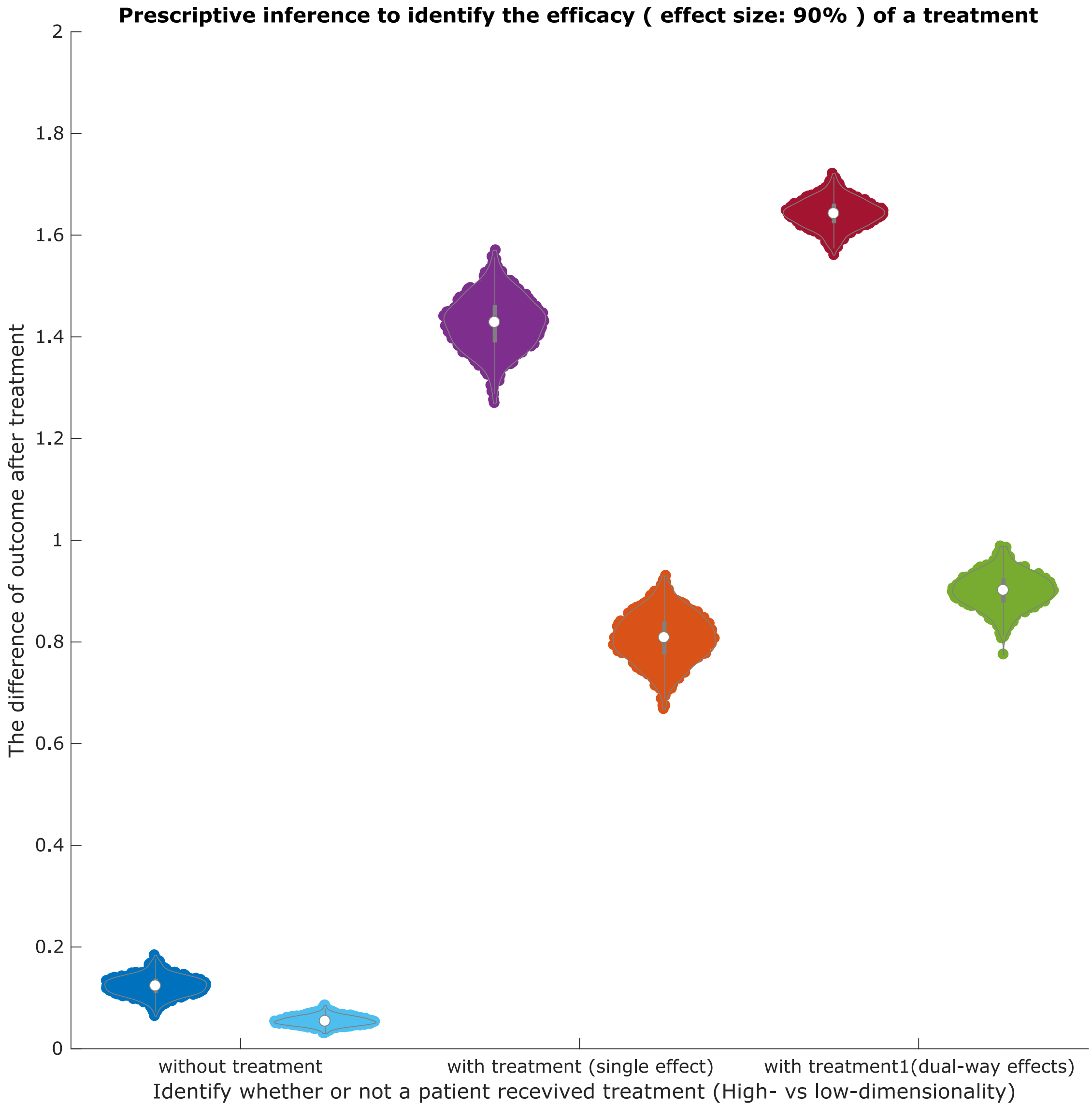


Figure 5.12 Identification of patients with treatment (efficacy ratio: 90%)

For the case of treatment efficacy at 90%, in both scenarios, high-dimensional approach shows large impact responding to the treatment comparing to the low-dimensional approach. Comparing to the case of 70% efficacy, the discrimination between two approaches was further extended. Moreover, the latter scenario where the treated patients might change their outcome from dual directions demonstrated further superior impact.

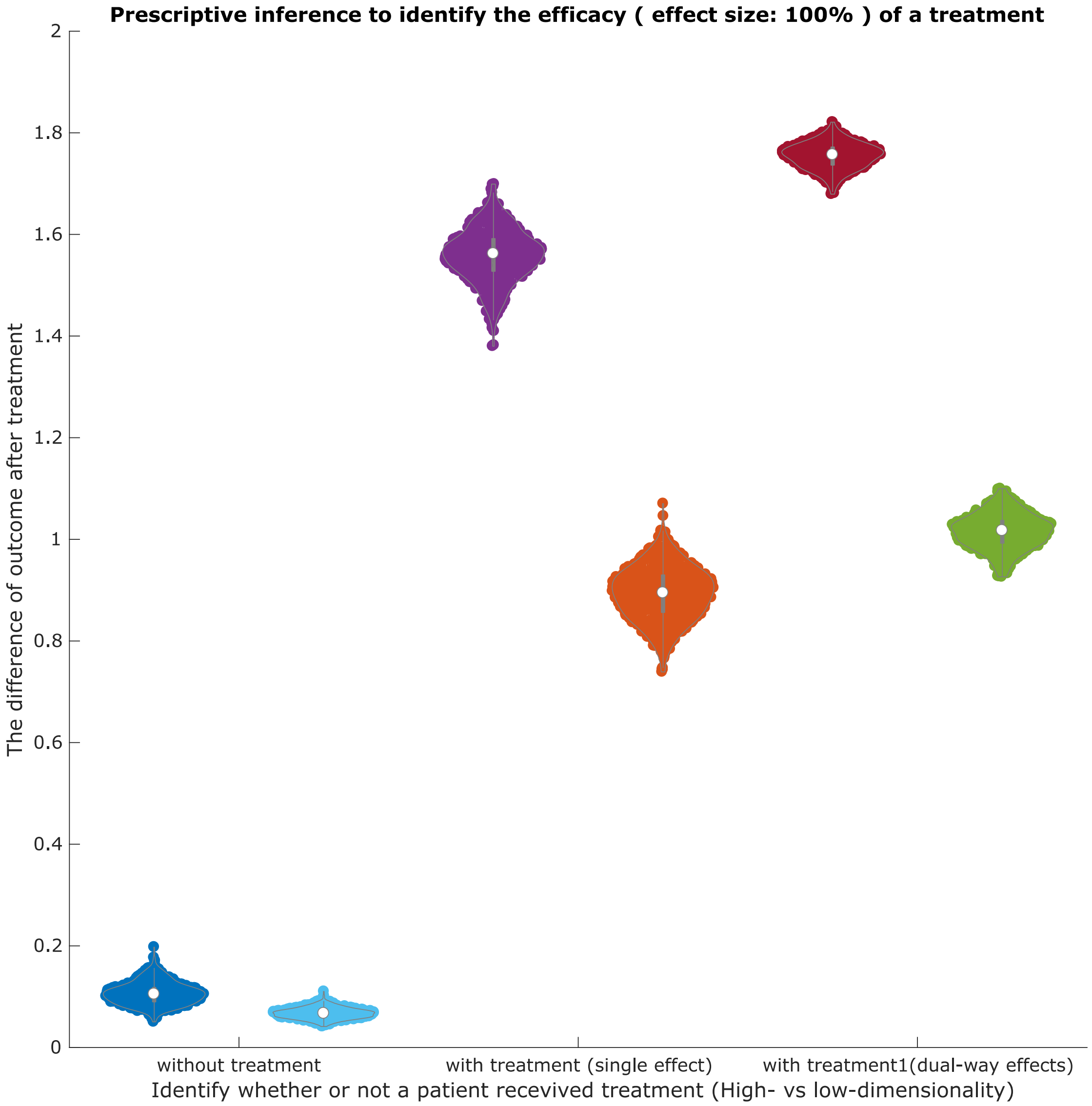


Figure 5.12 Identification of patients with treatment (efficacy ratio: 100%)

For the case of treatment efficacy ideally at 100%, in both scenarios, high-dimensional approach shows large impact responding to the treatment comparing to the low-dimensional approach. Comparing to the case of 90% efficacy, the discrimination between two approaches was maximised. Moreover, the latter scenario where the treated patients might change their outcome from dual directions demonstrated further superior impact.

5.4 Discussion

In medicine, the conventional criteria for prescribing a particular treatment to an individual tend to be low dimensional. Typically, an evidence-based criterion of treatment usually rely on a compact set of specifiable parameters. For example, when treating brain infarction, the guidance might be all patients under the age of 75 should have thrombolysis if they do not have middle cerebral artery (MCA) occlusion and mechanical thrombectomy if they do. However, given the intrinsically complex architecture of injured brain (Mah et al., 2015, Mah et al., 2014b, Xu et al., 2017a, Xu et al., 2017b) and its interaction with the complex pathological processes, the causal relation between a particular treatment and outcome should be individualised with a multiplicity of interacting factors rather than standardised with a small number of crude parameters.

We therefore modelled two series of interactions between treatment and outcome in low- and high-dimensionality, respectively. By contrast with the poor performance obtained in low-dimensional models, the inference modelled in high-dimensionality successfully distinguished between susceptible and non-susceptible patients. On the one hand, based on hypothetical ground truths, we established an inferential framework to reveal individual susceptiblity; on the other, we distinguished between the performance of low- and high-dimensional models of the lesion architecture.

We did not use observed outcome data to model such inference, because we would like to build an inferential framework that is readily translatable to any predictive model - a universally adaptive model, not constrained with a specified relation. Although the set of synthetic ground truths might be different from the fidelity in the context of a real treatment, our primary focus here is to contrast the models in low- and high-dimensionality. In essence, we opted for the notion of counterfactuals to model agnosticism. As the basis of causal inference, the concept of counterfactual gained increasing popularity, especially, in medicine (Höfler, 2005, Johansson et al., 2016). Counterfactual explanations can be clearly interpreted in causal terms. Moreover, the inferential framework implemented with counterfactuals can provide a wide range of statistical procedures to estimate the causality and demonstrate the limitations of factual data (Höfler, 2005). Additionally, in order to stabilise our framework, a relatively large number of randomisations (500) to assign the treatment (50% treated versus 50% non-treated) were iteratively conducted to minimise covariate shifts.

The causality between treatment and outcome was interpreted with counterfactual explanations. But, to quantify the impact models in low- and high-dimensionality, the set of hypothetical deficit and treatment must be evaluated with real damaged brain. The dimensionality of the model was commensurate with the scale of data. We employed vascular brain injury as a representative to establish the inference – on one hand, it was more feasible to obtain a relatively large-scale dataset; on the other, as one of the most prominent neurological disorder. A high-dimensional model that took into account numerous interacting factors across the entire brain maps implemented with computation intensive machine learning techniques improved our understanding of human brain mechanisms by uncovering the anatomical patterns of lesion data that were hardly detectable with a conventional low-dimensional models. High-dimensional analysis showed substantial advantages and is expected to supersede low-dimensional approaches, even if clinical adoption may not be straightforward. A key question is whether guidance on treatment derived from a high-dimensional inferential framework is too complicated to be intelligible for the clinician to follow, and equally if a clinician would ever defer to a mathematical model he or she cannot understand. Since the primary duty of clinicians directs them to seek the best possible outcome rather than the most perspicuous outcome, this obstacle is likely to be overcome in due course.

To optimise this high-dimensional framework and improve its adoption in clinical domain are our major targets in the future. Large-scale data is essential for modelling with high-dimensionality. It is clear we must grow our source data continuously so that we can enhance existing models, and more importantly, implement new models with greater flexibility. This will allow us to demonstrate the validity and necessity of high-dimensional analysis from a more comprehensive perspective. Technically, the notion of counterfactual explanations enables us to model causal inference with high adaptability. Yet, nothing is omnipotent. In some situations, the reference condition for assessing counterfactual causal effects is not always easy to choose. Moreover, each instance can be usually explained by multiple counterfactuals, which is inconvenient sometimes (Höfler, 2005). Hence, it is worth exploring new concept of modelling to enrich our framework.

5.5 Conclusion

By combining a particular treatment with a particular patient, in this chapter, two series of prescriptive inference in focally damaged human brain were conducted to identify individual susceptibility to treatment. To make the models aware of the complexity of lesion anatomy to varying degrees, we implemented the inference with low- and high-dimensionality, respectively. Low-dimensional models were based on a two-dimensional space derive from t-SNE (Maaten and Hinton, 2008); whereas, the parametrisations of high-dimensional models were dependent on a part-based decomposition in fifty dimensions derived from NMF (Lee and Seung, 1999), where the natural hierarchical structure of vascular injury was effectively retained. As predicted, in both series of counterfactuals, high-dimensional modelling was consistently superior in identifying individual-level susceptibility.

