Neurobehavioural Patterns of Alcohol Abuse in Adolescence

vorgelegt von Master of Science Caroline Matthis

von der Fakultät IV – Elektrotechnik und Informatik der Technischen Universität Berlin



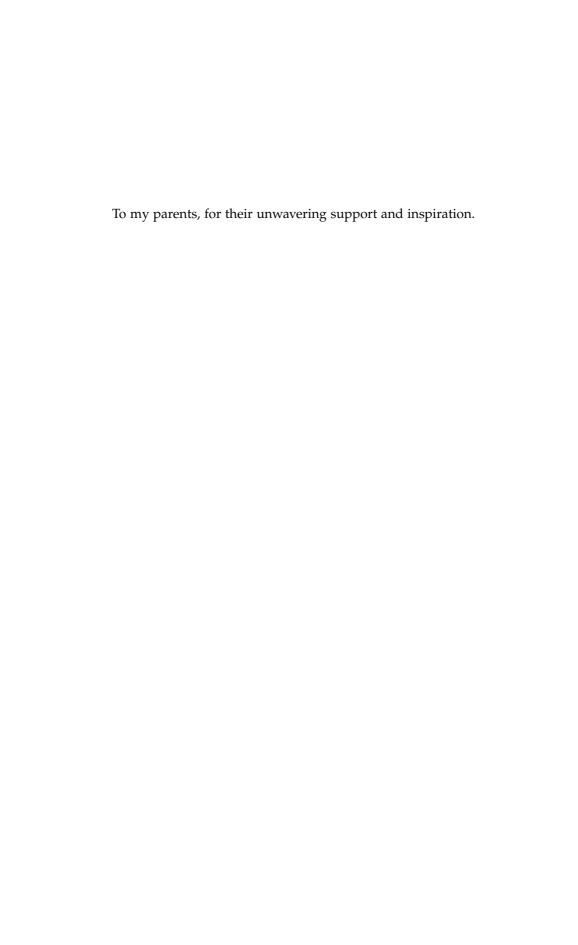
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
- Dr. rer. nat. genehmigte Dissertation

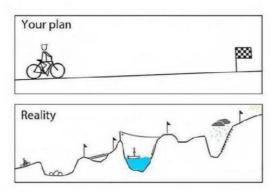
Promotionsausschuss:

Vorsitzende: Prof. Dr. Marianne Maertens Gutachter: Prof. Dr. Klaus Obermayer Gutachter: Prof. Dr. Dr. Andreas Heinz Gutachter: Dr. Matthias Guggenmos

Tag der wissenschaftlichen Aussprache: 11. Juni, 2019

Berlin 2019





@jpmillsphd

ACKNOWLEDGEMENTS

This doctoral thesis is the fruit of three and a half years of work under the supervision of Prof. Dr. Klaus Obermayer. None of it would have been possible without the support of a whole bunch of people. First and foremost I would like to thank Klaus for giving me the opportunity to write my thesis in his group and for suggesting such an exciting and interdisciplinary topic to me. His relentless critical advice and guidance have significantly formed my work and general approach to science.

Further, I am grateful to Prof. Dr. Dr. Andreas Heinz for his critical input on manuscripts, for his inspiring enthusiasm for psychiatric research and for agreeing to examine my doctoral thesis. Towards Dr. Matthias Guggenmos I am not only thankful for agreeing to be part of my examination committee, but also for his guidance, advice and encouragement, especially during the first part of my PhD.

I would moreover like to thank Dr. Sambu Seo for many reasons. We did not only share an office, but a scientific and personal journey throughout the last years. Her critical advice and input have very much shaped the outcome of my work. Thanks also to Alexander Genauck and Dr. Anne Beck for countless fruitful discussions and for their helpful guidance in all clinical questions. Further, I would like to thank both of them for the great collaboration in the drafting of manuscripts and all other work related to the eMed:SysMedAlc project.

I thank the BCCN Berlin and the GRK for financial support and for offering a fantastic environment for doing science. Special thanks go to Dr. Robert Martin and Margret Franke for their coordination magic and for the organisation of so many exciting talks, symposia and conferences. I further thank Camilla Groiß who made every bureaucratic hurdle much more manageable. My admiration and deep gratitude goes to Alexandra Elbakyan, the founder of Sci-Hub and a pioneer of open science.

Last, but not in any way least, I would like to thank my colleagues from the NI-group. Spending sheer infinite numbers of hours in the office would not have been bearable without all of you. Special shout-outs go to Veronika (for (non)-scientific discussions over coffee that were never limited to a 20 minute time window), to Ivo (for numerous long nights of science with Indian food, discussions over thesis formatting and more), to Franzi (for all the fun we had making hats, organising parties and complaining about life) and to Christian (for being an inspiration in efficiency).

ABSTRACT

Excessive alcohol consumption has a detrimental effect on public health. Alcohol abuse is a top-ranked disorder of the brain with respect to total costs to economy and is linked to an estimated 3.8 % of global deaths. Often, first experiences with alcohol are made during adolescence, the time of transition between childhood and adulthood. Adolescence marks a period of complex social, biological and psychological development; the interplay between alcohol consumption and these manifold developments are not yet fully understood. In my thesis, I use the IMAGEN database, a large-scale, longitudinal study of around 1000 healthy adolescents, to shed light on the neurobehavioural patterns of heavy drinking in adolescence.

In the first part of my thesis, I introduce a framework for the systematic analysis of predictive and explanatory quality of neurobehavioural features. I adapt a common cross-validation scheme to assess the predictive power in the presence of confounders. Moreover, I develop a novel method to estimate the influence of a single (neurobehavioural) feature on a binary outcome.

In the second part of my thesis, I build a comprehensive phenotype describing alcohol drinking behaviour for the subjects from the IMAGEN database. I show that the constructed phenotype outperforms all original drinking behaviour variables in terms of consistency.

In the third part, I apply the established methods on neurobehavioural features from the IMAGEN study to differentiate between drinking behaviour phenotypes. First, I focus on grey matter volume and psychosocial features. I show that heavy drinking in adolescence is associated with reduced grey matter volume across various cortical and subcortical structures, especially in females. Moreover, I observe that impulsivity and facets of novelty seeking are associated (also longitudinally) to heavy drinking. Then, I focus on functional imaging and cognitive features. I show that adolescent drinking is associated with various markers of impaired response inhibition, both neural and behavioural.

Overall, I shed light on both risk profiles and potential consequences of heavy drinking in adolescence. On the one hand, I contribute to the understanding of the detrimental effects of alcohol on the adolescent brain, calling for targeted intervention programs. On the other hand, I show that risk profiles based on personality traits may offer the potential for prevention procedures before the treacherous spiral of addiction begins.

7USAMMENEASSUNG

Übermäßiger Alkoholkonsum wirkt sich nachteilig auf die öffentliche Gesundheit aus. Alkoholmissbrauch ist eine der häufigsten Erkrankungen des Gehirns in Bezug auf die Gesamtkosten für die Wirtschaft und steht in Zusammenhang mit schätzungsweise 3,8 % der weltweiten Todesfälle. Häufig werden erste Erfahrungen mit Alkohol in der Adoleszenz gemacht, der Zeit des Übergangs zwischen Kindheit und Erwachsensein. Die Adoleszenz markiert eine Zeit komplexer sozialer, biologischer und psychologischer Entwicklungen; das Zusammenspiel zwischen Alkoholkonsum und diesen vielfältigen Entwicklungen ist noch nicht vollständig verstanden. In meiner Dissertation nutze ich die IMAGEN-Datenbank, eine groß angelegte Längsschnittstudie mit rund 1000 gesunden Jugendlichen, um die neurobehavioralen Muster des starken Alkoholkonsums in der Pubertät zu verstehen.

Im ersten Teil meiner Arbeit stelle ich eine Struktur für die systematische Analyse der prädiktiven und erklärenden Qualität von neurobehavioralen Merkmalen vor. Ich adaptiere ein gebräuchliches Kreuzvalidierungsschema, um die Vorhersagekraft in Gegenwart von Störfaktoren zu bewerten. Darüber hinaus entwickle ich eine neuartige Methode, um den Einfluss eines einzelnen (neurobehavioralen) Merkmals auf ein binäre abhängige Variable zu schätzen.

Im zweiten Teil meiner Arbeit erstelle ich einen Phänotyp, der das Alkoholkonsumverhalten der Probanden aus der IMAGEN-Datenbank beschreibt. Ich zeige, dass der konstruierte Phänotyp alle ursprünglichen Variablen des Trinkverhaltens in Bezug auf die Eindeutigkeit übertrifft.

Im dritten Teil wende ich die etablierten Methoden auf neurobehavioralen Daten aus der IMAGEN-Studie an, um zwischen den Phänotypgruppen zu unterscheiden. Zunächst konzentriere ich mich auf das Volumen der grauen Substanz und psychosoziale Eigenschaften. Ich zeige, dass starkes Trinken in der Adoleszenz mit einem reduzierten Volumen an grauer Substanz über verschiedene kortikale und subkortikale Strukturen hinweg assoziiert ist, insbesondere bei Frauen. Darüber hinaus beobachte ich, dass Impulsivität und weitere Persönlichkeitsmerkmale (auch in longitudinaler Weise) mit starkem Alkoholkonsum zusammenhängen. Dann konzentriere ich mich auf die funktionelle Bildgebung und kognitive Merkmale. Ich zeige, dass jugendliches Trinken mit verschiedenen Markern einer

beeinträchtigten Reaktionshemmung verbunden ist, sowohl auf neuronaler Ebene als auch im Verhalten.

Insgesamt beleuchte ich sowohl die Risikoprofile als auch die möglichen Folgen eines starken Alkoholkonsums in der Adoleszenz. Auf der einen Seite trage ich zum Verständnis der schädlichen Auswirkungen von Alkohol auf das jugendliche Gehirn bei und fordere gezielte Interventionsprogramme. Auf der anderen Seite zeige ich, dass Risikoprofile, die auf Persönlichkeitsmerkmalen basieren, das Potenzial für frühe Prävention bieten können.

This thesis was funded by GRK 1589/2.



CONTENTS

1	INT	RODUCI	10 N	1
	1.1	Alcoh	nol as a public health problem	1
	1.2	The c	ritical period of adolescence	2
	1.3	Aim a	and scope of thesis	3
	1.4	Struct	ture of thesis	3
I	Fo	undat	ions	5
2	MAC	HINE L	EARNING AND RELATED METHODS	7
	2.1		duction	7
	2.2		ocessing	9
		2.2.1	Outlier detection and treatment	9
		2.2.2	Imputation	10
	2.3	Predic	ction analysis	12
		2.3.1	General setup	12
		2.3.2	Logistic regression	13
		2.3.3	Naïve Bayes	14
		2.3.4	Random forests	15
		2.3.5	Support vector machines	16
		2.3.6	Learning vector quantisation	19
	2.4	Perfo	rmance assessment	20
		2.4.1	Specificity, Sensitivity and Balanced Accuracy	21
		2.4.2	Cross-validation	22
		2.4.3	Cross-validation accounting for confounders	24
		2.4.4	Significance of accuracies	27
	2.5	Assoc	ciation analysis	29
		2.5.1	Logistic regression for association analysis	29
		2.5.2	Repeated logistic lasso (ReLL) for association analysis	30
		2.5.3	Significance testing of ReLL	34
	2.6	Multi	ple testing correction	35
	2.7	Chap	ter summary	37

II	Th	e IMAGEN data base	36
3	THE	IMAGEN PROJECT	41
	3.1	Introduction	41
	3.2	Cross-sectional and longitudinal	42
	3.3	Our contribution	42
	3.4	Chapter summary	44
4	DEF	INING ALCOHOL DRINKING PHENOTYPES	45
	4.1	Introduction	45
	4.2	Available questionnaires	46
		4.2.1 ESPAD	46
		4.2.2 TLFB	47
		4.2.3 MAST	47
		4.2.4 AUDIT	48
	4.3	Construction of phenotype	49
	Τ.)	4.3.1 Grouping of drinking behaviour variables	49
		4.3.2 Establishing a comprehensive phenotype	50
		4.3.3 Dropping the moderate drinkers	52
		4.3.4 Validation of phenotype	53
		4.3.5 Confounding variables	53
	4.4	Definition of cumulative drinking score	52 54
	4.4	Chapter summary	
	4.5	Chapter summary	55
Ш	Pa	tterns of alcohol abuse in adolescence	59
5	GRE	Y MATTER VOLUME AND PSYCHOSOCIAL VARIABLES	61
	5.1	Introduction	61
	5.2	Data	63
		5.2.1 Structural brain imaging and preprocessing	6:
		5.2.2 Personality measures	63
		5.2.3 Life events questionnaire	62
	5.3	Methods	64
	<i>J</i> - <i>J</i>	5.3.1 Extraction of structural imaging features	64
		5.3.2 Extraction of personality features	6
		5.3.3 Extraction of life events features	65
		5.3.4 Summary of features	65
		5.3.5 Prediction analysis	
			70
		5.3.6 Association analysis	71

5.4	Results
	5.4.1 Prediction analysis
	5.4.2 Association analysis
	5.4.3 Further investigations 80
5.5	Discussion
5.6	Chapter summary
6 FU	NCTIONAL IMAGING AND COGNITIVE VARIABLES 87
6.1	Introduction
6.2	Data
	6.2.1 Wechsler intelligence scale 89
	6.2.2 Affective go-no/go task 89
	6.2.3 Stop signal task
	6.2.4 Preprocessing and first level analyses of functional
	brain images
6.3	Methods
	6.3.1 Extraction of cognitive features
	6.3.2 Extraction of functional imaging features 98
	6.3.3 Summary of features
	6.3.4 Prediction analysis
	6.3.5 Association analysis
6.4	Results
	6.4.1 Prediction analysis
	6.4.2 Association analysis
	6.4.3 Further investigations
6.5	Discussion
6.6	Chapter summary
N/ C	447
IV S	ynthesis 117
,	MMARY AND OUTLOOK 119
7.1	,
7.2	Outlook

ΑĮ	peno	dices		121
Α	SUP	PLEME	NTS FOR CHAPTER 5	123
	A.1	Site ef	ffect	123
	A.2	Histor	grams of features	124
		A.2.1		
		A.2.2	Life events features	
		A.2.3	Grey matter volume features	
	A.3		ts for prediction analysis	
	A.4		ts for association analysis	
	•		Cross-sectional setting	
		A.4.2		
		•		
В	SUP		NTS FOR CHAPTER 6	135
	B.1	Extrac	ction and modelling of SSRT	
		B.1.1	Sample inhibition functions	135
		B.1.2	Comparison of SSRT extraction methods	136
		B.1.3	Time-resolved SSRT	137
	B.2	Prepre	ocessing of functional imaging data	138
		B.2.1	Custom EPI template	138
		B.2.2	Quality control of imaging data	138
	в.3	Histog	grams of features	140
		B.3.1	Cognitive features	
		B.3.2	Functional features	
	B.4	Imput	tation investigation	146
	в.5		tigation on complete data	
	в.6		igation of other functional feature extraction methods.	
	В.7	Invest	igation of gender differences	150
	в.8		ts for association analysis	
		в.8.1	Cross-sectional setting	
		в.8.2	Longitudinal setting	
С	PUB	LICATIO	DNS	155

ABBREVIATIONS

AGN affective go/no-go task

AUD alcohol use disorder

balAcc balanced accuracy

BL baseline

BSSCV bootstrapped-stratified-cross-validation

CANTAB Cambridge Neuropsychological Test Automated Battery

CV cross-validation

DSM Diagnostic and Statistical Manual of Mental Disorders

FDR false discovery rate

fMRI functional magnetic resonance imaging

FU1 follow-up 1 FU2 follow-up 2

GNB Gaussian naïve Bayes

IMAGEN European research project collecting large-scale data

k-NN k-nearest neighbours

lasso least absolut shrinkage and selection operator

LVQ learning vector quantisation

ReLL repeated logistic lasso

ROI region of interest

RSLVQ robust soft learning vector quantisation

SPM statistical parametric mapping

SSD stop signal delay

SSRT stop signal reaction time

SST stop signal task

(l)-SVM (linear) support vector machine

1 INTRODUCTION

1.1 ALCOHOL AS A PUBLIC HEALTH PROBLEM

Humans began brewing beer as early as 13,000 years ago, as recent archaeological findings from the Raqefet Cave in Israel suggest (Liu et al. 2018). Ever since, alcohol has played a key role in human culture. We know from experience how having a drink or two feels relaxing and pleasurable. We may also know that having another drink (or two, or three) might make us feel unwell. In fact, already the first of those drinks may be harmful.

Social drinking is a slippery slope. The most recent version of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) merged the previously separate conditions of alcohol abuse and alcohol dependence into one holistic term – alcohol use disorders (AUDs). This emphasises the seamless transition between abuse and dependence.

Excessive alcohol consumption has a detrimental effect on public health; globally it is linked to an estimated 3.8 % of deaths and 4.6 % of disability-adjusted life years (Rehm et al. 2009). Furthermore, alcohol abuse is a topranked disorder of the brain with respect to total costs to economy (Effertz and Mann 2013).

The harmful effects of alcohol abuse are manifold. On the one hand, it is well established that acute intoxication leads to adverse social consequences, including drunk-driving and aggression towards self and others. For instance, the risk of being involved in a fatal car accident scales exponentially with the blood alcohol content (Naranjo and Bremner 1993). On the other hand, repeated alcohol intoxication may lead to the development of a chronic AUD. Many studies have shown the relationship of AUDs with widespread brain atrophy and functional irregularities (Bühler and Mann 2011; Chelune and Parker 1981; Heinz 2002; Sullivan et al. 2003). These effects are partly reversible, but only after detoxification and a prolonged period of abstinence (Carlen et al. 1978; Carlen et al. 1984; Trabert et al. 1995). Abstinence is rare, given relapse rates of up to 80 % within the first six months after detoxification (Boothby and Doering 2005). Hence, prospects are grim once the spiral of addiction begins.

1.2 THE CRITICAL PERIOD OF ADOLESCENCE

Adolescence, the transition between childhood and adulthood, marks a period of complex social, biological and psychological development. On a neural level, adolescence is associated with a reduction in grey matter volume, an effect presumably resulting from both synaptic pruning and myelination (Dosenbach et al. 2010). At the same time, the white matter volume is steadily increasing due to the myelin proliferation (Pfefferbaum et al. 1994). Synaptic pruning and myelination are thought to contribute to a more efficient communication between frontal and subcortical brain regions (Luna and Sweeney 2004). On a behavioural level, this increased efficiency may be the neural basis responsible for the widely observed improvement in inhibitory control during adolescence (Prencipe et al. 2011; Zelazo and Carlson 2012). On a personality level, continuous adolescent development occurs mainly on subscales related to overall maturity (Klimstra et al. 2009; Roberts et al. 2001).

First experiences with alcohol often occur during adolescence and may interfere with all three adolescent development levels – neural, behavioural and personality. On a neural level, it has been demonstrated that alcohol drinking during mid-to-late adolescence is related to variations in neurodevelopment across brain tissue classes (Luciana et al. 2013). On a behavioural level, impaired inhibitory control in adolescence has been associated with excessive alcohol consumption (López-Caneda et al. 2014). On a personality level, binge drinking has repeatedly been linked to various traits, such as decreased agreeableness (Kuntsche et al. 2006) and conscientiousness (Stewart and Devine 2000).

However, the association between various neurobehavioural aspects and alcohol consumption might not (only) be a one-directional interference, i.e. a negative effect of drinking. In fact, the association may also stem from an underlying propensity towards an increased alcohol consumption. Disentangling cause and effect is not easy and might be further complicated by an additional effect of experiences such as stressful life events on the drinking trajectory.

Besides having unfavourable interactions with various aspects of development, heavy drinking during adolescence may further pose a risk factor developing AUDs later in life (Grant et al. 2006). Understanding the effects and neurobehavioural profiles of excessive alcohol consumption are hence essential for the potential design of targeted intervention and prevention programs.

1.3 AIM AND SCOPE OF THESIS

In this thesis, I aim at identifying neurobehavioural patterns of alcohol abuse in adolescence. I use data from a large-scale, longitudinal study (IM-AGEN) to shed light on potential causes and consequences of early alcohol consumption. Based on structural and functional brain imaging, personality, life events and cognitive data and taking advantage of IMAGEN's longitudinal study design, I aim to find both pre-disposing markers and presumed consequences of heavy drinking in adolescence.

Note that this thesis is closely related to the project "e:Med Alcohol Addiction – A Systems Oriented Approach" (specifically to Subproject 6 – Mathematical Modeling I: Convergent Data Analysis and Statistics, see Spanagel et al. (2013)), which was financed through the German e:Med fund from the Federal Ministry of Education and Research (BMBF).

STRUCTURE OF THESIS 1.4

Part I: Foundations

In Chapter 2, I provide the reader with an overview and brief explanation of the machine learning methods applied in the subsequent parts. I motivate the two-stage procedure of my analysis and then take a closer look into both these stages. For the first stage, prediction analysis, I provide details of all the well-established machine learning classifiers used in Part III. I further introduce a novel method to estimate classification accuracies in a setting with confounders, as will be very relevant throughout the thesis. For the second stage, association analysis, I briefly introduce classical methods before proposing and briefly evaluating another novel method for assessing single-feature effects on labels in the presence of confounders. Moreover, I touch upon preprocessing methods and multiple testing correction.

Part II: The IMAGEN database

In Chapter 3, I give an overview of the general aims of the IMAGEN project and explain the prospective nature of the data acquisition. I further provide wordings related to the study design that will be used excessively throughout the thesis, including the definition of a "cross-sectional" and a "longitudinal" setting. Importantly, I also state our contributions for the acquisition, preprocessing and all further analyses.

In Chapter 4, I derive the drinking behaviour labels "light" and "heavy" for the adolescent subjects from the IMAGEN database. These labels are of central importance for all analyses in the subsequent material.

Part III: Patterns of alcohol abuse in adolescence

In this part, I apply the methods from Part I on the IMAGEN database, using the previously derived labels and following the two-stage procedure introduced in Chapter 2.

In Chapter 5, I concentrate on grey matter volume and psychosocial features for the prediction of drinking behaviour. Psychosocial features include various personality traits and stressful life events.

In Chapter 6, I focus on functional activation and cognitive features. Both sets of features are specifically trimmed to capture response inhibition.

Part IV: Synthesis

In Chapter 7, I draw overall conclusions and provide ideas for future work.

Part I Foundations

2 | MACHINE LEARNING AND RELATED METHODS

This chapter introduces the (old and new) machine learning and statistical methods that will be used in the rest of the thesis. Sections describing newly developed methods are marked by a star. Some aspects have previously been published (Kassraian-Fard et al. 2016; Seo et al. 2018). For author contributions, see Appendix \mathbb{C} .

2.1 INTRODUCTION

The analysis of neurobehavioural data with the goal of differentiating between distinct subject groups is popular. At the outset, individuals in a typical study are grouped (or deliberately recruited for participation) according to some criterion of interest. Common groupings are, for instance, patients versus controls, substance abusers versus non-abusing relatives or – and this will be our criterion of interest – light versus heavy drinking adolescents. In both subject groups, a set of neurobehavioural features is measured. The goal of the subsequent analysis is to detect differences in these features between the groups in order to shed light on their underlying characteristics. For instance, which neurobehavioural state characterises a relapser from an abstainer after treatment for alcohol dependency? Or, as in our setting, what marks a light versus heavy drinking adolescent?

Classical approaches involve testing for group differences for each measured feature. Slight extensions consider also multivariate profiles, i.e. the difference in interplay of several features. Importantly, classical approaches generally do not evaluate the generalisation strength, i.e. the validity of the discovered group differences on holdout data. As Shmueli (2009) and Yarkoni and Westfall (2017) nicely frame: the focus of traditional approaches is explanation, not prediction.

Fueled by the replication crisis (Ioannidis 2005; Open Science Collaboration 2015) and the trend towards gathering large-scale data sets (Jordan and Mitchell 2015), machine learning methods applied to problems in clinical psychiatry and psychology have gained popularity rapidly throughout the past years (see the reviews Dwyer et al. (2018) and Iniesta et al. (2016)).

Machine learning algorithms are statistical methods that learn to perform a specific task without relying on explicit instructions. They naturally deal with multivariate data and thereby allow to model intricate patterns of features. Moreover, a fundamental property of machine learning algorithms is the use of an out-of-sample error measure, i.e. the performance is evaluated on a different data set than the one used for the estimation of model parameters. In theory, a focus on such error measures yields transferable and, given data drawn from the same population, reproducible results (Kleinberg 1996). This is a highly desirable quality, especially in the framework of diagnosis or prediction of treatment outcome.

Machine learning methods have radically changed perspectives not only in clinical psychiatry and psychology but in the entire medical community (Darcy et al. 2016; Obermeyer and Emanuel 2016), early predictions of which can be found in Kononenko (2001). This revolution has not only led to positive resonance. A regularly raised concern is the apparent "black box" nature of machine learning (Cabitza et al. 2017; Chen and Asch 2017; Maddox et al. 2019), referring to the lack of straightforward model interpretations. Chen and Asch (2017) state that "an accurate prediction of a patient outcome does not tell us what to do if we want to change that outcome"; Cabitza et al. (2017) emphasise "the need to open the machine learning black box".

In this thesis, we hence propose an analysis pipeline that is a hybrid between prediction and explanation. We suggest and employ a two-stage procedure. First, we use machine learning methods and out-of-sample error measures to determine the predictive value of a set of neurobehavioural features for the discrimination of subject groups. If (and only if) the features contain significant predictive power, we proceed with the second stage, the post-hoc evaluation of individual feature differences. We thereby adhere to Yarkoni and Westfall (2017) who argue that "in a great many cases, research programs that emphasize prediction, and that treat explanation as a secondary goal, would be more fruitful", and refer to the first stage as prediction analysis and to the second stage as association analysis.

This chapter introduces the necessary foundations for this two-stage procedure. For preprocessing and prediction analysis, we rely on established methods and only briefly recapitulate the existing theory. For the evaluation of significance of prediction analysis in the presence of confounders, we develop "bootstrapped-stratified-cross-validation" (BSSCV). For association analysis, after reviewing classical approaches, we design a tailored adaption of logistic regression, namely the "repeated logistic lasso" (ReLL).

A hurried and informed reader is advised to skip the well-known content and head directly to the more novel parts, indicated by stars next to the section titles.

PREPROCESSING 2.2

Preprocessing of data is a crucial step to ensure the precision and accuracy of the subsequent modelling. Here, we provide only a short glimpse into the broad field of preprocessing, focusing on methods used in our analysis - namely outlier detection and imputation.

2.2.1 Outlier detection and treatment

Outliers in behavioural or neuroimaging variables can be present for various reasons, for instance misunderstood instructions by the tested subject, or measurement errors. Many methods exist to (i) detect and (ii) then treat them. Note that we perform outlier detection and treatment in an unsupervised fashion, i.e. only using the features $x \in \mathbb{R}^d$ and not the labels $y \in \{0, 1\}.$

The most basic approach to outlier detection (and one which, among others, is also used in this thesis) is to fully rely on hypothesisdriven thresholds. For instance, say an experiment involves reacting to stimuli and a subject turns out to have a response rate of 0, i.e. the subject never responds. Common sense would call this measurement an outlier, potentially caused by inattention of the subject or misunderstood instructions. Given data $(x_i, y_i)_{i=1,...,N}$, $x_i \in \mathbb{R}^d$, $y_i \in \{0,1\}$, this approach amounts to setting thresholds, c_i^{lower} and c_i^{upper} , for each feature j, and labeling each observation outside of these thresholds, i.e. if:

$$x_{ij} \notin [c_j^{lower}, c_j^{upper}],$$

then x_{ij} is an outlier.

A more data-driven approach is to consider the distribution of features x_i . One possibility is to use the z-score of a single feature to determine how many standard deviations a data point is from the sample's mean (assuming a Gaussian distribution). Common values for the number of standard deviations an observation is allowed to be away from the centre are values larger than 2.5. For instance, in Figure 2.1, the red triangle is

far away from the centre of feature x_2 and hence, the method would mark the observation's x_2 feature as an outlier. The method is univariate, since it does not take the distribution of x_1 into account when screening for outliers in x_2 .

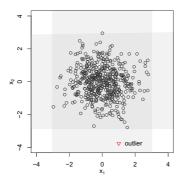


Figure 2.1: Simulated data of dimension d = 2 drawn from a multivariate normal distribution. The range of values that fall within 3 sample standard deviations for feature x_1 and x_2 . The red triangle would count as an outlier in x_2 but not x_1 .

The simplest reaction after the detection of an observation x_i with an outlying feature x_{ij} is to discard the entire observation. However, this can result in the removal of observations that only have a single outlying feature whereas all other features may still carry important (and accurate) information. We hence only remove the feature value x_{ij} and treat this feature entry as missing.

2.2.2 Imputation

To deal with the ubiquitous case of missing data, we make use of common imputation methods. Given data $(x_i, y_i)_{i=1,...,N}$, $x_i \in \mathbb{R}^d$ and $y_i \in \{0,1\}$ we consider the case of missing feature entries x_{ij} for a data sample i. If a label y_i is missing, the data sample is not considered for prediction and association analysis.

The most straightforward approach to missing values in data is to consider only the *complete cases*, i.e. to discard all samples (x_i, y_i) with at least one missing feature entry x_{ij} . However, if there are many missing values (and they are distributed over many subjects) the resulting data set may become too small for reasonable analyses. Furthermore, the acquisition

of data is often troublesome and costly. Discarding parts of the data may hence be a real pity.

A slightly more data-driven approach is univariate imputation. Given a sample index i* and feature index j*, and the corresponding missing feature $x_{i^*i^*}$, this approach makes use of the non-missing entries $\{x_{ij^*} \in \mathbb{R} | i = 1\}$ $1, \dots, N, i \neq i^*$ } to estimate a realistic value for $x_{i^*i^*}$. Common estimation approaches (all of which we use in this thesis) include (Gelman and Hill 2006):

- Mean imputation: replacing the missing feature by the sample mean of non-missing feature observations
- Median imputation: replacing the missing feature by the sample median of non-missing feature observations
- Sample imputation: replacing the missing feature by drawing a sample from the empirical distribution of non-missing feature observations

Multivariate imputation methods make use of potential interplay between features. Common approaches include (Gelman and Hill 2006):

- *kNN-imputation*: replacing the missing feature by the sample mean of the k-nearest neighbours according to some distance measure based on the scale of the features, e.g. Euclidean, see Zhang (2012) for further details
- Deterministic linear regression: replacing the missing feature by an estimated regression model using two steps (Gelman and Hill 2006):
 - 1. estimate a linear regression model, using the observed instances of the missing feature as target variable and the rest of the (observed) features as explanatory variables
 - 2. use the model to predict the response for the missing case
- Random forest imputation: similar to linear regression imputation, using a random forest model (Shah et al. 2014)

Extensions of these methods include multiple imputation by chained equations (MICE, see Gelman and Hill (2006)).

2.3 PREDICTION ANALYSIS

In our setting, prediction analysis serves to evaluate the predictive power of a set of neurobehavioural variables. We use a variety of linear and nonlinear classifiers, most of which are widely used. Several classifiers are used in order to evaluate both linear and non-linear models and to make full use of the advantages and disadvantages of different classifiers. In the following, we will briefly introduce the classifiers used in the subsequent chapters. A thorough description can be found, for instance, in Friedman et al. (2001).

2.3.1 General setup

The general problem setting for our purposes is a classification task. We are given a d-dimensional feature vector $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$ and an associated binary label $y \in \{0, 1\}$. (Note that the label could coded arbitrarily, $y \in \{0, 1\}$ is mere convention. In fact, for the description of support vector machines we will switch to $y \in \{-1, 1\}$, since this slightly simplifies the derivations.) The task is to learn the relationship between the features \mathbf{x} and the label y.

There are two distinct approaches for this task – generative and discriminative classifiers. Generative classifiers model the joint probability distribution p(x,y) and the distribution of the features x and y. They then use Bayes' rule to infer the posterior:

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})},$$

and lastly assign the most likely label to a feature vector x, according to this posterior, i.e.

$$\hat{y}(\mathbf{x}) = \underset{y}{\text{arg max}} p(y|\mathbf{x}).$$

On the other hand, discriminative classifiers model the posterior p(y|x)directly or learn a some other form of direct mapping between feature x and label y.

Discriminative classifiers approach the task more directly and were hence long assumed to be more efficient than generative classifiers. In fact, Ng and Jordan (2002) have shown that generative models do have a higher asymptotic error. However, Ng and Jordan (2002) have also demonstrated that generative models may approach the asymptotic error much faster. Hence, both approaches are legitimate and we will describe and use both generative and discriminative classifiers throughout this thesis.

2.3.2 Logistic regression

Logistic regression is one of the most basic linear approaches to classification (Cox 1958). Given the label $y \in \{0,1\}$, the set of features $x \in \mathbb{R}^d$ (usually augmented by $x_0 = 1$ to accommodate the intercept) and parameters $\beta = (\beta_0, ..., \beta_d)$ the following model is assumed:

$$\log\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right) = \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}.$$
 (2.1)

Since the conditional probability P(y = 1|x) is modeled directly, logistic regression counts as a discriminative classifier. Given observations $(x_i, y_i)_{i=1,\dots,N}$ with $x_i = (x_{i1}, ..., x_{id})$ as the feature values for observation i the model parameters can then be estimated by maximum likelihood estimation. The likelihood is derived as follows from Equation (2.1):

$$P(y_i = 1 | \mathbf{x}_i) = \frac{exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + exp(\mathbf{x}_i^T \boldsymbol{\beta})}.$$

Given a $y^* \in \{0, 1\}$:

$$\begin{split} P(y_i = y^* | \mathbf{x}_i) &= \left(\frac{P(y_i = 1 | \mathbf{x}_i)}{1 - P(y_i = 1 | \mathbf{x}_i)}\right)^{y^*} (1 - P(y_i = 1 | \mathbf{x}_i)) \\ &= \exp(\mathbf{x}_i^\mathsf{T} \boldsymbol{\beta})^{y^*} \frac{1}{1 + \exp(\mathbf{x}_i^\mathsf{T} \boldsymbol{\beta})} \\ &= \exp(y^* \mathbf{x}_i^\mathsf{T} \boldsymbol{\beta}) \exp\left(\log\left(\frac{1}{1 + \exp(\mathbf{x}_i^\mathsf{T} \boldsymbol{\beta})}\right)\right) \\ &= \exp(y^* \mathbf{x}_i^\mathsf{T} \boldsymbol{\beta} - \log(1 + \exp(\mathbf{x}_i^\mathsf{T} \boldsymbol{\beta}))). \end{split}$$

Assuming independent observations and $y_i^* \in \{0, 1\}, i = 1, ..., N$:

$$\begin{split} P\left((y_1, \dots, y_N) &= (y_1^*, \dots, y_N^*) \mid (x_1, \dots, x_d)\right) \\ &= \prod_{i=1}^N P(y_i = y_i^*) \\ &= \prod_{i=1}^N \exp(y_i^* \mathbf{x}_i^\mathsf{T} \boldsymbol{\beta} - \log(1 + \exp(\mathbf{x}_i^\mathsf{T} \boldsymbol{\beta}))) \\ &= \exp\left(\sum_{i=1}^N y_i^* \mathbf{x}_i^\mathsf{T} \boldsymbol{\beta} - \log(1 + \exp(\mathbf{x}_i^\mathsf{T} \boldsymbol{\beta}))\right). \end{split}$$

The log-likelihood is therefore given by:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_{i}^{*} \boldsymbol{x}_{i}^{\mathsf{T}} \boldsymbol{\beta} - log(1 + exp(\boldsymbol{x}_{i}^{\mathsf{T}} \boldsymbol{\beta})). \tag{2.2}$$

According to the general principle of maximum likelihood estimation, the parameter vector β is numerically estimated as the arg-max of Equation (2.2).