The complexity of the human brain interacts with the complexity of the pathological process determining the intrinsic dimensionality of brain lesions, across both prediction and prescription. Methodologically and conceptually, our counterfactual explanations would be applicable to a wide range of prediction and inference in medical domain.

Chapter 6 General discussion

The ultimate objective of understanding the human brain is to explain how it responds to various physiological and pathological perturbations. Though an extraordinarily complex system, the anatomically specified architecture underlying human brain opens a path to such understanding, at least at the macroscopic level of organisation.

Over the history of neuroscience, lesion-symptom mapping and functional magnetic resonance imaging (fMRI) are two major approaches to model the relations between brain anatomy and function. fMRI has been very popular during the past two decades as it provides excellent spatial resolution to localise brain activation. By identifying the activated regions in the brain during a particular task, specific brain regions are implicated in specific functions. However, given the complexity of large, densely connected neural networks, one primary limitation of fMRI is the activations induced by a specific task allow for multiple explanatory possibilities, leading to multiple interpretations. The other limitation is that fMRI merely identifies a change in neural activity related to a specific function, and cannot prove the function is critically dependent on the integrity of the activated regions. By contrast, the inferences generated by lesion-deficit mapping studies are based on the loss of function following focal damage. Lesion studies intrinsically model the causal relation between brain damage and function, enabling more robust prediction as well as illuminating the underlying mechanistic picture with greater intensity. The history of lesion studies can be traced back to 19th century -- from the period of post-mortem examinations (Broca, 1861, Wernicke, 1874) to *in vivo* imaging analysis in the contemporary era, the approaches to conducting lesion studies have evolved dramatically with the development of technology. Yet, the complexity hidden behind the architecture of brain injury is still ignored by conventional means of analysis. The behavioural outcome resulting from the complex interactions between the human brain and the pathological process plausibly requires complex anatomical patterns to interpret. Hence, high-dimensional modelling commensurate with complexity of brain damage is essentially necessary.

Crucially, sufficient parameterisation of lesion dimensionality is the foundation of lesion-function inference. The power of this foundation will be reflected in a wide range of contexts in medical research across advanced prediction of lesion-deficit and superior sensitivity of therapeutic and prescriptive inference.

6.1 The limits of lesion-deficit prediction

Although lesion-deficit mapping has a long history, the spatial patterns underlying the anatomical structure of brain injury have not received adequate attention. Instead, mass-univariate analysis, as a conventional approach, is widely employed in lesion mapping studies, where each volume unit, regardless of resolution, is modelled as an independent variable. This is true of the template overlay method, which implicitly treats each discrete brain region as spatially independent of any other (Frey et al., 1987). In recent years, voxel-based lesion symptom mapping (VLSM) (Bates et al., 2003) has become the technique of choice. It relies on a voxel-wise measure of confidence to quantify the contribution of each voxel independently. Some proponents of VLSM have proposed involving lesion volume as a covariate to marginally enhance the superficial performance of statistics (de Haan and Karnath, 2018, Sperber and Karnath, 2016, Wiesen et al., 2019).

The fundamental issue of mass-univariate analysis is the assumption of spatial independence, or at least simple spatial dependence such as Gaussian smoothness. This induces spatial biases arising from the complexity of the underlying lesion architecture. A previous study (Mah et al., 2014b) systematically explored this issue, showing substantial mislocalisation, for both very simple single-voxel ground truths and more plausible regional dependence, when mass-univariate mapping methods were used. By contrast, high-dimensional multivariate analysis was far less prone to mislocalisation. This study quantified the biases induced by complex spatial dependence and showed its potential remedy through high-dimensional techniques.

The mislocalisation was greatly remedied by a high-dimensional multivariate approach, which indicated the ability of high-dimensional modelling to cope with complexity. We therefore expected the validity of high-dimensional models could improve the performance of lesion-deficit prediction. In Chapter 2, we obtained an extended dataset including 1172 ischemic lesions normalised into a generic space at voxel size of 2mm. By collapsing the lesions to right hemisphere, a stack of 1172 lesion maps in binary including 66770 features was established. In order to validate the dimensionality of this largest reported vascular lesion set, t-SNE was used to generate a two-dimensional embedding of the lesion anatomy labelled by the volume of each lesion (Figure 2.1). The resultant clustering illustrated the distribution to be neither simply uniform nor wholly random, which implied the spatial structure of lesion anatomy could not be explained by any single or small number of factors. Instead, it was apprehensible in a complex pattern. We subsequently conducted two series of simulations to contrast the predictive power of low-dimensional models to the models in high-dimensionality (2.2.4.1 and 2.2.4.2). Mass-univariate based Fisher’s exact test was adopted as the technique in low-dimensional analysis; whereas, the technique employed in high-dimensional analysis was multivariate-based support vector machines (SVM). In both series of simulations, the prediction depends on the causal relation between lesion and function, so we adopted the notion of counterfactuals to involve a hypothetical functional deficit to construct causality. In addition, as the major objective of the simulations was to examine to what extent a high-dimensional multivariate analysis can lift the limit of lesion-deficit prediction, the benefit from the counterfactual explanations is the maximum flexibility for modelling. A hypothetical functional deficit allowed us to construct the prediction covering the whole brain which provided the optimal reliability to discriminate the performance between two approaches.

The first series of simulations was based on a single Brodmann region. For a total of 39 Brodmann regions, if 20% of a particular lesion’s volume was overlapped with a particular region, the patient was assumed to be affected by the hypothetic functional deficit. The models were complicated in the second series of simulations, which were based on a combinatorial pair of Brodmann regions. A total of 819 pairs covered the completed combination of Brodmann areas: a challenge of computational complexity. As predicted, in both series of simulations, high-dimensional models achieved significant superiority compared with the corresponding low-dimensional models. In the dual-region scenario, the extent of advantage was further magnified. For the mass-univariate analysis, the voxel-wise significant was derived from a *p*-value for each voxel independently; whereas, in SVM models, a weighted map took into account the all of the brain voxels and their interactions determined the significance. High-dimensional multivariate analysis mitigated the spatial biases induced in the mass-univariate analysis, hence improving predictive performance. Moreover, the issue of spatial independence was proportional to the complexity of predictive models. The more complex the causation, the more severe biases were induced. Although we were not able to further complicate the predictive model, such as by combining three or more Brodmann areas, owing to the constraints on computation, it is plausible that greater contrasts will be demonstrated with more complex ground truths, given sufficient data.

1172 has been the largest sample size reported in a lesion study. But, given the dimensionality of the feature space (66770 brain voxels were involved in the studies of Chapter 2 and 3), increased sample size can be expected to enhance the reliability and improve the performance of multivariate analysis. The distinct characteristics of clinical data determine the difficulties of rapid growth in scale. Taking maximum advantage of existing datasets is therefore worth exploring. As a variation of the original SVM, Transductive SVM (TSVM) -- a semi-supervised learning technique -- naturally extends learning to both labelled and unlabelled instances. Replicating our simulation approach, we thus sampled five single and ten combinatorial pairs of Brodmann regions to evaluate predictive models with supervised and semi-supervised learning methods (Chapter 3). In theory, by taking into account an extra set of unlabelled data, TSVM can adjust the optimised hyperplane with labelled data points alone towards the regions in lower data density. Hence, TSVM is effective in enhancing the classifier while reducing the probability of overfitting. In the case single Brodmann region ground truths, semi-supervised learning models improved prediction fidelity in comparison with purely supervised methods. In the more complicated dual-region case, semi-supervised learning models showed apparent superiority, but not significantly within the 95% confidence interval. The former case illustrated the fundamental value of including unlabelled data in such models, though this will interact with the complexity of the underlying relation. In the future, further investigation is required to clarify the limits of semi-supervised learning prediction.

The dimensionality of lesion anatomical structure compels a high-dimensional multivariate analysis. Indeed, we have to construct complex models to be commensurate with the complex anatomy of brain damage, which does not allow any compromise in terms of intensive computation and a large-scale dataset. In practice, the remarkably superior predictive power achieved by high-dimensional modelling will likely contribute to the precision of diagnoses substantially.

6.2 Enhanced detectability of therapeutic effects by high-dimensional modelling

We have seen that both conceptually and technically, awareness of brain lesion dimensionality could revolutionise the approach to modelling focal brain injury. High-dimensional multivariate analysis has the theoretical power to solve the systematic biases induced by complex patterns of spatial dependence present in lesions, even if this will always be dependent on sufficient data. High-dimensional multivariate analysis that is capable of taking into account the full complexity of lesion anatomical patterns has great potential to illuminate the causal relations between lesion and behaviours across an extensive range of contexts.

Therapeutic inference aims to identify the effectiveness of a drug across the target population. It is conventionally done with simple statistical models relying on a small number of crude parameters, such as age, sex, and basic lesion parameters that are too simple to capture individual variation across the population. Consequently, large positive effects may be entirely missed within a randomised controlled trial, a phenomenon that might explain the common failure of interventions at the human level already proven at the much simpler animal level. To address this question, in Chapter 4, we therefore evaluated therapeutic functions at low- and high-dimensionality to demonstrate to what extent the approach of high-dimensional multivariate analysis could uncover drug effects concealed by conventional studies. Counterfactually, we simulated non-lesion altering and lesion-altering simulations and quantified the sensitivities of low- and high-dimensional models in two series of simulations, showing substantial differences (62.90.0% vs. 56.00% and 78.40% vs. 55.00%). This combination of simulated ground truths with real lesion data allowed us to traverse the full space of possibility, quantifying the impact of modelled dimensionality with the greatest generalizability. Two series of simulations differing in the nature of the modelled intervention illuminated two cardinal modes of therapy in stroke: lesion-altering and non-lesion altering approaches.

Modelling therapeutic inference with these two contrasting approaches, we showed the relation between lesion and outcome to be best determined by a causal field of a multiplicity of collateral factors numerous in their scale and complex in their interactions. Simple low-dimensional models representing lesions merely by their volumes were shown to result in poorly sensitive models across all our simulations. The spatial dependencies of brain damage determine the complexity of valid models, which is proportional to the complexity of behavioural mechanism underlying the neural system in play and of the pathology that has deranged its operation. Here, a patient's neurological state was indicated by gaze control, whose complexity can be exceeded by many other behavioural mechanisms. In Chapters 2 and 3, we have elaborated the contrast between low- and high-dimensional models to a more complex context. We expect the difference in our ability to detect interventional effects will be more prominent the more complex the outcome, though we need to construct more complex explanatory models correspondingly.

Therapeutic inference not only to reveals the causal relation of lesion and behaviour; but also potentially the mechanism of the behaviour itself. Practically, the high-dimensional approach could contribute to the precision of targeted drug development; it will also be an opportunity to review the effects of interventions judged to be failures on the basis of low-dimensional trails that may well have been erroneous.

6.3 High-dimensional modelling of heterogeneous treatment effects

Our studies on lesion prediction and therapeutic inference have investigated the mapping between brain injury and function, and the effectiveness of a drug across the population. In the final section of this thesis, we modelled treatments with heterogeneous effects to investigate how a particular patient responded to a particular treatment. Strategically, this allowed us to explore achieving greater precision in tailoring treatment.

The concept of individualised medicine is not new. The critical question is how to develop valid approach to identifying the inherent differences across individuals so as to reveal individual susceptibility to a particular treatment. Fundamental research has shown convincingly that the complexity of the human brain is determined by the interactions between multiple physical and functional levels (Bassett and Gazzaniga, 2011). For example, at the physiological level, the selective response properties of each individual neuron give rise, in massive combinatorial interaction, to the complexity and diversity of functional neural networks (Yagi, 2013, Yoshimura et al., 2005); whilst, at the molecular level, neurobiological studies have revealed how genetic patterns individuate neurons (Baaré et al., 2001, Peper et al., 2007, Pol et al., 2006). All these studies imply the individual brain is bound to be unique, organised according to a complex order. To realise the goal of personalised treatment will thus ultimately require innovative approaches covering various levels. The conventional approach of relying on low-dimensional representations of the crude average of the population could never succeed outside niche, unusually simple cases.

In line with the models in preceding chapters, here we employed simulated ground truths to investigate prescriptive inference, because we are interested in isolating the contribution of the lesion architecture, for which we have empirical data, and which real-world ground truths of outcome would otherwise conceal. Simulating both functional deficits and treatment susceptibility, both dependent on complex features of the lesion architecture, we constructed low- and high-dimensional prescriptive models. In contrast with the models investigated in Chapter 2-4, we employed two techniques of dimensionality reduction to derive our dimensionally contrasting representations. In the high-dimensional case, non-negative matrix factorisation (NMF) was used to compact the lesion maps into a 50-dimensional space (Lee and Seung, 1999). The low-dimensional representation was derived from t-SNE – a nonlinear technique for dimensionality reduction (Maaten and Hinton, 2008).

The subsequent analysis showed, as predicted, that high-dimensional models were superior in identifying susceptibility to the given treatment. In the second series of simulations in particular (5.3.2), the corresponding proportion of patients responded to the treatment in a polarised way (outcome changed from persistence to recovery if susceptible, conversely if not) demonstrated more prominent discrimination that the other series (outcome was only affected from persistence to recovery if susceptible.) (5.3.1). This further revealed the ability of high-dimensional modelling to deal with complexity.

In essence, we demonstrated the potential feasibility of high-dimensional multivariate analysis in individualised prescription in the context of heterogeneous treatment effects. This further enriched the applicability of high-dimensional modelling to the neurological realm.

6.4 Limitations and future work

The spatial dependencies studied here stemmed from the intrinsic complexity of lesion anatomy and determined the foundation of lesion-function inference – a high-dimensional multivariate analysis. In this thesis, we demonstrated the approach of high-dimensional modelling in terms of ideas and methodology progressively – from lesion-deficit prediction about patients in general, therapeutic inference about a treatment in general, finally to prescriptive inference in the setting of heterogeneous treatment effects.

Lesion-deficit prediction contributes to our understanding of how individuals respond to focal brain injury. Practically, the superior predictive power achieved in high-dimensional approach allows for the greater precision in diagnosis and prognosis. Our models were based on a single and combinatorial pair of Brodmann regions, respectively. Given greater computational resource and a wider time frame, we could have constructed more complex models, such as combinations of three Brodmann regions, to enhance the exploration of the differences between low- and high-dimensional models. For the Brodmann atlas including 41 distinct regions, the number of three-region combinations will be 10660, which is indeed a great challenge. So, in the future, we may apply independent component analysis (ICA) to randomly sample a certain number of three-region combinations to demonstrate the predictive in a more complex context. On the other hand, besides the atlases derived from cytoarchitecture or anatomy, there have been some parcellations driven by brain connectivity (Glasser et al., 2016b) in recent years. It is worth establishing physiologically more plausible lesion-deficit mapping models with state-of-the-art brain parcellations in the future.

Therapeutic inference contributes to revealing the sensitivity of clinical trials, which enables greater precision in targeted drug development. In our study, the direction of gaze was measured by a semi-automated algorithm which required an experienced neurologist to localise the centre of the pupil before calculating the deviated degree automatically. Future work is therefore proposed to perform gaze segmentation in a full automated way. In addition, we are naturally motivated to explore other behaviours with plausibly much more complex mechanisms to see if the superiority of high-dimensional approach can be amplified.

Prescriptive inference contributes to identifying individual susceptibility to treatments. Inference capable of handling heterogeneous treatment effects will provide greater precision in therapeutic selection at the individual level. In our study, two sets of synthetic ground truths – affected or not by a functional deficit and susceptible or not to a give treatment – formed a 2 x 2 confusion matrix. Fifty- and two-dimensional representations were used for high- and low-dimensional feature space. The inferential models were smoothly constructed under a relatively simple context. In the next stage, we would like to extend the current model to a multi-label inference by involving a multiplicity of treatments because it is practically plausible for a patient to receive multiple treatments simultaneously. Evolving from the work we have done, future work could also model individual susceptibility to disease prevention.

In terms of concept and methodology, we demonstrated the intrinsic complex lesion architecture distributed in human brain which compels complex and intensive computational models to commensurate with. Certainly, this is the very much beginning to bind neuroimaging with state-of-art machine learning and artificial intelligence techniques for exploring the interactions and covariance of pathological processes. Regarding with the future phase, the works need to be improved and extended in serveral following aspects. First and foremost, the high-dimensional multivariate modelling can only release its power by feeding the sufficient data. Theoretically, the more complex and advanced computational model, the larger volume of data set is required. Therefore, we need data at massive scale, allied with improvements in algorithmics and computation. Moreover, a growing scale of dataset and developing machine learning algorithms are complementary with each other. Rapid advances in machine learning and artificial intelligence techniques will unlock new areas of understanding that were previously inaccessible within the high-dimensional multivariate analysis framework. Secondly, since we aim to continuously collect the data to approximate the population level as much as possible, a practical question is that it is necessary to establish a solid and efficient pipeline or framework to pre-process the heterogeneous clinical data, which is fundamental and prerequisite for any big data study. The manual and labour-intensive works are not feasible any more to prepare for a study involving a large-scale dataset, instead, we need a pipeline to be automatic and reliable. Ideally, it is worth of trying to build up an open source platform to generalise its functionality and benefit more research. Thirdly, the counterfactual explanations were broadly applied to model the hypothetical simulations. This is the initial step to demonstrate the concept and method derived from the high-dimensional modelling. In order to enhance and extend its superiorities, we need to upgrade the models by involving the real and complicated clinical outcomes. Fourthly, the translation of the complex computational models to the clinical applications is a crucial concern. We need to simplify the operation and optimise the user interface to fill the gap between the researchers and clinicians. Thereby, in the future, we need to closely work with the clinicians to see how the advances derived from the superior analysis models can actually benefit the routines in clinical environment. Last but not least, as aforementioned, the proposed method in this thesis is to establish the generalisability of relating the focal brain injury and its clinical behaviour, but not any subgroup study. We implemented a large-scale acute ischaemic stroke dataset in this thesis, because it is one of the most common and typical vascular damage not only in the UK but across the world. In future, we would like to translate our model to a broad range of neurological disorders, hence, establish the multidisciplinarily collaborations and impacts among multi-organisations.

Chapter 7 Appendix

7.1 Appendix A: the details of scanners involved into this thesis

Manufacturer	Model	Tesla
GE Medical System	Discovery MR 450	1.5
GE Medical System	Genesis Signa	1.5
GE Medical System	Signa Excite	1.5
GE Medical System	Signa HD	1.5
Philips Medical System	Achieva	1.5
Philips Medical System	Achieva	3
Philips Medical System	Ingenia	1.5
Philips Medical System	Ingenia	3
Philips Medical System	Intera	1.5
Siemens	Avanto	1.5
Siemens	Biograph mMR	1.5
Siemens	Espree	1.5
Siemens	Skyra	3
Siemens	Symphony	1.5
Siemens	Symphony Tim	1.5

Siemens	Trio Tim	3
Siemens	Verio	3

Table 7.1. The list of scanner details

7.2 Appendix B: demographic information

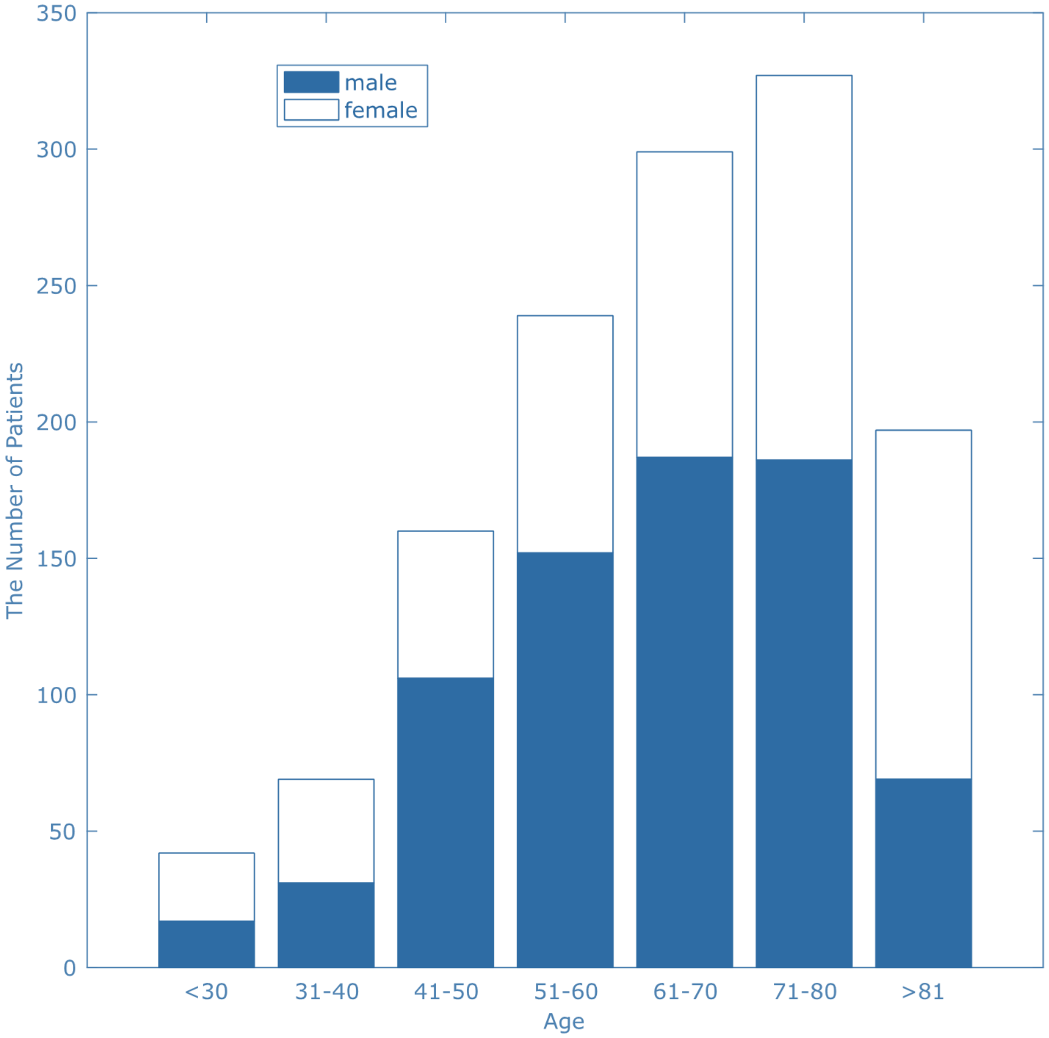


Figure 7.1 Age and sex distribution

The distribution of age was divided into seven bins to show the most involved age range. For each bin, the proportion with filled colour represented the ratio of male patients.

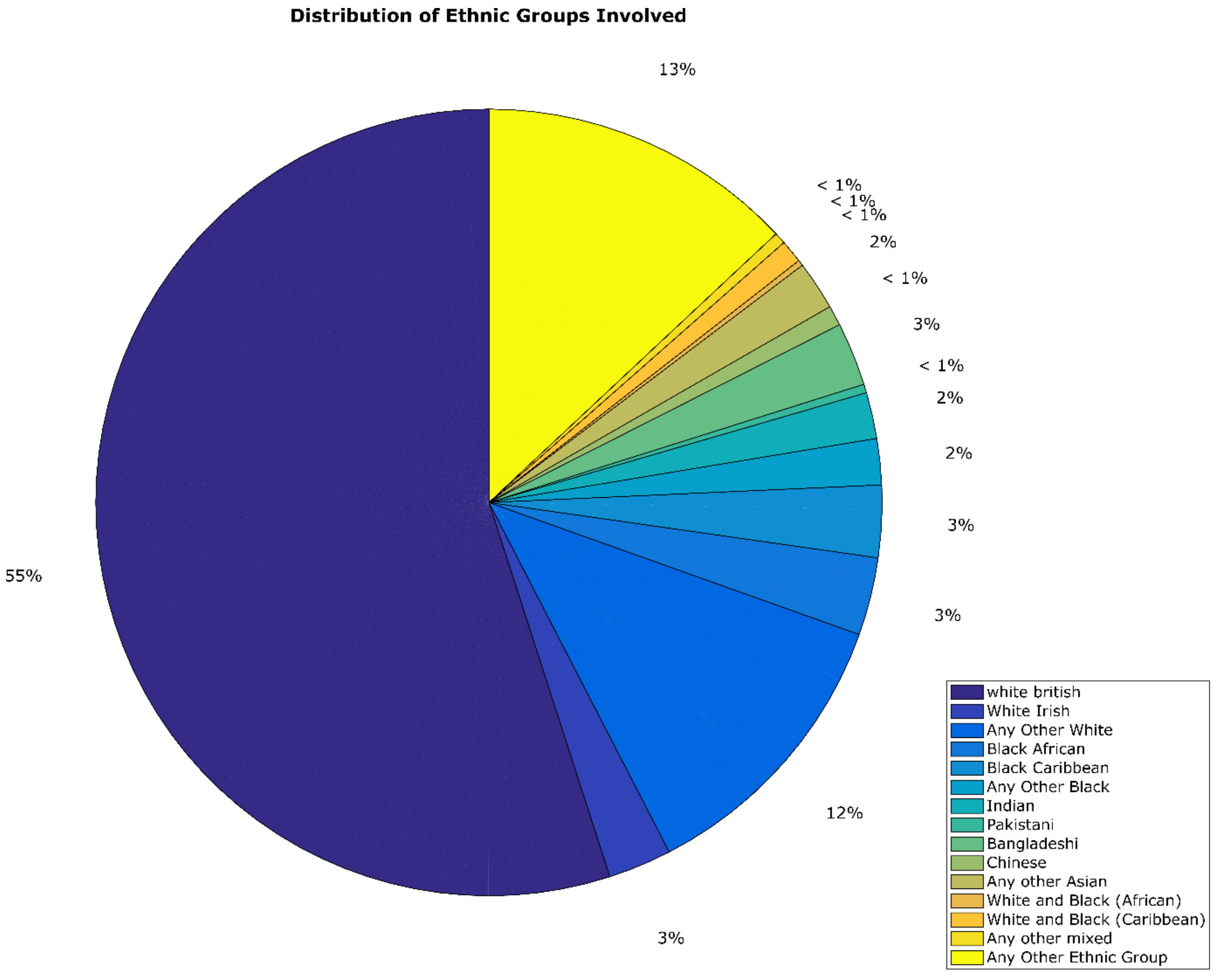


Figure 7.2 Ethnical group

The distribution of self-reported patient ethnicity is shown in this figure. 81.2% of the stack of 1172 patients was recorded.