Naïve Bayes

Naïve Bayes is a simple generative classifier that relies on Bayes' theorem. Using the latter we can rewrite:

$$P(y = y^*|\mathbf{x}) = \frac{p(\mathbf{x}|y = y^*)P(y = y^*)}{p(\mathbf{x})}$$

and then choose the class ŷ that maximizes this expression. This leads to the following:

$$\begin{split} \hat{y} &= \underset{y^* \in \{0,1\}}{\text{arg max}} \, P(y = y^* | \mathbf{x}) \\ &= \underset{y^* \in \{0,1\}}{\text{arg max}} \, p(\mathbf{x} | y = y^*) P(y = y^*) \\ &= \underset{y^* \in \{0,1\}}{\text{arg max}} \, p(\mathbf{x}, y = y^*). \end{split}$$

Now the strong assumption (hence the "naïve" in the name) is made that each feature is independent of each other feature, given the class label, i.e.

$$p(x_1,\ldots,x_d|y) = \prod_{j=1}^d p(x_j|y).$$

It then follows that

$$\hat{y} = P(y = y^*) \prod_{j=1}^{d} p(x_j|y).$$

Finally, a family of distributions is assumed to model each $p(x_i|y)$, Gaussians being the simplest choice. The fact that these distributions are modelled, and not directly $P(y = y^*|x)$ is the reason why naïve Bayes counts as a generative classifier. The parameters of a Gaussian distribution (mean and variance) are then estimated with the typical formulas for each feature and each class separately.

Random forests 2.3.4

Random forests (Ho 1995) are an extension of decision trees. A decision tree uses a top-down approach to recursively bucket data into disjunct sets. At each step it chooses a feature and a cut-off value for this feature that splits the data optimally according to a specific criterion. Typical criteria are the Gini impurity metric and information gain (Breiman 2017). The depth of the tree is a free parameter set by the user implicitly through the minimum node size, i.e. how many samples are minimally allowed to end up in one leaf. This depth parameter influences both computation time and regularisation strength.

A slight modification of a decision tree is to take only a random subset of features, drawn anew at every node instead of inspecting all features at every node as potential candidates for the next split. However, even with this modification, decision trees tend to overfit the data, especially if grown very deep and if no pruning is applied.

Random forests are one way of solving this problem. Simply said, instead of growing a single tree, a whole forest of trees is grown and the "average" result is used for classification. Formally, this is done by performing a variant of bootstrap aggregation (bagging, Breiman (1996)). Given data $\mathbb{D} = (\mathbf{x}_i, \mathbf{y}_i)_{i=1,...,N}$, do the following:

Algorithm 2.1: Random forest

for b = 1, ..., B **do**

- 1. Take a sample $\mathbb{D}_b \subseteq \mathbb{D}$ of size n with replacement
- 2. Train a modified decision tree T_b on \mathbb{D}_b

end

Classify a new data point x^* according to the majority vote of the B modified decision trees

As a suggestion by the inventors of random forests, the modified decision trees are usually grown to maximum depth (i.e. minimum node size = 1) and using random feature subsets of size \sqrt{d} . The modification decorrelates the individual decision trees and allows the averaging to truly reduce variance.

2.3.5 Support vector machines

The support vector machine (SVM) is a technique to construct an optimal linear decision boundary to separate two potentially overlapping classes, see e.g. Friedman et al. (2001, Chapter 12). Since it searches for this boundary directly and does not infer any class distributions, it counts as a discriminative classifier. Such a boundary is called a hyperplane and can be described mathematically as follows (in a d-dimensional space):

$$x_1\beta_1 + x_2\beta_2 + \dots + x_d\beta_d + \beta_0 = 0,$$
 (2.3)

where the β_i are normalised to guarantee a unique representation, i.e. $\|\beta\| = 1$. Setting $\mathbf{x} = (x_1, \dots, x_d)$ and $\beta = (\beta_1, \dots, \beta_d)$ one can express Equation (2.3) more compactly as

$$\mathbf{x}^{\mathsf{T}}\mathbf{\beta} + \mathbf{\beta}_{0} = \mathbf{0} \tag{2.4}$$

In order to easen visualisation, a closer look at the case d = 2 is worthwhile. In this case the hyperplane equation simplifies to:

$$x_1\beta_1 + x_2\beta_2 + \beta_0 = 0$$

which in turn can be rearranged to

$$x_2 = -\frac{\beta_1}{\beta_2} x_1 - \frac{\beta_0}{\beta_2}.$$

The accordance to the simple linear function equation y = ax + b now becomes apparent. A hyperplane can simply be considered as the natural extension of this concept for dimensions d > 2.

The hyperplane divides the feature space into two distinct areas. Mathematically, this corresponds to whether $\mathbf{x}^T \boldsymbol{\beta} + \beta_0$ is smaller or greater than 0. Given data points $(x_i, y_i)_{i=1,...,N}$, where the x_i 's are the d-dimensional features and the y_i 's are the labels as either -1 or +1 (recoded from 0 and 1), one would ideally want perfect classification. This can be expressed by requiring

$$y_i(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta} + \boldsymbol{\beta}_0) > 0, \tag{2.5}$$

since this is only the case if the sign of y_i corresponds to the sign of $x_i^T \beta$ + β_0 . In addition, one wants a maximal "distance" between the classes, the so-called margin M. Hence, in order to find the best separating hyperplane the following optimisation problem has to be solved:

$$\max_{\beta, |\beta_0|, |\beta|| = 1} M \quad \text{subject to } y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) \geqslant M \text{ for all } i = 1, \dots, N,$$

which can be written as

$$\min_{\beta,\beta_0} ||\beta||$$
 subject to $y_i(\mathbf{x}^T \beta + \beta_0) \geqslant 1$ for all $i = 1, ..., N$,

when dropping the constraint on $\|\beta\|$ and setting $M = \frac{1}{\|\beta\|}$.

However, since perfect classification is usually not possible, so-called slack-variables $\xi = (\xi_1, \dots, \xi_N)$ are introduced to allow for points on the wrong side of the margin. Hence, the constraint is modified to the following:

$$y_i(\mathbf{x}_i^T\boldsymbol{\beta} + \boldsymbol{\beta}_0) \geqslant 1 - \xi_i \text{ with } \xi_i \geqslant 0 \text{ for all } i = 1, \dots, N \text{ and } \sum_{i=1}^N \xi_i \leqslant C.$$
 (2.6)

Intuitively, ξ_i denotes the deviation of sample point i from the correct side of the margin. The constant C limits the number and gravity of misclassifications. The optimal parameter is usually determined by crossvalidation, see Section 2.4.2.

Computationally, it is beneficial to re-express Equation (2.6) as a quadratic problem:

$$\begin{split} \min_{\beta,\beta_0} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^N \xi_i & \text{subject to} \\ \begin{cases} y_i(\mathbf{x}^T \beta + \beta_0) \geqslant 1 - \xi_i & \text{for } i = 1, \dots, N \\ \xi_i \geqslant 0 & \text{for } i = 1, \dots, N \end{cases}. \end{split}$$

Then, the corresponding Lagrange primal function that is to be minimized w.r.t. β , β_0 and ξ is as follows:

$$L_{P} = \|\beta\|^{2} + C \sum_{i=1}^{N} \xi_{i} - \sum_{i=1}^{N} \alpha_{i} [y_{i}(\mathbf{x}^{T}\beta + \beta_{0}) - (1 - \xi_{i})] - \sum_{i=1}^{N} \mu_{i} \xi_{i},$$

with $\alpha_1, \ldots, \alpha_N$ and μ_1, \ldots, μ_N as the Lagrangian multipliers. The associated dual objective function is then the following:

$$L_{D} = \sum_{i=1}^{N} \alpha_{i} - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{i} \alpha_{j} y_{i} y_{j} \mathbf{x}_{i}^{\mathsf{T}} \mathbf{x}_{j}.$$
 (2.7)

Support vector machines can be used to model also non-linear boundaries by using the so-called "kernel-trick". The data is mapped into a higherdimensional space to achieve separability. Note that in Equation (2.7), $\mathbf{x_i^l x_i}$ can also be written as $\langle \mathbf{x}_i, \mathbf{x}_i \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner product. So-called basis functions $h_{\mathfrak{m}}:\mathbb{R}^{d}\to\mathbb{R}$ with $\mathfrak{m}=1,\ldots,M$ are used to map the features $\mathbf{h}(\mathbf{x}_i) = (h_1(\mathbf{x}_i), \dots, h_M(\mathbf{x}_i))$. This leads to the following modified dual objective function:

$$L_{D} = \sum_{i=1}^{N} \alpha_{i} - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_{i} \alpha_{j} y_{i} y_{j} \langle \mathbf{h}(\mathbf{x}_{i}), \mathbf{h}(\mathbf{x}_{j}) \rangle.$$

It can be demonstrated (see e.g. Friedman et al. (2001, Chapter 12)) that the estimated hyperplane is linked to h only through the inner product $\langle \mathbf{h}(\mathbf{x}), \mathbf{h}(\mathbf{y}) \rangle$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Hence, one can specify non-linear extensions to the support vector machine by defining a kernel function K(x, y) that computes inner products in the extended feature space. One of the most typical kernel functions is the radial basis kernel:

$$K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma ||\mathbf{x} - \mathbf{y}||^2),$$

with $\gamma > 0$.

2.3.6 Learning vector quantisation

Learning vector quantisation (LVQ) (Kohonen 1990) is a nearest prototype classifier. In contrast to the k-nearest-neighbour algorithm, where all data has to be accessed for classification of a new data point, prototype-based classifiers use only a set of representative data samples (the so-called prototypes) for each class. Learning involves adjusting the location of the prototypes using the labels. More specifically, given a data set $\mathbb{D} = (\mathbf{x}_i, \mathbf{y}_i)_{i=1,\dots,N}$ a prototype-based classification model consists of a set of M labeled proto types, $T = \{(\mathbf{m_j}, c_j)\}_{j=1}^M \subset \mathbb{D}$. For classification of a new sample, the distance to all prototypes m_i is measured according to some appropriate distance measure (e.g. Euclidean), and the sample is assigned to the class of the nearest prototype. During learning, the prototypes m_i are adjusted, but always remain in the labeled data set \mathbb{D} .

Robust soft learning vector quantisation (RSLVQ) is a variant of LVQ and was developed by Seo and Obermayer (2003). It finds an optimal set of prototypes based on the assumption that the data follows a Gaussian mixture model, making it into a generative classifier. It is assumed that the data $x \in \mathbb{R}^d$ is generated by a probability density p(x). Furthermore, it is now assumed that p(x) can be described as a mixture model and that every component j of this mixture generates data of only one class label c_i. Note, however, that more than one component could be generating this class label.

In our two-class setting, and for a set of prototypes T, the probability density is modeled as follows:

$$\begin{split} p(\textbf{x}|T) &= \sum_{\{j: c_j = 0\}} p(\textbf{x}, j) + \sum_{\{j: c_j = 1\}} p(\textbf{x}, j) \\ &= \sum_{k = 0}^{1} \sum_{\{j: c_j = k\}} p(\textbf{x}|j) p(j), \end{split}$$

where p(j) is the probability that data points are generated by component j and p(x|j) is the probability that a certain data point x is generated by component j.

RLSVQ maximises the likelihood ratio

$$L^{\text{ratio}}(\mathbb{D}) = \prod_{i=1}^{N} \frac{p(x_i, y_i | T)}{p(x_i | T)},$$

where $p(x_i, y_i|T)$ is the probability density that a data point x_i is generated by a mixture model for the correct class, i.e. the class referred to by label y_i . When inserting a d-dimensional Gaussian distribution for the components of the mixture model with identical width and strength, i.e. $\sigma_i = \sigma$ and p(j) = 1/M for j = 1, ..., M, one obtains:

$$p(\mathbf{x}|\mathbf{j}) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{(\mathbf{x} - \mathbf{m_j})^2}{2\sigma^2}\right),$$

and the following learning rule (Seo and Obermayer 2003):

$$\mathbf{m}_{\mathbf{j}}(\mathbf{t}+1) = \mathbf{m}_{\mathbf{j}}(\mathbf{t}) + \Delta \mathbf{m}_{\mathbf{j}}(\mathbf{t})$$

with

$$\Delta \mathbf{m}_{j}(t) = \alpha(t) \begin{cases} \left(P_{y}(j|\mathbf{x}) - P(j|\mathbf{x})\right)(\mathbf{x} - \mathbf{m}_{j}) & \text{if } c_{j} = y, \\ -P(j|\mathbf{x})(\mathbf{x} - \mathbf{m}_{j}) & \text{if } c_{j} \neq y \end{cases},$$

and

$$P_{y}(j|\textbf{x}) = \frac{exp\left(\frac{(\textbf{x} - \textbf{m}_{j})^{2}}{2\sigma^{2}}\right)}{\sum_{\{k:c_{k} = y\}} exp\left(-\frac{(\textbf{x} - \textbf{m}_{k})^{2}}{2\sigma^{2}}\right)},$$

and

$$P(j|\mathbf{x}) = \frac{exp\left(\frac{(\mathbf{x} - \mathbf{m_j})^2}{2\sigma^2}\right)}{\sum_{k=1}^{M} exp\left(-\frac{(\mathbf{x} - \mathbf{m_k})^2}{2\sigma^2}\right)},$$

where $\alpha(t)$ is the learning rate. Prototypes that match the label of data point x are attracted to the data point, proportional to the factor $P_{u}(j|x)$ – P(j|x). Prototypes that differ in label to data point x are repelled from it, proportional to the factor P(j|x).

PERFORMANCE ASSESSMENT 2.4

After having introduced several classifiers that are to be used throughout this thesis, note that it is of equal importance to be able to compare these. Hence, methods are needed to assess the quality of different classifiers on a given data set. A typical choice is to look at performance on unseen data

since generally, one is interested in generalisation performance and not in within-sample performance. One of the most typical methods is to use a classifier's accuracy, i.e. on to assess how often (on average) a classifier is correct on unseen data.

The data are i.i.d. realisations from random variables $(X,Y) \sim F$, where $X \in \mathbb{R}^d$, $Y \in \{0, 1\}$ and F some unknown distribution. Now, let

$$Q(y, \hat{G}(x)) = 1{y = \hat{G}(x)}$$

denote the *quality* of the estimate \hat{G} , where $\mathbb{I}\{\cdot\}$ is the indicator function. Let $\mathbb D$ denote the set of data and $\hat{\mathsf G}$ the estimate based on $\mathbb D$. The theoretical concept one is interested in is the test accuracy (Friedman et al. 2001), defined as:

$$Acc_{\mathbb{D}} = \mathbb{E}_{(\mathbf{X}^*,\mathbf{Y}^*)}[Q(\mathbf{Y}^*,\hat{G}(\mathbf{X}^*))|\mathbb{D}],$$

where (X^*, Y^*) is a new test data point, drawn from F and independent of D. This accuracy depends on the chosen (training) data set D. Averaging over \mathbb{D} then yields the *expected test accuracy*

$$Acc = \mathbb{E}_{\mathbb{D}}\mathbb{E}_{(\mathbf{X}^*,\mathbf{Y}^*)}[Q(\mathbf{Y}^*,\hat{\mathbf{G}}(\mathbf{X}^*))|\mathbb{D}]. \tag{2.8}$$

In practice, most methods effectively estimate Acc (Friedman et al. 2001).

Specificity, Sensitivity and Balanced Accuracy

The above theoretical accuracy Acc averages over all possible test samples (X^*, Y) . However, it is also crucial to consider performance separately for each class. Depending on the context, the performance on one class can have a different weight than the performance on the other class. Furthermore, unbalanced classes can give misleading accuracies, see e.g. Brodersen et al. (2010) for simulations. Say that, for instance, one has 1000 data samples, 200 of which belong to one class and the remaining 800 to the other class. Then, even a completely naive classifier predicting simply the label of the majority class, irrespective of the features, will achieve an accuracy of 80 %.

These separate performances are formalized in the following concepts (see e.g. (Fawcett 2006)): positive test samples (here corresponding to y = 1) and correctly identified as such are called truly positive (TP), test samples which are negative (here corresponding to label = 0) and correctly identified as such are called truly negative (TN), falsely classified positive and

negative test samples are called falsely positive (FP) and falsely negative (FN), respectively. Then, the so-called sensitivity or true positive rate is

$$SENS = \frac{TP}{TP + FN}'$$

and the so-called specificity or true negative rate is

$$SPEC = \frac{TN}{TN + FP}.$$

In fact, accuracy can be introduced in a similar manner as

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}.$$

Importantly, also the concept of balanced accuracy can be introduced (Brodersen et al. 2010). Balanced accuracy accounts for the potential imbalance of classes and is also symmetric in the label:

$$balAcc = \frac{SENS + SPEC}{2}.$$

Note that, similar to Equation (2.8), the theoretical notion behind the estimates of SENS and SPEC above can be written as follows:

$$Spec = \mathbb{E}_{\mathbb{D}}\mathbb{E}_{(\mathbf{X}^*,0)}[Q(\mathbf{0},\hat{G}(\mathbf{X}^*))|\mathbb{D}],$$

and

$$Sens = \mathbb{E}_{\mathbb{D}}\mathbb{E}_{(\boldsymbol{X}^*,1)}[Q(1,\hat{G}(\boldsymbol{X}^*))|\mathbb{D}].$$

2.4.2 Cross-validation

In practice, only finite amounts of data are available to estimate Acc. A very popular procedure for estimation is cross-validation.

Algorithm 2.2: Fold generation

K-fold cross-validation entails randomly partitioning the indices of the data set into K subsets \mathcal{B}_k of $\{1, \dots, N\}$ such that:

1.
$$\bigcup_{k=1}^{K} \mathcal{B}_k = \{1, \dots N\},\$$
2. $\mathcal{B}_j \cap \mathcal{B}_k = \emptyset \ (j \neq k),\$

2.
$$\mathcal{B}_{j} \cap \mathcal{B}_{k} = \emptyset \ (j \neq k)$$

3. the sets \mathcal{B}_k are (approximately) of equal size.

For a fixed k, all sample points with indices that are not in one of the sets \mathcal{B}_{k} , are used to train the classifier. The dependency of the estimate on the training set is made explicit by writing this as

$$\hat{G}^{(-B_k)}$$
.

Now, the expected test accuracy of Equation (2.8) is approximated with:

$$\widehat{Acc} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} Q(y_i, \hat{G}^{(-\mathcal{B}_k)}(\mathbf{x}_i)).$$
 (2.9)

By fixing k and iterating over different $i \in \mathcal{B}_k$ we are fixing the training set and giving the fixed estimate $\hat{G}^{(-B_k)}$ different test points (Y_i, X_i) . Hence, this step approximates $\mathbb{E}_{(\mathbf{X}^*,\mathbf{Y}^*)}[Q(\mathbf{Y}^*,\hat{\mathbf{G}}(\mathbf{X}^*))|\mathbb{D}]$ in Equation (2.8). The outer sum, i.e. the iteration over different values of k, adjusts the training set and hence approximates the outer expectation $\mathbb{E}_{\mathbb{D}}$ in Equation (2.8).

Similarly, the expected test specificity and sensitivity can be estimated. Let B_k^1 and B_k^0 denote the subsets of B_k corresponding to indices of data labeled as positive (1) and negative (0), respectively. Then the estimations are the following:

$$\widehat{Spec} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\mathcal{B}_{k}^{0}|} \sum_{i \in \mathcal{B}_{k}^{0}} Q(y_{i}, \hat{G}^{(-\mathcal{B}_{k})}(x_{i}))$$

and

$$\widehat{Sens} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|\mathcal{B}_{k}^{1}|} \sum_{i \in \mathcal{B}_{k}^{1}} Q(y_{i}, \hat{G}^{(-\mathcal{B}_{k})}(x_{i})).$$

This "classical" cross-validation procedure can be summarized as follows:

Algorithm 2.3: Cross-validation

- 1. Split data D into K folds
- 2. **for** k = 1, ..., K **do**
 - a) train model on data, excluding data from fold k
 - b) test model on data from fold k, computing balanced test accuracy balAcck

end

3. Average over the quality measures obtained for each fold, e.g. $\widehat{\text{balAcc}} = \frac{1}{K} \sum_{k=1}^{K} \widehat{\text{balAcc}}_k$

2.4.3 Cross-validation accounting for confounders

In practice, when wanting to analyse the relationship between two (sets of) variables, there are often so-called confounders that need to be considered. Fittingly to our setting, consider a set of neuroimaging variables with which one wants to predict the binary alcohol drinking behaviour. One can easily train and test a classifier for this problem and assess its quality by crossvalidation. However, now consider that there is an additional variable e.g. gender - that is closely linked to both the neuroimaging variables and the drinking behaviour (see Figure 2.2 for a schematic depiction). More specifically, let there be significant differences in the neuroimaging variables between genders and significantly more drinking males than drinking females. If the conventionally assessed classifier now yields balanced accuracies of 90 %, one cannot be sure of the source of this predictive power. It might well be that the classifier is in fact learning to predict gender and not drinking.

For a formal definition of confounders in the framework of causality, we refer to VanderWeele and Shpitser (2013).

In general, there are multiple approaches for accounting for confounders; we focus here on the specific setting of K-fold cross-validation. Consider as above the situation that we have data $\mathbb{D} = (x_i, y_i, z_i)_{i=1,...,N}, y \in \{0,1\}$

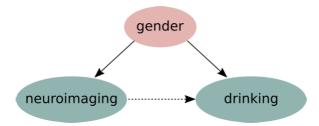


Figure 2.2: Gender as a confounder in the prediction of drinking behaviour from neuroimaging data.

being the binary label (encoding e.g. drinking behaviour) and $z \in \{f, m\}$ being a binary confounder (encoding e.g. gender). The idea is now that we artificially destroy the relationship between the label and the confounder in each fold, i.e. in each fold the confounder is balanced. In order to ensure this, we subsample the data and in order to use all available data, we repeat this procedure B times.

	z = f	z = m				z = m
y = 1	N_{1f}	N_{1m}	y	= 1	31	52
y = 0	N_{0f}	N_{0m}	y	= 0	59	48
	z = f	z = m			z = f	z = m
y = 1	$z = f$ \tilde{N}_1			= 1	z = f 30	z = m 30

Table 2.1: Distribution of data across labels ($y \in \{0, 1\}$) and confounder ($z \in \{f, m\}$). The top left table introduces the general notation, e.g. N_{1f} is the number of data samples with label y = 1 and confounder z = f. The bottom left table shows the situation after subsampling to balance for the confounder, in preparation for K-fold cross-validation. Now, the number of samples with confounder z = f equals the number of samples with z = m, whereas the distribution across labels may still be unbalanced. On the right we display showcase sample sizes.

More specifically, let the data be distributed across labels and confounder as shown in Table 2.1, where e.g. $N_{1f} = |\{(x_i, y_i, z_i) \in \mathbb{D} : y_i = 1, z_i = f\}|$.

Then set

$$\tilde{\mathfrak{n}}_1 = \lfloor \frac{min(N_{1f}, N_{1m})}{K} \rfloor \quad \text{ and } \quad \tilde{\mathfrak{n}}_0 = \lfloor \frac{min(N_{0f}, N_{0m})}{K} \rfloor,$$

where $|\cdot|$ rounds down to the next integer. Then let

$$\tilde{N}_1 = K\tilde{n}_1$$
 and $\tilde{N}_0 = K\tilde{n}_0$.

Now \tilde{N}_1 and \tilde{N}_0 data points are drawn (without replacement) from the data of the first and second row of Table 2.1, respectively. Each of these drawn data sets are split into K folds, according to the "normal procedure" of Algorithm 2.4 to obtain the following testing folds:

$$\{\mathcal{B}_k^{1f}\}_{k=1}^{\tilde{N}_1},\quad \{\mathcal{B}_k^{1m}\}_{k=1}^{\tilde{N}_1},\quad \{\mathcal{B}_k^{0f}\}_{k=1}^{\tilde{N}_0},\quad \text{and}\quad \{\mathcal{B}_k^{0m}\}_{k=1}^{\tilde{N}_0}.$$

Finally, these folds are combined to obtain the confounder-stratified folds $\mathcal{B}^{\text{strat}} = \{\mathcal{B}_{k}^{\text{strat}}\}_{k=1}^{K}$, i.e.

$$\mathfrak{B}_k^{\text{strat}} \coloneqq \mathfrak{B}_k^{\text{1f}} \cup \mathfrak{B}_k^{\text{1m}} \cup \mathfrak{B}_k^{\text{0f}} \cup \mathfrak{B}_k^{\text{0m}} \quad \text{ for } \quad k = 1, \dots, K$$

The adapted cross-validation procedure (that we call "bootstrapped-stratifiedcross-validation" or BSSCV) is now as follows:

Algorithm 2.4: BSSCV

for
$$b = 1, \ldots, B$$
 do

- 1. Subsample data D and prepare K confound-stratified folds
- 2. **for** k = 1, ..., K **do**
 - a) train model on data, excluding data from fold k
 - b) test model on data from fold k, computing e.g. balanced test accuracy balAcc₁

end

3. Average over the quality measures obtained for

each fold, e.g.
$$\widehat{\text{balAcc}}^b = \frac{1}{K} \sum_{k=1}^{K} \widehat{\text{balAcc}}_k^b$$

Compute the final quality measure as

e.g.
$$\widehat{\text{balAcc}} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\text{balAcc}}^b$$

Note that this way, it is easy to balance not only for the confounder but also for the label itself. This entails replacing \tilde{n}_1 and \tilde{n}_0 by

$$\tilde{n} = \lfloor \frac{\min(N_{1f}, N_{1m}, N_{0f}, N_{0m})}{K} \rfloor$$
 (2.10)

and \tilde{N}_1 and \tilde{N}_0 by $\tilde{N} = K\tilde{n}$.

Using balanced classes is often preferred in extremely unbalanced scenarios. This can hold even if a balanced quality measure, such as balanced accuracy, is used (Brodersen et al. 2010).

Furthermore, this procedure can be adapted to account for more than one (binary) confounder – with the cost of loosing predictive power due to even smaller subsetted data available for training.

Naturally, one cannot account for (or even be aware of) all possible confounders. In practice, one usually chooses a hypothesis-driven approach and checks for correlation structures of at least the most obvious potential confounders.

2.4.4 Significance of accuracies

Having obtained an estimation of accuracy (or another quality measure) from a cross-validation scheme, the question arises, whether this accuracy is significantly better than chance. If we have N = 10 data samples and an SVM yields a balanced accuracy of 55 % - is this better than chance? What would happen with N = 1000 samples? Statistical hypothesis testing offers a systematic answer. Say we are interested in the accuracy Acc of a classifier. Then null and alternative hypotheses could be as follows:

> H_0 : Acc = 0.5 (the classifier performs at chance), H_a : Acc > 0.5 (the classifier makes an informed choice).

Naively, one could consider a binomial test for this scenario. This would assume that, under the null-hypothesis, every single classification is identically and independently Bernoulli-distributed. However, due to the crossvalidation scheme, classifications are neither identically nor independently drawn from a Bernoulli-distribution. Since every classification decision can potentially arise from different fitted models (since the training data varies over folds), they are not identically distributed. Independence is violated due to the high correlation between training data, hence also between estimated models and classification decisions. For this reason, binomial testing is not a good approach for the evaluation of accuracy significance.

A better approach is to use permutation testing. This is a subfield of non-parametric statistics. In contrast to binomial testing, the distribution under the null-hypothesis is not assumed to come from a specific family of distributions. Instead, the null-distribution is estimated and the estimated accuracy is compared to this approximate null-distribution.

Roughly speaking, in permutation testing, the label is randomly permuted and the cross-validation scheme (if necessary accounting for confounders) is followed to estimate an accuracy. In theory, the relationship between label and features is destroyed by the permutation, hence the obtained accuracy is a "random" accuracy. By repeating this procedure many times (usually at least 1000) and recording the obtained accuracy, we get an empirical distribution under the null hypothesis.

More formally, the permutation testing follows the following scheme, given an estimate of balanced accuracy balacc and data $\mathbb{D} = (\mathbf{x}_i, \mathbf{y}_i)$:

```
Algorithm 2.5: Permutation testing for accuracies
     for j = 1, ..., nPerm do

    permute labels y<sub>i</sub>

             2. run a cross-validation scheme to obtain
                an estimated accuracy balAcciperm
     end
     Compute a p-value as:
           \hat{P}(balAcc \geqslant balAcc \mid H_0 : balAcc = 0.5)
                     = \frac{1}{nPerm} \sum_{j=1}^{nPerm} \delta(balAcc_j^{perm} \ge \widehat{balAcc})
```

Note that here we permute the labels. This will be different for the permutation test introduced in Section 2.5.3.

2.5 ASSOCIATION ANALYSIS

In the prediction analysis and the significance testing we assess the *predic*tive value of a set of variables. This does not naturally entail the information about single features. Hence, we introduce association analysis as a posthoc step to examine significantly predictive sets of variables more closely. As a result from association analysis, we would like an estimation of exactly how a (set of) variable(s) influences the label. For instance, given the size of a certain brain region, we would like to know whether it is, on average, larger or smaller in alcohol-drinking adolescents compared to non-drinking adolescents. As argued in Section 2.1, we crucially require the predictive setting to be significant in order to proceed with the post-hoc association analysis.

Importantly, since we already have an estimate of generalisation performance, we do not perform cross-validation for the association analysis. Instead, ideally we can use the entire available data set $\mathbb D$ at once for this post-hoc analysis.

Logistic regression for association analysis

Many different methods could be used to assess the so-called "association". For single variables, the most straightforward (and also most ubiquitous in the field of neuroimaging, see e.g. Nichols and Hayasaka (2003)) approach is using two-sample t-tests.

However, t-tests have several weaknesses, for instance that they cannot account for the presence of confounders. The natural extension is to resort to logistic regression models. As described in Section 2.3.2, logistic regression models can be used for classification. Also, they naturally output estimates of feature strength by β . Confounders can be included as additional variables of no interest.

Formally, let $\mathbf{x} = (x_1, \dots, x_p) \in \mathbb{R}^d$ and $\mathbf{y} \in \{0, 1\}$ be the features and label, respectively and let $\mathbf{z} = (z_1, \dots, z_q) \in \{0, 1\}^q$ be the confounders. Assume now that we want to assess the influence of a single feature x_k on the label y. Note that the following can easily be extended to non-binary confounders and to a set of features. Now let

$$\mathbf{w} = (\mathbf{x}_{k}, \mathbf{z}) \in \mathbb{R} \times \{0, 1\}^{\mathbf{q}} \tag{2.11}$$

be the augmented feature vector. Now, analogously to Section 2.3.2 we can formulate the logistic regression model as follows:

$$\log\left(\frac{P(y=1|\mathbf{w})}{1-P(y=1|\mathbf{w})}\right) = \mathbf{w}^{\mathsf{T}}\boldsymbol{\beta},\tag{2.12}$$

with $\beta \in \mathbb{R}^{p+1}$. Now β_1 is the coefficient of our feature of interest. With analogous methods as described in Section 2.3.2, we can estimate $\hat{\beta}_1$. With the usual distributional assumptions (see e.g. Friedman et al. (2001)) we can then perform inference on this estimate and obtain a p-value.

A typically employed statistic is the Wald-statistic, which is approximately normally distributed for large samples:

$$W = \frac{\hat{\beta}_1}{\sqrt{\widehat{\operatorname{Var}}(\hat{\beta}_1)}} \approx N(0,1), \tag{2.13}$$

where $\sqrt{\widehat{Var}(\hat{\beta}_1)}$ is an estimate of the standard error. The most straightforward interpretation of significant β_1 's is is to simply look at the sign: a positive β_1 means that an increase in variable x_k significantly increases the likelihood of belonging to the class y = 1 and vice-versa. Other means of interpretation can be done via odds ratios (Friedman et al. 2001).

Repeated logistic lasso (ReLL) for association analysis \star

As an extension of logistic regression for association analysis we have developed a method we call repeated logistic lasso (ReLL). As the name already states, the method relies heavily on the lasso for logistic regression, which we introduce here briefly.

The least-absolute shrinkage and selection operator (lasso) was developed by Robert Tibshirani (Tibshirani 1996) and is a regularisation and feature selection method. Recall the log-likelihood from Equation (2.2). For lasso regularised logistic regression, the maximisation of β is done over a penalised version of Equation (2.2), formulated as follows (with x_i , y_i and β defined as in Equation (2.2)):

$$l^{\text{lasso}}(\beta) = \sum_{i=1}^{N} \left[y_i x_i^\mathsf{T} \beta - \log(1 + \exp(x_i^\mathsf{T} \beta)) \right] - \lambda \sum_{j=1}^{d} |\beta_j|, \tag{2.14}$$

where $\lambda \in [0, 1]$ is the regularisation parameter. The regularisation by means of the L₁-norm entails that β_i are often estimated as exactly 0, if they do not add informative value to the model. This is a desirable property for our association analysis, since it automatically lets only "important" features survive.

The regularisation parameter λ is commonly chosen via K-fold crossvalidation, by optimising e.g. accuracy or deviance over a given set of potential λ 's (see e.g. Friedman et al. (2001) or Friedman et al. (2010)).

Now in order to adapt Equation (2.14) to the presence of confounders, we take again the notation that $\mathbf{w} = (x_k, \mathbf{z}) \in \mathbb{R} \times \{0, 1\}^q$ with x_k as the single feature we wish to assess. Given data $\mathbb{D} = \{(x_i, y_i, z_i)\}_{i=1}^N$ with $\mathbf{x}_i = (x_{i1}, \dots, x_i p)$ let $\mathbf{w}_i = (x_{ik}, \mathbf{z}_i)$.