Chapter 8 Bibliography

ADÈR, H. J. 2008. *Advising on research methods: A consultant's companion*, Johannes van Kessel Publishing.

ADOLPHS, R., TRANEL, D., DAMASIO, H. & DAMASIO, A. 1994. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature,* 372**,** 669-672.

ADOLPHS, R., TRANEL, D., DAMASIO, H. & DAMASIO, A. R. 1995. Fear and the human amygdala. *The Journal of neuroscience,* 15**,** 5879-5891.

ANDO, R. K. & ZHANG, T. A high-performance semi-supervised learning method for text chunking. Proceedings of the 43rd annual meeting on association for computational linguistics, 2005. Association for Computational Linguistics, 1-9.

ANTHONY, G., GREGG, H. & TSHILIDZI, M. 2007. Image classification using SVMs: one-against-one vs one-against-all. *arXiv preprint arXiv:0711.2914*.

ARLOT, S. & CELISSE, A. 2010. A survey of cross-validation procedures for model selection. *Statistics surveys,* 4**,** 40-79.

ASHBURNER, J. & FRISTON, K. J. 2000. Voxel-based morphometry—the methods. *Neuroimage,* 11**,** 805-821.

ASHBURNER, J. & FRISTON, K. J. 2005. Unified segmentation. *Neuroimage,* 26**,** 839-851.

BAARÉ, W. F., POL, H. E. H., BOOMSMA, D. I., POSTHUMA, D., DE GEUS, E. J., SCHNACK, H. G., VAN HAREN, N. E., VAN OEL, C. J. & KAHN, R. S. 2001. Quantitative genetic modeling of variation in human brain morphology. *Cerebral Cortex,* 11**,** 816-824.

BADIRU, A. B. 1992. Computational survey of univariate and multivariate learning curve models. *IEEE transactions on Engineering Management,* 39**,** 176-188.

BAHL, L., BROWN, P., DE SOUZA, P. & MERCER, R. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86., 1986. IEEE, 49-52.

BAKER, L. D. & MCCALLUM, A. K. Distributional clustering of words for text classification. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998. ACM, 96-103.

BARBEY, A. K., COLOM, R., SOLOMON, J., KRUEGER, F., FORBES, C. & GRAFMAN, J. 2012. An integrative architecture for general intelligence and executive function revealed by lesion mapping. *Brain,* 135**,** 1154-1164.

BARKER, A. T., JALINOUS, R. & FREESTON, I. L. 1985. Non-invasive magnetic stimulation of human motor cortex. *The Lancet,* 325**,** 1106-1107.

BARTLEY, A. J., JONES, D. W. & WEINBERGER, D. R. 1997. Genetic variability of human brain size and cortical gyral patterns. *Brain: a journal of neurology,* 120**,** 257-269.

BARUTCUOGLU, Z., SCHAPIRE, R. E. & TROYANSKAYA, O. G. 2006. Hierarchical multi-label prediction of gene function. *Bioinformatics,* 22**,** 830-836.

BASSETT, D. S. & GAZZANIGA, M. S. 2011. Understanding complexity in the human brain. *Trends in cognitive sciences,* 15**,** 200-209.

BATES, E., WILSON, S. M., SAYGIN, A. P., DICK, F., SERENO, M. I., KNIGHT, R. T. & DRONKERS, N. F. 2003. Voxel-based lesion–symptom mapping. *Nature neuroscience,* 6**,** 448-450.

BAUM, L. E., PETRIE, T., SOULES, G. & WEISS, N. 1970. A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics,* 41**,** 164-171.

BECHARA, A., DAMASIO, A. R., DAMASIO, H. & ANDERSON, S. W. 1994. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition,* 50**,** 7-15.

BECHARA, A., DAMASIO, H. & DAMASIO, A. R. 2000. Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex,* 10**,** 295-307.

BELKIN, M. & NIYOGI, P. 2004. Semi-supervised learning on Riemannian manifolds. *Machine learning,* 56**,** 209-239.

BELKIN, M., NIYOGI, P. & SINDHWANI, V. 2006. Manifold regularisation: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research,* 7**,** 2399-2434.

BELLMAN, R. E. 2015. *Adaptive control processes: a guided tour,* Princeton university press.

BEN-HUR, A., HORN, D., SIEGELMANN, H. T. & VAPNIK, V. 2001. Support vector clustering. *Journal of machine learning research,* 2**,** 125-137.

BEN-HUR, A. & WESTON, J. 2010. A user's guide to support vector machines. *Data mining techniques for the life sciences.* Springer.

BENNETT, K. P. & DEMIRIZ, A. Semi-supervised support vector machines. Advances in Neural Information processing systems, 1999. 368-374.

BERNARDO, J., BAYARRI, M., BERGER, J., DAWID, A., HECKERMAN, D., SMITH, A. & WEST, M. 2007. Generative or discriminative? getting the best of both worlds. *Bayesian statistics,* 8**,** 3-24.

BEYER, M. A. & LANEY, D. 2012. The importance of 'big data': a definition. *Stamford, CT: Gartner***,** 2014-2018.

BIANCHINI, M., BELAHCEN, A. & SCARSELLI, F. A comparative study of inductive and transductive learning with feedforward neural networks. Conference of the Italian Association for Artificial Intelligence, 2016. Springer, 283-293.

BIE, T. D. & CRISTIANINI, N. Convex methods for transduction. Advances in neural information processing systems, 2004. 73-80.

BLAND, J. M. & ALTMAN, D. G. 1995. Multiple significance tests: the Bonferroni method. *Bmj,* 310**,** 170.

BLUM, A. & MITCHELL, T. Combining labeled and unlabeled data with co-training. Proceedings of the eleventh annual conference on Computational learning theory, 1998. ACM, 92-100.

BONFERRONI, C. E. 1935. Il calcolo delle assicurazioni su gruppi di teste. *Studi in onore del professore salvatore ortu carboni***,** 13-60.

BONKHOFF, A., XU, T., JHA, A., JAGER, H. & NACHEV, P. 2019. Reclassifying stroke lesion anatomy.

BOSER, B. E., GUYON, I. M. & VAPNIK, V. N. A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory, 1992. ACM, 144-152.

BOTEV, Z. I., GROTOWSKI, J. F. & KROESE, D. P. 2010. Kernel density estimation via diffusion. *The Annals of Statistics,* 38**,** 2916-2957.

BOUTELL, M. R., LUO, J., SHEN, X. & BROWN, C. M. 2004. Learning multi-label scene classification. *Pattern recognition,* 37**,** 1757-1771.

BRACKEN, M. B. 2009. Why animal studies are often poor predictors of human reactions to exposure. *Journal of the royal society of medicine,* 102**,** 120-122.

BRAGA-NETO, U. M. & DOUGHERTY, E. R. J. B. 2004. Is cross-validation valid for small-sample microarray classification? 20**,** 374-380.

BREIMAN, L. & SPECTOR, P. 1992. Submodel selection and evaluation in regression. The X-random case. *International statistical review/revue internationale de Statistique***,** 291-319.

BREIMAN, L. J. M. L. 1996. Bagging predictors. 24**,** 123-140.

BRETT, M., JOHNSRUDE, I. S. & OWEN, A. M. 2002. The problem of functional localisation in the human brain. *Nature reviews neuroscience,* 3**,** 243-249.

BROCA, P. 1861. Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech). *Bulletin de la Société Anatomique,* 6**,** 330-57.

BRODERSEN, K. H., ONG, C. S., STEPHAN, K. E. & BUHMANN, J. M. The balanced accuracy and its posterior distribution. 2010 20th International Conference on Pattern Recognition, 2010. IEEE, 3121-3124.

BRODIE, B. B. 1962. Part VI. Difficulties in extrapolating data on metabolism of drugs from animal to man. *Clinical Pharmacology & Therapeutics,* 3**,** 374-380.

BRUNONI, A. R., NITSCHE, M. A., BOLOGNINI, N., BIKSON, M., WAGNER, T., MERABET, L., EDWARDS, D. J., VALERO-CABRE, A., ROTENBERG, A. & PASCUAL-LEONE, A. 2012. Clinical research with transcranial direct current stimulation (tDCS): challenges and future directions. *Brain stimulation,* 5**,** 175-195.

BULLMORE, E. & SPORNS, O. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience,* 10**,** 186-198.

BULLMORE, E. T. & BASSETT, D. S. 2011. Brain graphs: graphical models of the human brain connectome. *Annual review of clinical psychology,* 7**,** 113-140.

CARRERA, E. & TONONI, G. 2014. Diaschisis: past, present, future. *Brain,* 137**,** 2408-2422.

CARUANA, R., KARAMPATZIAKIS, N. & YESSENALINA, A. An empirical evaluation of supervised learning in high dimensions. Proceedings of the 25th international conference on Machine learning, 2008. ACM, 96-103.

CASTELLI, V. & COVER, T. M. 1996. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on information theory,* 42**,** 2102-2117.

CEVIKALP, H. & FRANC, V. 2017. Large-scale robust transductive support vector machines. *Neurocomputing,* 235**,** 199-209.

CHANG, C.-C. & LIN, C.-J. 2011. LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST),* 2**,** 27.

CHAPELLE, O., SCHOLKOPF, B. & ZIEN, A. 2009. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks,* 20**,** 542-542.

CHAPELLE, O., SINDHWANI, V. & KEERTHI, S. S. 2008. Optimisation techniques for semi-supervised support vector machines. *Journal of Machine Learning Research,* 9**,** 203-233.

CHAPELLE, O. & ZIEN, A. Semi-supervised classification by low density separation. AISTATS, 2005. Citeseer, 57-64.

CHECHLACZ, M., ROTSHTEIN, P. & HANSEN, P. 2018. Mapping functional brain organisation: Rethinking lesion symptom mapping and advanced neuroimaging methods in the understanding of human cognition. Elsevier.

CHEN, J., LIU, J., CALHOUN, V. D., ARIAS-VASQUEZ, A., ZWIERS, M. P., GUPTA, C. N., FRANKE, B. & TURNER, J. A. 2014. Exploration of scanning effects in multi-site structural MRI studies. *Journal of neuroscience methods,* 230**,** 37-50.

CHEN, L. 2009. Curse of dimensionality. *Encyclopedia of Database Systems.* Springer.

CLARE, A. & KING, R. D. Knowledge discovery in multi-label phenotype data. European Conference on Principles of Data Mining and Knowledge Discovery, 2001. Springer, 42-53.

COLLINS, F. S. & VARMUS, H. 2015. A new initiative on precision medicine. *New England Journal of Medicine,* 372**,** 793-795.

COLLOBERT, R., SINZ, F., WESTON, J. & BOTTOU, L. Trading convexity for scalability. Proceedings of the 23rd international conference on Machine learning, 2006. ACM, 201-208.

CORBETTA, M., AKBUDAK, E., CONTURO, T. E., SNYDER, A. Z., OLLINGER, J. M., DRURY, H. A., LINENWEBER, M. R., PETERSEN, S. E., RAICHLE, M. E. & VAN ESSEN, D. C. 1998. A common network of functional areas for attention and eye movements. *Neuron,* 21**,** 761-773.

CORDUNEANU, A. A. D. 2002. *Stable mixing of complete and incomplete information.* Massachusetts Institute of Technology.

CORTES, C. & VAPNIK, V. 1995. Support-vector networks. *Machine learning,* 20**,** 273-297.

CULP, M. & MICHAILIDIS, G. 2008. An iterative algorithm for extending learners to a semi-supervised setting. *Journal of Computational and Graphical Statistics,* 17**,** 545-571.

CUMMINGS, J. L., MORSTORF, T. & ZHONG, K. 2014. Alzheimer’s disease drug-development pipeline: few candidates, frequent failures. *Alzheimer's research & therapy,* 6**,** 1.

DAFFNER, K. R., AHERN, G. L., WEINTRAUB, S. & MESULAM, M. M. J. A. O. N. 1990. Dissociated neglect behavior following sequential strokes in the right hemisphere. 28**,** 97-101.

DAMASIO, H. & DAMASIO, A. R. 1989. *Lesion analysis in neuropsychology,* Oxford University Press, USA.

DAYAN, P. & HINTON, G. E. 1997. Using expectation-maximisation for reinforcement learning. *Neural Computation,* 9**,** 271-278.

DE HAAN, B. & KARNATH, H.-O. 2018. A hitchhiker's guide to lesion-behaviour mapping. *Neuropsychologia,* 115**,** 5-16.

DEMBCZYNSKI, K., CHENG, W. & HÜLLERMEIER, E. Bayes Optimal Multilabel Classification via Probabilistic Classifier Chains. ICML, 2010. 279-286.

DEMIRIZ, A. & BENNETT, K. P. 2001. Optimisation approaches to semi-supervised learning. *Complementarity: Applications, Algorithms and Extensions.* Springer.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)***,** 1-38.

DRONKERS, N. F., PLAISANT, O., IBA-ZIZEN, M. T. & CABANIS, E. A. 2007. Paul Broca's historic cases: high resolution MR imaging of the brains of Leborgne and Lelong. *Brain,* 130**,** 1432-1441.

DRUCKER, H., BURGES, C. J., KAUFMAN, L., SMOLA, A. J. & VAPNIK, V. Support vector regression machines. Advances in neural information processing systems, 1997. 155-161.

EFRON, B. 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association,* 78**,** 316-331.

EFRON, B. & GONG, G. 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician,* 37**,** 36-48.

EFRON, B. & TIBSHIRANI, R. 1997. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association,* 92**,** 548-560.

EFRON, B. J. J. O. T. A. S. A. 1986. How biased is the apparent error rate of a prediction rule? 81**,** 461-470.

ELISSEEFF, A. & WESTON, J. A kernel method for multi-labelled classification. Advances in neural information processing systems, 2002. 681-687.

FEENEY, D. M. & BARON, J.-C. 1986. Diaschisis. *Stroke,* 17**,** 817-830.

FISHER, R. A. 2006. *Statistical methods for research workers*, Genesis Publishing Pvt Ltd.

FOCKE, N. K., HELMS, G., KASPAR, S., DIEDERICH, C., TÓTH, V., DECHENT, P., MOHR, A. & PAULUS, W. 2011. Multi-site voxel-based morphometry—not quite there yet. *Neuroimage,* 56**,** 1164-1170.

FORTMANN-ROE, S. 2012. Accurately measuring model prediction error.

FOX, M. D., SNYDER, A. Z., VINCENT, J. L., CORBETTA, M., VAN ESSEN, D. C. & RAICHLE, M. E. 2005. The human brain is intrinsically organised into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 9673-9678.

FOX, M. D. J. N. E. J. O. M. 2018. Mapping symptoms to brain networks with the human connectome. 379**,** 2237-2245.

FREY, R., WOODS, D., KNIGHT, R., SCABINI, D. & CLAYWORTH, C. Defining functional areas with averaged CT scans. Society for Neuroscience Abstracts, 1987. 1266.

FRISTON, K., CHU, C., MOURÃO-MIRANDA, J., HULME, O., REES, G., PENNY, W. & ASHBURNER, J. 2008. Bayesian decoding of brain images. *Neuroimage,* 39**,** 181-205.

FRISTON, K. J., HARRISON, L. & PENNY, W. 2003. Dynamic causal modelling. *Neuroimage,* 19**,** 1273-1302.

FRISTON, K. J., HOLMES, A. P., WORSLEY, K. J., POLINE, J. P., FRITH, C. D. & FRACKOWIAK, R. S. 1994. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping,* 2**,** 189-210.

FUNG, G. & MANGASARIAN, O. L. 2001. Semi-superyised support vector machines for unlabeled data classification. *Optimisation methods and software,* 15**,** 29-44.

FÜRNKRANZ, J., HÜLLERMEIER, E., MENCÍA, E. L. & BRINKER, K. J. M. L. 2008. Multilabel classification via calibrated label ranking. 73**,** 133-153.

GAMMERMAN, A., VOVK, V. & VAPNIK, V. Learning by transduction. Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, 1998. Morgan Kaufmann Publishers Inc., 148-155.

GANDOMI, A. & HAIDER, M. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management,* 35**,** 137-144.

GE, T., MÜLLER-LENKE, N., BENDFELDT, K., NICHOLS, T. E. & JOHNSON, T. D. 2014. Analysis of multiple sclerosis lesions via spatially varying coefficients. *The annals of applied statistics,* 8**,** 1095.

GEERTS, H. 2009. Of Mice and Men. *CNS drugs,* 23**,** 915-926.

GHAMRAWI, N. & MCCALLUM, A. Collective multi-label classification. Proceedings of the 14th ACM international conference on Information and knowledge management, 2005. ACM, 195-200.

GLÄSCHER, J., RUDRAUF, D., COLOM, R., PAUL, L. K., TRANEL, D., DAMASIO, H. & ADOLPHS, R. 2010. Distributed neural system for general intelligence revealed by lesion mapping. *Proceedings of the National Academy of Sciences,* 107**,** 4705-4709.

GLASSER, M. F., COALSON, T. S., ROBINSON, E. C., HACKER, C. D., HARWELL, J., YACOUB, E., UGURBIL, K., ANDERSSON, J., BECKMANN, C. F., JENKINSON, M., SMITH, S. M. & VAN ESSEN, D. C. 2016a. A multi-modal parcellation of human cerebral cortex. *Nature,* 536**,** 171-8.

GLASSER, M. F., SMITH, S. M., MARCUS, D. S., ANDERSSON, J. L., AUERBACH, E. J., BEHRENS, T. E., COALSON, T. S., HARMS, M. P., JENKINSON, M. & MOELLER, S. 2016b. The human connectome project's neuroimaging approach. *Nature Neuroscience,* 19**,** 1175-1187.

GLEICHGERRCHT, E., FRIDRIKSSON, J., RORDEN, C. & BONILHA, L. 2017. Connectome-based lesion-symptom mapping (CLSM): a novel approach to map neurological function. *NeuroImage: Clinical,* 16**,** 461-467.

GODBOLE, S. & SARAWAGI, S. Discriminative methods for multi-labeled classification. Pacific-Asia conference on knowledge discovery and data mining, 2004. Springer, 22-30.

GOLDBERG, A., ZHU, X., SINGH, A., XU, Z. & NOWAK, R. Multi-manifold semi-supervised learning. Artificial Intelligence and Statistics, 2009. 169-176.

GOOD, C. D., JOHNSRUDE, I. S., ASHBURNER, J., HENSON, R. N., FRISTON, K. J. & FRACKOWIAK, R. S. 2001. A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage,* 14**,** 21-36.

GRANDVALET, Y. & BENGIO, Y. Semi-supervised learning by entropy minimisation. Advances in neural information processing systems, 2005. 529-536.

GROPPE, D. M., URBACH, T. P. & KUTAS, M. 2011. Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology,* 48**,** 1711-1725.

GROUP, N. T.-P. S. S. 1997. Generalised efficacy of t-PA for acute stroke: subgroup analysis of the NINDS t-PA Stroke Trial. *Stroke,* 28**,** 2119-2125.

HADLEY, S. W., RENDL, F. & WOLKOWICZ, H. Bounds for the Quadratic Assignment Problems Using Continuous Optimisation Techniques. IPCO, 1990. 237-248.

HALLIGAN, P. W., FINK, G. R., MARSHALL, J. C. & VALLAR, G. J. T. I. C. S. 2003. Spatial cognition: evidence from visual neglect. 7**,** 125-133.

HAMBURG, M. A. & COLLINS, F. S. 2010. The path to personalised medicine. *New England Journal of Medicine,* 363**,** 301-304.

HAN, X., JOVICICH, J., SALAT, D., VAN DER KOUWE, A., QUINN, B., CZANNER, S., BUSA, E., PACHECO, J., ALBERT, M. & KILLIANY, R. 2006. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *Neuroimage,* 32**,** 180-194.

HANLEY, J. A. & MCNEIL, B. J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* 143**,** 29-36.

HAWRYLYCZ, M. J., LEIN, E. S., GUILLOZET-BONGAARTS, A. L., SHEN, E. H., NG, L., MILLER, J. A., VAN DE LAGEMAAT, L. N., SMITH, K. A., EBBERT, A. & RILEY, Z. L. 2012. An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature,* 489**,** 391.

HEDNA, V. S., BODHIT, A. N., ANSARI, S., FALCHOOK, A. D., STEAD, L., HEILMAN, K. M. & WATERS, M. F. 2013. Hemispheric differences in ischemic stroke: is left-hemisphere stroke more common? *Journal of Clinical Neurology,* 9**,** 97-102.

HEILMAN, K. M. & VALENSTEIN, E. J. N. 1972. Frontal lobe neglect in man. 22**,** 660-660.

HERBET, G., LAFARGUE, G. & DUFFAU, H. 2014. Rethinking voxel-wise lesion-deficit analysis: A new challenge for computational neuropsychology. *cortex,* 30**,** e4.

HINTZE, J. L. & NELSON, R. D. 1998. Violin plots: a box plot-density trace synergism. *The American Statistician,* 52**,** 181-184.

HÖFLER, M. 2005. Causal inference based on counterfactuals. *BMC medical research methodology,* 5**,** 28.

HOUNSFIELD, G. N. 1973. Computerized transverse axial scanning (tomography): Part 1. Description of system. *The British journal of radiology,* 46**,** 1016-1022.

HSU, C.-W., CHANG, C.-C. & LIN, C.-J. 2003. A practical guide to support vector classification.

HSU, C.-W. & LIN, C.-J. 2002. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks,* 13**,** 415-425.

HUANG, X. D., ARIKI, Y. & JACK, M. A. 1990. Hidden Markov models for speech recognition.

HUETTEL, S. A. 2017. Functional MRI (fMRI).

HUGDAHL, K. 2005. Symmetry and asymmetry in the human brain. *European Review,* 13**,** 119-133.

HUSAIN, M. & KENNARD, C. J. J. O. N. 1996. Visual neglect associated with frontal lobe infarction. 243**,** 652-657.

HUSAIN, M. & NACHEV, P. 2007. Space and the parietal cortex. *Trends in cognitive sciences,* 11**,** 30-36.

INDYK, P. & MOTWANI, R. Approximate nearest neighbors: towards removing the curse of dimensionality. Proceedings of the thirtieth annual ACM symposium on Theory of computing, 1998. ACM, 604-613.

JAAKKOLA, T. 2001. 10 tutorial on variational approximation methods. *Advanced mean field methods: theory and practice***,** 129.

JELONEK, J. & STEFANOWSKI, J. 1997. Feature subset selection for classification of histological images. *Artificial Intelligence in Medicine,* 9**,** 227-239.

JHA, A., DIEHL, B., SCOTT, C., MCEVOY, A. W. & NACHEV, P. 2016. Reversed procrastination by focal disruption of medial frontal cortex. *Current Biology,* 26**,** 2893-2898.

JI, S., SUN, L., JIN, R. & YE, J. Multi-label multiple kernel learning. Advances in neural information processing systems, 2009. 777-784.

JIANG, A., WANG, C. & ZHU, Y. Calibrated rank-svm for multi-label image categorisation. Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, 2008. IEEE, 1450-1455.

JOACHIMS, T. Text categorisation with support vector machines: Learning with many relevant features. European conference on machine learning, 1998. Springer, 137-142.

JOACHIMS, T. 1999a. Svmlight: Support vector machine. *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund,* 19.

JOACHIMS, T. Transductive inference for text classification using support vector machines. ICML, 1999b. 200-209.

JOACHIMS, T. 2006. Transductive support vector machines. *Chapelle et al.(2006)***,** 105-118.

JOHANSSON, B. B. 2000. Brain plasticity and stroke rehabilitation: the Willis lecture. *Stroke,* 31**,** 223-230.

JOHANSSON, F., SHALIT, U. & SONTAG, D. Learning representations for counterfactual inference. International conference on machine learning, 2016. 3020-3029.

JOHN, G. H., KOHAVI, R. & PFLEGER, K. 1994. Irrelevant features and the subset selection problem. *Machine Learning Proceedings 1994.* Elsevier.

JOSHI, A., KOEPPE, R. A. & FESSLER, J. A. 2009. Reducing between scanner differences in multi-center PET studies. *Neuroimage,* 46**,** 154-159.