We introduce the following adapted likelihood for ReLL:

$$l^{ReLL}(\beta) = \sum_{i=1}^{N} \left[y_i \mathbf{w}_i^{\mathsf{T}} \beta - \log(1 + \exp(\mathbf{w}_i^{\mathsf{T}} \beta)) \right] - \lambda \beta_1.$$
 (2.15)

Note that the penalisation is performed only on the coefficient corresponding to our feature of interest and not on the coefficients of the confounders. This enforces that the confounders are never de-selected. In turn this means that a "surviving" feature xk must survive in presence of the confounders.

Note that (again) this can be formulated analogously for non-binary confounders and/or a set of features.

Due to the randomness of K-fold cross-validation, the chosen λ and hence also the estimated model may vary over different runs. This is problematic for our purposes: we would not wish a certain feature to be "chosen" as informative in one run and then not chosen in the next.

Hence, we propose to repeat the entire model fitting (including the optimisation of λ) N_{rep} times. As a final estimate of β we propose to use the median. More formally, we suggest the following algorithm:

```
Algorithm 2.6: ReLL for association analysis
     for r = 1, ..., N_{rep} do
             • fit a lasso-regularised regression model on the
                augmented data \mathbb{D} = \{(\mathbf{w}_i, y_i)\}_{i=1}^N
                maximising Equation (2.15)
             • choose λ by K-fold cross-validation

 obtain estimate β<sup>r</sup><sub>1</sub>
```

end

Obtain a final estimate by:

$$\hat{\beta}_1^{ReLL} \coloneqq median(\beta_1^1, \dots, \beta_1^{N_{rep}})$$

Interpretation of $\hat{\beta}_1^{ReLL}$ is analogous to interpretation in logistic regression, i.e. a positive β_1 implies that an increase in variable x_k increases the likelihood of belonging to the class y = 1 and vice-versa.

To check if our developed method has the desired properties, we conducted a small simulation study. We simulate a feature x_k that is linked to a label y with a true association coefficient $\beta_1 \in [-0.03, 0.03]$, emulating very weak (or even 0) associations. Furthermore, we add a confounder $z \in \{0,1\}$ that is related to the feature x_k and to the label y, based loosely on the setting described in Figure 2.2. We then let the association β_1 be estimated by logistic regression, as described in Section 2.5.1 and with the ReLL method for each simulated scenario. Figure 2.3 shows that our ReLL method is always more conservative in its estimation of association than logistic regression, since $|\hat{\beta}_1^{ReLL}| < |\hat{\beta}_1|$. Furthermore, the simulation suggests that the ReLL method seems to give a less biased estimate of β_1 , since $|\hat{\beta}_1^{Rell} - \beta_1| < |\hat{\beta}_1 - \beta_1|$. However, these simulation results are only indicative. A more extensive analysis would be required to draw definite conclusions. This could also include an analytic derivation of the distribution of the ReLL estimator. We comment in more detail on this idea in Section 7.2.

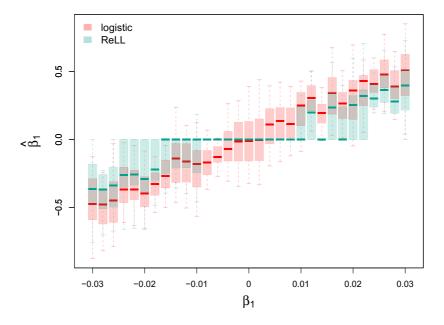


Figure 2.3: Simulation study to compare logistic regression to ReLL for association analysis. We simulate observations of a feature $x_k \in \mathbb{R}$ that is related both to a label $y \in \{0,1\}$ and a confounder $z \in \{0,1\}$, based loosely on the setting described in Figure 2.2. More precisely, we draw the first 500 samples i.i.d. from a Gaussian distribution $\mathcal{N}(\mu_{z=0}, \sigma_{z=0})$ and the another 500 samples i.i.d. from $\mathcal{N}(\mu_{z=1}, \sigma_{z=1})$, where $\mu_{z=0} = 2.1$, $\mu_{z=0} = -1.2$, $\sigma_{z=1} = 0.9$ and $\sigma_{z=1} = 0.4$. This gives us a total of N = 1000 samples. The difference in distributions of x_k is supposed to emulate the relationship between the confounder z and the feature x_k . We then simulate labels y by the model described in Equation (2.12), using $\beta_z = -1.9$ and an additional noise term, drawn from $\mathcal{N}(0,0.1)$. The negative β_z results in a negative effect of the confound on the label, simulating the negative effect of being female on the probability of being a drinker. The noise term with a mean of 0 lets the data follow the model on average, but with some fluctuation. The simulation of data is repeated for different coefficients of interest $\beta_1 \in [-0.03, 0.03]$. The weak association strengths are meant to emulate the weak associations typically found in real neurobehavioural features. For each simulated data sets corresponding to a different true association β_1 , we estimate the association with (i) logistic regression, as described in Section 2.5.1 and (ii) with the ReLL method. Boxplots represent the distribution of respective estimates over 100 different noise realisations, displaying the 95% coverage.

2.5.3 Significance testing of ReLL

Whether the $\hat{\beta}_1^{ReLL}$ are significant is not a trivial problem, since significance testing for the lasso is not straightforward (Lockhart et al. 2014). This makes significance testing for the ReLL all the less straightforward. We propose to resort to permutation testing – i.e. to compare true estimates to an empirically obtained null-distribution.

In the version of permutation testing we use for testing prediction accuracies, see Section 2.4.4, we permute labels. Here, we instead permute the feature-of-interest column $(x_{ij})_{i=1}^{N}$. This has been proposed by Leo Breiman in the framework of random forests (Breiman 2001). Permuting the feature x_i breaks any association between x_i and the label y. However, the relationship between the confounders z and the label y is preserved. Hence, the resulting significance test assesses the specific influence of the feature-of-interest x_i .

Formally, we propose the following algorithm, given an estimate $\hat{\beta}_{1}^{Rell}$:

Algorithm 2.7: Permutation testing for ReLL

for $k = 1, ..., N_{perm}$ do

- 1. permute the feature-of-interest x_{ij}
- 2. run Algorithm 2.7 to obtain an estimated feature weight:

βReLL,perm

end

Compute a p-value as:

$$\begin{split} \hat{P}(\beta_1^{ReLL} \geqslant |\hat{\beta}_1^{ReLL}| \mid H_0: \beta_1^{ReLL} = 0) \\ &= \frac{1}{N_{perm}} \sum_{k=1}^{N_{perm}} \delta(\hat{\beta}_{1,k}^{ReLL,perm} \geqslant \hat{\beta}_1^{ReLL}) \end{split}$$

2.6 MULTIPLE TESTING CORRECTION

We have introduced the testing of balanced accuracy and the testing of the ReLL-association method. They have to be performed for various classifiers and features-of-interest, resulting in M different null- and alternative hypotheses and M associated p-values. For instance, when testing ReLLassociation we have the following:

$$\begin{split} H_{0,m}: \quad \beta_{1,m}^{ReLL} &= 0 \\ H_{\alpha,m}: \quad \beta_{n1,m}^{ReLL} &\neq 0 \end{split}$$

for $\mathfrak{m}=1,\dots M.$ From significance testing we obtain M associated p-values p_m^{ReLL} . A null-hypothesis, $H_{0,m}$, is rejected at level α if $p_m^{ReLL} < \alpha$. Then, by construction:

$$P(H_{0,m} \text{ rejected} \mid H_{0,m} \text{ true}) = \alpha.$$

Let further $A_{\rm m}$ be the event that we falsely reject the null-hypothesis, i.e.

$$A_m = H_{0,m}$$
 rejected $\mid H_{0,m}$ true,

then the family wise error rate is the probability of at least one false rejection, i.e.:

$$FWER = \bigcup_{m=1}^{M} P(A_m).$$

For M independent tests, it can easily be shown that

$$FWER = 1 - (1 - \alpha)^{M}.$$

Hence, already for M = 2 (independent) tests, FWER $> \alpha$ and it becomes more likely to reject null-hypotheses that are in fact true.

The Bonferroni-method (Friedman et al. 2001) accounts for the multiple tests by simply adjusting the significance level $\tilde{\alpha} = \frac{\alpha}{M}$. It can then easily be shown that FWER $\leq \alpha$. However, for many tests, the Bonferroni-method is known to be too strict (Friedman et al. 2001).

Hence, it is a typical procedure to instead control the so-called *false dis*covery rate (FDR). If we are given an underlying scenario as described in Table 2.2, then:

$$FDR = \mathbb{E}\left[\frac{FR}{rej}\right]$$
,

where FR is the number of falsely rejected null hypotheses TR is the number of correctly rejected null hypotheses and rej = FR + TR.

Table 2.2: Overview of multiple testing setting. TA (FA) is the number of correctly (incorrectly) non-rejected null hypotheses, and FR (TR) the number of incorrectly (correctly) rejected null hypotheses. M denotes the total number of tests and rej the number of rejected null hypotheses.

	not rejected	rejected
H ₀ true	TA	FR
$H_{\mathfrak{a}}$ true	FA	TR
	M – rej	rej := FR + TR

The FDR can then be controlled by the so-called Benjamini-Hochberg method (Friedman et al. 2001) in the following way:

Algorithm 2.8: Benjamini-Hochberg correction

 $H_1 \dots, H_M$ tests yielding p-values p_1, \dots, p_M

- 1. order p-values: $p_{(1)} \leqslant \cdots \leqslant p_{(M)}$
- 2. set $K = \max\{m = 1, ..., M \mid p_{(m)} \leqslant \frac{m}{M} \alpha^*\}$
- 3. reject all $H_{(i)}$ with i = 1, ..., K

Throughout this thesis, we will be referring to the Benjamini-Hochberg method when mentioning FDR-correction.

2.7 CHAPTER SUMMARY

In this chapter, we have introduced all the (old and new) methods used in the subsequent chapters.

- We motivated our two-step procedure (prediction analysis, then association analysis). If (and only if) a set of features is declared significant by the prediction analysis, we will move on to post-hoc association analysis.
- We further gave a brief overview of outlier detection and treatment and of the various imputation methods used in this thesis.
- We then dived into prediction analysis, briefly recapitulating wellestablished classifiers (logistic regression, naïve Bayes, random forests, support vector machines, RSLVQ). We commented on the different ways to assess the performance of a classifier (accuracy, specificity, sensitivity), both theoretically and practically using cross-validation. We further introduced a novel adapted version of cross-validation – bootstrapped-stratified-cross-validation (BSSCV) – for the assessment of classifier performance in the presence of confounders.
- Next, we took a closer look at association analysis. We recapitulated the classical methods for association analysis, namely t-tests and logistic regression. We then introduced a novel method for association analysis - repeated logistic lasso (ReLL) - and showed in a small simulation study that it can outperform logistic regression for association analysis.

Part II The IMAGEN data base

3 | THE IMAGEN PROJECT

This chapter gives an overview of the IMAGEN project and our specific contribution. The remaining chapters will be based on the data of this study.

3.1 INTRODUCTION

The IMAGEN project (Schumann et al. 2010) is a European, large-scale, multi-centre and multidisciplinary collaboration that was initiated to shed light how neurobiological, psychological and environmental aspects during adolescence may or may not influence brain development and, in the long-term, mental health.

More specifically, the IMAGEN consortium describes their aim as follows (Schumann et al. 2010):

...to identify the genetic and neurobiological basis of individual variability in quantitative psychological traits, and to determine their predictive value for the development of frequent neuropsychiatric disorders.

In order to pursue this aim, the IMAGEN study combines (i) behavioural and neuropsychological characterisation, (ii) functional and structural neuroimaging and (iii) genome-wide association analysis. Furthermore, the study has a longitudinal design, starting off with the assessment of around 2000 14 year-old adolescents (Baseline, BL) and following up on them at approximate age 16 (Follow-Up 1, FU1) and 19 (Follow-Up 2, FU2). This design enables drawing trajectories of adolescent development.

The data was acquired at eight sites across four European countries (United Kingdom: London, Nottingham; Germany: Berlin, Dresden, Hamburg, Mannheim; France: Paris; Ireland: Dublin). Subjects were recruited at high schools; in order to obtain a heterogeneous (and representative) sample in terms of socio-economic status and cognitive development, private, statefunded and special units were all targeted. After written consent, personal-

ity and cognitive testing was done via home assessments. In a study centre visit, adolescents and their parents underwent various further testing.

Standard operating procedures and a full list of administered instruments can be found online (IMAGEN SOP 2019). The IMAGEN project has resulted in more than 100 publications and currently a third follow-up is being conducted.

CROSS-SECTIONAL AND LONGITUDINAL 3.2

The general longitudinal design of the IMAGEN study plays an important role throughout this thesis. We hence introduce the main wordings related to the study design here. Figure 3.1 shows the setup of the IMAGEN study at three time-points: Baseline (BL), Follow-Up 1 (FU1) and Follow-Up 2 (FU₂).

Wordings specific to our work are the so-called cross-sectional and longitudinal setting. By the cross-sectional setting, we mean using various features from FU2 to associate with drinking behaviour at the same time point. By the longitudinal setting, we mean using features from BL to associate with drinking behaviour at FU2.

3.3 OUR CONTRIBUTION

We focussed our research on only a sub-domain of the broad aims of the IMAGEN project. More specifically, we use data from two of the three general data domains:

- (i) behavioural and neuropsychological characterisation (personality, life events and cognitive assessments),
- (ii) functional and structural neuroimaging (contrast images from an fMRI paradigm and grey matter volume).

In this thesis, we do not examine data from the third domain (genome-wide association). Note that we comment in 7.2 on possible extensions.

In the style of the aims of IMAGEN, our aims in this thesis can be described as follows:

...to identify the neurobiological basis of individual variability in alcohol drinking behaviour to determine the predictive value for the development of later alcohol abuse.

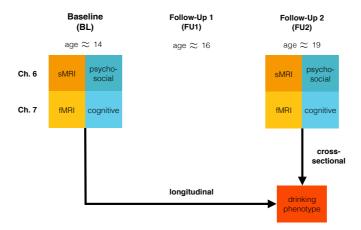


Figure 3.1: Overview of longitudinal study design of IMAGEN, explaining the meaning of cross-sectional and longitudinal settings for all subsequent analyses. Chapter 6 considers structural magnetic resonance imaging (sMRI) and psychosocial data, whereas Chapter 7 treats functional MRI (fMRI) and cognitive data. We generally focus on the drinking behaviour at Follow-Up 2 (FU2), whereas the rest of the data can also originate from FU2 - resulting in the cross-sectional setting - or from Baseline (BL) - resulting in the longitudinal setting. Data from Follow-Up 1 (FU1) is used only in some additional analyses and will be mentioned explicitly.

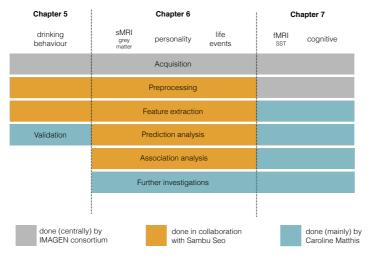


Figure 3.2: The pipeline of the analysis, separated into chapters of this thesis and marked with the main contributors.

Furthermore, parts of the general "pipeline" of analysis (that will be explained in more detail throughout this thesis) were done centrally by experts from the IMAGEN consortium or together with local collaborators. Figure 3.2 displays this pipeline and the main contributors to every step. Importantly, note that we were not in any way involved in the acquisition of the data and that some preprocessing was also performed centrally.

CHAPTER SUMMARY 3.4

In this chapter, we gave an overview of the aims and design of the IM-AGEN study. We further introduced the cross-sectional and longitudinal settings. Finally, we gave a detailed description of our contribution within the IMAGEN framework.

4 DEFINING ALCOHOL DRINKING PHENOTYPES

This chapter introduces the label – light and heavy drinkers – that will be used throughout the rest of the thesis.

4.1 INTRODUCTION

The fifth version of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V, American Psychiatric Association (2014)) indicates an alcohol use disorder (AUD) when at least two of a list of 11 criteria were met within the past year. Criteria include having

- had times when you ended up drinking more, or longer, than intended,
- continued to drink even though it is causing trouble with your family or friends,
- more than once wanted to cut down or stop drinking, or tried to, but couldn't.

The number of fulfilled criteria gives a classification into mild (2-3 symptoms), moderate (4-5 symptoms) and severe (6 or more symptoms) AUD. Note than none of the criteria relate to thresholds of quantity of consumed alcohol.

In contrast to the plentiful studies comparing alcohol addicted patients to a control group, population studies such as IMAGEN with per se healthy samples (Schumann et al. 2010) do not offer such a pre-defined split. Typical information that is recorded in such population studies is average amount of alcohol consumed within a certain period (see e.g. Leon et al. (2007) and McCambridge et al. (2011)) or, especially for adolescent subjects, the frequency of binge drinking events, see e.g. Hill et al. (2000), Mashhoon et al. (2014), Petit et al. (2014), Squeglia et al. (2011), and Squeglia et al. (2012).

Within the IMAGEN study several different questionnaires concerning alcohol consumption were administered. Instead of focusing on a single

answer to a single question of a single questionnaire, we chose to establish a more comprehensive alcohol drinking behaviour variable, explained in detail in the subsequent sections.

AVAILABLE QUESTIONNAIRES 4.2

In the following, we briefly describe the available questionnaires of the IM-AGEN study regarding alcohol consumption and behaviour. Note that (except for Section 4.4) we will consider only the data from FU2, i.e. when subjects are around 19 years old.

4.2.1 ESPAD

The version of The European School Survey Project on Alcohol and other Drugs (ESPAD) (Hibell et al. 1997; Hibell et al. 2004) administered in the IMAGEN sample includes a set of more than 100 questions on legal and illegal substance use within several past time frames. Questions that we focus on are as follows:

Frequency of drinking: On how many occasions in your whole lifetime / over the last 12 months / over the last 30 days have you had any alcoholic beverage to drink?

Frequency of bingeing: On how many occasions in your whole lifetime / over the last 12 months / over the last 30 days have you had five or more drinks in a row?

Frequency of drunkenness: On how many occasions in your whole lifetime / over the last 12 months / over the last 30 days have you been drunk from drinking alcoholic beverages?

To any of the three variants (lifetime, 12 months, 30 days) of each question, subjects could choose an answer coded as an ordinal scale ranging from o to 6 (for frequency of drinking and drunkenness) or from o to 5 (for frequency of bingeing). The corresponding values are shown in Table 4.1.

Table 4.1: Correspondence of answer possibilities in the ESPAD questionnaire to actual values. Answer possibilities range from o to a maximum of 6. Values indicate occasions over whole lifetime, last 12 months or last 30 days.

Ordinal scale	О	1	2	3	4	5	6
Frequency of drinking	o	1-2	3-5	6-9	10-19	20-39	≥ 40
Frequency of bingeing	О	1	2	3-5	6-9	≥ 10	
Frequency of drunkenness	О	1-2	3-5	6-9	10-19	20-39	≥ 40

4.2.2 TLFB

The Timeline Followback (TLFB) is an assessment instrument to obtain retrospective daily estimates of substance consumption across a certain time span (Sobell and Sobell 1992; Sobell et al. 1996). The relevant (since standardised and specific to alcohol) question of the TLFB administered in the IMAGEN study is as follows:

Quantity: Total number of Alcohol Drink Units in past 30 days?

One Alcohol Drink Unit hereby refers to the equivalent of 8g of absolute ethanol. Note that in contrast to the ESPAD, the TLFB only considers the past 30 days.

4.2.3 MAST

The Michigan Alcoholism Screening Test (MAST) is a structured interview instrument for detecting alcoholism (Selzer 1971). In contrast to the ESPAD and the TLFB, the MAST does not include questions on exact amounts or frequencies of alcohol consumptions. Rather, it is aligned more closely to the DSM-V criteria for alcohol use disorders, asking for the relationship to drinking and symptomatic behaviours. More precisely, binary answers (yes/no) to 25 questions such as the following had to be given:

- Can you stop drinking without a struggle after one or two drinks?
- Do you ever feel bad about your drinking?
- Have you ever lost a job because of drinking?
- Have you ever been arrested for drunk driving or driving after drinking?

Modified from Selzer (1971), a positive alcoholism screening is defined as follows:

MAST-flag: Set as true if at least 5 of the 25 criteria from the MAST are met.

This MAST-flag is a relatively strict measure of alcohol drinking behaviour.

4.2.4 AUDIT

The Alcohol Use Disorders Identification Test (AUDIT) is a 10-item questionnaire covering alcohol consumption, drinking behaviour and alcoholrelated problems (Saunders et al. 1993). Similar to the MAST, it focuses more on early detection of alcohol use disorders than on assessing general drinking habits. Questions asked include the following:

- How often do you have a drink containing alcohol?
- How often do you have six or more drinks on one occasion?
- How often during the last year have you had a feeling of guilt or remorse after drinking?
- How often during the last year have you been unable to remember what happened the night before because you had been drinking?

All 10 questions are answered on an ordinal scale ranging from o ("never") to 4 ("daily or almost daily"), giving a maximum score of 40 over the entire test.

According to Saunders et al. (1993), a flag for likely alcohol abuse is defined as follows:

AUDIT-flag: Set as true if a score of at least 8 is reached.

In comparison to the MAST, the AUDIT is slightly less strict.

CONSTRUCTION OF PHENOTYPE 4.3

We use all subjects for which at least all three ESPAD variables were nonmissing, which amounts to a total of 1472 subjects.

Grouping of drinking behaviour variables

We grouped each of the four variables (Frequency of drinking, bingeing and drunkenness from ESPAD and Quantity from TLFB) into three groups - light, moderate and heavy. This grouping was performed based on a trade-off between clinical representation of drinking patterns at age 19 and having roughly balanced group sizes for the IMAGEN sample. Furthermore, having three groups is parallel to the grouping of AUD into mild, moderate and severe. The histograms and groupings are shown in Figure 4.1. The number of subjects allocated to each group for each variable is shown in Table 4.2.

Importantly, we can observe that a consistent allocation across the four variables (i.e. that all four variables agree on the classification of a subject into either light, moderate or heavy drinker) happens only in 339 subjects of the 1147 where all four variables are available, see Table 4.2.

Thus, all further analysis crucially depends on the drinking variable that we choose to focus on. We decided to take a more systematic approach, by using the information from all four variables for the construction of one comprehensive drinking behaviour phenotype.

Table 4.2: Number of subjects belonging to light, moderate and heavy groups (according to the grouping depicted in Figure 4.1) for the four drinking variables. The last row shows the number of subjects for which the grouping of all four phenotypes agree.

	light	moderate	heavy	total
Frequency of drinking	447	633	392	1472
Frequency of bingeing	471	545	456	1472
Frequency of drunkenness	650	368	454	1472
Quantity	253	431	463	1147
All four agree	96	66	177	339

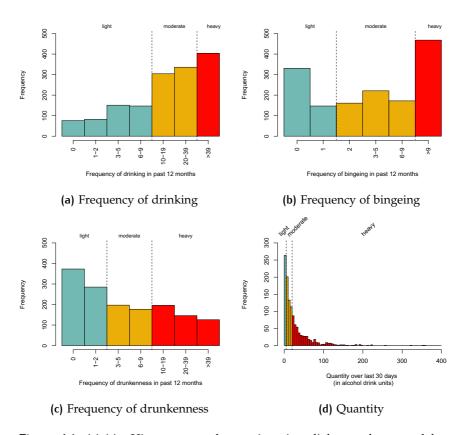


Figure 4.1: (a)-(c): Histograms and groupings into light, moderate and heavy drinkers of the three drinking variables from ESPAD, assessed at FU2 and counting drinking occasions of the past 12 months. (d): Histogram and grouping into light, moderate and heavy drinkers of the Quantity of alcohol from TLFB, assessed at FU2 and measuring the total number of Alcohol Drink Units consumed in the past 30 days.

4.3.2 Establishing a comprehensive phenotype

In order to construct a comprehensive phenotype from the four grouped drinking variables, we used a pairwise clustering algorithm (Seo et al. 2009; Seo and Obermayer 2003), setting the number of clusters a priori to three. Pairwise clustering uses the pairwise distance between all pairs of subjects. The four grouped drinking variables are on an ordinal scale – light, moderate and heavy, which we code as 0, 1 and 2, respectively. On ordinal scales,

the only operation that one can use is counting events of one entry being smaller, larger, or equal to another entry.

For computing the pairwise distance between pairs of subjects we hence use the following distance measure for ordinal data from Walesiak (1999):

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{m} \alpha_{ikj} \alpha_{kij} + \sum_{j=1}^{m} \sum_{l=1, l \neq i, k}^{n} \alpha_{ilj} \alpha_{klj}}{2 \left[\left(\sum_{j=1}^{m} \sum_{l=1, l \neq i}^{n} \alpha_{ilj}^{2} \right) \left(\sum_{j=1}^{m} \sum_{l=1, l \neq k}^{n} \alpha_{klj}^{2} \right) \right]^{\frac{1}{2}}},$$

where

$$a_{ikj} = \begin{cases} 1, & \text{if} & x_{ij} > x_{kj} \\ 0, & \text{if} & x_{ij} = x_{kj}, \\ -1, & \text{if} & x_{ij} < x_{kj} \end{cases}$$

n denoting the total number of subjects and m the number of ordinal variables (in our case four). The dik denotes the distance between subject i and subject k and x_i is the m-dimensional variable vector of subject i and

 $x_{ij} \in \{0,1,2\}$ is the j-th variable of subject i. Note that $\sum_{j=1}^m \alpha_{ikj} \alpha_{kij} \leqslant 0$, since $\alpha_{ijk} = -\alpha_{kij}$. Note further that it counts the number of variables j for which subject i differs from subject k. The larger the "disparity" between x_i and x_k , the more negative $\sum_{i=1}^{m} a_{ikj} a_{kij}$ and hence the larger d_{ik} .

Furthermore, $\sum_{j=1}^{m} \sum_{l=1, l \neq i, k}^{n} a_{ilj} a_{klj}$ is the disparity between x_i and x_k , relative to the similarity to all other subjects.

Lastly, it holds that $\sum_{j=1}^{m} \sum_{l=1, l \neq i}^{n} \alpha_{ilj}^{2} \geqslant 0$ and note that it counts the number of variables j for which subject i differs from all other subjects (analogously for $\sum_{j=1}^{m} \sum_{l=1, l \neq k}^{n} \alpha_{klj}^2$).

Intuitively, the distance between two subjects is 0, if the two subjects have equal relation (smaller or larger) to all other subjects. Note that dik does not define a metric, since from $d_{ik} = 0$ it does not follow that i = k.

In general, we used all four variables for calculating the pairwise distance between subjects. However, since some subjects were missing the variable Quantity from the TLFB (see Table 4.2), we used only the three available variables for these cases.

Figure 4.2 shows the three resulting clusters and the distribution of the four input variables over each cluster. The three clusters show clear correspondence to the three original groups (light, moderate and heavy drinkers). A total of 550 subjects are assigned to the first cluster and most of the input variables assign subjects in this cluster to the "light" drinker group. We hence refer to this cluster as the new light drinker group. The second

cluster contains a total of 458 subjects and most of the input variables assign subjects in this cluster to the "moderate" drinker group. Therefore, we refer to this cluster as the new moderate drinker group. Finally, a total of 464 subjects are assigned to the third cluster and most of the input variables assign subjects in this cluster to the "heavy" drinker group. We thus refer to this cluster as the new heavy drinker group.

4.3.3 Dropping the moderate drinkers

In order to obtain more disparate groups, we decided to focus all subsequent analyses only on the two extreme groups – light (N = 550) and heavy (N = 464) drinkers, i.e. we dropped the moderate drinkers; this results in a binary label.

The disparity of the two resulting groups can also be observed in the original variables, see Table 4.3. Here, we can see that the new groups correspond nicely to the former division into light and heavy drinkers. For instance, the mean value for Frequency of drinking (2.6) falls into the former light drinker group for this variable, see Figure 4.1. The only exception is the Quantity variable, where the mean of the subjects falling into the new light drinker group (5.8) is slightly higher than the threshold for the light drinker group on the original scale (4.99, see Figure 4.1).

Table 4.3: Mean values of the original	variables for the new	light and heavy drinker
groups.		

	New phenotype					
	(Mean values)					
	Light drinkers Heavy drinker					ers
Original variable	female	male	all	female	male	all
Frequency of drinking	2.8	2.4	2.6	5.6	5.7	5.6
Frequency of bingeing	1.4	1.5	1.4	4.8	4.8	4.8
Frequency of drunkenness	0.8	0.8	0.8	4.4	4.5	4.5
Quantity	5.6	6.1	5.8	40.0	60.0	52.1

Table 4.3 shows further that for the three variables originating from the ESPAD questionnaire the differences between genders within the light and heavy groups are minimal (not significant). Only for the variable Quantity, males score significantly higher than females in both the light and the heavy drinker group. This is expected, since men generally drink larger alcohol quantities than women (Wilsnack et al. 2000; Wilsnack et al. 2009).

4.3.4 Validation of phenotype

For further validation of the new phenotype, we compared it to the flags from the MAST and AUDIT. For the MAST, 45 subjects were screened as positive for alcoholism out of the 816 subjects that completed the questionnaire at age 19. For the AUDIT, 399 subjects were screened as likely for alcohol abuse out of the 1511 subjects that completed the questionnaire at age 19.

Table 4.4 shows the distribution of the 45 and 399 positively screened subjects across the four original variables and the new drinking phenotype, for MAST and AUDIT, respectively. Importantly, the new phenotype shows the best correspondence to the screening results. Note that the construction of the new drinking phenotype did not, however, include any of the information from the MAST and the AUDIT.

This indicates that the new phenotype reflects abusive drinking behaviour better than the four original drinking variables and therefore gives a comprehensive and consistent assessment.

Table 4.4: The distribution across the light, moderate and heavy drinker groupings of the positively screened subjects of MAST and AUDIT, for the four original drinking variables and the new phenotype.

		MAST		AUDIT			
		(45 of 816		(399 of 1511			
	screened positive)			screened positive)			
	light	light moderate heavy			moderate	heavy	
Freq. of drinking	1	15	29	11	146	242	
Freq. of bingeing	2	8	35	12	74	313	
Freq. of drunkenness	3	9	33	25	79	295	
Quantity	1	8	29	20	75	263	
New phenotype	О	10	35	5	79	315	

4.3.5 Confounding variables

An advantage of the IMAGEN data base is the large SITE IMBALANCE sample size that was achieved by the collaboration of eight different European sites. However, a challenging aspect is revealed when looking at the distribution of the new drinking phenotype across these sites. Figure 4.3 shows the imbalance of the heavy and light drinker groups across sites - especially the sites London, Nottingham and Dublin have disproportionally many heavy drinkers, whereas the sites Berlin, Hamburg, Mannheim, Paris and Dresden are dominated by the light drinkers. A χ^2 -test on the contingency table confirms the visual impression and delivers a p-value $< 10^{-15}$ for the null-hypothesis of statistical independence between site and group. Note that this imbalance was already present for the original variables and is not an "artefact" of the new phenotype.

The differences in drinking behaviour across gender GENDER IMBALANCE were already visible to some extent in Table 4.3 and are still present in the new phenotype. Figure 4.3 depicts the distribution of heavy and light drinkers across genders. There are more female light than heavy drinkers and vice-versa for males. This is confirmed by a χ^2 -test, giving a p-value $< 10^{-12}$. Note again that this imbalance across gender was already present for the original variables and is not an "artefact" of the new phenotype.

4.4 DEFINITION OF CUMULATIVE DRINKING SCORE

So far, we have focussed on drinking behaviour assessed at FU2, i.e. at an approximate age of 19, covering the past year. A typically explored quantity is the cumulative drinking across the entire life span. This quantity is often assessed using the Lifetime Drinking History (LDH) interview-based procedure (Koenig et al. 2009; Skinner and Sheu 1982).

The LDH was not administered in the IMAGEN study. In order to nevertheless obtain some impression of the cumulative drinking for each subject, one possibility would be to resort back to the ESPAD questionnaire. In addition to assessing the drinking behaviour during the past year, it also asks for different time spans, namely the past 30 days and the whole lifetime (see Section 4.2.1). However, the disadvantage is that the ordinal scale for all three time spans (lifetime, 12 months and 30 days) is exactly the same - see Table 4.1. Hence, for the lifetime span, most subjects fall into the most extreme category (6 for Frequency of drinking and drunkenness and 5 for Frequency of bingeing, corresponding to \geq 40 and \geq 10 occasions). This does not offer a satisfactory level of differentiation to serve as a good "proxy" for life time drinking.

Thus, we construct a different "proxy" for cumulative drinking. In addition to the ESPAD assessment at FU2, we take the assessments at FU1 (approximate age of 16) and at BL (approximate age of 14) into account.

We define cumulative scores as follows:

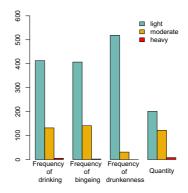
Cumulative bingeing: Sum of raw scores of Frequency of bingeing at BL, FU1 and FU2.

Cumulative drunkenness: Sum of raw scores of Frequency of drunkenness at BL, FU1 and FU2.

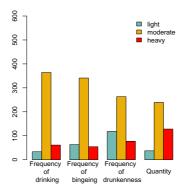
Figure 4.4 shows histograms of the resulting cumulative scores.

4.5 CHAPTER SUMMARY

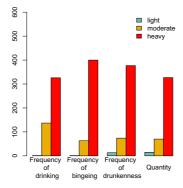
In this chapter, we introduced the various alcohol-related instruments administered in the IMAGEN study. We showed how we use a pairwise clustering algorithm to build a comprehensive binary drinking behaviour phenotype (light and heavy drinkers). Furthermore, we validated the new phenotype. Finally, we defined continuous cumulative drinking scores.



(a) New light drinker group.



(b) New moderate drinker group.



(c) New heavy drinker group.

Figure 4.2: The distribution of the new light, moderate and heavy drinker groups over the previous groupings according the the four original drinking variables.

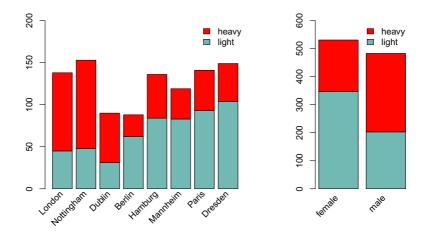


Figure 4.3: Distribution of the new drinking phenotype across sites and genders.

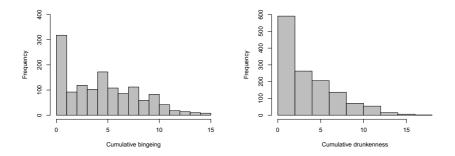


Figure 4.4: Histograms of the cumulative bingeing (left) and drunkenness (right) score.

Part III

Patterns of alcohol abuse in adolescence

5 GREY MATTER VOLUME AND PSYCHOSOCIAL VARIABLES

This chapter treats the analysis of grey matter volume and psychosocial data from the IMAGEN study, using the methods introduced in Chapter 2. Most of the contents has been published (Seo et al. 2018). For author contributions, see Appendix \mathbb{C} .

5.1 INTRODUCTION

Alcohol abuse is a top-ranked disorder of the brain with respect to total costs to economy and human suffering (Effertz and Mann 2013). Chronic alcohol use disorders (AUDs) are known to be linked with widespread brain atrophy and functional irregularities (Bühler and Mann 2011; Chelune and Parker 1981; Heinz 2002; Sullivan et al. 2003). Furthermore, adolescent-onset AUDs have also been associated to grey matter atrophy in young adulthood, specifically in the prefrontal cortex (Bellis et al. 2005).

Heavy drinking during adolescence may pose a risk factor for developing AUDs later in life (Grant et al. 2006). Furthermore, early alcohol consumption may interfere significantly with the crucial brain maturation during adolescent development (Luna et al. 2010). In fact, it has been demonstrated that even subclinical alcohol drinking during mid-to-late adolescence is related to variations in neurodevelopment across brain tissue classes and specifically to decreases in grey matter volume in several brain regions (Luciana et al. 2013).

The cited studies underline the relevance of establishing neurobehavioural risk profiles in adolescents for later alcohol abuse. These profiles can be used in order to better devise, plan and target early intervention programs with the ultimate goal of diminishing the probability of high-risk adolescents to spiral into a full-blown AUD. Importantly, in the identification of such profiles, predisposing effects have to be disentangled from alcohol-related effects, both an neural and psychosocial level.

Gender (with which we here mean biological sex) may be a pivotal aspect of the desired risk profiles. Firstly, adolescent males have shown to be at greater general risk for heavy drinking and AUDs, see e.g. Schulte et al. (2009). Further, AUDs typically begin later but then advance faster in women than in men – this is the so-called "telescoping effect", i.e. the faster progression of negative somatic alcohol-related effects (Mann et al. 2005).

On a neural level, stimulant drug addiction has been linked with increased (striatum, amygdala) and decreased (insula) grey matter volume (Ersche et al. 2010), presumptively even as a predisposing feature of addiction vulnerability. The mentioned brain regions, as well as further brain areas (e.g. anterior cingulate cortex (ACC), medial prefrontal cortex (MPFC)) are furthermore well established in the altered processing of alcohol-related cues (termed cue-reactivity) in AUD (Beck et al. 2012; Kühn and Gallinat 2011; Schacht et al. 2013; Zilberman et al. 2019). Moreover, prefrontal grey matter volume was shown to be associated to (smoking) cue reactivity (Zhang et al. 2011). Grey matter volume together with functional cue reactivity in the MPFC have been observed to be predictive of relapse in detoxified patients (Seo et al. 2015). Hence, both grey matter volume and functional activity in cue reactivity regions have been shown to be relevant in both subclinical and clinical AUD settings.

When inspecting the neural level separately for genders, it has been shown that male drinkers seem to display more cue reactivity than female drinkers (Nesic and Duka 2006). Moreover, differential gender effects have been found in local brain volume and cortical thickness between binge drinkers and controls (Squeglia et al. 2012). Variation in gray matter volume specifically in regions previously linked with a function pertinent to AUDs, such as cue reactivity, may hence be highly relevant both for establishing pre-dispoising factors to develop an AUD and also for a thorough grasp of the effects of excessive alcohol consumption in adolescence. We thus chose to focus on a cue reactivity network (Kalivas and Volkow 2005; Schacht et al. 2013) to understand the relationship between grey matter volume and heavy drinking in adolescence.

On a psychosocial level, stressful life events have been linked with an increased risk for harmful alcohol consumption and AUDs (Keyes et al. 2011). Furthermore, anxiety and impulsivity have been associated to the development of an AUD (Heinz et al. 2011). The relationship between personality traits and the risk for AUD does not seem to differ between genders (Schulte et al. 2009). However, gender differences can be found in the social context associated to alcohol consumption: males seem drink more in order to fit in with male peers and perceived gender roles, whereas females seem to do exactly the opposite (Schulte et al. 2009).

In this chapter, we aimed at identifying a profile of grey matter volume and psychosocial features associated with heavy alcohol drinking, separately for males and females using the IMAGEN study (see Chapter 3 for details). We used grey matter volume in areas previously linked with (alcohol) cue reactivity (Kühn and Gallinat 2011; Schacht et al. 2013) as well as psychosocial variables at age 14 and 19 (separately) to predict drinking behaviour at age 19. The combination of both a cross-sectional (inspecting differences between heavy and light drinkers at age 19 with respect to their neurobehavioural features at the same age) and a longitudinal (inspecting pre-existing differences between heavy and light drinkers at age 19 based on features at age 14) setting facilitated distinguishing between potentially pre-disposing markers from neurotoxic effects.

5.2 DATA

Structural brain imaging and preprocessing

Structural brain images were acquired by the eight sites of the IMAGEN study on 3-Tesla scanners. Comprehensive details and standard operating procedures have been reported previously (Schumann et al. 2010).

We preprocessed the three-dimensional T₁-weighted images using the Statistical Parametric Mapping 8 (SPM8) and the Voxel-Based Morphometry (VBM8) toolbox. We segmented the images into grey matter, white matter and cerebrospinal fluid and then transferred them into MNI space, making use of an affine registration with the International Consortium for Brain Mapping template for European brains (Template_1_IXI550_MNI152.nii). Using the high-dimensional DARTEL normalization and non-linear-only modulation, the segmented and registered images were finally normalized and modulated.

5.2.2 Personality measures

We use data from three personality inventories that were administered within the IMAGEN study: the revised self-administered 60-item NEO Five-Factor-Inventory (NEO-FFI, Costa and McCrae (1992)), the revised Temperament and Character Inventory (TCI-R, Cloninger et al. (1994)) and the Substance Use Risk Profile Scale (SURPS, Woicik et al. (2009)).

5.2.3 Life events questionnaire

To evaluate the history of stressful life events per subject, we use data from the Life Events Questionnaire (LEQ, adapted from Newcomb et al. (1981)). Events such as "parents divorced" or "changed schools" are evaluated both for happening within a certain time frame (past year or ever) and valence. Valence is assessed per item on a rating scale (-2: "very unhappy"; 1: "unhappy"; o: "neutral"; 1: "happy"; 2: "very happy").

METHODS 5.3

5.3.1 Extraction of structural imaging features

We pre-select 24 regions of interest that have been linked with cue reactivity processes specific to alcohol cues (Kühn and Gallinat 2011; Schacht et al. 2013). Based on these meta-studies, we focus on the anterior cingulate cortex, amygdala, caudate nucleus, medial prefrontal cortex, inferior-, medial-, middle- and superior orbitofrontal cortex, putamen, and thalamus from the Automated Anatomical Labeling (AAL) atlas (Tzourio-Mazoyer et al. 2002) and the anterior and posterior insula from SPM12's Neuromorphometrics Inc. atlas. For each region (left and right separately), we extract the mean value of gray matter volume as features.

We resort to this cue-reactivity network since (i) the CHOICE OF ROIS associated regions are specific to the substance of abuse (Kühn and Gallinat 2011; Schacht et al. 2013), (ii) the network has been shown to reliably predict relapse in detoxified alcohol-dependent patients, based on both functional and structural data (Seo et al. 2015) and (iii) differential gender effects have been shown in alcohol cue-reactivity (Nesic and Duka 2006).

We intentionally focused our analysis of structural (grey matter) data on regions that were chosen based on functional properties. We assume that changes in grey matter volume in areas previously associated in a function related to alcohol abuse, such as cue reactivity, might be highly relevant both for the identification of pre-disposing factors to develop alcohol use disorders, as well as for the sound understanding of the effects of excessive alcohol use. Using a priori defined (instead of data-derived) regions of interest means that the full data set can be used for further analyses, without risking double-dipping. Note that there is always a trade-off between (i) how small (and hence locally precise) we choose the regions to be and (ii) statistical power. The more fine-grained the mask, the more features would have to then be controlled for multiple testing. The more coarse the mask, the more prone we are to "averaging out" interesting local effects. We hence decided for an intermediate option, using the AAL atlas. We additionally split the insula in potentially functionally different posterior and anterior parts.

5.3.2 Extraction of personality features

From the NEO-FFI we extracted the five dimensions neuroticism, extraversion, openness to experience (in short: openness), agreeableness and conscientiousness (Costa and McCrae 1992). The data from the TCI-R was used to extracted the four sub-dimensions of the Novelty-Seeking scale. More specifically, we extracted exploratory excitability impulsiveness, extravagance and disorderliness (Cloninger et al. 1994). From the SURPS we extracted the four dimensions hopelessness, anxiety sensitivity, impulsivity and sensation seeking (Woicik et al. 2009).

5.3.3 Extraction of life events features

From the LEQ we extracted the mean frequency and valence of the summary scores accident/illness (e.g. a death in the family), sexuality (e.g. lost virginity), autonomy (e.g. found a new group of friends), deviance (e.g. got in trouble at school), distress (e.g. gained a lot of weight), and relocation (e.g. family moved).

5.3.4 Summary of features

For an overview of number of subjects available see Table 5.1. The number in brackets indicates the number of female subjects available for each cell. Note that there are much fewer female heavy drinkers compared to male heavy drinkers. This imbalance is present in general for the drinking phenotype defined on the IMAGEN data set, see Chapter 4 for more details. We dealt with this imbalance in prediction analysis by using a bootstrapped version of cross-validation (BSSCV), see Section 2.4.3. For association analysis, we included gender (and also the scanning site) as additional covariate(s), see Section 2.5.2 for details. For a list of features, see Tables 5.2, 5.3 and 5.4. All missing values were mean-imputed, see Section 2.2.2 for brief methodological details.

Table 5.1: Sample sizes for each data modality – grey matter and psychosocial – and
time point. Values in brackets indicate the number of female subjects.

	Phenotype	Grey matter	Grey matter	Psychosocial	Psychosocial
	age ≈ 19	age ≈ 14	age ≈ 19	age ≈ 14	age ≈ 14
All	1472	2073	1387	2128	2166
	(772)	(1054)	(713)	(1087)	(1075)
Light	550	539	498	550	550
drinkers	(347)	(341)	(316)	(347)	(347)
Heavy	464	412	452	464	464
drinkers	(184)	(153)	(178)	(184)	(184)

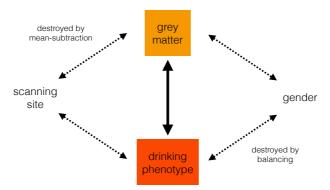


Figure 5.1: Overview of how we removed the effect of the confounders site and gender in prediction analysis.

Table 5.2: Summary of personality variables and their meanings and origins. NEO-FFI: NEO Five-Factor-Inventory; TCI-R: revised Temparament and Character Inventory; SURPS: Substance Use Risk Profile Scale.

Name	Meaning	Origin
neo_neuroticism	Tendency to experience negative emotions	NEO-FFI
neo_extraversion	Engagement with the outside world	NEO-FFI
neo_openness	Openness to variety of experiences	NEO-FFI
neo_agreeableness	Concern for social harmony	NEO-FFI
neo_conscientiousness	Tendency to be organised and dependable	NEO-FFI
tci_impulsivity	Tendency to act with little forethought	TCI-R
tci_disorderliness	Tendency for lack of organisation	TCI-R
tci_extravagance	Engagement with the outside world	TCI-R
tci_excitability	Tendency to be readily roused into action	TCI-R
surps_anxiety_sensitivity	Tendency to be anxious easily	SURPS
surps_hopelessness	Tendency to feel hopeless	SURPS
surps_impulsivity	Tendency to act with little forethought	SURPS
surps_sensation_seeking	Tendency to search for new experiences	SURPS

Table 5.3: Summary of life events variables and their meanings, origin and percentage of missing values. LEQ: Life Events Questionnaire.

Name	Meaning	Origin	% Missing BL	% Missing FU2
leq_accident_freq	Frequency of accidents and illnesses in past year	LEQ	0.21	0.00
leq_accident_valence	Valence of accidents and illnesses	LEQ	0.62	0.21
leq_autonomy_freq	Frequency of autonomous behaviours in past year	LEQ	0.00	0.00
leq_autonomy_valence	Valence of autonomous behaviours	LEQ	0.00	0.00
leq_deviance_freq	Frequency of deviant behaviours in past year	LEQ	0.00	9.42
leq_deviance_valence	Valence of deviant behaviours	LEQ	0.41	8.17
leq_distress_freq	Frequency of stressful behaviours in past year	LEQ	0.00	0.21
leq_distress_valence	Valence of stressful behaviours	LEQ	0.00	0.00
leq_relocation_freq	Frequency of relocations in past year	LEQ	0.93	0.31
leq_relocation_valence	Valence of relocations	LEQ	0.83	0.31
leq_sexuality_freq	Frequency of sexual experiences in past year	LEQ	0.00	0.00
leq_sexuality_valence	Valence of sexual experiences	LEQ	0.00	0.00

Table 5.4: Summary of grey matter volume variables and their meanings. AAL: Automated Anatomical Labeling atlas.

Name	Meaning	Atlas	Name in atlas
ACC_L/R	Anterior cingulate cortex (left / right)	AAL	Cingulum_Ant_L/R
Amygdala_L/R	Amygdala (left / right)	AAL	Amygdala_L/R
Caudate_L/R	Caudate nucleus (left / right)	AAL	Caudate_L/R
Insula_Ant_L/R	Anterior insula (left / right)	Neuro- morpho- metrics	Insula_Ant_L/R
Insula_Post_L/R	Posterior insula (left / right)	Neuro- morpho- metrics	Insula_Post_L/R
MPFC_L/R	Medial prefrontal cortex (left / right)	AAL	Frontal_Sup_Medial_L/R
OFC_Inf_L/R	Inferior orbitofrontal cortex (left / right)	AAL	Frontal_Inf_Orb_L/R
OFC_Med_L/R	Middle orbitofrontal cortex (left / right)	AAL	Frontal_Med_Orb_L/R
OFC_Mid_L/R	Middle orbitofrontal cortex (left / right)	AAL	Frontal_Mid_Orb_L/R
OFC_Sup_L/R	Posterior orbitofrontal cortex (left / right)	AAL	Frontal_Sup_Orb_L/R
Putamen_L/R	Putamen (left / right)	AAL	Putamen_L/R
Thalamus_L/R	Thalamus (left / right)	AAL	Thalamus_L/R

SITE AND GENDER EFFECTS Note that we found the above regions of grey matter volume to be highly predictive of both the scanning site and gender. Furthermore, we have seen that there is an imbalance of the drinking behaviour phenotype and gender across the eight different scanning sites (see Chapter 4 for details). If blindly performing classification for the drinking phenotype on the data (without accounting for these counfounders), the obtained accuracies would be inflated. The intrinsic site and gender information contained in the grey matter volume data would be used by the classifier. It would associate a subject with a site and gender and would then classify the subject to the majority class of this site-gender combination.

We account for these effects as follows (also depicted in Figure 5.1):

- (i) For prediction analysis in scenarios with both genders, we balanced for gender in each training and testing fold using BSSCV, see Section 2.4.3. In order to account for the confound scanning site, we chose a different approach. Subsampling for site as done for gender by BSSCV is computationally intensive and would lead to much smaller data sets due to the imbalance of label across sites. Instead, we attempted to destroy the intrinsic site information in the grey matter data by subtracting the respective site-mean of grey matter for each subject for each feature. We showed that such a linear model suffices to remove virtually all of the intrinsic site information (see Appendix A.1 for details).
- (ii) For association analysis we added site (and gender for scenarios with both genders) as additional covariates into the ReLL model (see Appendix 2.5.2).

5.3.5 Prediction analysis

For prediction analysis we relied on the methods described in Section 2.3. As classifiers we used linear support vector machines (see Section 2.3.5), naïve Bayes (see Section 2.3.3) and robust soft learning quantisation (RSLVQ) with one prototype per class (see Section 2.3.6). We used K = 10 folds for cross-validation. On each training and test fold we balanced for class and in scenarios including both genders we additionally balanced the confound gender using BSSCV (see Section 2.4.3) with B = 10 bootstraps. For significance testing of accuracies (see Section 2.4.4) we used $N_{perm} = 10000$. We corrected for multiple testing via the Benjamini-Hochberg method at level $\alpha = 0.05$ (see Section 2.6). The p-values of non-selected features in the association analysis (i.e. if $\bar{\beta} = 0$) were set to 1.0. Furthermore, to ensure equal weighting of features, we z-scored each feature on the training folds and applied the resulting z-scoring to the respective test folds.

5.3.6 Association analysis

For association analysis, we used our ReLL-method (see Section 2.5.2) with $N_{rep} = 100$. We again correct for multiple testing via the Benjamini-Hochberg method at level $\alpha = 0.05$ (see Section 2.6). Furthermore, we z-scored each feature prior to the association analysis.

RESULTS 5.4

Prediction analysis

In the prediction analysis, we investigated the predictive power of grey matter volume and psychosocial data, respectively, for the discrimination of heavy versus light drinkers.

GREY MATTER DATA Figures 5.2 and 5.3 show the prediction accuracies (plotted as yellow triangles) for the grey matter volume features in the crosssectional and longitudinal setting, respectively.

In the cross-sectional setting (i.e. using features assessed at approximate age 19 to predict the drinking behaviour at the same age), using all subjects all three classifiers yield low, yet significant prediction accuracies. In the gender separate cases, the predictive power of grey matter volume in female subjects seems to be higher than in male subjects – all three classifiers show significant accuracies for females and only one for males (naïve Bayes). The highest accuracy is obtained by the linear SVM for the female only case (balAcc = 0.59).

In the longitudinal setting (i.e. using features assessed at approximate age 14 to predict the drinking behaviour at approximate age 19), none of the classifier in none of the cases (all, only females, only males) give significant accuracies.

Our analysis indicates that cue-reactivity ROIs from grey matter volume data is predictive for drinking behaviour only in the cross-sectional and not in the longitudinal setting. Hence, according to the framework proposed in Section 2.1, we infer that only the cross-sectional setting provides generalisable discrimination between heavy and light drinkers. We thus take a closer look at this setting in the association analysis.

PSYCHOSOCIAL DATA Figures 5.4 and 5.5 show the prediction accuracies for the grey matter volume features in the cross-sectional and longitudinal setting, respectively. In the cross-sectional setting (i.e. using features assessed at approximate age 19 to predict the drinking behaviour at the same age), the classifiers yield much higher prediction accuracies than for the grey matter case. All cases are significant. Prediction for males seems to be slightly better than for females. The highest accuracy is obtained by the linear SVM for the male only case (balAcc = 0.70).

In the longitudinal setting (i.e. using features assessed at approximate age 14 to predict the drinking behaviour at approximate age 19) using all subjects, all three classifiers yield lower, yet still significant prediction accuracies. The only non-significant case is the RSLVQ for males only. Prediction accuracies do not differ much between genders. The highest accuracy is obtained by naïve Bayes for the male only case (balAcc = 0.59).

We thus conclude that psychosocial data is predictive for drinking behaviour in both the cross-sectional and the longitudinal setting. Hence, according to the framework proposed in Section 2.1, we take a closer look at both settings in the association analysis.

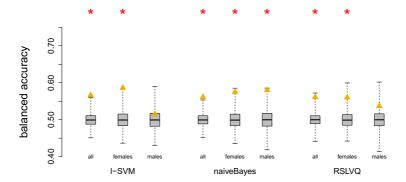


Figure 5.2: Balanced accuracies (as yellow triangles) for prediction of heavy versus light drinkers from grey matter features for the cross-sectional setting, based on three classifiers: linear support vector machine (I-SVM), naïve Bayes and robust soft learning vector quantisation (RSLVQ) and for all subjects, only females and only males. Grey box plots (spanning the entire range) indicate prediction accuracies obtained by label permutation. The farther away the yellow triangles (true accuracies) are from the grey box plots (permuted accuracies), the more likely it will survive hypothesis testing and correction for multiple testing. Significant cases after controlling for FDR at 0.05 are indicated by a red star.

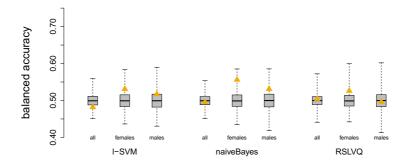


Figure 5.3: Balanced accuracies (as yellow triangles) for prediction of heavy versus light drinkers from greymatter features for the longitudinal setting, analogous to Figure 5.2.

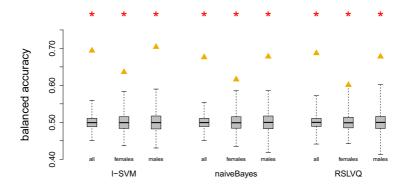


Figure 5.4: Balanced accuracies (as yellow triangles) for prediction of heavy versus light drinkers from psychosocial features for the cross-sectional setting, analogous to Figure 5.2.

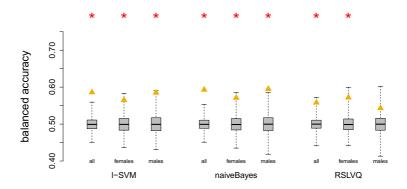


Figure 5.5: Balanced accuracies (as yellow triangles) for prediction of heavy versus light drinkers from psychosocial features for the longitudinal setting, analogous to Figure 5.2.

5.4.2 Association analysis

In the association analysis, we examined predictive settings more closely. More specifically, we evaluated the influence of single variables on the drinking behaviour using our ReLL-method 2.5.2.

GREY MATTER DATA Figure 5.6 shows the association between grey matter volume in cue-reactivity ROIs and drinking behaviour in the cross-sectional setting. All non-zero parameter estimates are negative. This indicates extensive gray matter reduction in the heavy compared to the light drinker class. For the case with all subjects, the bilateral ACC, MPFC, median, middle and superior OFC, thalamus as well as the left amygdala and anterior insula and the right inferior OFC present significantly lower grey matter volume in the heavy drinker compared to the light drinker class.

Inspecting the results separately per gender, it becomes apparent that the significance in all-subject case is almost solely driven by the female subjects. The female heavy drinker class shows significantly lower grey matter volume in the bilateral ACC, medial OFC and thalamus, as well as the left amygdala, MPFC, inferior and superior OFC. For the male subjects, much fewer features yielded non-zero parameter estimates and they show significantly lower grey matter volume only in the bilateral medial OFC.

Note that e.g. for the right MPFC the parameter estimate for the female only-group is larger than for the analysis including all subjects. Counterintuitively, the case including all subjects is deemed significant, whereas the female-only case is not significant. This may happen due to the difference in sample sizes, see Table 5.1.

Figures 5.7 and 5.8 show the association between psy-PERSONALITY DATA chological features and drinking behaviour in the cross-sectional and longitudinal setting, respectively.

In the cross-sectional setting for all three subject cases (all, only males, only females), six personality traits (NEO-extraversion, TCI-impulsivity, extravagance and -excitability, SURPS-impulsivity and -sensation-seeking), are significantly positively associated with heavy drinking and the male group shows a stronger association than the female group. Furthermore, NEO-agreeableness and NEO-conscientiousness are significantly associated to heavy drinking in a negative direction. NEO-agreeableness is no longer significant when assessed separately for each gender.

In the longitudinal setting, the pattern is similar. The male group again shows more and stronger significant association between personality traits and drinking behaviour. Note that SURPS-hopelessness is positively associated for all three subject cases, whereas there was no association in the cross-sectional setting.

LIFE EVENTS DATA Figures 5.9 and 5.10 show the association between stressful life event features and drinking behaviour in the cross-sectional and longitudinal setting, respectively.

In the cross-sectional setting and for all subject groups, there is significant positive association of frequency of autonomy, frequency of deviance, valence of deviance and frequency of sexuality to heavy drinking. Valence of autonomy positively associated for all subjects and male subjects only. Furthermore, valence of sexuality shows significant positive association to heavy drinking for female subjects only.

In the longitudinal setting there is no significant association between any of the stressful life event features and drinking behaviour.

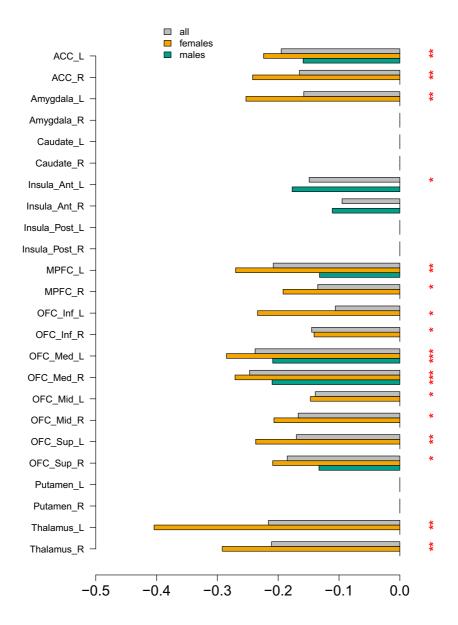


Figure 5.6: Median parameter estimates from association analysis on **grey matter** features using the ReLL-method in the **cross-sectional** setting. Non-zero estimates are obtained if the median of the parameter estimates over 100 runs is non-zero. Significant cases after correction for multiple testing are indicated by a red star.

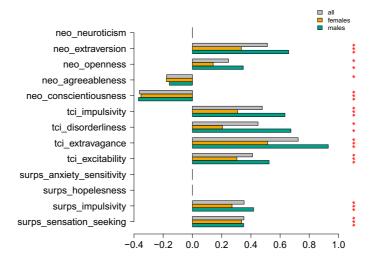


Figure 5.7: Median parameter estimates from association analysis on personality features in the cross-sectional setting, analogous to Figure 5.6.

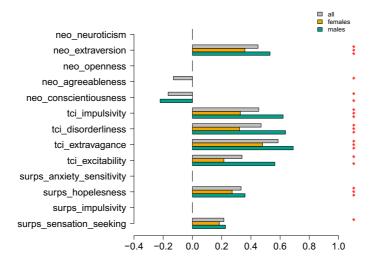


Figure 5.8: Median parameter estimates from association analysis on personality features in the longitudinal setting, analogous to Figure 5.6.

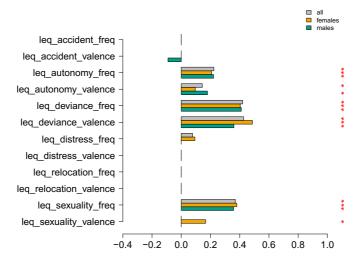


Figure 5.9: Median parameter estimates from association analysis on **life events** features in the **cross-sectional** setting, analogous to Figure 5.6.

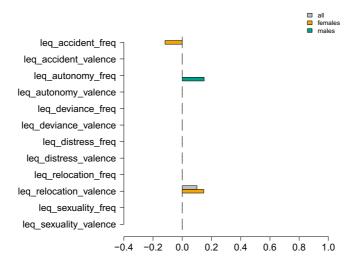


Figure 5.10: Median parameter estimates from association analysis on **life events** features in the **longitudinal** setting, analogous to Figure 5.6.

5.4.3 Further investigations

CUMULATIVE DRINKING BEHAVIOUR In an additional (post-hoc) investigation, we explored the relationship between a cumulative drinking score and grey matter volume in regions that were shown to be significantly associated to heavy drinking (see Section 5.4.2) at FU2. We did this for all subjects, and for females and males separately - looking in each case only at the subset of significant regions from the association analysis. For details on how we computed the cumulative drinking score, see Chapter 4.

Figures 5.11, 5.12 and 5.13 show the results from separate linear regression analyses between the cumulative bingeing score (target variable) and each grey matter volume (explanatory variables), corrected for FDR at $\alpha =$ 0.05 over all 28 tests (16 for all subjects, 10 for females and 2 for males), including site and (for the case of all subjects) gender as covariates. In Figures 5.11, 5.12 and 5.13 we plotted the estimated coefficients and confidence intervals of the grey matter volumes resulting from each linear regression model estimation. Significant cases after multiple testing correction are indicated by a red star.

We see in all cases only negative relationships, indicating that more life time drinking is associated to lower grey matter volume. For all subjects, we see significant reduction for the bilateral ACC, anterior insula, MPFC, median OFC, for the right superior OFC and for the left thalamus. For only females, we see significant reduction again in the bilateral ACC and median OFC and furthermore in the left MPFC and thalamus. For only males, we see significantly negative correlation between cumulative bingeing and grey matter volume in the bilateral median OFC – the two only regions identified as significant for only males in the association analysis.



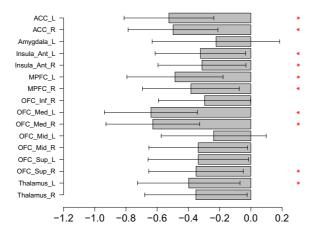


Figure 5.11: Relationship of significant grey matter features to the cumulative bingeing score. Coefficient estimates and confidence intervals from a linear regression analysis between the cumulative drunkenness score (target variable) and each cognitive feature (explanatory variables), including gender and site as covariates. Significant estimates after correction for multiple testing are indicated by a red star.

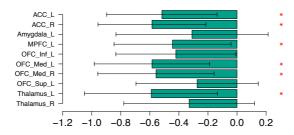


Figure 5.12: Relationship of significant grey matter features for female subjects to the cumulative bingeing score. Analogous to Figure 5.11.

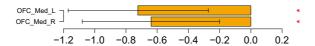


Figure 5.13: Relationship of significant grey matter features for male subjects to the cumulative bingeing score. Analogous to Figure 5.11.

CORRELATION STRUCTURE In an additional analysis, we investigated the correlation structure between psychosocial features and grey matter volume, both at FU2 and BL. In order to account for site effects, we use sitemean corrected grey matter volume (see Figure 5.1). We controlled the FDR at level $\alpha = 0.05$.

Figure 5.14 shows the correlation at FU2 between personality features (see Table 5.2 for details) and site-corrected grey matter volume for all subjects. Significant negative correlation is observed between TCI-disorganization with grey matter volume in various brain regions (bilateral MPFC and median OFC, left inferior OFC and right ACC and amygdala). Furthermore, the analysis shows significant negative correlation between TCI-extravagance and grey matter volume in the left median OFC. NEO-agreeableness shows significant negative correlation to the bilateral putamen and positive correlation to the bilateral thalamus.

Correlations between grey matter volume and personality features were not significant at BL. Moreover, no correlations between grey matter volume and life events were significant for FU2 or BL.

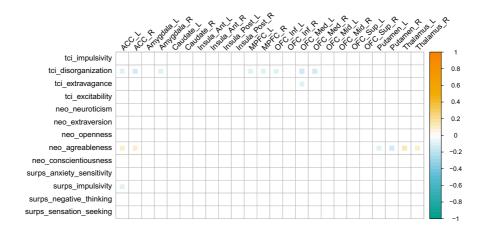


Figure 5.14: Pearson's correlations between personality and site-corrected grey matter volume features, both assessed at FU2. Only significant correlations (after FDRcorrection) are shown.

DISCUSSION 5.5

Problematic drinking regularly begins in adolescence (Petit et al. 2014), an especially crucial (Luna et al. 2010) and vulnerable (Squeglia et al. 2009) phase of brain development. Furthermore, early onset of heavy drinking may constitute a risk factor for the development of alcohol use disorders later in life (Grant et al. 2006).

In this chapter, we explored the predictive and associative relationship of neural and psychosocial variables to heavy drinking in a large adolescent

Our main findings are that heavy drinking at age 19 is significantly associated with:

- (i) grey matter reduction in cue reactivity relevant brain areas at age 19 (but not at age 14), especially for female subjects,
- (ii) psychological variables like impulsivity and extravagance at age 14 and at age 19, in both genders,
- (iii) and stressful life events assessed at age 19 but not at age 14, in both genders.

GREY MATTER VOLUME We have shown that grey matter volume of 19 year olds in areas relevant to cue reactivity was predictive of drinking behaviour at the same age (cross-sectionally) for all subjects, and for each gender separately. Longitudinally, the same areas' grey matter volume at age 14 was not predictive for drinking behaviour at age 19. The evaluation of cross-sectional association between single grey matter volume features and drinking behaviour revealed reduced grey matter volume in bilateral ACC, medial, middle and superior OFC, MPFC, and thalamus, and further in the left anterior insula and right inferior OFC in heavy drinking subjects. This effect was driven mainly by the female subjects. Since the longitudinal setting, we did not find any relationship between grey matter volume at age 14 and heavy drinking at age 19, we suggest that the reduced grey matter volume in cue reactivity related areas of heavy drinking adolescents might be a consequence of their alcohol drinking behaviour, i.e. a potential neurotoxic effect. Neurotoxic effects have been reported frequently in alcoholism (Harper et al. 1985; Jernigan et al. 1991; Pfefferbaum et al. 1992). Also for sub-clinical (binge) drinking, studies have previously observed this effect in cingulate cortex (Mashhoon et al. 2014), the caudate nucleus (Squeglia et al. 2014), and frontal, temporal, cerebellum and brain stem areas (Heikkinen et al. 2017; Luciana et al. 2013; Squeglia et al. 2014; Squeglia et al. 2015). Only few studies have reported increased grey matter volume or cortical thickness associated to binge drinking (Loheswaran et al. 2016).

Importantly, we have seen in our study that the reduction in brain volume of heavy compared to light drinkers was much more pronounced in female than in male subjects at age 19. This is potentially similar to the telescoping effect of the development of alcohol use disorders in female adults, where alcohol-dependent women spiral faster from their onset of drinking to alcohol-related problems and finally into treatment (Hernandez-Avila et al. 2004; Johnson et al. 2005; Piazza et al. 1989). It has previously been found that also atrophies seem to develop faster in alcohol-dependent women than in men (Mann et al. 2005), in line with our findings.

Note that less grey matter volume reduction in males in response to heavy drinking does not necessarily imply that they are less susceptible for developing an alcohol use disorder later in life. In fact, the opposite might be the case, since a high tolerance for alcohol may be a risk factor for the development of alcohol use disorders (Schuckit et al. 2017; Schulte et al. 2009).

PSYCHOSOCIAL DATA We found that psychosocial variables can significantly predict drinking behaviour, both cross-sectionally and longitudinally.

Various personality traits at age 14 and at age 19 were significantly associated with drinking behaviour at age 19. Specifically, subjects that were more extroverted, hopeless, impulsive, disordered, and extravagant at age 19 were shown to have a higher chance of being heavy drinkers at age 19 across both genders. A similar pattern could be seen at age 14, except for hopelessness. On the other hand, reduced conscientiousness was found to be associated to heavy drinking at age 19 across both genders. Our results are in line with many studies that have found relationships, both crosssectionally and longitudinally, of harmful alcohol drinking in adolescents to sensation seeking, agreeableness and neuroticism (Kuntsche et al. 2006) and to conscientiousness and extraversion (Stewart and Devine 2000).

Furthermore, our results agree with the notion that sub-dimensions of novelty seeking (excitability, disorderliness, extravagance) differentiate heavy drug abusers from controls that show only similarly high scores on impulsivity (Ersche et al. 2010), whereas we also observe this effect longitudinally.