JOVICICH, J., CZANNER, S., HAN, X., SALAT, D., VAN DER KOUWE, A., QUINN, B., PACHECO, J., ALBERT, M., KILLIANY, R. & BLACKER, D. 2009. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *Neuroimage,* 46**,** 177-192.

JOVICICH, J., MARIZZONI, M., SALA-LLONCH, R., BOSCH, B., BARTRÉS-FAZ, D., ARNOLD, J., BENNINGHOFF, J., WILTFANG, J., ROCCATAGLIATA, L. & NOBILI, F. 2013. Brain morphometry reproducibility in multi-center 3T MRI studies: a comparison of cross-sectional and longitudinal segmentations. *Neuroimage,* 83**,** 472-484.

JUCKER, M. 2010. The benefits and limitations of animal models for translational research in neurodegenerative diseases. *Nature medicine,* 16**,** 1210-1214.

KARNATH, H.-O., FERBER, S. & HIMMELBACH, M. 2001a. Spatial awareness is a function of the temporal not the posterior parietal lobe. *Nature,* 411**,** 950.

KARNATH, H.-O., FERBER, S. & HIMMELBACH, M. J. N. 2001b. Spatial awareness is a function of the temporal not the posterior parietal lobe. 411**,** 950.

KARNATH, H.-O., FRUHMANN BERGER, M., KÜKER, W. & RORDEN, C. 2004. The anatomy of spatial neglect based on voxelwise statistical analysis: a study of 140 patients. *Cerebral Cortex,* 14**,** 1164-1172.

KARNATH, H.-O. & SMITH, D. V. 2014. The next step in modern brain lesion analysis: multivariate pattern analysis. *Brain,* 137**,** 2405-2407.

KARNATH, H. O., HIMMELBACH, M. & RORDEN, C. 2002. The subcortical anatomy of human spatial neglect: putamen, caudate nucleus and pulvinar. *Brain,* 125**,** 350-360.

KATSIOS, C. & ROUKOS, D. H. 2010. Individual genomes and personalised medicine: life diversity and complexity. *Personalised Medicine,* 7**,** 347-350.

KEERTHI, S. S. & DECOSTE, D. 2005. A modified finite Newton method for fast solution of large scale linear SVMs. *Journal of Machine Learning Research,* 6**,** 341-361.

KEERTHI, S. S. & LIN, C.-J. 2003. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural computation,* 15**,** 1667-1689.

KIBRIYA, A. M., FRANK, E., PFAHRINGER, B. & HOLMES, G. Multinomial naive bayes for text categorisation revisited. Australasian Joint Conference on Artificial Intelligence, 2004. Springer, 488-499.

KIM, S. C. & KANG, T. J. 2007. Texture classification and segmentation using wavelet packet frame and Gaussian mixture model. *Pattern Recognition,* 40**,** 1207-1221.

KIMBERG, D. Y., COSLETT, H. B. & SCHWARTZ, M. F. 2007. Power in voxel-based lesion-symptom mapping. *Journal of Cognitive Neuroscience,* 19**,** 1067-1080.

KINKINGNÉHUN, S., VOLLE, E., PÉLÉGRINI-ISSAC, M., GOLMARD, J.-L., LEHÉRICY, S., DU BOISGUÉHENEUC, F., ZHANG-NUNES, S., SOSSON, D., DUFFAU, H. & SAMSON, Y. 2007. A novel approach to clinical–radiological correlations: Anatomo-Clinical Overlapping Maps (AnaCOM): Method and validation. *Neuroimage,* 37**,** 1237-1249.

KLOFT, M., BREFELD, U., LASKOV, P., MÜLLER, K.-R., ZIEN, A. & SONNENBURG, S. Efficient and accurate lp-norm multiple kernel learning. Advances in neural information processing systems, 2009. 997-1005.

KLOMJAI, W., KATZ, R. & LACKMY-VALLÉE, A. 2015. Basic principles of transcranial magnetic stimulation (TMS) and repetitive TMS (rTMS). *Annals of physical and rehabilitation medicine,* 58**,** 208-213.

KLÖPPEL, S., ABDULKADIR, A., JACK JR, C. R., KOUTSOULERIS, N., MOURÃO-MIRANDA, J. & VEMURI, P. 2012. Diagnostic neuroimaging across diseases. *Neuroimage,* 61**,** 457-463.

KOCEV, D., VENS, C., STRUYF, J. & DŽEROSKI, S. Ensembles of multi-objective decision trees. European conference on machine learning, 2007. Springer, 624-631.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai, 1995. Montreal, Canada, 1137-1145.

KOHAVI, R. Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. KDD, 1996. Citeseer, 202-207.

KOHAVI, R. & JOHN, G. H. 1997. Wrappers for feature subset selection. *Artificial intelligence,* 97**,** 273-324.

KRAVITZ, D. J., SALEEM, K. S., BAKER, C. I. & MISHKIN, M. 2011. A new neural framework for visuospatial processing. *Nature Reviews Neuroscience,* 12**,** 217.

KROPOTOV, J. D. 2016. *Functional neuromarkers for psychiatry: Applications for diagnosis and treatment*, Academic Press.

KÜMMERER, D., HARTWIGSEN, G., KELLMEYER, P., GLAUCHE, V., MADER, I., KLÖPPEL, S., SUCHAN, J., KARNATH, H.-O., WEILLER, C. & SAUR, D. 2013. Damage to ventral and dorsal language pathways in acute aphasia. *Brain,* 136**,** 619-629.

LANEY, D. 2001. 3D data management: Controlling data volume, velocity and variety. *META Group Research Note,* 6**,** 70.

LAUTERBUR, P. C. 1973. Image formation by induced local interactions: examples employing nuclear magnetic resonance.

LEE, C. H. & YOON, H.-J. 2017. Medical big data: promise and challenges. *Kidney research and clinical practice,* 36**,** 3.

LEE, D. D. & SEUNG, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature,* 401**,** 788.

LESSOV-SCHLAGGAR, C. N., RUBIN, J. B. & SCHLAGGAR, B. L. 2016. The fallacy of univariate solutions to complex systems problems. *Frontiers in neuroscience,* 10**,** 267.

LI, C., EBRAHIMI, A. & SCHLUESENER, H. 2013. Drug pipeline in neurodegeneration based on transgenic mice models of Alzheimer's disease. *Ageing research reviews,* 12**,** 116-140.

LIN, H.-T. & LIN, C.-J. 2003. A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. *submitted to Neural Computation,* 3**,** 1-32.

LIU, H.-J., WANG, Y.-N. & LU, X.-F. A method to choose kernel function and its parameters for support vector machines. Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on, 2005. IEEE, 4277-4280.

LOMBER, S. G. 1999. The advantages and limitations of permanent or reversible deactivation techniques in the assessment of neural function. *Journal of neuroscience methods,* 86**,** 109-117.

LÖUVBLD, K. O., BAIRD, A. E., SCHLAUG, G., BENFIELD, A., SIEWERT, B., VOETSCH, B., CONNOR, A., BURZYNSKI, C., EDELMAN, R. R. & WARACH, S. 1997. Ischemic lesion volumes in acute stroke by diffusion-weighted magnetic resonance imaging correlate with clinical outcome. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society,* 42**,** 164-170.

LÖVBLAD, K.-O., LAUBACH, H.-J., BAIRD, A. E., CURTIN, F., SCHLAUG, G., EDELMAN, R. R. & WARACH, S. 1998. Clinical experience with diffusion-weighted MR in patients with acute stroke. *American Journal of Neuroradiology,* 19**,** 1061-1066.

M WILCOCK, D. 2010. The usefulness and challenges of transgenic mouse models in the study of Alzheimer's disease. *CNS & Neurological Disorders-Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders),* 9**,** 386-394.

MAATEN, L. V. D. & HINTON, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research,* 9**,** 2579-2605.

MADJAROV, G., GJORGJEVIKJ, D. & DELEV, T. Efficient two stage voting architecture for pairwise multi-label classification. Australasian Joint Conference on Artificial Intelligence, 2010. Springer, 164-173.

MADJAROV, G., GJORGJEVIKJ, D. & DŽEROSKI, S. 2012a. Two stage architecture for multi-label learning. *Pattern Recognition,* 45**,** 1019-1034.

MADJAROV, G., KOCEV, D., GJORGJEVIKJ, D. & DŽEROSKI, S. 2012b. An extensive experimental comparison of methods for multi-label learning. *Pattern recognition,* 45**,** 3084-3104.

MAH, Y.-H., HUSAIN, M., REES, G. & NACHEV, P. 2014a. The complexities of lesion-deficit inference in the human brain: Reply to Herbet et al. *Cortex; a journal devoted to the study of the nervous system and behavior.*

MAH, Y. 2014a. *The foundations of lesion-function inference in the human brain.* UCL (University College London).

MAH, Y. 2014b. The foundations of lesion-function inference in the human brain. *PhD Thesis*.

MAH, Y. H., HUSAIN, M., REES, G. & NACHEV, P. 2014b. Human brain lesion-deficit inference remapped. *Brain,* 137**,** 2522-31.

MAH, Y. H., HUSAIN, M., REES, G. & NACHEV, P. 2015. The complexities of lesion-deficit inference in the human brain: reply to Herbet et al. *Cortex,* 64**,** 417-9.

MAH, Y. H., JAGER, R., KENNARD, C., HUSAIN, M. & NACHEV, P. 2014c. A new method for automated high-dimensional lesion segmentation evaluated in vascular injury and applied to the human occipital lobe. *Cortex,* 56**,** 51-63.

MAK, I. W., EVANIEW, N. & GHERT, M. 2014. Lost in translation: animal models and clinical trials in cancer treatment. *American journal of translational research,* 6**,** 114.

MALDONADO, M. A., ALLRED, R. P., FELTHAUSER, E. L. & JONES, T. A. 2008. Motor skill training, but not voluntary exercise, improves skilled reaching after unilateral ischemic lesions of the sensorimotor cortex in rats. *Neurorehabilitation and neural repair,* 22**,** 250-261.

MANJÓN, J. V., COUPÉ, P., BUADES, A., COLLINS, D. L. & ROBLES, M. 2012. New methods for MRI denoising based on sparseness and self-similarity. *Medical image analysis,* 16**,** 18-27.

MCCALLUM, A. Multi-label text classification with a mixture model trained by EM. AAAI workshop on Text Learning, 1999. 1-7.

MCCALLUM, A. & NIGAM, K. A comparison of event models for naive bayes text classification. AAAI-98 workshop on learning for text categorisation, 1998. Citeseer, 41-48.

MCINNES, L., HEALY, J. & MELVILLE, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

MCINTOSH, A. R. & LOBAUGH, N. J. 2004. Partial least squares analysis of neuroimaging data: applications and advances. *Neuroimage,* 23**,** S250-S263.

MEDINA, J., KIMBERG, D. Y., CHATTERJEE, A. & COSLETT, H. B. 2010. Inappropriate usage of the Brunner–Munzel test in recent voxel-based lesion-symptom mapping studies. *Neuropsychologia,* 48**,** 341-343.

MENCÍA, E. L., PARK, S.-H. & FÜRNKRANZ, J. 2010. Efficient voting prediction for pairwise multilabel classification. *Neurocomputing,* 73**,** 1164-1176.

MEYER-LINDENBERG, A. 1996. The evolution of complexity in human brain development: an EEG study. *Clinical Neurophysiology,* 99**,** 405-411.

MILLER, R. G. 1974. The jackknife-a review. *Biometrika,* 61**,** 1-15.

MIRNEZAMI, R., NICHOLSON, J. & DARZI, A. 2012. Preparing for precision medicine. *New England Journal of Medicine,* 366**,** 489-491.

MITCHELL, T. M. 2004. The role of unlabeled data in supervised learning. *Language, Knowledge, and Representation.* Springer.

MOURAO-MIRANDA, J., BOKDE, A. L., BORN, C., HAMPEL, H. & STETTER, M. 2005. Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. *NeuroImage,* 28**,** 980-995.

MUIR, K. W., WEIR, C. J., MURRAY, G. D., POVEY, C. & LEES, K. R. 1996. Comparison of neurological scales and scoring systems for acute stroke prognosis. *Stroke,* 27**,** 1817-1820.

NACHEV, P. 2015. The first step in modern lesion-deficit analysis. *Brain,* 138**,** e354.

NACHEV, P., COULTHARD, E., JAGER, H. R., KENNARD, C. & HUSAIN, M. 2008. Enantiomorphic normalisation of focally lesioned brains. *Neuroimage,* 39**,** 1215-26.

NANDY, R. R. & CORDES, D. 2003. Novel nonparametric approach to canonical correlation analysis with applications to low CNR functional MRI data. *Magnetic Resonance in Medicine,* 50**,** 354-365.

NASRABADI, N. M. 2007. Pattern recognition and machine learning. *Journal of electronic imaging,* 16**,** 049901.