Interestingly, while more hopeless subjects at age 14 are more likely to be heavy drinkers at 19 (longitudinally), this effect is not present crosssectionally, i.e. hopelessness at age 19 is not associated to drinking behaviour at the same age. Hopelessness in combination with sensation seeking at ages between 11 and 15 have been linked to higher drinking probability at the same age (Malmberg et al. 2010). In combination with our new results, this may suggest a lasting effect of hopelessness in early adolescence, influencing the likelihood of drinking also later in life.

The association between various personality traits and heavy drinking underscores the relevance of personality-based intervention programs (Conrod et al. 2010).

Stressful life events showed significant association to drinking behaviour only in the cross-sectional setting. In females, the subscales autonomy, deviance and sexuality showed positive association to heavy drinking, both in frequency and valence. For males, the same pattern was observed, except for valence of sexuality, which did not show a significant relationship to the drinking behaviour at age 19.

Interestingly, all these significantly associated subscales (autonomy, deviance and sexuality) have an internal locus of control, i.e. they are active stressors. The remaining subscales that were not shown to be associated to heavy drinking (accident, distress relocation) have an external locus of control, i.e. they are passive stressors. A limited influence of passive stressors on alcohol drinking behaviour in adolescents has been found previously (Hoffmann 2016). A further study did find some effects of stressful life events on the likelihood of following a more harmful drinking trajectory. However, the stressful life events score in that study comprised both active and passive stressors (Windle et al. 2005).

In the longitudinal setting, none of the stressful life events assessed at age 14 were associated with drinking behaviour at age 19. However, this negative finding does not rule out that more detailed prospective studies may find an influence of stressful life events (Hoffmann 2016; King and Chassin 2008; Windle et al. 2005).

We conclude that there is no simple influence of stressful life events on harmful drinking. Mediating and moderating effects like external or internal loci of control, externalising and internalising symptoms (King and Chassin 2008) and coping strategies have to be incorporated to shed more light the effects of stressful life events (Corbin et al. 2013).

In a post-hoc investigation we ob-CUMULATIVE DRINKING BEHAVIOUR served negative correlations between a score for cumulative drinking (incorporating information from drinking behaviour at 14, 16 and 19) and the areas of grey matter volume that were significantly associated with the phenotype capturing the drinking behaviour at age 19. This finding supports our interpretation that the reduction in grey matter volume may be a consequence of heavy drinking, i.e. a neurotoxic effect as seen in other studies

investigating subclinical drinking (Heikkinen et al. 2017; Luciana et al. 2013; Mashhoon et al. 2014; Squeglia et al. 2014; Squeglia et al. 2015).

5.6 CHAPTER SUMMARY

The results from the analyses on the IMAGEN data base indicate a reduction in grey matter volume through heavy drinking in adolescence in areas relevant to cue reactivity. Interestingly, this reduction seems to be stronger in females than in males. Further studies should explore how these grey matter reductions can be linked to and potentially augment the diagnosis of alcohol-related cognitive deficits. We have moreover found that several personality traits at age 14 and 19 could be seen as a risk factor for heavy drinking at age 19. We thereby add to a large corpus of research indicating similar results. The administration of personality questionnaires is easy and comparatively cheap. Early interventions could be aimed at adolescents at age 14 who consider themselves being impulsive, disorderly, extraverted or hopeless to prevent them from alcohol abuse later in adolescence. Moreover, the communication of coping strategies for stressful life situations, specifically those with an internal locus of control, may be an additional approach to avert heavy drinking in adolescence.

6 FUNCTIONAL IMAGING AND COGNITIVE VARIABLES

This chapter treats the analysis of functional imaging and cognitive data from the IMAGEN study, using the methods introduced in Chapter 2. We focus especially on features related to response inhibition.

6.1 INTRODUCTION

Excessive alcohol consumption has a detrimental effect on public health; globally, it is linked to an estimated 3.8 % of deaths and 4.6 % of disability-adjusted life years (Rehm et al. 2009). Furthermore, alcohol abuse is a top-ranked disorder of the brain with respect to total costs to economy (Effertz and Mann 2013). Many of the costs and damages are driven by the dangerous behaviour during acute intoxication, e.g. involvement in fatal traffic accidents. Such risky behaviour under the influence of alcohol is also a consequence of loosened inhibitory control (Field et al. 2010; Weafer and Fillmore 2012).

Inhibitory control, i.e. the executive function allowing for inhibition of natural or habitual responses to stimuli with the effect of choosing a more appropriate ("sensible") behaviour, is not only affected by alcohol consumption, but may also play a key role in the initiation of a harmful drinking trajectory. Such trajectories often begin in adolescence and young adulthood, a period of immature inhibitory functioning (Field et al. 2010; López-Caneda et al. 2014) and a time during which heavy drinking poses a risk factor for the development of alcohol use disorders (AUDs) (Grant et al. 2006).

On a cognitive level, response inhibition, i.e. the ability to suppress responses that are already initiated, is commonly measured by the stop signal task (SST) and the go/no-go task (Murphy et al. 1999; Verbruggen et al. 2008). Both tasks require fast and repeated responses to stimuli, but demand a suppression of those responses in a fraction of trials. While the go/no-go task presents two different groups of stimuli, one for which subjects are instructed to respond and one for which they should withhold their response, the SST requires subjects to inhibit a response that has already been initiated. For the go/no-go task, the number of inappropriate

responses to a no-go stimulus, and for the SST, the time needed to stop a response once it has been initiated (stop signal reaction time, SSRT), are popular markers for inhibitory control, see Schulz et al. (2007) and López-Caneda et al. (2014) for a review. Importantly, impaired response inhibition has also been linked to intellectual disabilities (Bexkens et al. 2014). It is hence important to integrate measurements of intelligence as potential mediators of response inhibition.

On a neural level, inhibitory control is often measured by functional brain activity during the SST (Aron and Poldrack 2006; Chevrier et al. 2015; Li et al. 2006; Sharp et al. 2010; Verbruggen and Logan 2008). The frontostriatal network seems to be particularly involved in the inhibition of initiated responses (Aron et al. 2007; Chambers et al. 2009) and there are indications for abnormalities in this network in alcohol abuse (Bednarski et al. 2012; Hu et al. 2015; Li et al. 2009; Noël et al. 2001; Pfefferbaum et al. 2001). Furthermore, an established finding is the inverse scaling of the SSRT with activity in various brain regions during successful stopping (Aron and Poldrack 2006; Li et al. 2006; Rubia et al. 2007).

A central issue, comprising both cognitive and neural levels, is whether poor inhibitory control is a cause or a consequence of excessive alcohol use. Identifying weakened inhibition as cause or risk factor may allow for more tailored prevention strategies, while establishing it as a consequence would possibly call for targeted intervention and training (Houben et al. 2011). A recent review (López-Caneda et al. 2014) suggests that poor inhibitory control may be both cause and consequence of alcohol abuse; in the worst case this interplay may trap heavy drinkers in a vicious cycle. The authors encourage further longitudinal research on adolescent subjects to properly investigate the interaction between impaired inhibitory control and alcohol use.

In this chapter, we used the large-scale, longitudinal IMAGEN database (see Chapter 3 for details) to identify the link between heavy alcohol drinking in adolescence and a range of cognitive and functional imaging markers of inhibitory control. We used features at age 14 and 19 (separately) to predict drinking behaviour at age 19. The combination of both a crosssectional (inspecting differences between heavy and light drinkers at age 19 with respect to their markers of inhibitory control at the same age) and a longitudinal (inspecting pre-existing differences between heavy and light drinkers at age 19 based on features at age 14) setting facilitated distinguishing between cause and effect of excessive alcohol consumption.

6.2 DATA

6.2.1 Wechsler intelligence scale

The Wechsler intelligence scale for children was developed by David Wechsler as an intelligence test for children between age 6 and 16 (Wechsler 1949). In the IMAGEN study, subtests of the fourth version (WISC-IV) were administered (Wechsler 2003). More specifically, two indices were tested – Verbal Comprehension Index (VCI) and the Perceptual Organization Index (PCI). Of the VCI, two subscales were administered – "Vocabulary" and "Similarities". Of the PCI, also two subscales were administered – "Block Design" and "Matrix Reasoning". For details on these tasks, see e.g. Wechsler (2003). Note that the Wechsler intelligence scale was administered only once in the IMAGEN study at an approximate age of 14.

6.2.2 Affective qo-no/qo task

The affective go/no-go (AGN) task is a variant of the classical go/no-go task and measures signatures of response inhibition (Murphy et al. 1999). In the classical version, participants are instructed to perform a motor response, such as a button press, when a stimulus from a certain "target" class is shown, and to withhold the motor response if the stimulus is of another, "non-target", class. Typically, the number of trials showing stimuli the from target class is much larger than the number of "non-target" trials. Hence, participants are generally focussed on a quick reaction and thereby make many false responses in "non-target" trials. The number of false responses can be used as a signature of response inhibition.

The affective go/no-go task was originally proposed as the affective shifting task by Murphy et al. (1999). The stimuli are words of either positive or negative nature. The alteration to the classical version is that the "target" class switches between blocks, hence requiring more mental flexibility from participants. One of the advantages of this more difficult setup, is that fewer trials are needed for obtaining the same number of false responses.

In the IMAGEN study, an implementation of the affective go/no-task from the Cambridge Neuropsychological Test Automated Battery (CANTAB 2018) was used.

6.2.3 Stop signal task

The stop signal task (SST) is an extension of the go/no-go task and was first introduced by Lappin and Eriksen (1966). Since, there have been extensive developments and refinements to the SST paradigm (see Logan (1994), Logan et al. (1984), Verbruggen and Logan (2008), Verbruggen and Logan (2009), and Verbruggen et al. (2008)). The paradigm requires subjects to react as fast as possible to a "go"-stimulus. In a fraction of trials, the "go"stimulus is followed by a "stop"-stimulus, and subjects have to inhibit their previously initiated go-response.

The IMAGEN study uses left and right arrows as go-stimuli and participants are required to react with corresponding button presses (a left and a right button press). The stop signal is displayed as an arrow pointing upwards. The time between go-stimulus and stop-signal presentation (the so-called stop-signal-delay) is adjusted according to the subjects' performance. If a subject inhibits a go-response on a stop-trial successfully, the next stop-signal-delay is prolonged by 50 ms (making stopping more difficult). If a subjects falsely responds in a stop-trial, the next stop-signal-delay is reduced by 50 ms (making stopping easier). This tracking-procedure produces stop-success rates of around 50% (Levitt 2005).

A block contained 400 go-trials and approximately 80 stop-trials (with varying stop-signal-delay) for the BL-acquisition and 300 go-trials and approximately 60 stop-trials for the FU2-acquisition. If a subject responded before the stop-stimulus was shown, then that particular stop trial (with the set stop-signal-delay) was repeated up to seven times. Subjects were instructed not to wait for the upwards arrow, but to try to respond as accurately and quickly as possible to the presented stimuli. The maximum time window for response was 1000 ms.

Note that this task was performed in an fMRI-scanner (see Section 6.2.4 for details). Subjects were previously familiarised with the task in a practice session of 60 trials outside of the scanner.

6.2.4 Preprocessing and first level analyses of functional brain images

Functional brain images were acquired by the various sites of the IMAGEN study on 3-Tesla scanners using Echo Planar Imaging (EPI). Comprehensive details, standard operating procedures and quality checks have been reported previously (Schumann et al. 2010).

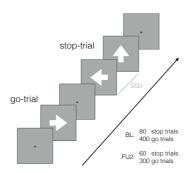


Figure 6.1: Schematic representation of the stop signal task. Subjects were required to respond as quickly and as correctly as possible to the go-stimulus (arrows pointing right or left). During a stop trial, a stop signal (an arrow pointing upwards) was presented at a certain stop-signal-delay (SSD) after the onset of the go-stimulus. Subjects had to stop the already initiated motor response.

PREPROCESSING Preprocessing was performed centrally by the IMAGEN consortium (Schumann et al. 2010) as follows: The pre-processing of the EPI data were performed with Statistical Parametric Mapping 12 (SPM12) software. Time series data were first corrected for slice-timing, then corrected for movement (spatial realignment), non-linearly warped onto standard Montreal Neurological Institute (MNI) space, and smoothed with a Gaussian kernel at 5mm full-width-at-half-maximum (FWHM). Estimated movement parameters (three translations, three rotations, three translations shifted one volume acquisition before and three translations shifted one volume acquisition later) were added as nuisance variables. Furthermore, automatic spike detection was applied to each fMRI time series (see Appendix B.2 for details).

FIRST LEVEL ANALYSES Activation maps were computed with SPM12, and regressed using a general linear model with an autoregressive noise model. Based on behavioural records, each subject's design matrix included regressors for stop success trials, stop failure trials, trials on which the go response was too late, trials on which the go response was wrong (if any) and the nuisance variables. The regressors modeling the experimental conditions were convolved using SPM's default hemodynamic response function. No motor responses are modeled since the timings involved would not yield a regressor orthogonal enough to the other stimuli.

We investigated two of the contrasts centrally computed and suggested by the IMAGEN consortium Schumann et al. 2010:

- stopSuccess: stop success minus an implicit baseline
- stopSuccess-stopFail: stop success minus stop failure

Feature extraction from these contrasts is described in Section 6.3.2.

6.3 METHODS

6.3.1 Extraction of cognitive features

Various features were extracted from the tasks described in Section 6.2, separately for BL and FU2 (except for the IQ tasks since the Wechsler intelligence scale was administered only at BL). Some were available in a straightforward fashion (as the reaction time in the AGN-task), others (as the stop-signal-reaction-time) involved modelling, see below. Note that all features were extracted on a per subject basis.

EXTRACTION OF IQS From the results of the Wechsler intelligence scale two features were extracted: an estimation of the performance IQ (wechsler_piq) and an estimation of the verbal IQ (wechsler_viq). Since only a subset of the full Wechsler scale was measured, the respective subsets were used as an approximation of the performance and verbal IQ, as is common practice (Ward 1990; Whelan et al. 2012). Scores for the single subtests (Vocabulary, Similarities, Block Design and Matrix Reasoning) were z-scored across subjects. As an estimation for verbal IQ, the mean of the z-scored Vocabulary and Similarities subscales was used. As an estimation for the performance IQ, the mean of the z-scored Block Design and Matrix Reasoning subscales was used. Age-normalisation was not performed, since all subjects were of very similar age at time of assessment.

EXTRACTION OF BEHAVIOURAL DATA FROM THE AGN-TASK From the AGNtask four features were included:

- *agn_latency_positive*: reaction time to in trials with positive words,
- *agn_latency_negative*: reaction time to in trials with negative words,
- agn_omissions_positive: total number of false negative responses in trials with positive words,
- agn_omissions_negative: total number of false negative responses in trials with negative words.

Unfortunately, the number of commissions (i.e. false positives - the participant falsely reacted to a non-target word) was not recorded in the IMAGEN study.

EXTRACTION OF BEHAVIOURAL DATA FROM THE SST Three features can easily be extracted from the behavioural data from the Stop-signal-task:

- sst_go_reaction_time: mean reaction time over all successful go-trials,
- sst_go_success_rate: average of the ratio of number of correct go-responses (given in time) divided by the total number of correct go-responses (including the responses given too late),
- sst_ssd: stop-signal-delay, i.e. the mean delay between go-stimulus and stop-stimulus presentation.

Since the stop-signal-delay is adjusted according to the subjects' behaviour, a longer stop-signal-delay indicates a better task performance, i.e. a more successful inhibition.

MODELLING OF THE SSRT IN THE SST The SSRT is the analogy to the go reaction time. It describes how long a participant takes (on average) to "react" to a stop signal. However, since a correct "reaction" is in the case of an appearing stop signal one that is inhibited, the stop signal reaction time is not measured directly.

A common modelling approach is the horse-race model (see Verbruggen and Logan (2008) for a review). It asserts that the go and stop processes are independently racing each other - the go process usually with a head start. If the stop process manages to "catch up", the action is successfully inhibited. Otherwise, the go process (and hence the response) "escapes" from successful inhibition.

If we stay within the metaphorical setting of a horse-race, we have two horses - a go-horse and a stop-horse. The go-horse runs at a certain pace, subject to some variability. Observing the go-horse for a number of runs, we can infer its finishing-time distribution. The second horse, the stophorse, runs completely independently of the go-horse (e.g. equipped with blinkers). This stop-horse is generally a faster runner than the go-horse and always starts a (for now: fixed) delay period later than the go-horse. However, for technical reasons we cannot measure the time when the stop-horse crosses the finishing line. All we observe is the information whether the stop-horse managed to catch up with the go-horse or not. After many such races (with the fixed delay period), we know that the stop-horse catches up

with the go-horse in x % of cases. The question we now ask is: how long does the stop-horse need on average to get to the finishing line, i.e. what is the *stop-horse finishing-time*?

The so-called mean-method (Verbruggen and Logan 2008; Verbruggen and Logan 2009; Verbruggen et al. 2008) assumes a chance of 50 % for the stop-horse to win the race. (In fact, when dropping the fixed delay period constraint and when using the tracking algorithm as described in Section 6.2.3, the stop-success rates do truly approach 50 %.) Hence, in an rough approximation, the stop-horse finishes at the same time as the gohorse finishes on average over all its runs, so we can estimate the stop-horse finishing time as the average go-horse finishing time.

However, the mean-method does not take the variability of the go-horse finishing time and the variability of stop-success rates into account. The so-called integration method incorporates both aspects. Say now that the stop-success rate is 30 %. We can infer that in 70 % of the trials, the go-horse has already reached the finishing line by the time the stop-horse crosses it. Since we assume that go and stop-horses run entirely independent of one another, we can check when the go-horse has already reached the finishing line in 70 % of its runs. This time we can then take as the estimate of the finishing time of the stop-horse.

Note that for a symmetric finishing-time-distribution of the go-horse and for stop-success rates of exactly 50 %, the mean-method and the integration method give the same results.

Finally, the actual SSRT is calculated in both cases as the estimated stophorse-finishing time minus the delay period (the time when the stop-signal is presented after the go-stimulus) - i.e. it is the time the stop-horse is actually running. In the case of a non-fixed delay period, the mean-method subtracts the mean of the delays from the mean go-horse finishing time for an estimation of the stop signal reaction time. The integration-method estimates the SSRT separately for each delay period and then takes the average over these as the final estimate.

Moving to a more formal setting, let goRT(t) be the probability density function of go-reaction times and

$$F_{goRT}(t) = \int_{-\infty}^{t} goRT(t')dt'$$

the cumulative distribution function. The corresponding quantile function is then:

$$Q_{\text{goRT}}(p) = \inf\{t \in \mathbb{R}_{\geqslant 0} : p \leqslant F_{\text{goRT}}(t)\}.$$

With N_{qo} go-trials, each yielding a reaction time of $goRT_i$, $i = 1, ..., N_{qo}$, we can estimate the empirical cumulative distribution function as follows:

$$\hat{F}_{\text{goRT}}(t) = \frac{1}{N_{\text{go}}} \sum_{i=1}^{N_{\text{go}}} \mathbb{1}\{\text{goRT}_i \leqslant t\},$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. This cumulative distribution is a step function.

The empirical quantile function can then be estimated by:

$$\hat{Q}_{\text{goRT}}(\mathfrak{p}) = \inf\{t \in \mathbb{R}_{\geqslant 0} : \mathfrak{p} \leqslant \hat{F}_{\text{goRT}}(t)\}.$$

Furthermore, let $SSR(SSD) \in [0, 1]$ be the probability of (falsely) reacting in a stop-trial with a stop-signal delay SSD. Note that this is usually called the inhibition function in literature (Verbruggen and Logan 2008). We can estimate this inhibition function as follows:

$$\widehat{SSR}(SSD) = \frac{1}{N_{stop,SSD}} \sum_{i=1}^{N_{stop,SSD}} \mathbb{1} \left\{ \begin{array}{l} \text{go-response in a stop-trial with} \\ \text{stop signal delay} = SSD \end{array} \right\},$$

where N_{stop,SSD} is the number of stop-trials with stop-signal delay SSD. Note that the average of \$SR(SSD) over different SSDs in a setup with a tracking procedure (as we have it) should approach 50 %. However, certain SSR(SSD) might deviate strongly.

Let us further set N_{SSD} as the number of unique stop-signal delays used over all stop-trials and let SSD_i , $i = 1, ..., N_{stop}$ be the respective delays. Having introduced the necessary definitions, we can now formalise the mean-method estimate of the stop signal reaction time:

$$\widehat{SSRT}_{mean} := \frac{1}{N_{go}} \sum_{i=1}^{N_{go}} goRT_i - \frac{1}{N_{stop}} \sum_{i=1}^{N_{stop}} SSD_i.$$
 (6.1)

The integration-method estimates the stop-signal reaction time separately for each delay:

$$\widehat{SSRT}_{in}(SSD) := \widehat{Q}_{goRT}(\widehat{SSR}(SSD)) - SSD$$
 (6.2)

for $SSD \in \{SSD_1, ..., SSD_{N_{SSD}}\}$, and then takes an average for the final estimate:

$$\widehat{SSRT}_{in} := \frac{1}{N_{SSD}} \sum_{i=1}^{N_{SSD}} \widehat{SSRT}_{in}(SSD_i).$$
 (6.3)

Even though the mean-method seemed superior to the integration-method for many years (Band et al. 2003; Verbruggen and Logan 2008), recent findings have shown that the mean-method is very susceptible to skewed goreaction time distributions and gradual slowing of the go-reaction times over trials (Boehler et al. 2012; Verbruggen et al. 2013). Especially for positive skewness, the mean-method was shown to consistently overestimate the stop-signal reaction times. This overestimation is especially problematic, since it may lead to spurious SSRT differences between subject groups that in fact differ only in the shape of the go-reaction time distribution. The integration method was shown to be more robust against these variations and was hence recommended for further use (Verbruggen et al. 2013). Hence, we used the integration-method for the main analyses and only performed basic comparisons to the mean-method (see Appendix B.1 for details).

Note that within the integration-method for the computation of the final estimate (6.3) we exclude all $SSD^* \in SSD_1, ..., SSD_{N_{SSD}}$ that lead to $\widehat{SSRT}(SSD^*) < 0$. This occurs for instance when $\widehat{SSR}(SSD^*) = 0$, i.e. the participant never responded falsely for that certain delay. Then, it holds that:

$$\hat{Q}_{goRT}(\widehat{SSR}(SSD^*)) = min\{goRT_1, \dots, goRT_{N_{go}}\}\$$

and hence if $min\{goRT_1, ..., goRT_{N_{go}}\}\$ < SSD* the estimate $\widehat{SSRT}(SSD^*)$ will be negative.

Note that detailed analyses for the extraction of the SSRT can be found in Appendix B.1.

OUTLIER DETECTION IN THE SST In order to ensure that we consider only data of subjects who have understood the task correctly and to ensure a stable estimation of the SSRT, we removed outliers. The necessity of an outlier treatment for the stop-signal-task has recently been demonstrated (Congdon et al. 2012). We specifically excluded data from subjects that have an average go success rate lower than 80 % or an average stop success rate larger than 80 %. Due to the tracking procedure the average stop success rate should be around 50 % and we wanted to exclude subjects that (almost) never reacted, even in go-trials. Such crude thresholding is common in literature (Congdon et al. 2012; Whelan et al. 2012). Furthermore, it has been shown that such rather lenient criteria are superior to more conservative approaches in the trade-off between reliability and sample size (Congdon et al. 2012). All four extracted features for the SST were replaced by respective mean values across all subjects for subjects considered outliers, analogous to mean-imputation for missing values (see Section 2.2.2).

Furthermore, we required the subjects to have performed the full experiment. Specifically, we required $60 \leqslant N_{stop} \leqslant 67$ for FU2-data and $80 \leqslant N_{stop} \leqslant 87$ for BL-data.

Note that we have also performed a more sophisticated outlier detection, based not only on the stop signal behavioural features but on all cognitive features (as described in Section 2.2.1). However, this did not improve accuracies, hence we kept with the simple approach.

Figure 6.2 shows the outlier removal procedure for BL and FU2 data. Based on their combined stop and go success rates, a total of 31 and 28 subjects are considered as outliers for BL and FU2 data, respectively. Both values are low in comparison to the total sample size (see Table 6.1). Furthermore, no clear patterns regarding the distribution of light and heavy drinkers are visible from Figure 6.2, other than potentially a tendency for light drinkers having slightly higher (outlying) stop success rates and slightly lower (outlying) go success rates.

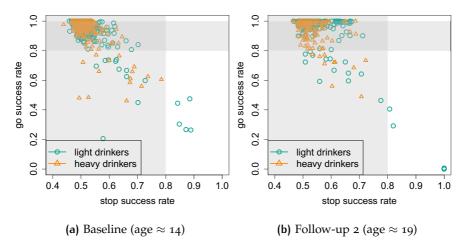


Figure 6.2: Outlier removal for the behavioural data from the SST. The dark grey shaded area (the overlap of the two light grey shaded areas) show the region where subjects are not treated as outliers. More specifically, subjects that either had a stop success rate larger than 0.8 or a go success rate smaller than 0.8 were considered an outlier.

6.3.2 Extraction of functional imaging features

We pre-selected 12 regions of interest that have been linked to response inhibition in the stop signal task (see Verbruggen and Logan (2008) for a review). Based on this review, we focused on the frontal-thalamic-basalganglia network, specifically the inferior frontal gyrus (separated into the pars opercularis, pars triangularis and the pars orbitalis), the caudate nucleus, the putamen and the thalamus from the Automated Anatomical Labeling (AAL) atlas (Tzourio-Mazover et al. 2002). For each region, left and right separately, we extracted the mean value of the stopSuccess and the stopSuccess-stopFail contrasts. We also performed comparisons to a different single-value summary measure instead of the mean, based on recommendations given in Tong et al. (2016). Details and results are given in Appendix B.6.

CHOICE OF REGIONS We resort to the frontal-thalamic-basal-ganglia network since

- (i) there are clear indications for its involvement in response inhibition:
 - inferior frontal gyrus, see Aron et al. (2007), Aron and Poldrack (2006), Chevrier et al. (2007), Li et al. (2006), Rubia et al. (2003), and Whelan et al. (2012),
 - basal ganglia, see Aron et al. (2007), Aron and Poldrack (2006), and Chevrier et al. (2007),
 - thalamus, see López-Caneda et al. (2014) and Rubia et al. (2007),
- (ii) significant substance use effects have been found for the network:
 - subclinical, see Bednarski et al. (2012),
 - clinical, see Hu et al. (2015) and Li et al. (2009),
- (iii) there is extensive evidence for an inverse scaling of activation in the network with the SSRT, see Aron and Poldrack (2006), Li et al. (2006), Rubia et al. (2007), and Whelan et al. (2012),
- (iv) and this scaling may be different between substance abusers and controls.

We intentionally did not include the pre-supplementary motor area (pre-SMA). Even though there is evidence for its general involvement in response inhibition, it seems to be linked more to task-monitoring than to actual stopping processes (Aron et al. 2007; Aron and Poldrack 2006).

Furthermore, note that our arguments from Section 5.3.1 for (i) using a priori defined (instead of data derived) regions of interest and (ii) using the AAL atlas as a intermediate option between fine-grainedness and coarseness apply here, too.

OUTLIER TREATMENT We excluded data from all subjects that were considered outliers according to the lenient criteria described in 6.3.1.

6.3.3 Summary of features

For an overview of number of subjects available see Table 6.1. The number in brackets indicates the number of female subjects available for each cell. Note that there are much fewer female heavy drinkers compared to male heavy drinkers. This imbalance is present in general for the drinking phenotype defined on the IMAGEN data set, see 4 for more details. We deal with this imbalance in prediction analysis by using a bootstrapped version of cross-validation (BSSCV), see Section 2.4.3. For association analysis we include gender (and also the scanning site) as a additional covariates, see 2.5.2 for details.

For a list of features including a short description, see Tables 6.2 and 6.3. All missing values were mean-imputed for prediction analysis (see Section 2.2.2 for brief methodological details). We have performed several more sophisticated imputation methods, both univariate and multivariate. Furthermore, we also performed all prediction analyses additionally on the subset of subjects for which complete features were available. For further details and comparative results, see Appendix B.5. For histograms of all features, see Appedix B.3.

Table 6.1: Sample sizes for each data modality – functional imaging (SST fMRI) and cognitive variables - and time point. Values in brackets indicate the number of female subjects.

	Phenotype age \approx 19 FU2	SST fMRI age \approx 14 BL	SST fMRI age \approx 19 FU2	Cognitive age \approx 14 BL	Cognitive age \approx 14 FU2
All	1472	1995	1378	1929	1094
	(772)	(1027)	(713)	(990)	(498)
Light	550	534	491	517	395
drinkers	(347)	(339)	(311)	(329)	(256)
Heavy	464	440	408	421	328
drinkers	(184)	(179)	(153)	(169)	(125)

Table 6.2: Summary of cognitive variables and their meanings, origin and percentage of missing values. AGN: affective go/no-go task; SST: stop signal task.

Name	Meaning	Origin	Missing age ≈ 19 BL (%)	Missing age \approx 19 FU2 (%)
agn_mean_correct _latency_negative	Mean reaction time for negative words	AGN	13.14	0.55
agn_mean_correct _latency_positive	Mean reaction time for positive words	AGN	12.93	0.42
agn_total_omissions _negative	Number of false negatives for negative words	AGN	10.9	30.29
agn_total_omissions _positive	Number of false negatives for positive words	AGN	10.9	30.29
wechsler_piq	Performance IQ	Wechsler	0.00	0.28
wechsler_viq	Verbal IQ	Wechsler	0.00	0.28
sst_ssrt	Stop signal reaction time from integration method	SST	0.00	0.28
sst_mean_go _reaction_time	Mean reaction time over all successful go-trials	SST	0.00	0.28
sst_mean_ssd	Mean delay between go- and stop-stimulus	SST	0.00	0.00
sst_mean_go _success_rate	Average ratio of successful go-trials	SST	0.00	0.00

Table 6.3: Summary of functional imaging variables and their meaning.

Name	Meaning	Atlas
Caudate_L	Left caudate nucleus	AAL
Caudate_R	Right caudate nucleus	AAL
Putamen_L	Left putamen	AAL
Putamen_R	Right putamen	AAL
Frontal_Inf_Oper_L	Left inferior frontal gyrus pars opercularis	AAL
Frontal_Inf_Oper_R	Right inferior frontal gyrus pars opercularis	AAL
Frontal_Inf_Tri_L	Left inferior frontal gyrus pars triangularis	AAL
Frontal_Inf_Tri_R	Right inferior frontal gyrus pars triangularis	AAL
Frontal_Inf_Orb_L	Left inferior frontal gyrus pars orbitalis	AAL
Frontal_Inf_Orb_R	Right inferior frontal gyrus pars orbitalis	AAL
Thalamus_R	Right thalamus	AAL
Thalamus_L	Left thalamus	AAL

6.3.4 Prediction analysis

For prediction analysis, we rely on the methods described in Section 2.3. As classifiers we use logistic regression (see Section 2.3.2), random forests (see Section 2.3.2) and naïve Bayes (see Section 2.3.3). We use K = 10 folds for cross-validation and account for the confound gender using BSSCV (see 2.4.3) with B = 10 bootstraps. Class labels are not balanced. See Appendix B.7 for an investigation using balanced class labels. For significance testing of accuracies (see Section 2.4.4), we use nPerm = 1000. Multiple testing is corrected via the Benjamini-Hochberg method at level $\alpha = 0.05$ (see Section 2.6). Furthermore, to ensure equal weighting of features, we z-score each feature on the training folds and apply the resulting z-scoring to the respective test folds.

6.3.5 Association analysis

For association analysis, we use our ReLL-method (see Section 2.5.2) with number of repetitions nRep = 100. We correct for multiple testing via the Benjamini-Hochberg method at level $\alpha = 0.05$ (see Section 2.6). We perform the association analysis only on the existing (missing data is removed instead of mean-imputed). Each feature is z-scored prior to association analysis.

6.4 **RESULTS**

6.4.1 Prediction analysis

In the prediction analysis, we investigated the predictive power of cognitive variables and fMRI data from the stop signal task for the discrimination of heavy versus light drinkers.

Figure 6.3 shows the prediction accuracies for the COGNITIVE FEATURES cognitive features in the cross-sectional and longitudinal setting. In the cross-sectional setting (i.e. using features assessed at approximate age 19 to predict the drinking behaviour at the same age), all three classifiers yield significant prediction accuracies. The highest accuracy is obtained by naïve Bayes (balAcc = 0.61). In the longitudinal setting (i.e. using features assessed at approximate age 14 to predict the drinking behaviour at approximate age 19), all accuracies are lower, yet are still significant. The highest accuracy is obtained by Random Forest (balAcc = 0.56).

Hence, according to the framework proposed in Section 2.1, we infer that both the cross-sectional and the longitudinal setting provides generalisable discrimination between heavy and light drinkers. We thus take a closer look at both settings in the association analysis.

Note that, analogously to the analyses in Chapter 5, we also considered the gender-separate case. We did not observe any large differences between genders (see Appendix B.7 for further details).

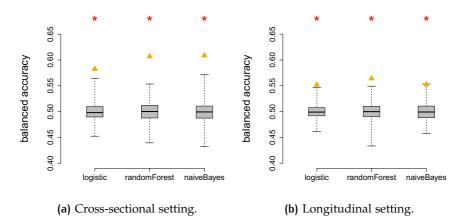


Figure 6.3: Balanced accuracies (yellow triangles) for prediction of heavy versus light drinkers from cognitive features, based on three classifiers: logistic regression, random forest and naïve Bayes, and for (a) the cross-sectional and (b) the longitudinal setting. Grey box plots (spanning the entire range) indicate prediction accuracies obtained by label permutation. The farther away the yellow triangles (true accuracies) are from the grey box plots (permuted accuracies), the more likely it will survive hypothesis testing and correction for multiple testing. Significant cases, after controlling for FDR at $\alpha = 0.05$, are indicated by a red star.

FUNCTIONAL IMAGING FEATURES Figures 6.4 and 6.5 show the prediction accuracies based on functional imaging features from the stopSuccessstopFail and stopSuccess contrast, respectively. The prediction analysis reveals that only the cross-sectional setting for the stopSuccess-stopFail contrast yields significant (yet low) accuracies. The highest accuracy is obtained by Random Forest (balAcc = 0.55).

Hence, according to the framework proposed in Section 2.1, we infer that only the cross-sectional setting for the stopSuccess-stopFail contrast provides generalisable discrimination between heavy and light drinkers. We thus take a closer look at this setting in the association analysis.

Note that (analogously to the analyses in Chapter 5) we also considered the gender-separate case. We did not observe any large differences between genders (see Appendix B.7 for further details).

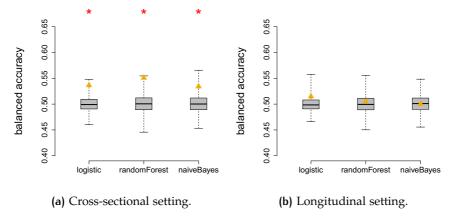


Figure 6.4: Balanced accuracies for prediction of heavy versus light drinkers from functional imaging features from the stopSuccess-stopFail contrast, based on three classifiers: logistic regression, random forest and naïve Bayes and for (a) the crosssectional and **(b)** the longitudinal setting. Grey box plots (spanning the entire range) indicate prediction accuracies obtained by label permutation. The farther away the yellow triangles (true accuracies) are from the grey box plots (permuted accuracies), the more likely it will survive hypothesis testing and correction for multiple testing. Significant cases, after controlling for FDR at $\alpha = 0.05$, are indicated by a red star.

6.4.2 Association analysis

In the association analysis, we examined predictive settings more closely. More specifically, we evaluated the influence of single variables on the drinking behaviour using the ReLL-method, see Section 2.5.2.

COGNITIVE FEATURES Figure 6.6 shows the association between cognitive features and drinking behaviour in the cross-sectional and longitudinal setting. In the cross-sectional setting, both the negative and positive total omissions in the AGN task are significantly negatively associated with heavy drinking, i.e. heavy drinkers gave less false negative responses in both negative and positive word categories. Moreover, heavy drinking is

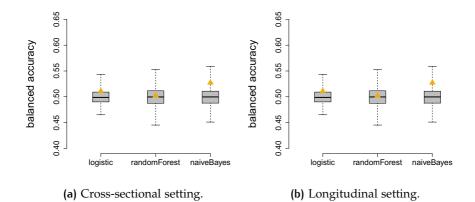
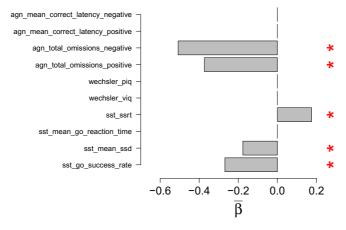


Figure 6.5: Balanced accuracies for prediction of heavy versus light drinkers from functional imaging features from the stopSuccess contrast, based on three classifiers: logistic regression, random forest and naïve Bayes, and for (a) the cross-sectional and (b) the longitudinal setting. Grey box plots (spanning the entire range) indicate prediction accuracies obtained by label permutation. The farther away the yellow triangles (true accuracies) are from the grey box plots (permuted accuracies), the more likely it will survive hypothesis testing and correction for multiple testing. Significant cases after controlling for FDR at $\alpha = 0.05$ are indicated by a red star.

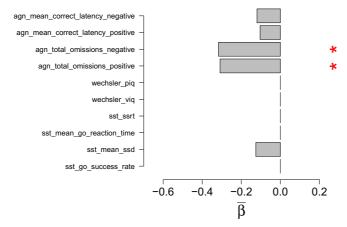
associated positively with the SSRT, meaning that heavy drinkers showed a slower reaction to stop signal signs compared to light drinkers. Further, both the mean SSD and the go-success-rate in the SST display a significant negative association to heavy drinking, see Figure 6.6. This implies that heavy drinkers were generally less successful in the SST.

In the longitudinal setting, only two features survive significance testing and multiple testing correction, namely the negative and positive total omissions in the AGN task. As in the cross-sectional setting, they are negatively associated with heavy drinking.

FUNCTIONAL IMAGING FEATURES Figure 6.7 shows the association between drinking behaviour and activity of regions relevant for response inhibition in the **stopSuccess-stopFail** contrast, in the cross-sectional setting. Only three features are estimated as non-zero and only the left and right thalamus survive significance testing and correction for multiple testing. Both of these features are associated negatively to heavy drinking, i.e. the less activity specific to successful stopping a subject shows in the bilateral thalamus, the more likely the subject is a heavy drinker.



(a) Cross-sectional setting.



(b) Longitudinal setting.

Figure 6.6: Median parameter estimates from association analysis on cognitive features using the ReLL-method. Non-zero estimates are obtained if the median of the parameter estimates over 100 runs is non-zero. Significant cases after correction for multiple testing are indicated by a red star. If the $\bar{\beta}$ for a score is negative, then this means that heavy drinkers showed a smaller score compared to light drinkers (and vice versa).

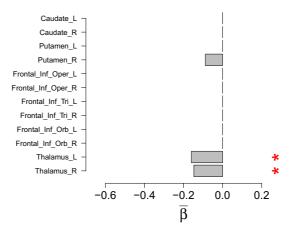
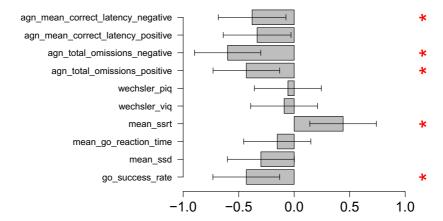


Figure 6.7: Median parameter estimates from association analysis on functional imaging features from the stopSuccess-stopFail contrast using the ReLL-method. Non-zero estimates are obtained if the median of the parameter estimates over 100 runs is non-zero. Significant cases after correction for multiple testing are indicated by a red star. If the $\bar{\beta}$ for a score is negative, then this means that heavy drinkers showed a smaller score compared to light drinkers (and vice versa).

6.4.3 Further investigations

CUMULATIVE DRINKING BEHAVIOUR In an additional (post-hoc) investigation, we explored the relationship between a cumulative drinking score (cumulative drunkenness) and the cognitive features at FU₂. For details on how we computed the cumulative drinking score, see Chapter 4.

Figure 6.8 shows a the results of this exploration and reveals a similar pattern as for the cross-sectional association analysis (see Figure 6.6): significant negative relationship between the cumulative drunkenness score and (i) negative total omissions and (ii) positive total omissions in the AGN task and (iii) the go-success rate in the SST. Furthermore, we see a significant positive relationship between the cumulative drunkenness score and the SSRT, again in line with the cross-sectional association analysis.



relationship to cumulative drunkenness

Figure 6.8: Relationship of the cognitive features to a cumulative drinking score. The plot shows coefficient estimates and confidence intervals from a linear regression analysis between the cumulative drunkenness score (target variable) and each cognitive feature (explanatory variables), including gender as a covariate. Significant estimates after correction for multiple testing are indicated by a red star.

CORRELATION STRUCTURE In another analysis, we investigated the correlation structure between the sets of variables investigated in this chapter, namely:

- (i) cognitive features and functional imaging features from the stopSuccess contrast,
- (ii) cognitive features and functional imaging features from the stopSuccessstopFail contrast,

For these sets of variables, we considered both the data from FU2 and the data from BL. In order to account for potential site effects, we use sitemean corrected functional imaging features. We controlled the FDR at level $\alpha = 0.05$. Figure 6.9 shows all the results.

The main observations are as follows:

(i) cognitive features and stopSuccess features:

- negative correlations between the SSRT and bilateral caudate nucleus and putamen,
- contrast positive correlations between both go-reaction-time and SSD and bilateral caudate nucleus, putamen, and thalamus,

(ii) cognitive features and stopSuccess-stopFail features:

- positive correlations between the go-success-rate in the SST and bilateral thalamus.
- only for FU2 negative correlations between go-reaction-time, SSD and various brain regions,
- no significant correlations between functional activity and SSRT.

Generally it is noticeable that the correlation structure for the **stopSuccess** contrast looks quite different to the stopSuccess-stopFail contrast. Furthermore, BL and FU2 correlation structures are overall quite consistent.

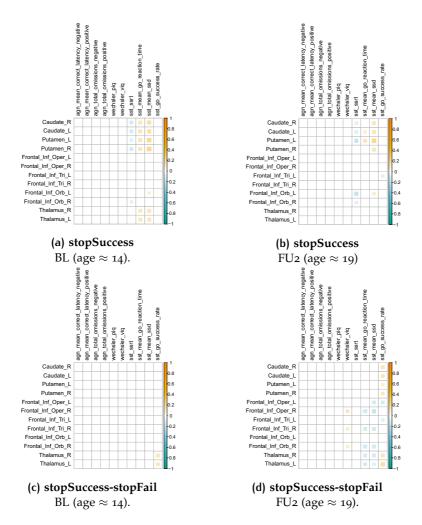


Figure 6.9: Correlations between different sets of features with the cognitive features at BL and FU2. Note that only significant (after FDR-correction at level $\alpha = 0.05$) are displayed.

A CLOSER LOOK INTO THE SSRT Further, we took a closer look into the correlates of the SSRT. Figures 6.10 and 6.12 show the correlation between the **stopSuccess** contrast and the SSRT for all subjects, for BL and FU₂, respectively. Figures 6.11 and 6.13 show the same correlations, but separately for light and heavy drinkers.

The bottom two plots in each Figure show the same correlations separately for light and heavy drinkers. For FU2 (i.e. Figure 6.12) we used the usual phenotype, as constructed and described in Chapter 4. For BL (i.e. Figure 6.10), we show in the bottom two plots again the results for the usual phenotype, that means the separation in light and heavy drinkers assessed at FU2. In addition, in the middle two plots, we show results separately for drinking behaviour groups assessed at BL. As usual, significant cases after correction for multiple testing are indicated by a red star.

In general, the results over all subjects (top plots) show significant negative correlations between the SSRT and the activity in successful stop trials in almost all selected regions. This indicates that the longer the SSRT, the less active these regions are during successful stop trials. Interestingly, when looking at the correlations separately for light and heavy drinkers, it becomes apparent that the significant correlations for all subjects are driven exclusively by the light drinkers. For both BL (using the BL drinking behaviour grouping and the usual drinking behaviour phenotype from FU2) and FU2 the correlations are much smaller and not significant anymore for the heavy drinkers.

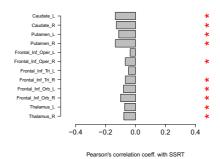


Figure 6.10: Pearsons's correlations between the SSRT and site-corrected imaging features from the stopSuccess contrast, assessed at BL (approximate age of 14). Significant estimates after correction for multiple testing are indicated by a red star.

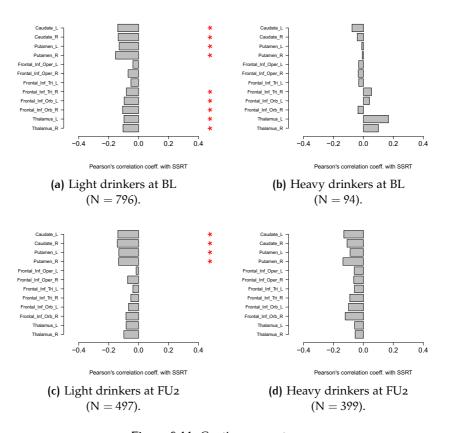


Figure 6.11: Caption on next page.

Figure 6.11: Pearson's correlations between the SSRT and imaging features from the stopSuccess contrast, all assessed at BL (approximate age of 14). Significant estimates after correction for multiple testing are indicated by a red star. (a): Correlations for BL light drinkers only; (b): Correlations for BL heavy drinkers only; (c): Correlations for BL light drinkers only; (d): Correlations for FU2 heavy drinkers only. For (a) and (b), we defined the light drinkers as subjects falling into the "light" category of the "Frequency of drinking" variable, as described in Chapter 4 and depicted in Figure 4.1, and the heavy drinkers as the remaining subjects. Note that the results look similar when choosing a different variable for the grouping. For (c) and (d), the separation into light and heavy drinkers was performed according to the usual drinking behaviour phenotype, as constructed and described in Chapter 4.

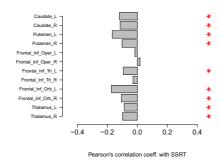


Figure 6.12: Analogous to Figure 6.10, here for FU2.

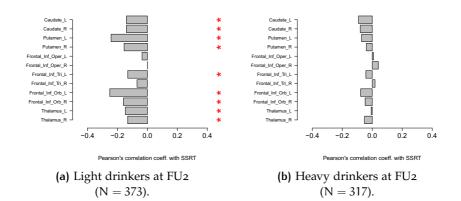


Figure 6.13: Analagous to Figure 6.11, here for FU2 and (a): Correlations for FU2 light drinkers only; **(b)**: Correlations for FU₂ heavy drinkers only.

6.5 DISCUSSION

In this chapter, we explored the predictive and associative relationship of functional imaging and cognitive variables to heavy drinking in a large adolescent sample.

Our main findings are that heavy drinking at age 19 is significantly associated with:

- (i) decreased functional activation in the bilateral thalamus at age 19 (but not age 14),
- (ii) decreased behavioural markers of inhibitory control, e.g. prolonged SSRT and worse performance in the SST at age 19 (but not age 14),
- (iii) reduced false positive reactions (omissions) in the AGN task both at age 14 and age 19.

Interestingly, we also found that the SSRT scales inversely with activation during successful stop trials only for light drinkers and not at all for heavy drinkers. This effect appears consistently for ages 14 and 19.

COGNITIVE FEATURES We found that direct cognitive markers of inhibitory control can significantly predict drinking behaviour cross-sectionally but not longitudinally.

Several inhibition markers from the SST at age 19 were significantly associated with drinking behaviour at age 19. Specifically, subjects with a longer SSRT, a shorter SSD (i.e. a worse task performance) and a lower go-success rate were shown to have a higher chance of being heavy drinkers at age 19. Our results are in line with many studies that have found that impaired inhibitory control is related to binge drinking in adolescence and to the risk of later alcohol dependence (Ahmadi et al. 2013; Czapla et al. 2015; Henges and Marczinski 2012; Nigg et al. 2006; Rubio et al. 2008).

Furthermore, a lower number of omission errors in negative and positive word categories of the AGN at both age 14 and 19 were associated to heavy drinking at age 19. This appears rather surprising, since we expected a worse (and not better) task performance in heavy compared to light drinkers. This finding is also not in line with literature. Ahmadi et al. (2013) have for instance observed increased reaction times in heavy compared to light college drinkers during a go/no-go task. Notably, most other studies concentrate on the more natural marker of response inhibition in go/no-go tasks, namely commission, i.e. false positive errors, and almost unanimously report increased commission errors in heavy drinkers and alcoholics (Henges and Marczinski 2012; Noël et al. 2007; Petit et al. 2012); note that Ahmadi et al. (2013) did not find group differences. Omission errors have been associated to inattention, however, their exact interpretation has been debated (Bezdjian et al. 2009; Kertzman et al. 2008). We speculate that the heavy drinking subjects may have generally responded more, both to go and to no-go stimuli, hence making less omission errors but thereby also more commission errors. However, we cannot make definite conclusions due to the lack of direct measurement of commission errors.

Importantly, we observed no association between intelligence and drinking behaviour. Thus, we assume the differences in inhibitory control between heavy and light drinkers do not stem from discrepancies in general cognitive abilities.

Overall, we find that a relation between definite markers of response inhibition and drinking behaviour are present cross-sectionally but not longitudinally. Further, the magnitude of impaired inhibitory control seems to be correlated with cumulative drinking. Hence, our findings point in the direction of poor response inhibition being more likely a consequence than a direct cause of heavy drinking in adolescence and is thereby in line with e.g. the findings of a prospective study (Goudriaan et al. 2011).

FUNCTIONAL IMAGING FEATURES We showed that functional activity of 19 year olds in the stopSuccess-stopFail contrast in areas relevant to response inhibition were predictive of drinking behaviour at the same age (cross-sectionally). Longitudinally, the same areas' functional activity at age 14 was not predictive for drinking behaviour at age 19. The evaluation of cross-sectional association between functional activity in single regions and drinking behaviour revealed reduced activity of heavy drinkers in the bilateral thalamus. Differences in the frontal-basal-ganglia network between substance abusers and controls have been found before (Bednarski et al. 2012; Hu et al. 2015; Li et al. 2009), yet the thalamus was usually not in the search space. The central role of the thalamus as a relay station between different subcortical areas and the cerebral cortex is also played in the context of response inhibition. It is postulated that during response inhibition, the inferior frontal cortex sends a stopping command, indirectly resulting in the inhibition of the thalamus, which then in turn inhibits thalamocortical projections (López-Caneda et al. 2014). Diminished thalamic activity in heavy drinkers may hence indicate a less efficient inhibition process.

Since the link between functional activity and drinking behaviour are not present in the longitudinal setting, we suggest anew that they may be a consequence of heavy drinking. Furthermore, the altered functional activity may even constitute the neural basis for the impaired response inhibition that we have observed cross-sectionally in the cognitive features. Note that Norman et al. (2011) have found reduced neural activity during response inhibition to be related to future alcohol and other substance abuse, yet for a much smaller sample size (N = 38).

SCALING OF THE SSRT We have further found a significant inverse scaling of the SSRT with functional activity during successful stopping across various regions, in agreement with a corpus of literature (Aron and Poldrack 2006; Li et al. 2006; Rubia et al. 2007; Whelan et al. 2012). An Importantand to our knowledge novel - finding is that this inverse scaling is present only for light drinkers and vanished completely for heavy drinkers. The cited studies either did not focus on alcohol drinking behaviour or did not consider the of correlation between functional activity and the SSRT separately for light and heavy drinkers. We speculate that this finding may indicate that heavy drinkers encode the inhibition process less efficiently. Furthermore, since the presence of the scaling of SSRT with functional activity at age 14 is able to distinguish drinking behaviour at age 19, the neural efficiency of response inhibition may even be an indirect risk factor for a harmful alcohol drinking trajectory.

6.6 CHAPTER SUMMARY

The results from the analyses on the IMAGEN database tentatively indicate impaired response inhibition in adolescence as a consequence of excessive alcohol consumption, both on a neural and a behavioural level. Interestingly, heavy drinkers seem to neurally encode response inhibition in a less efficient manner than light drinkers.

Part IV Synthesis

SUMMARY AND OUTLOOK

7.1 SUMMARY

Part I: Foundations

In the first part of the thesis, I introduced all the (old and new) methods used in the subsequent chapters. I motivated the two-step procedure (prediction analysis, then association analysis). If (and only if) a set of features is declared significant by the prediction analysis, I would move on to post-hoc association analysis. I then dived into prediction analysis, briefly recapitulating well-established classifiers (logistic regression, naïve Bayes, random forests, support vector machines, RSLVQ). I commented on the different ways to assess the performance of a classifier (accuracy, specificity, sensitivity), both theoretically and practically using cross-validation. I further introduced a novel adapted version of cross-validation – bootstrappedstratified-cross-validation (BSSCV) – for the assessment of classifier performance in the presence of confounders. Next, I took a closer look at association analysis. I recapitulated the classical methods for association analysis, namely t-tests and logistic regression. I then introduced a novel method for association analysis - repeated logistic lasso (ReLL) - and showed in a small simulation study that it can outperform logistic regression for association analysis.

Part II: The IMAGEN database

In the second part of the thesis, I gave an overview of the aims and design of the IMAGEN study. I further introduced the cross-sectional and longitudinal settings. I gave a detailed description of my contribution within the IMAGEN framework. Moreover, I introduced the various alcohol-related instruments administered in the IMAGEN study. I showed how a pairwise clustering algorithm was used to build a comprehensive binary drinking behaviour phenotype (light and heavy drinkers). Furthermore, I validated the new phenotype and showed that the constructed phenotype outperforms all original drinking behaviour variables in terms of consistency.

Part III: Patterns of alcohol abuse in adolescence

In the third part of the thesis, I applied the established methods on various neurobehavioural features from the IMAGEN study to differentiate between drinking behaviour phenotypes. First, I focussed on grey matter volume and psychosocial features. I showed that heavy drinking in adolescence is associated with reduced grey matter volume across various cortical and subcortical structures, especially in females. Moreover, I observed that impulsivity and facets of novelty seeking are associated (also longitudinally) to heavy drinking. Then, I focussed on functional imaging and cognitive features. I showed that adolescent drinking is associated with various markers of impaired response inhibition, both neural and behavioural.

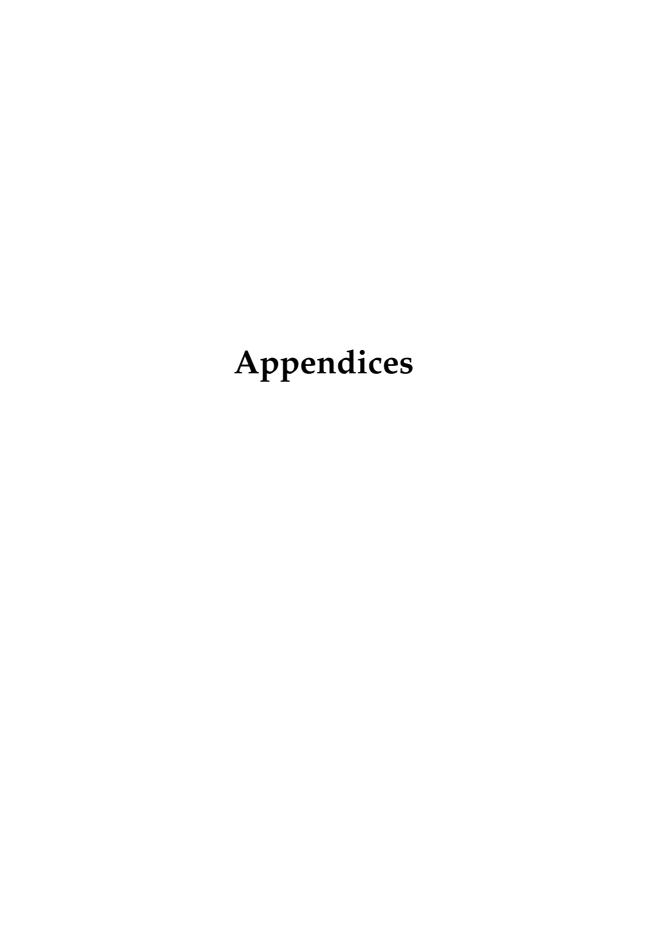
7.2 OUTLOOK

Theory

The ReLL method for association analysis introduced in Chapter 2 showed promising performance in a simulation study and in the application to neurobehavioural data. It would be interesting to explore the propertiers of the ReLL in more complex simulated settings. For instance, one could add more confounders or analyse association based on multivariate features. Furthermore, it may be possible to derive an approximate distribution of the ReLL estimator, based on the work of Buja and Brown (2014) and Lockhart et al. (2014). Finding such an (approximate) distribution would make the time-intense permutation testing obsolete.

Application

In this thesis, I have analysed a wide range of neurobehavioural features with respect to their predictive and associative value for distinguishing light from heavy drinkers. The rich IMAGEN study offers even more data that was in parts studied in the framework of the e:Med project. An aspect that was not yet covered but seems promising given my results is the detailed study of connectivity networks from resting state fMRI data. While wholebrain connectivity yielded only mediocre predictive power, the presented differences in response inhibition may be rooted in altered functional connectivity in e.g. the executive control network.





A.1 SITE EFFECT

As mentioned in Section 5.3.4, a challenging aspect of the data is the correlation of site and grey matter volume data and the imbalance of the drinking behaviour and gender across sites.

For a more exact determination of the correlation structure between site and brain imaging data we performed a classification task predicting the eight sites based on grey matter volume data. We used all 116 regions of interest from the AAL atlas (Tzourio-Mazoyer et al. 2002) and used the RSLVQ as classifier (see Section 2.3.6 for more details)). The prediction accuracy was estimated using 100 training-test scenarios. In each one of these scenarios, a class-balanced subset of maximum size was randomly selected and split into a class-balanced test set (10 % of the total size) and a class-balanced training (90 % of the total size). The final prediction accuracy was calculated as the mean of the 100 test accuracies.

The linear RSLVQ (using only one prototype per class) achieved a prediction accuracy of 78.3 %. This means, the site could be predicted from the grey matter volume data with a high accuracy. Due to these strong interactions between grey matter volume data and site, one obtains misleading accuracies if a model is trained on the grey matter volume data to predict the drinking behaviour without accounting for the site, as explained at length in Section 2.4.3.

As we have explained in Section 5.3.4, we approach the removal of intrinsic site information in the imaging data by subtracting the respective site-mean of the grey matter volume data for each subject. We show that this linear model suffices to remove most of the intrinsic site information. The same setup as above was used on the site-corrected grey matter volume data. The best achieving version of the RSLVQ was one with four prototypes per class and it achieved an accuracy of 15.8 %. This is only slightly higher than chance (12.5%). We hence assume that we have removed most of the site information from the data.

HISTOGRAMS OF FEATURES A.2

Personality features A.2.1

Figures A.1 and A.2 show the histograms per personality feature, for BL and FU2, respectively.

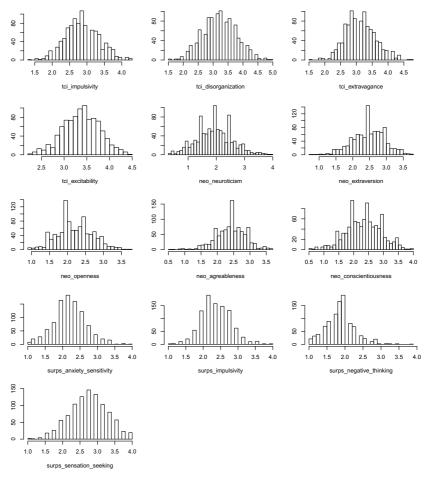


Figure A.1: Personality features for BL (approximate age of 14).

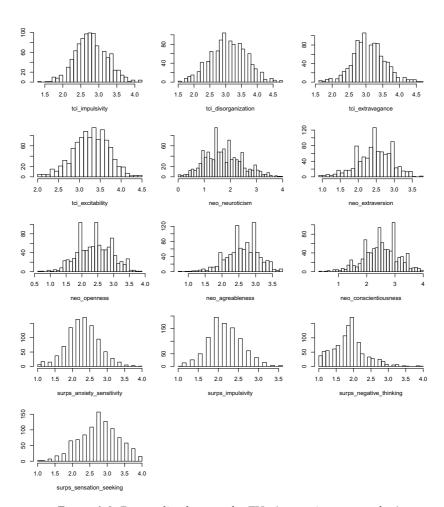


Figure A.2: Personality features for FU2 (approximate age of 19).

A.2.2 Life events features

Figures A.3 and A.4 show the histograms per life events feature, for BL and FU2, respectively.

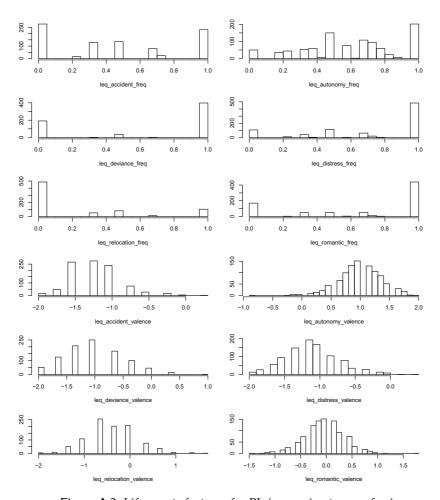


Figure A.3: Life events features for BL (approximate age of 14).

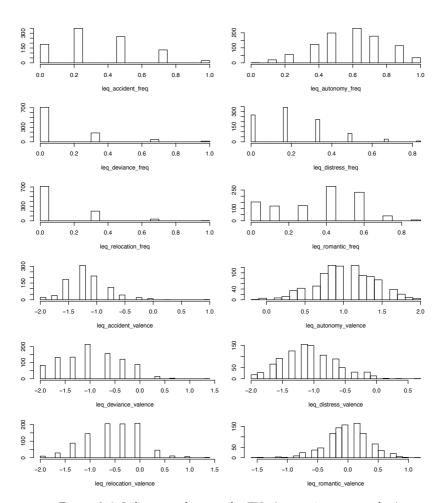


Figure A.4: Life events features for FU2 (approximate age of 19).

A.2.3 Grey matter volume features

Figures A.5 and A.6 show the histograms per grey matter volume feature, for BL and FU2, respectively.

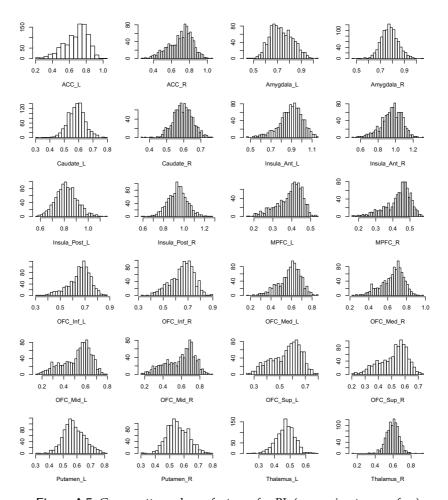


Figure A.5: Grey matter volume features for BL (approximate age of 14).

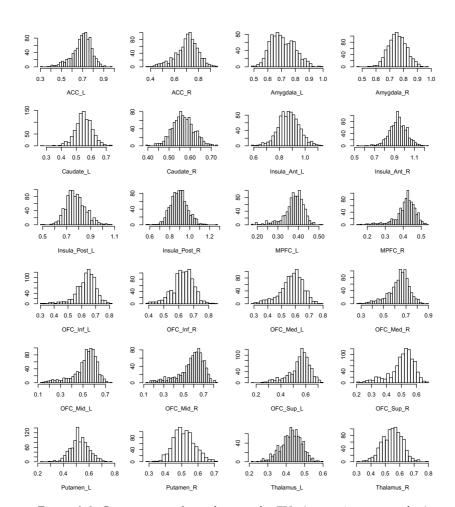


Figure A.6: Grey matter volume features for FU2 (approximate age of 19).

A.3 RESULTS FOR PREDICTION ANALYSIS

Tables A.1 and A.2 show more detailed results of the prediction analysis in both the cross-sectional and the longitudinal setting for grey matter volume and psychosocial features, respectively. These are the values on which Figures 5.2, 5.3, 5.4 and 5.5 are based.

Table A.1: Balanced accuracies for prediction of drinking behaviour at 19 with grey matter volume features from age 19 (cross-sectional setting) and from age 14 (longitudinal setting) using three different classifiers, linear support vector machine (I-SVM), Naïve Bayes (naïveBayes), and Robust Soft Learning Vector Quantisation (RSLVQ). The classifiers use either all, only female or only male subjects. The star indicates significance evaluated by label permutation and after correction for multiple testing at level $\alpha = 0.05$.

	Cross-sectional			Longitudinal		
	all	female	male	all	female	male
l-SVM naïveBayes RSLVQ	0.566* 0.561* 0.561*	0.586* 0.575* 0.560*	0.515 0.580* 0.537	0.482 0.495 0.504	0.531 0.556 0.526	0.517 0.531 0.496

Table A.2: Balanced accuracies for prediction of drinking behaviour at 19 with psychosocial features from age 19 (cross-sectional setting) and from age 14 (longitudinal setting) using three different classifiers, linear support vector machine (I-SVM), Naïve Bayes (naïveBayes), and Robust Soft Learning Vector Quantisation (RSLVQ). The classifiers use either all, only female or only male subjects. The star indicates significance evaluated by label permutation and after correction for multiple testing at level $\alpha = 0.05$.

	Cross-sectional			Longitudinal		
	all	female	male	all	female	male
l-SVM naïveBayes RSLVQ	0.694* 0.676* 0.687*	0.636* 0.616* 0.601*	0.704* 0.678* 0.678*	0.586* 0.593* 0.558*	0.565* 0.571* 0.572*	0.585* 0.595* 0.534

A.4 RESULTS FOR ASSOCIATION ANALYSIS

Tables A.3, A.4 and A.5 show more detailed results for the association analysis, as displayed in Figures 5.6, 5.7, 5.9, 5.8 and 5.10.