NG, A. Y. & JORDAN, M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in neural information processing systems, 2002. 841-848.

NIGAM, K. & GHANI, R. Analyzing the effectiveness and applicability of co-training. Proceedings of the ninth international conference on Information and knowledge management, 2000. ACM, 86-93.

NITSCHE, M., LIEBETANZ, D., TERGAU, F. & PAULUS, W. 2002. Modulation of cortical excitability by transcranial direct current stimulation. *Der Nervenarzt,* 73**,** 332-335.

NITSCHE, M. A., COHEN, L. G., WASSERMANN, E. M., PRIORI, A., LANG, N., ANTAL, A., PAULUS, W., HUMMEL, F., BOGGIO, P. S. & FREGNI, F. 2008. Transcranial direct current stimulation: state of the art 2008. *Brain stimulation,* 1**,** 206-223.

NIYOGI, P. 2013. Manifold regularisation and semi-supervised learning: Some theoretical analyses. *The Journal of Machine Learning Research,* 14**,** 1229-1250.

NOBLE, W. S. 2006. What is a support vector machine? *Nature biotechnology,* 24**,** 1565.

NORMAN, K. A., POLYN, S. M., DETRE, G. J. & HAXBY, J. V. 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences,* 10**,** 424-430.

OGAWA, S., LEE, T.-M., KAY, A. R. & TANK, D. W. 1990a. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences,* 87**,** 9868-9872.

OGAWA, S. & LEE, T. M. 1990. Magnetic resonance imaging of blood vessels at high fields: in vivo and in vitro measurements and image simulation. *Magnetic resonance in medicine,* 16**,** 9-18.

OGAWA, S., LEE, T. M., NAYAK, A. S. & GLYNN, P. 1990b. Oxygenation‐sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic resonance in medicine,* 14**,** 68-78.

OGAWA, S., TANK, D. W., MENON, R., ELLERMANN, J. M., KIM, S. G., MERKLE, H. & UGURBIL, K. 1992. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences,* 89**,** 5951-5955.

PALMER, M. W. 1993. Putting things in even better order: the advantages of canonical correspondence analysis. *Ecology,* 74**,** 2215-2230.

PARDOE, H., PELL, G. S., ABBOTT, D. F., BERG, A. T. & JACKSON, G. D. 2008. Multi-site voxel-based morphometry: methods and a feasibility demonstration with childhood absence epilepsy. *Neuroimage,* 42**,** 611-616.

PARK, S., CHOI, D., KIM, M., CHA, W., KIM, C. & MOON, I.-C. 2017. Identifying prescription patterns with a topic model of diseases and medications. *Journal of biomedical informatics,* 75**,** 35-47.

PAUKER, S. G. & KASSIRER, J. P. 1975. Therapeutic decision making: a cost-benefit analysis. *New England Journal of Medicine,* 293**,** 229-234.

PEIKARI, M., SALAMA, S., NOFECH-MOZES, S. & MARTEL, A. L. 2018. A cluster-then-label semi-supervised learning approach for pathology image classification. *Scientific reports,* 8**,** 7193.

PELLEGRINO, J. W. & GLASER, R. 1979. *Components of inductive reasoning*, Learning Research and Development Center, University of Pittsburgh.

PENFIELD, W. & MILNER, B. 1958. Memory deficit produced by bilateral lesions in the hippocampal zone. *AMA Archives of Neurology & Psychiatry,* 79**,** 475-497.

PENNY, W. D., FRISTON, K. J., ASHBURNER, J. T., KIEBEL, S. J. & NICHOLS, T. E. 2011. *Statistical parametric mapping: the analysis of functional brain images*, Elsevier.

PEPER, J. S., BROUWER, R. M., BOOMSMA, D. I., KAHN, R. S., POL, H. & HILLEKE, E. 2007. Genetic influences on human brain structure: a review of brain imaging studies in twins. *Human brain mapping,* 28**,** 464-473.

PEREL, P., ROBERTS, I., SENA, E., WHEBLE, P., BRISCOE, C., SANDERCOCK, P., MACLEOD, M., MIGNINI, L. E., JAYARAM, P. & KHAN, K. S. 2007. Comparison of treatment effects between animal experiments and clinical trials: systematic review. *Bmj,* 334**,** 197.

PERMUTER, H., FRANCOS, J. & JERMYN, I. 2006. A study of Gaussian mixture models of colour and texture features for image classification and segmentation. *Pattern Recognition,* 39**,** 695-706.

PLATT, J. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers,* 10**,** 61-74.

POL, H. E. H., SCHNACK, H. G., POSTHUMA, D., MANDL, R. C., BAARÉ, W. F., VAN OEL, C., VAN HAREN, N. E., COLLINS, D. L., EVANS, A. C. & AMUNTS, K. 2006. Genetic contributions to human brain morphology and intelligence. *Journal of Neuroscience,* 26**,** 10235-10242.

POTTER, K., HAGEN, H., KERREN, A. & DANNENMANN, P. 2006. Methods for presenting statistical information: The box plot. *Visualization of large and unstructured data sets,* 4**,** 97-106.

POWER, J. D., COHEN, A. L., NELSON, S. M., WIG, G. S., BARNES, K. A., CHURCH, J. A., VOGEL, A. C., LAUMANN, T. O., MIEZIN, F. M. & SCHLAGGAR, B. L. 2011. Functional network organisation of the human brain. *Neuron,* 72**,** 665-678.

PRAKASH, V. J. & NITHYA, D. L. 2014. A survey on semi-supervised learning techniques. *arXiv preprint arXiv:1402.4645*.

PRICE, C., WARBURTON, E., MOORE, C., FRACKOWIAK, R. & FRISTON, K. 2001. Dynamic diaschisis: anatomically remote and context-sensitive human brain lesions. *Journal of Cognitive Neuroscience,* 13**,** 419-429.

PRIORI, A., HALLETT, M. & ROTHWELL, J. C. 2009. Repetitive transcranial magnetic stimulation or transcranial direct current stimulation? *Brain stimulation,* 2**,** 241-245.

PUSTINA, D., AVANTS, B., FASEYITAN, O. K., MEDAGLIA, J. D. & COSLETT, H. B. 2018. Improved accuracy of lesion to symptom mapping with multivariate sparse canonical correlations. *Neuropsychologia,* 115**,** 154-166.

QUINLAN, J. R. 1986. Induction of decision trees. *Machine learning,* 1**,** 81-106.

QUINLAN, J. R. 2014. *C4. 5: programs for machine learning,* Elsevier.

RABINER, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE,* 77**,** 257-286.

RAICHLE, M. E. 1998. Behind the scenes of functional brain imaging: a historical and physiological perspective. *Proceedings of the National Academy of Sciences,* 95**,** 765-772.

RAICHLE, M. E., MACLEOD, A. M., SNYDER, A. Z., POWERS, W. J., GUSNARD, D. A. & SHULMAN, G. L. 2001. A default mode of brain function. *Proceedings of the National Academy of Sciences,* 98**,** 676-682.

RAMAT, S., LEIGH, R. J., ZEE, D. S. & OPTICAN, L. M. 2006. What clinical disorders tell us about the neural control of saccadic eye movements. *Brain,* 130**,** 10-35.

READ, J., PFAHRINGER, B., HOLMES, G. & FRANK, E. Classifier chains for multi-label classification. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2009. Springer, 254-269.

READ, J., PFAHRINGER, B., HOLMES, G. & FRANK, E. 2011. Classifier chains for multi-label classification. *Machine learning,* 85**,** 333.

RIDDING, M. C. & ROTHWELL, J. C. 2007. Is there a future for therapeutic use of transcranial magnetic stimulation? *Nature Reviews Neuroscience,* 8**,** 559.

ROBERTSON, L., LAMB, M. & KNIGHT, R. 1988. Effects of lesions of temporal-parietal junction on perceptual and attentional processing in humans. *Journal of Neuroscience,* 8**,** 3757-3769.

ROBERTSON, L. C. & LAMB, M. R. 1991. Neuropsychological contributions to theories of part/whole organisation. *Cognitive psychology,* 23**,** 299-330.

ROHRER, M., BAUER, H., MINTOROVITCH, J., REQUARDT, M. & WEINMANN, H.-J. 2005. Comparison of magnetic properties of MRI contrast media solutions at different magnetic field strengths. *Investigative radiology,* 40**,** 715-724.

RORDEN, C. & BRETT, M. 2000. Stereotaxic display of brain lesions. *Behavioural neurology,* 12**,** 191-200.

RORDEN, C., FRIDRIKSSON, J. & KARNATH, H.-O. 2009. An evaluation of traditional and novel tools for lesion behavior mapping. *Neuroimage,* 44**,** 1355-1362.

RORDEN, C. & KARNATH, H.-O. 2004. Using human brain lesions to infer function: a relic from a past era in the fMRI age? *Nature Reviews Neuroscience,* 5**,** 812-819.

RORDEN, C., KARNATH, H.-O. & BONILHA, L. 2007. Improving lesion-symptom mapping. *Journal of cognitive neuroscience,* 19**,** 1081-1088.

ROSENBERG, C., HEBERT, M. & SCHNEIDERMAN, H. Semi-Supervised Self-Training of Object Detection Models. WACV/MOTION, 2005. 29-36.

RUFF, C. C., DRIVER, J. & BESTMANN, S. 2009. Combining TMS and fMRI: from ‘virtual lesions’ to functional-network accounts of cognition. *Cortex,* 45**,** 1043-1049.

SABBAGH, J. J., KINNEY, J. W. & CUMMINGS, J. L. 2013. Animal systems in the development of treatments for Alzheimer's disease: challenges, methods, and implications. *Neurobiology of aging,* 34**,** 169-183.

SACK, A. T. & LINDEN, D. E. 2003. Combining transcranial magnetic stimulation and functional imaging in cognitive brain research: possibilities and limitations. *Brain Research Reviews,* 43**,** 41-56.

SAUR, D., RONNEBERGER, O., KÜMMERER, D., MADER, I., WEILLER, C. & KLÖPPEL, S. 2010. Early functional magnetic resonance imaging activations predict language outcome after stroke. *Brain,* 133**,** 1252-1264.

SCHAPIRE, R. E. & SINGER, Y. 2000. BoosTexter: A boosting-based system for text categorisation. *Machine learning,* 39**,** 135-168.

SCHWARTZ, M. F., FASEYITAN, O., KIM, J. & COSLETT, H. B. 2012. The dorsal stream contribution to phonological retrieval in object naming. *Brain,* 135**,** 3799-3814.

SCOVILLE, W. B. & MILNER, B. 1957. Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery & Psychiatry,* 20**,** 11-21.

SEEGER, M. 2000. Learning with labeled and unlabeled data.

SINDHWANI, V. & KEERTHI, S. S. Large scale semi-supervised linear SVMs. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006. ACM, 477-484.

SINDHWANI, V. & KEERTHI, S. S. 2007. Newton methods for fast solution of semi-supervised linear SVMs. *Large scale kernel machines***,** 155-174.

SINDHWANI, V., NIYOGI, P. & BELKIN, M. Beyond the point cloud: from transductive to semi-supervised learning. Proceedings of the 22nd international conference on Machine learning, 2005. ACM, 824-831.

SINHA, A., HRIPCSAK, G. & MARKATOU, M. 2009. Large datasets in biomedicine: a discussion of salient analytic issues. *Journal of the American Medical Informatics Association,* 16**,** 759-767.

SMITH, D. V., CLITHERO, J. A., RORDEN, C. & KARNATH, H.-O. 2013. Decoding the anatomical network of spatial attention. *Proceedings of the National Academy of Sciences,* 110**,** 1518-1523.

SMITH, S. M. & NICHOLS, T. E. 2018. Statistical Challenges in "Big Data" Human Neuroimaging. *Neuron,* 97**,** 263-268.

SOROWER, M. S. 2010. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis,* 18.

SOSA, M. A. G., DE GASPERI, R., JANSSEN, P. L., YUK, F. J., ANAZODO, P. C., PRICOP, P. E., PAULINO, A. J., WICINSKI, B., SHAUGHNESS, M. C. & MAUDLIN-JERONIMO, E. 2014. Selective vulnerability of the cerebral vasculature to blast injury in a rat model of mild traumatic brain injury. *Acta neuropathologica communications,* 2**,** 67.

SOUZA, C. R. 2010. Kernel functions for machine learning applications. *Creative Commons Attribution-Noncommercial-Share Alike,* 3**,** 29.

SPERBER, C. & KARNATH, H. O. 2016. Impact of correction factors in human brain lesion-behavior inference. *Human Brain Mapping.*

SPORNS, O. 2011. The human connectome: a complex network. *Annals of the New York Academy of Sciences,* 1224**,** 109-125.