A.4.1 Cross-sectional setting

Table A.3: Detailed results from the ReLL method 2.5.2 for association analysis on grey matter volume features in the cross-sectional setting using $N_{rep} = 100$, as displayed in Figure 5.6. $\bar{\beta}$ is the median of N_{rep} coefficients of the respective feature, resulting from N_{rep} repetitions of lasso-regularised logistic regression with the binary drinking behaviour label as dependent variable and the respective feature and the confounds gender (for the case with all subjects) and site as independent variables. If the feature is survives lasso-regularisation in at least $N_{rep}/2$ repetitions, then the median estimate will be non-zero. Significance testing is done according to 2.5.3 and corrected for multiple testing at $\alpha = 0.05$.

		all	fe	emale	1	male	
	β	selected (max: 100)	β	selected (max: 100)	β	selected (max: 100)	
ACC_L	-0.195*	100	-0.224*	100	-0.159	89	
ACC_R	-0.165*	100	-0.242*	100	0.000	13	
Amygdala_L	-0.158*	91	-0.253*	98	0.000	О	
Amygdala_R	0.000	О	0.000	О	0.000	О	
Caudate_L	0.000	4	0.000	О	0.000	О	
Caudate_R	0.000	13	0.000	46	0.000	О	
Insula_Ant_L	-0.149*	100	0.000	43	-0.177	97	
Insula_Ant_R	-0.095	86	0.000	О	-0.111	58	
Insula_Post_L	0.000	О	0.000	2	0.000	2	
Insula_Post_R	0.000	О	0.000	17	0.000	16	
MPFC_L	-0.208*	100	-0.270*	100	-0.132	70	
MPFC_R	-0.135*	99	-0.192	96	0.000	5	
OFC_Inf_L	-0.106	79	-0.234*	100	0.000	О	
OFC_Inf_R	-0.145*	100	-0.141	81	0.000	43	
OFC_Med_L	-0.238*	100	-0.285*	100	-0.209*	99	
OFC_Med_R	-0.247*	100	-0.271*	100	-0.210*	99	
OFC_Mid_L	-0.139*	97	-0.147	64	0.000	34	
OFC_Mid_R	-0.167*	100	-0.207	98	0.000	29	
OFC_Sup_L	-0.170*	100	-0.237*	100	0.000	19	
OFC_Sup_R	-0.185*	100	-0.209	99	-0.133	73	
Putamen_L	0.000	О	0.000	38	0.000	33	
Putamen_R	0.000	О	0.000	О	0.000	9	
Thalamus_L	-0.216*	100	-0.404*	100	0.000	О	
Thalamus_R	-0.211*	100	-0.292*	100	0.000	40	

Table A.4: Detailed results from the ReLL method 2.5.2 for association analysis on psychosocial features in the cross-sectional setting using $N_{rep} = 100$, as displayed in Figures 5.7 and 5.9. $\bar{\beta}$ is the median of N_{rep} coefficients of the respective feature, resulting from N_{rep} repetitions of lasso-regularised logistic regression with the binary drinking behaviour label as dependent variable and the respective feature and the confounds gender (for the case with all subjects) and site as independent variables. If the feature is survives lasso-regularisation in at least $N_{rep}/2$ repetitions, then the median estimate will be non-zero. Significance testing is done according to 2.5.3 and corrected for multiple testing at $\alpha = 0.05$.

		all	fe	emale	male	
	β	selected (max: 100)	β	selected (max: 100)	β	selected (max: 100)
neo_neuroticism	0.000	О	0.000	2	0.000	1
neo_extraversion	0.513*	100	0.335*	100	0.659*	100
neo_openness	0.247*	100	0.142	85	0.348*	100
neo_agreeableness	-0.177*	100	-0.178	89	-0.158	88
neo_conscientiousness	-0.363*	100	-0.352*	100	-0.369*	100
tci_impulsivity	0.478*	100	0.309*	100	0.633*	100
tci_disorderliness	0.449*	100	0.205	100	0.674*	100
tci_extravagance	0.724*	100	0.515*	100	0.930*	100
tci_excitability	0.411*	100	0.304*	100	0.525*	100
surps_anxiety_sensitivity	0.000	4	0.000	4	0.000	О
surps_hopelessness	0.000	О	0.000	24	0.000	40
surps_impulsivity	0.345*	100	0.271*	100	0.419*	100
surps_sensation_seeking	0.352*	100	0.336*	100	0.351*	100
leq_accident_freq	0.000	1	0.000	5	0.000	О
leq_accident_valence	0.000	О	0.000	О	-0.091	59
leq_autonomy_freq	0.223*	100	0.207*	100	0.221*	100
leq_autonomy_valence	0.144*	100	0.096	57	0.179*	100
leq_deviance_freq	0.420*	100	0.404*	100	0.411*	100
leq_deviance_valence	0.427*	100	0.487*	100	0.360*	100
leq_distress_freq	0.077	93	0.093	80	0.000	О
leq_distress_valence	0.000	11	0.000	4	0.000	20
leq_relocation_freq	0.000	О	0.000	o	0.000	О
leq_relocation_valence	0.000	О	0.000	О	0.000	О
leq_sexuality_freq	0.370*	100	0.380*	100	0.358*	100
leq_sexuality_valence	0.000	О	0.167*	100	0.000	9

A.4.2 Longitudinal setting

Table A.5: Detailed results from the ReLL method 2.5.2 for association analysis on psychosocial features in the longitudinal setting using $N_{rep} = 100$, as displayed in Figures 5.8 and 5.10. $\bar{\beta}$ is the median of N_{rep} coefficients of the respective feature, resulting from N_{rep} repetitions of lasso-regularised logistic regression with the binary drinking behaviour label as dependent variable and the respective feature and the confounds gender (for the case with all subjects) and site as independent variables. If the feature is survives lasso-regularisation in at least $N_{rep}/2$ repetitions, then the median estimate will be non-zero. Significance testing is done according to 2.5.3 and corrected for multiple testing at $\alpha = 0.05$.

		all	f	emale		male
	β	selected (max: 100)	β	selected (max: 100)	β	selected (max: 100)
neo_neuroticism	0.000	О	0.000	1	0.000	О
neo_extraversion	0.449*	100	0.359*	100	0.531*	100
neo_openness	0.000	2	0.000	1	0.000	О
neo_agreeableness	-0.130*	91	0.000	37	0.000	47
neo_conscientiousness	-0.166*	100	0.000	24	-0.221	99
tci_impulsivity	0.454*	100	0.328*	100	0.621*	100
tci_disorderliness	0.470*	100	0.322*	100	0.637*	100
tci_extravagance	0.586*	100	0.480*	100	0.690*	100
tci_excitability	0.339*	100	0.215	93	0.564*	100
surps_anxiety_sensitivity	0.000	31	0.000	О	0.000	34
surps_hopelessness	0.333*	100	0.272*	99	0.360*	100
surps_impulsivity	0.000	38	0.000	26	0.000	2
surps_sensation_seeking	0.215*	100	0.184	89	0.226	98
leq_accident_freq	0.000	2	-0.118	61	0.000	О
leq_accident_valence	0.000	О	0.000	О	0.000	О
leq_autonomy_freq	0.000	О	0.000	35	0.150	97
leq_autonomy_valence	0.000	О	0.000	О	0.000	1
leq_deviance_freq	0.000	4	0.000	47	0.000	О
leq_deviance_valence	0.000	48	0.000	46	0.000	О
leq_distress_freq	0.000	О	0.000	О	0.000	2
leq_distress_valence	0.000	О	0.000	О	0.000	1
leq_relocation_freq	0.000	1	0.000	13	0.000	О
leq_relocation_valence	0.101	94	0.147	92	0.000	12
leq_sexuality_freq	0.000	29	0.000	24	0.000	1
leq_sexuality_valence	0.000	0	0.000	1	0.000	0

B | SUPPLEMENTS FOR CHAPTER 6

B.1 EXTRACTION AND MODELLING OF SSRT

B.1.1 Sample inhibition functions

Figure B.1 shows estimated inhibition functions (using piecewise polynomial splines) of two different subjects. One example is a "good" example – one where the inhibition function is increasing, as expected – and one is a "bad" example – one where the inhibition function fluctuates.

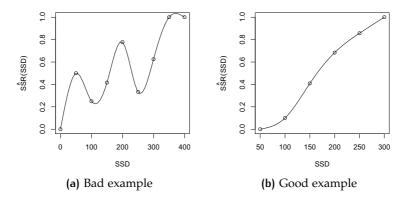


Figure B.1: Sample inhibition functions for two different subjects. SSD: stop signal delay, in ms. $\widehat{SSR}(SSD)$: probability of stopping for stop signal delay SSD, estimated by relative frequency of successful stopping.

B.1.2 Comparison of SSRT extraction methods

In Figure B.2 we show comparisons of two SSRT extraction methods described in Section 6.3.1 – the mean method and the integration method. We show the comparison both for BL and for FU2 data. Generally, the integration method produces slightly higher SSRTs, yet they are very correlated (Person's correlation for BL: $\hat{\rho} = 0.51$, p-value $< 10^{-15}$, for FU2: $\hat{\rho} = 0.76$, p-value $< 10^{-15}$).

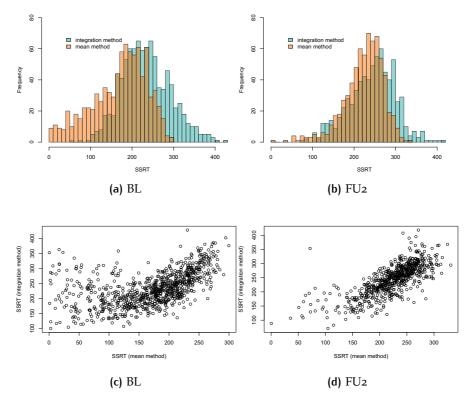


Figure B.2: Comparisons two methods (mean method and integration method) of SSRT extraction. For the mean method we excluded negative SSRTs. Outliers were removed according to the thresholding described in Section 6.3.2.

B.1.3 Time-resolved SSRT

In Figure B.3 we show the average SSRT across subjects computed in a time-resolved manner, i.e. for each time point and each subject we took the 100 trials around the time point to compute the SSRT with the integration method (see Section 6.3.2 for details on the integration method). The differences that were found in the SSRT cross-sectionally in the association analysis (see Section 6.4.2) are present also in a time-resolved manner heavy drinkers seem to consistently show a longer SSRT compared to light drinkers. This difference is not present for the longitudinal setting, i.e. for the SSRTs at BL separated into light and heavy drinkers from FU2. This is again consistent with the findings from Section 6.4.2. Note, however, that here we do not account for the gender confound.

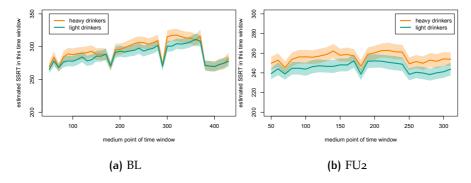


Figure B.3: SSRT extracted by the integration method, time-resolved for the length of the experiment with 95% confidence intervals. For each time point and each subject we took the 100 trials around the time point to compute the SSRT. We furthermore depict the SSRTs separately for the light and heavy drinkers (where drinking behaviour is as usual assessed at FU2, i.e. at approximate age 19).

B.2 PREPROCESSING OF FUNCTIONAL IMAGING DATA

B.2.1 Custom EPI template

A set of 240 randomly-selected subjects (30 for each of the eight acquisition sites) EPI sequences for two protocols (SST and Monetary Incetive Delay (MID), i.e. one from each session) were spatially-realigned, and their temporal-mean image were rigidly co-registered to their respective anatomical image. Those anatomical images were warped to the MNI space using the SPM8 non-linear normalisation algorithm, and the corresponding transformations parameters were applied on the mean-EPI images. From those 480 MNI-warped mean EPI images, a simple xor-based metric is computed to quantify the overlap quality between individual EPI masks and the MNI mask. Using this metric we discarded 17 % outliers, and the remaining 400 mean-images are averaged, then smoothed with a 5mm FWHM (full-width at half maximum) Gaussian kernel. The final template image is 53 x 63 x 46 voxels, with 3mm x 3mm x 3mm voxel resolution.

B.2.2 Quality control of imaging data

The following was performed centrally by the IMAGEN consortium Schumann et al. 2010:

In order to detect and remove blatant outliers, an automatic procedure has been used. In this context, outliers refers mainly to data with obvious problem, such as missing values or heavy misregistration, but also subject maps which exhibit largely disparate activations profile for expected contrasts, which usually reflects task behavioural misdoing. Those had been applied blindly database-wide, and is therefore not biased to traits of interest of the current study.

First, at the time series level, a mean-square based metric identifies spikes as specific slice- and time-limited burst, and those are corrected for using temporal linear interpolation. Second, following the SPM realignment procedure, another metric flags remaining time-points where the volume mean-square difference to the (robust) average raises above a threshold precalibrated on 180 random subjects, and the subject data is discarded if more than one such bad volume occurs per minute.

At the activations maps level, an automatic procedure ranks the maps which departs unreasonably from the group.

The procedure rely on a distance metric from each individual activation profile to the average one, computed as follows. In MNI space, a set of

Region of Interest (ROI) is defined based either on anatomical or functional knowledge for the relevant contrast, and every subject's average effect over each of those ROI is gathered. Then \times p resulting matrix, where n is the number of subjects and p is the number of ROIs, is used to estimate to p * p covariance matrix of the ROIs signal accross the subjects. To improve the robustness of the covariance measurement, the 0.5 % upper and lowest values had been trimmed out each component before the estimation; Additionally, the covariance matrix is inverted using an Eigen-decomposition for which the 25 % least informative component are nullified. This allows to further compute the Mahalanobis distance of every subjects to the group mean. As it had been checked using subgroups that the results does not depend much on the group size anymore after a certain size, the number of subjects involved to estimate to metric has been fixed to 1000. The ability to extract this multivariate, Mahalanobis distance from every subject to the fixed group average allows to quickly compute their score and compare it to a relevant non-parametric threshold, to flag them as outliers. A visual checking did actually confirm that the more extreme scores corresponds to obvious artefacts.

B.3 HISTOGRAMS OF FEATURES

Cognitive features B.3.1

Figures B.4 and B.5 show the histograms per feature, for BL and FU2, respectively. Outliers for the SST are already removed, according to the thresholding introduced in Section 6.3.1.

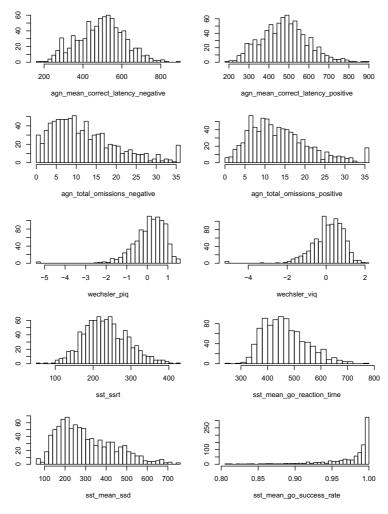


Figure B.4: Cognitive features for BL (approximate age of 14).

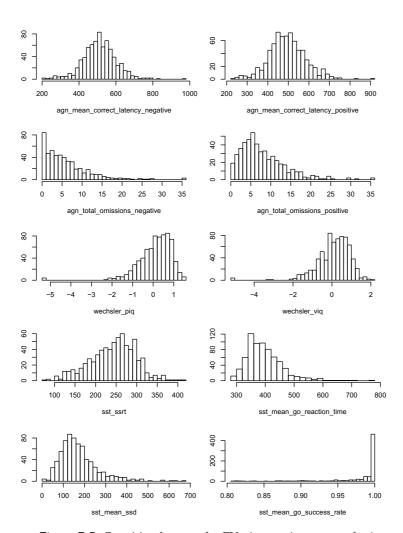


Figure B.5: Cognitive features for FU2 (approximate age of 19).

B.3.2 Functional features

Figures B.6 and B.7 show the histograms per feature of the functional imaging features from the stopSuccess contrast, for BL and FU2, respectively. Figures B.8 and B.9 show the histograms per feature of the functional imaging features from the stopSuccess-stopFail contrast, for BL and FU2, respectively. Outliers for the SST are removed, according to the thresholding introduced in Section 6.3.1.

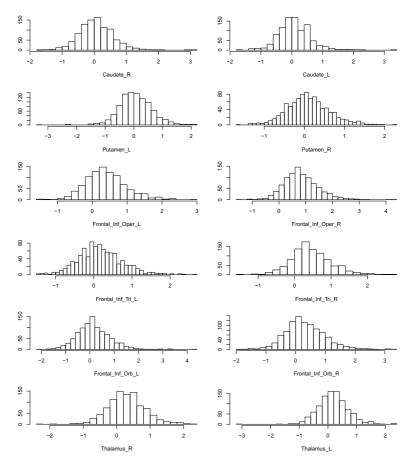


Figure B.6: Functional imaging features from the stopSuccess contrast for BL (approximate age of 14).

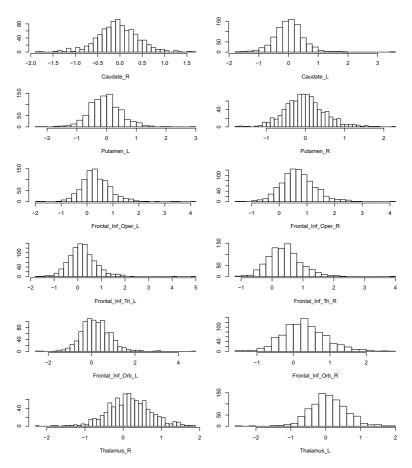


Figure B.7: Functional imaging features from the stopSuccess contrast for FU2 (approximate age of 19).

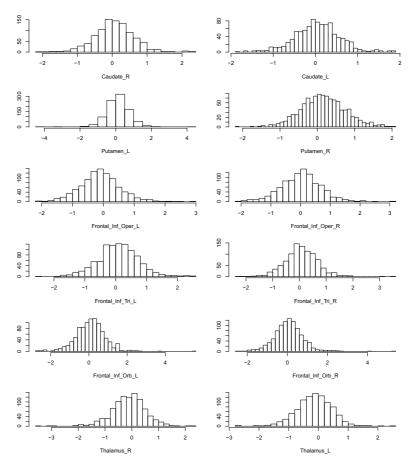


Figure B.8: Functional imaging features from the stopSuccess-stopFail contrast for BL (approximate age of 14).

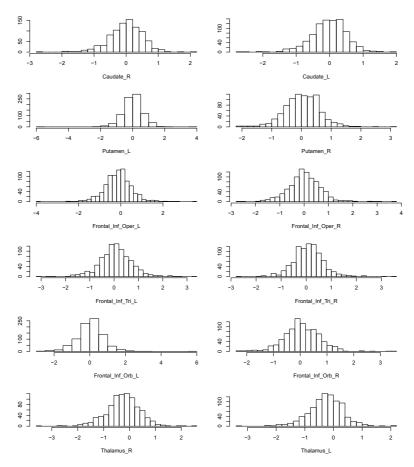


Figure B.9: Functional imaging features from the stopSuccess-stopFail contrast for FU2 (approximate age of 19).

B.4 IMPUTATION INVESTIGATION

Tables B.1 and B.2 show the prediction accuracies obtained for different imputation methods with the same parameters as described in Section 6.3.4, for the cross-sectional and longitudinal setting, respectively. Note that the results for mean imputation are exactly the results shown in Figure 6.3.

Note that the different imputation methods produce very similar prediction accuracies. We hence opted for the simplest approach - mean imputation – in our main analyses.

Table B.1: Balanced accuracies for prediction of drinking behaviour at 19 with cognitive features from the same age (cross-sectional setting) for different imputation schemes (see Section 2.2.2 for a brief description); mean: mean imputation, median: median imputation, sample: sample imputation, k-NN: k-nearest neighbour imputation. linReg: imputation by deterministic linear regression, rF: random forest imputation; Logistic: logistic regression classifier, Random Forest: random forest classifier, naïve Bayes: naïve Bayes classifier.

Imputation method	Logistic	Random Forest	naive Bayes
mean	0.5823	0.6081	0.6087
median	0.5749	0.6095	0.6055
sample	0.5892	0.5953	0.5996
3-NN	0.6019	0.6010	0.6117
5-NN	0.6069	0.6171	0.6123
10-NN	0.6199	0.6189	0.6169
linReg	0.5884	0.5975	0.6120
rF	0.5825	0.5804	0.5982

Table B.2: Balanced accuracies for prediction of drinking behaviour at 19 with cognitive features from age 14 (longitudinal setting) for different imputation schemes (see Section 2.2.2 for a brief description); mean: mean imputation, median: median imputation, sample: sample imputation, k-NN: k-nearest neighbour imputation. linReg: imputation by deterministic linear regression, rF: random forest imputation; Logistic: logistic regression classifier, Random Forest: random forest classifier, naive Bayes: naive Bayes classifier.

Imputation method	Logistic	Random Forest	naive Bayes
mean	0.5518	0.5642	0.5526
median	0.5526	0.5652	0.5531
sample	0.5609	0.5412	0.5493
3-NN	0.5604	0.5453	0.5528
5-NN	0.5578	0.5507	0.5468
10-NN	0.5565	0.5639	0.5473
linReg	0.5492	0.5567	0.5498
rF	0.5521	0.5566	0.5500

B.5 INVESTIGATION ON COMPLETE DATA

Table B.3 shows the prediction accuracies on cognitive data in the crosssectional and longitudinal setting using mean-imputed data and subsetting on subjects for which there are no missing values. The mean-imputed cases correspond exactly to the accuracies in the first lines of Tables B.2 and B.1 and also to the ones shown in Figure 6.3.

Table B.3: Balanced accuracies for prediction of drinking behaviour at 19 with cognitive features from age 19 (cross-sectional setting) and from age 14 (longitudinal setting) using complete data, i.e. only using subjects with no missing values and for mean imputed data; N: sample size, Logistic: logistic regression classifier, Random Forest: random forest classifier, naive Bayes: naive Bayes classifier.

Setting	Data	Ν	Logistic	Random Forest	naive Bayes
cross-sectional	complete	481	0.6144	0.5878	0.6095
cross-sectional	mean-imputed	723	0.5823	0.6068	0.6087
longitudinal	complete	735	0.5488	0.5582	0.5692
longitudinal	mean-imputed	938	0.5518	0.5642	0.5526

в.6 INVESTIGATION OF OTHER FUNCTIONAL FEA-TURE EXTRACTION METHODS

In addition to using the mean value per region of interest for the analysis of functional data from the SST, we performed comparisons to another summary measure. Specifically, we followed the recommendations given in Tong et al. 2016 to use single value summary measures representing values from the top most-activated voxels within a region of interest. Tong et al. 2016 show that these types of measures are more powerful at detecting group differences between e.g. patients and controls. We hence use the mean value from the top 10 % voxels within one region of interest.

Table B.4 shows the results from prediction analysis on both the **stopSuccess**stopFail and the stopSsuccess contrast, in the cross-sectional and longitudinal setting, for the usual (mean) measure and the new measures. Prediction accuracies are very similar.

Table B.4: Balanced accuracies for prediction of drinking behaviour at 19 with functional imaging features from age 19 (cross-sectional setting) an from age 14 (longitudinal setting) for two different summary measure. Mean: mean of all voxels within each of the 12 regions of interest described in Section 6.3.2, top10%: mean of top 10 % of voxels within each of the 12 regions of interest. Logistic: logistic regression classifier, Random Forest: random forest classifier, naive Bayes: naive Bayes classifier.

Setting	Summary measure	Data	Logistic	Random Forest	Naive Bayes
cross-sectional	mean	stopSuccess-stopFail	0.5364	0.5510	0.5341
cross-sectional	top10%	stopSuccess-stopFail	0.5321	0.5604	0.5378
cross-sectional	mean	stopSuccess	0.5109	0.5022	0.5273
cross-sectional	top10%	stopSuccess	0.5034	0.5436	0.5343
longitudinal	mean	stopSuccess-stopFail	0.5149	0.5050	0.5001
longitudinal	top10%	stopSuccess-stopFail	0.5164	0.5057	0.4870
longitudinal	mean	stopSuccess	0.5119	0.5044	0.4854
longitudinal	top10%	stopSuccess	0.5042	0.4851	0.4986

B.7 INVESTIGATION OF GENDER DIFFERENCES

We investigated the predictive power of cognitive and functional imaging features, separately for genders. Tables B.5, B.6 and B.7 show the balanced accuracies for prediction of drinking behaviour at 19 with features from age 19 (cross-sectional setting) and age 14 (longitudinal setting). All accuracies are very similar. Note that the slight differences of the scenarios with all subjects compared to the results presented in Section 6.4.1 are due to the fact that we additionally balanced for label here. A balanced label is important for the gender-separate scenarios, since otherwise the classes are too imbalanced.

Table B.5: Balanced accuracies for prediction of drinking behaviour at 19 with functional imaging features from age 19 (cross-sectional setting) an from age 14 (longitudinal setting) from cognitive features for all subjects and for each gender separately. For all scenarios, label is balanced. For the scenario with both genders, we further balance for gender. N: sample size, Logistic: logistic regression classifier, Random Forest: random forest classifier, naive Bayes: naive Bayes classifier.

Subjects	Setting	N	Logistic	Random Forest	Naive Bayes
all	cross-sectional	723	0.5779	0.5979	0.6044
female	cross-sectional	381	0.5571	0.5608	0.5688
male	cross-sectional	342	0.5779	0.5979	0.6044
all	longitudinal	938	0.5561	0.5589	0.5508
female	longitudinal	498	0.5622	0.5528	0.5575
male	longitudinal	440	0.5508	0.5144	0.5425

Table B.6: Balanced accuracies for prediction of drinking behaviour at 19 with functional imaging features from age 19 (cross-sectional setting) an from age 14 (longitudinal setting) from functional imaging features from stopSuccess-stopFail contrast for all subjects and for each gender separately. For all scenarios, label is balanced. For the scenario with both genders, we further balance for gender. N: sample size, Logistic: logistic regression classifier, Random Forest: random forest classifier, naive Bayes: naive Bayes classifier.

Subjects	Setting	Ν	Logistic	Random Forest	Naive Bayes
all female male all female male	cross-sectional cross-sectional cross-sectional longitudinal longitudinal	873 450 423 943 501 442	0.5316 0.5207 0.5356 0.5050 0.4856 0.5031	0.5546 0.5629 0.5465 0.5067 0.4738 0.5264	0.5341 0.5643 0.5441 0.5055 0.4875 0.4861

Table B.7: Balanced accuracies for prediction of drinking behaviour at 19 with functional imaging features from age 19 (cross-sectional setting) an from age 14 (longitudinal setting) from functional imaging features from stopSuccess contrast for all subjects and for each gender separately. For all scenarios, label is balanced. For the scenario with both genders, we further balance for gender. N: sample size, Logistic: logistic regression classifier, Random Forest: random forest classifier, naive Bayes: naive Bayes classifier.

Subjects	Setting	N	Logistic	Random Forest	Naive Bayes
all	cross-sectional	873	0.4923	0.4975	0.5146
female	cross-sectional	450	0.5039	0.5482	0.5246
male	cross-sectional	423	0.5282	0.4679	0.5121
all	longitudinal	944	0.5117	0.4955	0.4769
female	longitudinal	501	0.4822	0.4900	0.4819
male	longitudinal	443	0.5111	0.5039	0.4997

в.8 RESULTS FOR ASSOCIATION ANALYSIS

We show in Tables B.8, B.9 and B.10 more detailed results for the association analysis performed on cognitive and functional imaging features in both cross-sectional and longitudinal settings. Note that in Figures 6.6 and 6.7 we have plotted the median values from these tables.

B.8.1 Cross-sectional setting

Table B.8: Detailed results from the ReLL method (see Section2.5.2) for association analysis on **cognitive features** in the cross-sectional setting using $N_{rep} = 100$, as displayed in Figure 6.6. Mean (median) is the mean (median) of N_{rep} coefficients of the respective feature, resulting from N_{rep} repetitions of lasso-regularised logistic regression with the binary drinking behaviour label as dependent variable and the respective feature and the confound gender as independent variable. If the feature is survives lasso-regularisation in at least $N_{rep}/2$ repetitions, then the median estimate will be non-zero. Significance testing is done according to Section 2.5.3 and corrected for multiple testing at $\alpha = 0.05$. N denotes the number of non-missing (see Table 6.1) and non-outlying (see Section 6.3.1) subjects for the respective feature and hence the number of subjects used for association analysis for the feature.

Name of feature	N	mean β	median β	Selected (max: 100)	Significance
agn_mean_correct _latency_negative	719	0.0005	0.0000	1	no
agn_mean_correct _latency_positive	720	0.0000	0.0000	O	no
agn_total_omissions _negative	504	-0.4996	-0.5080	100	yes
agn_total_omissions _positive	504	-0.3710	-0.3740	100	yes
wechsler_piq	721	0.0000	0.0000	O	no
wechsler_viq	721	0.0000	0.0000	O	no
sst_ssrt	695	0.1747	0.1750	100	yes
sst_mean_go _reaction_time	695	-0.0182	0.0000	23	no
sst_mean_ssd	695	-0.1750	-0.1770	99	yes
sst_mean_go _success_rate	695	-0.2659	-0.2690	100	yes

Table B.9: Detailed results from the ReLL method (see Section 2.5.2) for association analysis on functional imaging features from the stopSuccess-stopFail contrast in the cross-sectional setting using $N_{rep} = 100$, as displayed in Figure 6.7. Mean (median) is the mean (median) of N_{rep} coefficients of the respective feature, resulting from N_{rep} repetitions of lasso-regularised logistic regression with the binary drinking behaviour label as dependent variable and the respective feature and the confounds gender and site as independent variables. If the feature is survives lasso-regularisation in at least $N_{rep}/2$ repetitions, then the median estimate will be non-zero. Significance testing is done according to Section 2.5.3 and corrected for multiple testing at $\alpha = 0.05$. N denotes the number of non-missing (see Table 6.1) and non-outlying (see Section 6.3.1) subjects for the respective feature and hence the number of subjects used for association analysis for the feature.

Name of feature	N	mean β	median β	Selected (max: 100)	Significance
Caudate_L	873	-0.0103	0.0000	22	no
Caudate_R	873	-0.0300	0.0000	48	no
Putamen_L	873	-0.0167	0.0000	27	no
Putamen_R	873	-0.0768	-0.0082	87	no
Frontal_Inf_Oper_L	873	0.0000	0.0000	О	no
Frontal_Inf_Oper_R	873	0.0009	0.0000	3	no
Frontal_Inf_Tri_L	873	0.0000	0.0000	О	no
Frontal_Inf_Tri_R	873	0.0000	0.0000	О	no
Frontal_Inf_Orb_L	873	0.0068	0.0000	14	no
Frontal_Inf_Orb_R	873	0.0010	0.0000	4	no
Thalamus_L	873	-0.1598	-0.1598	100	yes
Thalamus_R	873	-0.1474	-0.1474	100	yes

B.8.2 Longitudinal setting

Table B.10: Detailed results from the ReLL method (see Section 2.5.2) for association analysis on **cognitive features** in the longitudinal setting using $N_{rep} = 100$, as displayed in Figure 6.6. Mean (median) is the mean (median) of N_{rep} coefficients of the respective feature, resulting from N_{rep} repetitions of lasso-regularised logistic regression with the binary drinking behaviour label as dependent variable and the respective feature and the confound gender as independent variable. If the feature is survives lasso-regularisation in at least N_{rep}/2 repetitions, then the median estimate will be non-zero. Significance testing is done according to Section 2.5.3 and corrected for multiple testing at $\alpha = 0.05$. N denotes the number of non-missing (see Table 6.1) and non-outlying (see Section 6.3.1) subjects for the respective feature and hence the number of subjects used for association analysis for the feature.

Name of feature	N	mean β	median β	Selected (max: 100)	Significance
agn_mean_correct _latency_negative	814	-0.1119	-0.1200	93	no
agn_mean_correct _latency_positive	816	-0.0823	-0.1040	79	no
agn_total_omissions _negative	835	-0.3173	-0.3170	100	yes
agn_total_omissions _positive	835	-0.3093	-0.3090	100	yes
wechsler_piq	936	0.0000	0.0000	O	no
wechsler_viq	936	0.0011	0.0000	2	no
sst_ssrt	907	0.0000	0.0000	O	no
sst_mean_go _reaction_time	907	-0.0094	0.0000	15	no
sst_mean_ssd	907	-0.1197	-0.1260	95	no
sst_mean_go _success_rate	907	-0.0047	-0.000	8	no

C | PUBLICATIONS

"Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example"

Pegah Kassraian-Fard, Caroline Matthis, Joshua H Balsters, Marloes H Maathuis, and Nicole Wenderoth (2016). "Promises, Pitfalls, and Basic Guidelines for Applying Machine Learning Classifiers to Psychiatric Imaging Data, with Autism as an Example". In: *Frontiers in Psychiatry* 7, p. 177

AUTHOR CONTRIBUTIONS PF: main contribution in drafting the article, code implementation, data analysis and interpretation, as well as contributions to the modeling of data. CM and JB: contributions to the modeling of data, code implementation, data analysis and interpretation, and critical revision of draft. MM and NW: contributions to the modeling of data, data interpretation, and critical revision of draft.

"Risk profiles for heavy drinking in adolescence: Differential effects of gender"

Sambu Seo, Anne Beck, Caroline Matthis, Alexander Genauck, Tobias Banaschewski, Arun L.W. Bokde, Uli Bromberg, Christian Büchel, Erin Burke Quinlan, Herta Flor, Vincent Frouin, Hugh Garavan, Penny Gowland, Bernd Ittermann, Jean Luc Martinot, Marie Laure Paillère Martinot, Frauke Nees, Dimitri Papadopoulos Orfanos, Luise Poustka, Sarah Hohmann, Juliane H. Fröhner, Michael N. Smolka, Henrik Walter, Robert Whelan, Sylvane Desrivières, Andreas Heinz, Gunter Schumann, and Klaus Obermayer (2018). "Risk profiles for heavy drinking in adolescence: Differential effects of gender". In: *Addiction Biology*

AUTHOR CONTRIBUTIONS TB, AB, UB, CB, EQ, HF, VF, HG, PG, BI, JM, MM, FN, DO, LP, SH, JF, MS, HW, RW, SV, GS and AH collected the data. SS pre-processed the imaging data and designed the analysis method and evaluated the prediction and association analyses, CM and KO contributed to these analyses. SS, AG, CM and AB drafted the manuscript. KO, AH, GS

and SD designed the study, discussed the results and thoroughly revised the manuscript for important intellectual content.

"Cue-induced effects on decision-making distinguish subjects with gambling disorder from healthy controls"

Alexander Genauck, Caroline Matthis, Milan Andrejevic, Lukas Ballon, Francesca Chiarello, Katharina Duecker, Andreas Heinz, Norbert Kathmann, and Nina Romanczuk-Seiferth (Dec. 2018). "Neural correlates of cue-induced changes in decision-making distinguish subjects with gambling disorder from healthy controls". In: bioRxiv, p. 498725

AUTHOR CONTRIBUTIONS AG designed the experiment, collected the data, analyzed the data, wrote the manuscript. CM reviewed the machine-learning algorithm and revised the manuscript. MA collected data, revised manuscript. AH revised the manuscript, oversaw manuscript drafting. NK revised the manuscript, advised first author. FC analyzed data, revised manuscript. KD collected data, analysed data, revised manuscript. NRS designed and supervised study and experiment, oversaw manuscript drafting and data analysis.

"Neural correlates of cue-induced changes in decision-making distinguish subjects with gambling disorder from healthy controls"

Alexander Genauck, Milan Andrejevic, Katharina Brehm, Caroline Matthis, Andreas Heinz, Andre Weinreich, Norbert Kathmann, and Nina Romanczuk-Seiferth (Mar. 2019). "Cue-induced effects on decision-making distinguish subjects with gambling disorder from healthy controls". In: bioRxiv, p. 564781

AUTHOR CONTRIBUTIONS AG designed the experiment, collected the data, analyzed the data, and wrote the manuscript. MA implemented the ratings and questionnaire electronically, analysed the ratings data, and revised the manuscript. KB collected data and revised the manuscript. CM reviewed the machine-learning algorithm and revised the manuscript. AH revised the manuscript, and oversaw manuscript drafting and data analyses. AW revised the manuscript and oversaw implementation of experiment in the lab. NK revised the manuscript and, advised first author. NRS designed and supervised study and experiment, and oversaw manuscript drafting and data analyses.

BIBLIOGRAPHY

- Ahmadi, Aral, Godfrey D Pearlson, Shashwath A Meda, Alecia Dager, Marc N Potenza, Rivkah Rosen, Carol S Austad, Sarah A Raskin, Carolyn R Fallahi, Howard Tennen, Rebecca M Wood, and Michael C Stevens (Oct. 2013). "Influence of Alcohol Use on Neural Response to Go/No-Go Task in College Drinkers". In: *Neuropsychopharmacology* 38.11, pp. 2197–2208.
- American Psychiatric Association (2014). *Desk reference to the diagnostic crite- ria from DSM-5* (®). American Psychiatric Pub.
- Aron, Adam R, Tim E Behrens, Steve Smith, Michael J Frank, and Russell A Poldrack (Apr. 2007). "Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (MRI) and functional MRI." In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 27.14, pp. 3743–52.
- Aron, Adam R and Russell A Poldrack (2006). "Cortical and subcortical contributions to stop signal response inhibition: role of the subthalamic nucleus". In: *Journal of Neuroscience* 26.9, pp. 2424–2433.
- Band, Guido P.H., Maurits W. van der Molen, and Gordon D. Logan (Feb. 2003). "Horse-race model simulations of the stop-signal procedure". In: *Acta Psychologica* 112.2, pp. 105–142.
- Beck, Anne, Torsten Wüstenberg, Alexander Genauck, Jana Wrase, Florian Schlagenhauf, Michael N. Smolka, Karl Mann, and Andreas Heinz (Aug. 2012). "Effect of Brain Structure, Brain Function, and Brain Connectivity on Relapse in Alcohol-Dependent Patients". In: *Archives of General Psychiatry* 69.8, p. 842.
- Bednarski, Sarah R., Emily Erdman, Xi Luo, Sheng Zhang, Sien Hu, and Chiang-Shan R. Li (May 2012). "Neural Processes of an Indirect Analog of Risk Taking in Young Nondependent Adult Alcohol Drinkers-An fMRI Study of the Stop Signal Task". In: *Alcoholism: Clinical and Experimental Research* 36.5, pp. 768–779.
- Bellis, Michael D., Anandhi Narasimhan, Dawn L. Thatcher, Matcheri S. Keshavan, Paul Soloff, and Duncan B. Clark (Sept. 2005). "Prefrontal Cortex, Thalamus, and Cerebellar Volumes in Adolescents and Young Adults with Adolescent-Onset Alcohol Use Disorders and Comorbid Mental Disorders". In: *Alcoholism: Clinical and Experimental Research* 29.9, pp. 1590–1600.