SPORNS, O. 2013. Structure and function of complex brain networks. *Dialogues Clin Neurosci,* 15**,** 247-262.

SPORNS, O., TONONI, G. & KÖTTER, R. 2005. The human connectome: a structural description of the human brain. *PLoS computational biology,* 1**,** e42.

STERNBERG, P., DE JUAN, E., MICHELS, R. G. & AUER, C. 1984. Multivariate analysis of prognostic factors in penetrating ocular injuries. *American journal of ophthalmology,* 98**,** 467-472.

SUYKENS, J. A. & VANDEWALLE, J. 1999. Least squares support vector machine classifiers. *Neural processing letters,* 9**,** 293-300.

SWETS, J. A. 1973. The relative operating characteristic in psychology: a technique for isolating effects of response bias finds wide use in the study of perception and cognition. *Science,* 182**,** 990-1000.

THOMPSON, P., CANNON, T. D. & TOGA, A. W. 2002. Mapping genetic influences on human brain structure. *Annals of medicine,* 34**,** 523-536.

THOMPSON, P. M., CANNON, T. D., NARR, K. L., VAN ERP, T., POUTANEN, V.-P., HUTTUNEN, M., LÖNNQVIST, J., STANDERTSKJÖLD-NORDENSTAM, C.-G., KAPRIO, J. & KHALEDY, M. 2001. Genetic influences on brain structure. *Nature neuroscience,* 4**,** 1253-1258.

THORNBURY, J. R., FRYBACK, D. G. & EDWARDS, W. 1975. Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information. *Radiology,* 114**,** 561-565.

TIMMANN, D., ERNST, T. M., ILG, W. & DONCHIN, O. 2016. Lesion-Symptom Mapping. *Essentials of Cerebellum and Cerebellar Disorders.* Springer.

TOGA, A. W. & THOMPSON, P. M. 2003. Mapping brain asymmetry. *Nature Reviews Neuroscience,* 4**,** 37.

TONG, S. & KOLLER, D. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research,* 2**,** 45-66.

TSOCHANTARIDIS, I., HOFMANN, T., JOACHIMS, T. & ALTUN, Y. Support vector machine learning for interdependent and structured output spaces. Proceedings of the twenty-first international conference on Machine learning, 2004. ACM, 104.

TSOUMAKAS, G. & KATAKIS, I. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM),* 3**,** 1-13.

TSOUMAKAS, G., KATAKIS, I. & VLAHAVAS, I. 2011. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering,* 23**,** 1079-1089.

TSOUMAKAS, G. & VLAHAVAS, I. Random k-labelsets: An ensemble method for multilabel classification. European conference on machine learning, 2007. Springer, 406-417.

TU, J. V. 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology,* 49**,** 1225-1231.

TURIAN, J., RATINOV, L. & BENGIO, Y. Word representations: a simple and general method for semi-supervised learning. Proceedings of the 48th annual meeting of the association for computational linguistics, 2010. Association for Computational Linguistics, 384-394.

UTGOFF, P. E. 1989. Incremental induction of decision trees. *Machine learning,* 4**,** 161-186.

VALLAR, G. J. N. 2001. Extrapersonal visual unilateral spatial neglect and its neuroanatomy. 14**,** S52-S58.

VAN DER WORP, H. B., HOWELLS, D. W., SENA, E. S., PORRITT, M. J., REWELL, S., O'COLLINS, V. & MACLEOD, M. R. 2010. Can animal models of disease reliably inform human studies? *PLoS medicine,* 7**,** e1000245.

VAN ENGELEN, J. E. & HOOS, H. H. 2020. A survey on semi-supervised learning. *Machine Learning,* 109**,** 373-440.

VAN ESSEN, D. C., UGURBIL, K., AUERBACH, E., BARCH, D., BEHRENS, T., BUCHOLZ, R., CHANG, A., CHEN, L., CORBETTA, M. & CURTISS, S. W. 2012. The Human Connectome Project: a data acquisition perspective. *Neuroimage,* 62**,** 2222-2231.

VAPNIK, V. 1998. *Statistical learning theory. 1998*, Wiley, New York.

VAPNIK, V. 2006. *Estimation of dependences based on empirical data,* Springer Science & Business Media.

VAPNIK, V. 2013. *The nature of statistical learning theory,* Springer science & business media.

VAPNIK, V. N. & CHERVONENKIS, A. J. 1974. Theory of pattern recognition.

VAPNIK, V. N. & CHERVONENKIS, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. Soviet Math. Dokl, 1968. 915-918.

VENS, C., STRUYF, J., SCHIETGAT, L., DŽEROSKI, S. & BLOCKEEL, H. 2008. Decision trees for hierarchical multi-label classification. *Machine learning,* 73**,** 185.

VERLEYSEN, M. & FRANÇOIS, D. The curse of dimensionality in data mining and time series prediction. International Work-Conference on Artificial Neural Networks, 2005. Springer, 758-770.

VINKEN, P. & BRUYN, G. 1969. Handbook of clinical neurology.

VUILLEUMIER, P., HESTER, D., ASSAL, G. & REGLI, F. J. N. 1996. Unilateral spatial neglect recovery after sequential strokes. 46**,** 184-189.

WAGNER, T., FREGNI, F., FECTEAU, S., GRODZINSKY, A., ZAHN, M. & PASCUAL-LEONE, A. 2007. Transcranial direct current stimulation: a computer-based human model study. *Neuroimage,* 35**,** 1113-1124.

WALSH, V. & COWEY, A. 2000. Transcranial magnetic stimulation and cognitive neuroscience. *Nature Reviews Neuroscience,* 1**,** 73.

WANG, J., FENG, J., SUN, X., CHEN, S.-S. & CHEN, B. Simplified constraints Rank-SVM for multi-label classification. Chinese Conference on Pattern Recognition, 2014. Springer, 229-236.

WANG, J. & SHEN, X. 2007. Large margin semi-supervised learning. *Journal of Machine Learning Research,* 8**,** 1867-1891.

WANG, L. H. & JOHNSON, E. M. 2008. Mixed lineage kinase inhibitor CEP-1347 fails to delay disability in early Parkinson disease. *Neurology,* 71**,** 462-463.

WANG, W. & KRISHNAN, E. 2014. Big data and clinicians: a review on the state of the science. *JMIR medical informatics,* 2.

WARACH, S., GAA, J., SIEWERT, B., WIELOPOLSKI, P. & EDELMAN, R. R. 1995. Acute human stroke studied by whole brain echo planar diffusion‐weighted magnetic resonance imaging. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society,* 37**,** 231-241.

WARDLAW, J., WARLOW, C. & COUNSELL, C. 1997. Systematic review of evidence on thrombolytic therapy for acute ischaemic stroke. *The Lancet,* 350**,** 607-614.

WASSERMANN, E. M. & LISANBY, S. H. 2001. Therapeutic application of repetitive transcranial magnetic stimulation: a review. *Clinical Neurophysiology,* 112**,** 1367-1377.

WEI, Y., XIA, W., HUANG, J., NI, B., DONG, J., ZHAO, Y. & YAN, S. 2014. CNN: single-label to multi-label. *arXiv preprint arXiv:1406.5726*.

WEILLER, C., WILLMES, K., REICHE, W., THRON, A., ISENSEE, C., BUELL, U. & RINGELSTEIN, E. B. 1993. The case of aphasia or neglect after striatocapsular infarction. *Brain,* 116**,** 1509-1525.

WERNICKE, C. 1874. *Der aphasische Symptomencomplex: eine psychologische Studie auf anatomischer Basis,* Cohn.

WHITNEY, A. W. 1971. A direct method of nonparametric measurement selection. *IEEE Transactions on Computers,* 100**,** 1100-1103.

WIESEN, D., SPERBER, C., YOURGANOV, G., RORDEN, C. & KARNATH, H.-O. 2019. Using machine learning-based lesion behavior mapping to identify anatomical networks of cognitive dysfunction: spatial neglect and attention. *bioRxiv***,** 556753.

WINTERS-HILT, S. & MERAT, S. SVM clustering. BMC bioinformatics, 2007. BioMed Central, S18.

WRIGHT, I. C., SHAM, P., MURRAY, R. M., WEINBERGER, D. R. & BULLMORE, E. T. 2002. Genetic contributions to regional variability in human brain structure: methods and preliminary results. *Neuroimage,* 17**,** 256-271.

WRIGHT, S. N., KOCHUNOV, P., MUT, F., BERGAMINO, M., BROWN, K. M., MAZZIOTTA, J. C., TOGA, A. W., CEBRAL, J. R. & ASCOLI, G. A. 2013. Digital reconstruction and morphometric analysis of human brain arterial vasculature from magnetic resonance angiography. *Neuroimage,* 82**,** 170-181.

WU, F., HAN, Y., TIAN, Q. & ZHUANG, Y. Multi-label boosting for image annotation by structural grouping sparsity. Proceedings of the 18th ACM international conference on Multimedia, 2010. ACM, 15-24.

XU, J. 2011. An extended one-versus-rest support vector machine for multi-label classification. *Neurocomputing,* 74**,** 3114-3124.

XU, S.-Y. & PAN, S.-Y. 2013. The failure of animal models of neuroprotection in acute ischemic stroke to translate to clinical efficacy. *Medical science monitor basic research,* 19**,** 37.

XU, T., JHA, A. & NACHEV, P. 2017a. The dimensionalities of lesion-deficit mapping. *Neuropsychologia*.

XU, T., ROLF JÄGER, H., HUSAIN, M., REES, G. & NACHEV, P. 2017b. High-dimensional therapeutic inference in the focally damaged human brain. *Brain*.

XU, Z., JIN, R., ZHU, J., KING, I. & LYU, M. Efficient convex relaxation for transductive support vector machine. Advances in neural information processing systems, 2008. 1641-1648.

YAGI, T. 2013. Genetic basis of neuronal individuality in the mammalian brain. *Journal of neurogenetics,* 27**,** 97-105.

YAMAMOTO, K., KOH, H., SHIMADA, H., TAKEUCHI, J., YAMAKAWA, Y., KAWAMURA, M. & MIKI, T. 2014. Cerebral infarction in the left hemisphere compared with the right hemisphere increases the risk of aspiration pneumonia. *Osaka city medical journal,* 60**,** 81-86.

YAROWSKY, D. Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd annual meeting on Association for Computational Linguistics, 1995. Association for Computational Linguistics, 189-196.

YOSHIMURA, Y., DANTZKER, J. L. & CALLAWAY, E. M. 2005. Excitatory cortical neurons form fine-scale functional networks. *Nature,* 433**,** 868.

YOURGANOV, G., FRIDRIKSSON, J., RORDEN, C., GLEICHGERRCHT, E. & BONILHA, L. 2016. Multivariate connectome-based symptom mapping in post-stroke patients: networks supporting language and speech. *Journal of Neuroscience,* 36**,** 6668-6679.

ZHANG, M.-L. 2009. M l-rbf: RBF Neural Networks for Multi-Label Learning. *Neural Processing Letters,* 29**,** 61-74.

ZHANG, M.-L., PEÑA, J. M. & ROBLES, V. 2009. Feature selection for multi-label naive Bayes classification. *Information Sciences,* 179**,** 3218-3229.

ZHANG, M.-L. & ZHANG, K. Multi-label learning by exploiting label dependency. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010. ACM, 999-1008.

ZHANG, M.-L. & ZHOU, Z.-H. A k-nearest neighbor based algorithm for multi-label classification. Granular Computing, 2005 IEEE International Conference on, 2005. IEEE, 718-721.

ZHANG, M.-L. & ZHOU, Z.-H. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern recognition,* 40**,** 2038-2048.

ZHANG, M.-L. & ZHOU, Z.-H. 2014. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering,* 26**,** 1819-1837.

ZHANG, P. J. T. A. O. S. 1993. Model selection via multifold cross validation. 299-313.

ZHANG, Y., KIMBERG, D. Y., COSLETT, H. B., SCHWARTZ, M. F. & WANG, Z. 2014. Multivariate lesion-symptom mapping using support vector regression. *Human brain mapping,* 35**,** 5861-5876.

ZHAO, L., BIESBROEK, J. M., SHI, L., LIU, W., KUIJF, H. J., CHU, W. W., ABRIGO, J. M., LEE, R. K., LEUNG, T. W. & LAU, A. Y. 2017. Strategic infarct location for post-stroke cognitive impairment: A multivariate lesion-symptom mapping study. *Journal of Cerebral Blood Flow & Metabolism***,** 0271678X17728162.

ZHU, X. 2006. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison,* 2**,** 4.

ZHU, X., GHAHRAMANI, Z. & LAFFERTY, J. D. Semi-supervised learning using gaussian fields and harmonic functions. Proceedings of the 20th International conference on Machine learning (ICML-03), 2003. 912-919.

ZHU, X. & GOLDBERG, A. B. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning,* 3**,** 1-130.