- Bexkens, A., L. Ruzzano, A. M. L. Collot d' Escury-Koenigs, M. W. Van der Molen, and H. M. Huizenga (Jan. 2014). "Inhibition deficits in individuals with intellectual disability: a meta-regression analysis". In: Journal of Intellectual Disability Research 58.1, pp. 3–16.
- Bezdjian, Serena, Laura A Baker, Dora Isabel Lozano, and Adrian Raine (June 2009). "Assessing inattention and impulsivity in children during the Go/NoGo task." In: The British journal of developmental psychology 27.Pt 2, pp. 365-83.
- Boehler, C. Nicolas, L. Gregory Appelbaum, Ruth M. Krebs, Jens-Max Hopf, and Marty G. Woldorff (Apr. 2012). "The influence of different Stopsignal response time estimation procedures on behavior-behavior and brain-behavior correlations". In: Behavioural Brain Research 229.1, pp. 123-
- Boothby, Lisa A. and Paul L. Doering (June 2005). "Acamprosate for the treatment of alcohol dependence". In: Clinical Therapeutics 27.6, pp. 695-
- Breiman, Leo (1996). "Bagging predictors". In: Machine learning 24.2, pp. 123–
- Breiman, Leo (Oct. 2001). "Random Forests". In: Machine Learning 45.1, pp. 5–
- Breiman, Leo (2017). Classification and regression trees. Routledge.
- Brodersen, Kay Henning, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann (Aug. 2010). "The Balanced Accuracy and Its Posterior Distribution". In: 2010 20th International Conference on Pattern Recognition. IEEE, pp. 3121-3124.
- Bühler, Mira and Karl Mann (Oct. 2011). "Alcohol and the Human Brain: A Systematic Review of Different Neuroimaging Methods". In: Alcoholism: Clinical and Experimental Research 35.10, pp. 1771–1793.
- Buja, A. and L. Brown (2014). Discussion: "A Significance test for the lasso".
- Cabitza, Federico, Raffaele Rasoini, and Gian Franco Gensini (Aug. 2017). "Unintended Consequences of Machine Learning in Medicine". In: JAMA 318.6, p. 517.
- CANTAB (2018). *CANTAB*(R) [Cognitive assessment software].
- Carlen, P L, G Wortzman, R C Holgate, D A Wilkinson, and J C Rankin (June 1978). "Reversible cerebral atrophy in recently abstinent chronic alcoholics measured by computed tomography scans." In: Science (New York, N.Y.) 200.4345, pp. 1076-8.
- Carlen, P.L., D.A. Wilkinson, G. Wortzman, and R. Holgate (Nov. 1984). "Partially Reversible Cerebral Atrophy and Functional Improvement

- in Recently Abstinent Alcoholics". In: Canadian Journal of Neurological Sciences / Journal Canadien des Sciences Neurologiques 11.04, pp. 441–446.
- Chambers, Christopher D, Hugh Garavan, and Mark A Bellgrove (2009). "Insights into the neural basis of response inhibition from cognitive and clinical neuroscience". In: Neuroscience & biobehavioral reviews 33.5, pp. 631-646.
- Chelune, Gordon J. and James B. Parker (Jan. 1981). "Neuropsychological deficits associated with chronic alcohol abuse". In: Clinical Psychology Review 1.2, pp. 181–195.
- Chen, Jonathan H and Steven M Asch (June 2017). "Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations." In: The New England journal of medicine 376.26, pp. 2507–2509.
- Chevrier, Andre D., Michael D. Noseworthy, and Russell Schachar (Dec. 2007). "Dissociation of response inhibition and performance monitoring in the stop signal task using event-related fMRI". In: Human Brain Mapping 28.12, pp. 1347-1358.
- Chevrier, Andre, Douglas Cheyne, Simon Graham, and Russell Schachar (2015). "Dissociating Two Stages of Preparation in the Stop Signal Task Using fMRI." In: PloS one 10.6, e0130992.
- Cloninger, C Robert, Thomas R Przybeck, Dragan M Svrakic, and Richard D Wetzel (1994). "The Temperament and Character Inventory (TCI): A guide to its development and use". In:
- Congdon, Eliza, Jeanette A. Mumford, Jessica R. Cohen, Adriana Galvan, Turhan Canli, and Russell A. Poldrack (Feb. 2012). "Measurement and Reliability of Response Inhibition". In: Frontiers in Psychology 3, p. 37.
- Conrod, Patricia J., Natalie Castellanos-Ryan, and John Strang (Jan. 2010). "Brief, Personality-Targeted Coping Skills Interventions and Survival as a Non-Drug User Over a 2-Year Period During Adolescence". In: Archives of General Psychiatry 67.1, p. 85.
- Corbin, William R., Nicole M. Farmer, and Susan Nolen-Hoekesma (Apr. 2013). "Relations among stress, coping strategies, coping motives, alcohol consumption and related problems: A mediated moderation model". In: *Addictive Behaviors* 38.4, pp. 1912–1919.
- Costa, Paul T. and Robert R. McCrae (Dec. 1992). "The Five-Factor Model of Personality and Its Relevance to Personality Disorders". In: Journal of Personality Disorders 6.4, pp. 343-359.
- Cox, D. R. (July 1958). "The Regression Analysis of Binary Sequences". In: Journal of the Royal Statistical Society: Series B (Methodological) 20.2, pp. 215-232.

- Czapla, Marta, Joe J Simon, Hans-Christoph Friederich, Sabine C Herpertz, Peter Zimmermann, and Sabine Loeber (2015). "Is binge drinking in young adults associated with an alcohol-specific impairment of response inhibition?" In: European addiction research 21.2, pp. 105–113.
- Darcy, Alison M., Alan K. Louie, and Laura Weiss Roberts (Feb. 2016). "Machine Learning and the Profession of Medicine". In: JAMA 315.6, p. 551.
- Dosenbach, Nico U F, Binyam Nardos, Alexander L Cohen, Damien A Fair, Jonathan D Power, Jessica A Church, Steven M Nelson, Gagan S Wig, Alecia C Vogel, Christina N Lessov-Schlaggar, et al. (2010). "Prediction of individual brain maturity using fMRI". In: Science 329.5997, pp. 1358-1361.
- Dwyer, Dominic B., Peter Falkai, and Nikolaos Koutsouleris (May 2018). "Machine Learning Approaches for Clinical Psychology and Psychiatry". In: Annual Review of Clinical Psychology 14.1, pp. 91–118.
- Effertz, Tobias and Karl Mann (July 2013). "The burden and cost of disorders of the brain in Europe with the inclusion of harmful alcohol use and nicotine addiction". In: European Neuropsychopharmacology 23.7, pp. 742-748.
- Ersche, Karen D., Abigail J. Turton, Shachi Pradhan, Edward T. Bullmore, and Trevor W. Robbins (Oct. 2010). "Drug Addiction Endophenotypes: Impulsive Versus Sensation-Seeking Personality Traits". In: Biological Psychiatry 68.8, pp. 770-773.
- Fawcett, Tom (June 2006). "An introduction to ROC analysis". In: Pattern Recognition Letters 27.8, pp. 861-874.
- Field, Matt, Reinout W Wiers, Paul Christiansen, Mark T Fillmore, and Joris C Verster (Aug. 2010). "Acute alcohol effects on inhibitory control and implicit cognition: implications for loss of control over drinking." In: Alcoholism, clinical and experimental research 34.8, pp. 1346-52.
- Friedman, Jerome, Trevor Hastie, and Rob Tibshirani (2010). "Regularization paths for generalized linear models via coordinate descent". In: Journal of statistical software 33.1, p. 1.
- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). The elements of statistical learning. Vol. 1. 10. Springer series in statistics New York.
- Gelman, Andrew and Jennifer Hill (2006). "Missing-data imputation". In: Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press, pp. 529-544.
- Genauck, Alexander, Milan Andrejevic, Katharina Brehm, Caroline Matthis, Andreas Heinz, Andre Weinreich, Norbert Kathmann, and Nina Romanczuk-Seiferth (Mar. 2019). "Cue-induced effects on decision-making distin-

- guish subjects with gambling disorder from healthy controls". In: bioRxiv, p. 564781.
- Genauck, Alexander, Caroline Matthis, Milan Andrejevic, Lukas Ballon, Francesca Chiarello, Katharina Duecker, Andreas Heinz, Norbert Kathmann, and Nina Romanczuk-Seiferth (Dec. 2018). "Neural correlates of cue-induced changes in decision-making distinguish subjects with gambling disorder from healthy controls". In: bioRxiv, p. 498725.
- Goudriaan, Anna E., Emily R. Grekin, and Kenneth J. Sher (June 2011). "Decision Making and Response Inhibition as Predictors of Heavy Alcohol Use: A Prospective Study". In: Alcoholism: Clinical and Experimental Research 35.6, pp. 1050-1057.
- Grant, Juila D, Jeffrey F Scherrer, Michael T Lynskey, Michael J Lyons, Seth A Eisen, Ming T Tsuang, William R True, and Kathleen K Bucholz (2006). "Adolescent alcohol use is a risk factor for adult alcohol and drug dependence: evidence from a twin design". In: Psychological *Medicine* 36.1, pp. 109–118.
- Harper, C G, J J Kril, and R L Holloway (1985). "Brain shrinkage in chronic alcoholics: a pathological study." In: Br Med J (Clin Res Ed) 290.6467, pp. 501-504.
- Heikkinen, Noora, Eini Niskanen, Mervi Könönen, Tommi Tolmunen, Virve Kekkonen, Petri Kivimäki, Heikki Tanila, Eila Laukkanen, and Ritva Vanninen (Apr. 2017). "Alcohol consumption during adolescence is associated with reduced grey matter volumes". In: Addiction 112.4, pp. 604-613.
- Heinz, A (Mar. 2002). "Dopaminergic dysfunction in alcoholism and schizophrenia – psychopathological and behavioral correlates". In: European Psychiatry 17.1, pp. 9-16.
- Heinz, Adrienne J., Anne Beck, Andreas Meyer-Lindenberg, Philipp Sterzer, and Andreas Heinz (June 2011). "Cognitive and neurobiological mechanisms of alcohol-related aggression". In: Nature Reviews Neuroscience 12.7, pp. 400-413.
- Henges, Amy L. and Cecile A. Marczinski (Feb. 2012). "Impulsivity and alcohol consumption in young social drinkers". In: Addictive Behaviors 37.2, pp. 217-220.
- Hernandez-Avila, Carlos A, Bruce J Rounsaville, and Henry R Kranzler (June 2004). "Opioid-, cannabis- and alcohol-dependent women show more rapid progression to substance abuse treatment". In: Drug and Alcohol Dependence 74.3, pp. 265-272.
- Hibell, Björn, Barbro Andersson, Thoroddur Bjarnason, Anna Kokkevi, Mark Morgan, and Anu Narusk (1997). The 1995 ESPAD report: Alcohol and

- other drug use among students in 26 European countries. Swedish council for information on alcohol and other drugs, CAN.
- Hibell, Björn, Barbro Andersson, Thoruddur Bjarnason, Salme Ahlström, Olga Balakireva, Anna Kokkevi, and Mark Morgan (2004). "The ESPAD report 2003". In:
- Hill, Karl, Helene White, Ick-Joong Chung, David Hawkins, and Richard Catalano (June 2000). "Early Adult Outcomes of Adolescent Binge Drinking: Person- and Variable-Centered Analyses of Binge Drinking Trajectories". In: Alcoholism: Clinical and Experimental Research 24.6, pp. 892-
- Ho, Tin Kam (1995). "Random Decision Forests". In: Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1. ICDAR '95. Washington, DC, USA: IEEE Computer Society, pp. 278-.
- Hoffmann, John P. (July 2016). "Cumulative Stress and Substance Use From Early Adolescence to Emerging Adulthood". In: Journal of Drug Issues 46.3, pp. 267-288.
- Houben, Katrijn, Chantal Nederkoorn, Reinout W. Wiers, and Anita Jansen (July 2011). "Resisting temptation: Decreasing alcohol-related affect and drinking behavior by training response inhibition". In: Drug and Alcohol Dependence 116.1-3, pp. 132-136.
- Hu, Sien, Jaime S. Ide, Sheng Zhang, Rajita Sinha, and Chiang-shan R. Li (Jan. 2015). "Conflict anticipation in alcohol dependence — A modelbased fMRI study of stop signal task". In: NeuroImage: Clinical 8, pp. 39-
- IMAGEN SOP (2019).
- Iniesta, R., D. Stahl, and P. McGuffin (Sept. 2016). "Machine learning, statistical learning and the future of biological research in psychiatry". In: Psychological Medicine 46.12, pp. 2455–2465.
- Ioannidis, John P. A. (Aug. 2005). "Why Most Published Research Findings Are False". In: PLoS Medicine 2.8, e124.
- Jernigan, Terry L, Nelson Butters, Gina DiTraglia, Kimberly Schafer, Tom Smith, Michael Irwin, Igor Grant, Marc Schuckit, and Laird S Cermak (1991). "Reduced cerebral grey matter observed in alcoholics using magnetic resonance imaging". In: Alcoholism: Clinical and Experimental Research 15.3, pp. 418-427.
- Johnson, Patrick B., Linda Richter, Herbert D. Kleber, A. Thomas McLellan, and Deni Carise (Jan. 2005). "Telescoping of Drinking-Related Behaviors: Gender, Racial/Ethnic, and Age Comparisons". In: Substance Use & Misuse 40.8, pp. 1139-1151.

- Jordan, M I and T M Mitchell (July 2015). "Machine learning: Trends, perspectives, and prospects." In: Science (New York, N.Y.) 349.6245, pp. 255-60.
- Kalivas, Peter W. and Nora D. Volkow (Aug. 2005). "The Neural Basis of Addiction: A Pathology of Motivation and Choice". In: American Journal of Psychiatry 162.8, pp. 1403-1413.
- Kassraian-Fard, Pegah, Caroline Matthis, Joshua H Balsters, Marloes H Maathuis, and Nicole Wenderoth (2016). "Promises, Pitfalls, and Basic Guidelines for Applying Machine Learning Classifiers to Psychiatric Imaging Data, with Autism as an Example". In: Frontiers in Psychiatry 7, p. 177.
- Kertzman, Semion, Katherine Lowengrub, Anat Aizer, Michael Vainder, Moshe Kotler, and Pinhas N Dannon (2008). "Go-no-go performance in pathological gamblers". In: Psychiatry research 161.1, pp. 1–10.
- Keyes, Katherine M., Mark L. Hatzenbuehler, and Deborah S. Hasin (Nov. 2011). "Stressful life experiences, alcohol consumption, and alcohol use disorders: the epidemiologic evidence for four main types of stressors". In: Psychopharmacology 218.1, pp. 1–17.
- King, Kevin M. and Laurie Chassin (Sept. 2008). "Adolescent Stressors, Psychopathology, and Young Adult Substance Dependence: A Prospective Study". In: Journal of Studies on Alcohol and Drugs 69.5, pp. 629-638.
- Kleinberg, E. M. (Dec. 1996). "An overtraining-resistant stochastic modeling method for pattern recognition". In: The Annals of Statistics 24.6, pp. 2319-2349.
- Klimstra, Theo A., William W. Hale, Quinten A. W. Raaijmakers, Susan J. T. Branje, and Wim H. J. Meeus (Apr. 2009). "Maturation of personality in adolescence." In: Journal of Personality and Social Psychology 96.4, pp. 898-912.
- Koenig, Laura B, Theodore Jacob, and Jon Randolph Haber (Mar. 2009). "Validity of the lifetime drinking history: a comparison of retrospective and prospective quantity-frequency measures." In: Journal of studies on alcohol and drugs 70.2, pp. 296-303.
- Kohonen, T. (1990). "Improved versions of learning vector quantization". In: 1990 IJCNN International Joint Conference on Neural Networks. IEEE, 545-550 vol.1.
- Kononenko, Igor (Aug. 2001). "Machine learning for medical diagnosis: history, state of the art and perspective". In: Artificial Intelligence in Medicine 23.1, pp. 89-109.
- Kühn, Simone and Jürgen Gallinat (2011). "Common biology of craving across legal and illegal drugs - a quantitative meta-analysis of cue-

- reactivity brain response". In: European Journal of Neuroscience 33.7, pp. 1318– 1326.
- Kuntsche, Emmanuel, Ronald Knibbe, Gerhard Gmel, and Rutger Engels (Oct. 2006). "Who drinks and why? A review of socio-demographic, personality, and contextual issues behind the drinking motives in young people". In: Addictive Behaviors 31.10, pp. 1844–1857.
- Lappin, Joseph S. and Charles W. Eriksen (1966). "Use of a delayed signal to stop a visual reaction-time response." In: Journal of Experimental Psychology 72.6, pp. 805-811.
- Leon, David A, Lyudmila Saburova, Susannah Tomkins, Evgueny Andreev, Nikolay Kiryanov, Martin McKee, and Vladimir M Shkolnikov (June 2007). "Hazardous alcohol drinking and premature mortality in Russia: a population based case-control study". In: The Lancet 369.9578, pp. 2001-2009.
- Levitt, H. (2005). "Transformed Up-Down Methods in Psychoacoustics". In: The Journal of the Acoustical Society of America 49.2B, pp. 467–477.
- Li, Chiang-shan Ray, Cong Huang, R Todd Constable, and Rajita Sinha (2006). "Imaging response inhibition in a stop-signal task: neural correlates independent of signal monitoring and post-response processing". In: Journal of Neuroscience 26.1, pp. 186-192.
- Li, Chiang-shan Ray, Xi Luo, Peisi Yan, Keri Bergquist, and Rajita Sinha (Apr. 2009). "Altered Impulse Control in Alcohol Dependence: Neural Measures of Stop Signal Performance". In: Alcoholism: Clinical and Experimental Research 33.4, pp. 740-750.
- Liu, Li, Jiajing Wang, Danny Rosenberg, Hao Zhao, Gyorgy Lengyel, and Dani Nadel (Oct. 2018). "Fermented beverage and food storage in 13,000 y-old stone mortars at Raqefet Cave, Israel: Investigating Natufian ritual feasting". In: Journal of Archaeological Science: Reports 21, pp. 783-
- Lockhart, Richard, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani (Apr. 2014). "A significance test for the lasso". In: Annals of statistics 42.2, pp. 413-468.
- Logan, Gordon D (1994). "On the ability to inhibit thought and action: A users' guide to the stop signal paradigm." In: Inhibitory processes in attention, memory, and language. San Diego, CA, US: Academic Press, pp. 189-239.
- Logan, Gordon D., William B. Cowan, and Kenneth A. Davis (1984). "On the ability to inhibit simple and choice reaction time responses: A model and a method." In: Journal of Experimental Psychology: Human Perception and Performance 10.2, pp. 276-291.

- Loheswaran, Genane, Mera S Barr, Tarek K Rajji, Daniel M Blumberger, Bernard Le Foll, and Zafiris J Daskalakis (2016). "Alcohol intoxication by binge drinking impairs neuroplasticity". In: Brain stimulation 9.1, pp. 27-32.
- López-Caneda, E., S. Rodríguez Holguín, F. Cadaveira, M. Corral, and S. Doallo (2014). "Impact of Alcohol Use on Inhibitory Control (and Vice Versa) During Adolescence and Young Adulthood: A Review". In: Alcohol and Alcoholism 49.2, pp. 173-181.
- Luciana, Monica, Paul F. Collins, Ryan L. Muetzel, and Kelvin O. Lim (Nov. 2013). "Effects of alcohol use initiation on brain structure in typically developing adolescents". In: The American Journal of Drug and Alcohol Abuse 39.6, pp. 345-355.
- Luna, Beatriz, Aarthi Padmanabhan, and Kirsten O'Hearn (Feb. 2010). "What has fMRI told us about the development of cognitive control through adolescence?" In: Brain and cognition 72.1, pp. 101–13.
- Luna, Beatriz and John A. Sweeney (June 2004). "The Emergence of Collaborative Brain Function: fMRI Studies of the Development of Response Inhibition". In: Annals of the New York Academy of Sciences 1021.1, pp. 296-
- Maddox, Thomas M., John S. Rumsfeld, and Philip R. O. Payne (Jan. 2019). "Questions for Artificial Intelligence in Health Care". In: JAMA 321.1,
- Malmberg, Monique, Geertjan Overbeek, Karin Monshouwer, Jeroen Lammers, Wilma A. M. Vollebergh, and Rutger C. M. E. Engels (Dec. 2010). "Substance use risk profiles and associations with early substance use in adolescence". In: Journal of Behavioral Medicine 33.6, pp. 474-485.
- Mann, K, K Ackermann, B Croissant, G Mundle, H Nakovics, and A Diehl (May 2005). "Neuroimaging of Gender Differences in Alcohol Dependence: Are Women More Vulnerable?" In: Alcoholism: Clinical & Experimental Research 29.5, pp. 896-901.
- Mashhoon, Yasmin, Charles Czerkawski, David J. Crowley, Julia E. Cohen-Gilbert, Jennifer T. Sneider, and Marisa M. Silveri (July 2014). "Binge Alcohol Consumption in Emerging Adults: Anterior Cingulate Cortical "Thinness" Is Associated with Alcohol Use Patterns". In: Alcoholism: Clinical and Experimental Research 38.7, pp. 1955–1964.
- McCambridge, Jim, John McAlaney, and Richard Rowe (Feb. 2011). "Adult Consequences of Late Adolescent Alcohol Consumption: A Systematic Review of Cohort Studies". In: PLoS Medicine 8.2. Ed. by Bruce P. Lanphear, e1000413.

- Murphy, F C, B J Sahakian, J S Rubinsztein, A Michael, R D Rogers, T W Robbins, and E S Paykel (Nov. 1999). "Emotional bias and inhibitory control processes in mania and depression." In: Psychological medicine 29.6, pp. 1307-21.
- Naranjo, C. A. and K. E. Bremner (Jan. 1993). "Behavioural correlates of alcohol intoxication". In: Addiction 88.1, pp. 31-41.
- Nesic, J. and T. Duka (Feb. 2006). "Gender specific effects of a mild stressor on alcohol cue reactivity in heavy social drinkers". In: Pharmacology Biochemistry and Behavior 83.2, pp. 239-248.
- Newcomb, Michael D., George J. Huba, and Peter M. Bentler (Dec. 1981). "A Multidimensional Assessment of Stressful Life Events among Adolescents: Derivation and Correlates". In: Journal of Health and Social Behavior 22.4, p. 400.
- Ng, Andrew Y and Michael I Jordan (2002). "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes". In: Advances in neural information processing systems, pp. 841–848.
- Nichols, Thomas and Satoru Hayasaka (Oct. 2003). "Controlling the familywise error rate in functional neuroimaging: a comparative review". In: Statistical Methods in Medical Research 12.5, pp. 419–446.
- Nigg, Joel T, Maria M Wong, Michelle M Martel, Jennifer M Jester, Leon I Puttler, Jennifer M Glass, Kenneth M Adams, Hiram E Fitzgerald, and Robert A Zucker (2006). "Poor response inhibition as a predictor of problem drinking and illicit drug use in adolescents at risk for alcoholism and other substance use disorders". In: Journal of the American Academy of Child & Adolescent Psychiatry 45.4, pp. 468-475.
- Noël, Xavier, Jacques Paternot, Martial der Linden, Rita Sferrazza, Michel Verhas, Catherine Hanak, Charles Kornreich, Philippe Martin, Jacques De Mol, Isidore Pelc, et al. (2001). "Correlation between inhibition, working memory and delimited frontal area blood flow measured by 99MTc-Bicisate SPECT in alcohol-dependent patients". In: Alcohol and alcoholism 36.6, pp. 556-563.
- Noël, Xavier, Martial Van der Linden, Mathieu D'Acremont, Antoine Bechara, Bernard Dan, Catherine Hanak, and Paul Verbanck (June 2007). "Alcohol cues increase cognitive impulsivity in individuals with alcoholism". In: Psychopharmacology 192.2, pp. 291–298.
- Norman, Andria L., Carmen Pulido, Lindsay M. Squeglia, Andrea D. Spadoni, Martin P. Paulus, and Susan F. Tapert (Dec. 2011). "Neural activation during inhibition predicts initiation of substance use in adolescence". In: Drug and Alcohol Dependence 119.3, pp. 216–223.

- Obermeyer, Ziad and Ezekiel J Emanuel (Sept. 2016). "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine." In: The New England journal of medicine 375.13, pp. 1216–9.
- Open Science Collaboration, Open Science (Aug. 2015). "Estimating the reproducibility of psychological science". In: Science 349.6251, aac4716aac4716.
- Petit, Géraldine, Charles Kornreich, Xavier Noël, Paul Verbanck, and Salvatore Campanella (May 2012). "Alcohol-Related Context Modulates Performance of Social Drinkers in a Visual Go/No-Go Task: A Preliminary Assessment of Event-Related Potentials". In: PLoS ONE 7.5. Ed. by Antonio Verdejo García, e37466.
- Petit, Géraldine, Pierre Maurage, Charles Kornreich, Paul Verbanck, and Salvatore Campanella (Mar. 2014). "Binge Drinking in Adolescents: A Review of Neurophysiological and Neuroimaging Research". In: Alcohol and Alcoholism 49.2, pp. 198-206.
- Pfefferbaum, Adolf, John E Desmond, Christopher Galloway, Vinod Menon, Gary H Glover, and Edith V Sullivan (2001). "Reorganization of frontal systems used by alcoholics for spatial working memory: an fMRI study". In: Neuroimage 14.1, pp. 7-20.
- Pfefferbaum, Adolf, Kelvin O Lim, Robert B Zipursky, Daniel H Mathalon, Margaret J Rosenbloom, Barton Lane, Chung Nim Ha, and Edith V Sullivan (1992). "Brain gray and white matter volume loss accelerates with aging in chronic alcoholics: a quantitative MRI study". In: Alcoholism: Clinical and Experimental Research 16.6, pp. 1078–1089.
- Pfefferbaum, Adolf, Daniel H Mathalon, Edith V Sullivan, Jody M Rawles, Robert B Zipursky, and Kelvin O Lim (1994). "A quantitative magnetic resonance imaging study of changes in brain morphology from infancy to late adulthood". In: Archives of neurology 51.9, pp. 874-887.
- Piazza, Nick J., Jean L. Vrbka, and Rebecca D. Yeager (1989). "Telescoping of alcoholism in women alcoholics". In: Substance Use and Misuse 24.1, pp. 19-28.
- Prencipe, Angela, Amanda Kesek, Julia Cohen, Connie Lamm, Marc D. Lewis, and Philip David Zelazo (Mar. 2011). "Development of hot and cool executive function during the transition to adolescence". In: Journal of Experimental Child Psychology 108.3, pp. 621-637.
- Rehm, Jürgen, Colin Mathers, Svetlana Popova, Montarat Thavorncharoensap, Yot Teerawattananon, and Jayadeep Patra (June 2009). "Global burden of disease and injury and economic cost attributable to alcohol use and alcohol-use disorders". In: The Lancet 373.9682, pp. 2223-2233.

- Roberts, Brent W., Avshalom Caspi, and Terrie E. Moffitt (2001). "The kids are alright: Growth and stability in personality development from adolescence to adulthood." In: Journal of Personality and Social Psychology 81.4, pp. 670-683.
- Rubia, Katya, Anna B Smith, Michael J Brammer, and Eric Taylor (Sept. 2003). "Right inferior prefrontal cortex mediates response inhibition while mesial prefrontal cortex is responsible for error detection". In: NeuroImage 20.1, pp. 351-358.
- Rubia, Katya, Anna B Smith, Eric Taylor, and Michael Brammer (2007). "Linear age-correlated functional development of right inferior frontostriato-cerebellar networks during response inhibition and anterior cingulate during error-related processes". In: Human brain mapping 28.11, pp. 1163-1177.
- Rubio, Gabriel, Mónica Jiménez, Roberto Rodríguez-Jiménez, Isabel Martínez, César Ávila, Francisco Ferre, Miguel Angel Jiménez-Arriero, Guillermo Ponce, and Tomás Palomo (Sept. 2008). "The Role of Behavioral Impulsivity in the Development of Alcohol Dependence: A 4-Year Follow-Up Study". In: Alcoholism: Clinical and Experimental Research 32.9, pp. 1681-
- Saunders, John, Olaf Aasland, Thomas Babor, Juan De la Fuente, and Marcus Grant (June 1993). "Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption-II". In: Addiction 88.6, pp. 791-804.
- Schacht, Joseph P., Raymond F. Anton, and Hugh Myrick (2013). "Functional neuroimaging studies of alcohol cue reactivity: A quantitative meta-analysis and systematic review". In: Addiction Biology 18.1, pp. 121-133.
- Schuckit, Marc A., Tom L. Smith, George Danko, Robert Anthenelli, Lara Schoen, Mari Kawamura, John Kramer, Danielle M. Dick, Zoe Neale, Samuel Kuperman, Vivia McCutcheon, Andrey P. Anokhin, Victor Hesselbrock, Michie Hesselbrock, and Kathleen Bucholz (July 2017). "A Prospective Comparison of How the Level of Response to Alcohol and Impulsivity Relate to Future DSM-IV Alcohol Problems in the COGA Youth Panel". In: Alcoholism: Clinical and Experimental Research 41.7, pp. 1329-1339.
- Schulte, Marya T, Danielle Ramo, and Sandra A Brown (2009). "Clinical Psychology Review Gender differences in factors in fl uencing alcohol use and drinking progression among adolescents". In: Clinical Psychology Review 29.6, pp. 535-547.

- Schulz, Kurt P., Jin Fan, Olga Magidina, David J. Marks, Bella Hahn, and Jeffrey M. Halperin (Feb. 2007). "Does the emotional go/no-go task really measure behavioral inhibition?: Convergence with measures on a non-emotional analog". In: Archives of Clinical Neuropsychology 22.2, pp. 151-160.
- Schumann, G, E Loth, T Banaschewski, A Barbot, G Barker, C Büchel, P J Conrod, J W Dalley, H Flor, J Gallinat, H Garavan, A Heinz, B Itterman, M Lathrop, C Mallik, K Mann, J-L Martinot, T Paus, J-B Poline, T W Robbins, M Rietschel, L Reed, M Smolka, R Spanagel, C Speiser, D N Stephens, A Ströhle, and M Struve (Dec. 2010). "The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology". In: Molecular Psychiatry 15.12, pp. 1128–1139.
- Selzer, Melvin (June 1971). "The Michigan Alcoholism Screening Test: The Quest for a New Diagnostic Instrument". In: American Journal of Psychiatry 127.12, pp. 1653-1658.
- Seo, S, J Mohr, and K Obermayer (Dec. 2009). "A New Incremental Pairwise Clustering Algorithm". In: 2009 International Conference on Machine Learning and Applications, pp. 223–228.
- Seo, Sambu, Anne Beck, Caroline Matthis, Alexander Genauck, Tobias Banaschewski, Arun L.W. Bokde, Uli Bromberg, Christian Büchel, Erin Burke Quinlan, Herta Flor, Vincent Frouin, Hugh Garavan, Penny Gowland, Bernd Ittermann, Jean Luc Martinot, Marie Laure Paillère Martinot, Frauke Nees, Dimitri Papadopoulos Orfanos, Luise Poustka, Sarah Hohmann, Juliane H. Fröhner, Michael N. Smolka, Henrik Walter, Robert Whelan, Sylvane Desrivières, Andreas Heinz, Gunter Schumann, and Klaus Obermayer (2018). "Risk profiles for heavy drinking in adolescence: Differential effects of gender". In: Addiction Biology.
- Seo, Sambu, Johannes Mohr, Anne Beck, Torsten Wüstenberg, Andreas Heinz, and Klaus Obermayer (Nov. 2015). "Predicting the future relapse of alcohol-dependent patients from structural and functional brain images". In: Addiction Biology 20.6, pp. 1042-1055.
- Seo, Sambu and Klaus Obermayer (2003). "Soft Learning Vector Quantization". In: Neural Computation 15.7, pp. 1589–1604.
- Shah, Anoop D., Jonathan W. Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway (Mar. 2014). "Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study". In: American Journal of Epidemiology 179.6, pp. 764-774.
- Sharp, D. J., V. Bonnelle, X. De Boissezon, C. F. Beckmann, S. G. James, M. C. Patel, and M. A. Mehta (2010). "Distinct frontal systems for response

- inhibition, attentional capture, and error processing". In: Proceedings of the National Academy of Sciences 107.13, pp. 6106-6111. arXiv: arXiv: 1408.1149.
- Shmueli, Galit (2009). "To Explain or To Predict?" In: Ssrn 25.3, pp. 289–310. arXiv: 1101.0891.
- Skinner, H A and W J Sheu (Nov. 1982). "Reliability of alcohol use indices. The Lifetime Drinking History and the MAST." eng. In: Journal of studies on alcohol 43.11, pp. 1157-1170.
- Sobell, Linda C. and Mark B. Sobell (1992). "Timeline Follow-Back". In: *Measuring Alcohol Consumption*. Totowa, NJ: Humana Press, pp. 41–72.
- Sobell, Linda, Joanne Brown, Gloria Leo, and Mark Sobell (Sept. 1996). "The reliability of the Alcohol Timeline Followback when administered by telephone and by computer". In: Drug and Alcohol Dependence 42.1, pp. 49-54.
- Spanagel, Rainer, Daniel Durstewitz, Anita Hansson, Andreas Heinz, Falk Kiefer, Georg Köhr, Franziska Matthäus, Markus M. Nöthen, Hamid R. Noori, Klaus Obermayer, Marcella Rietschel, Patrick Schloss, Henrike Scholz, Gunter Schumann, Michael Smolka, Wolfgang Sommer, Valentina Vengeliene, Henrik Walter, Wolfgang Wurst, Uli S. Zimmermann, Sven Stringer, Yannick Smits, Eske M. Derks, and Eske M. Derks (Nov. 2013). "A systems medicine research approach for studying alcohol addiction". In: Addiction Biology 18.6, pp. 883-896.
- Squeglia, L. M., J. Jacobus, and S. F. Tapert (2009). "The influence of substance use on adolescent brain development". In: Clinical EEG and Neuroscience 40.1, pp. 31-38. arXiv: NIHMS150003.
- Squeglia, Lindsay M., Daniel A. Rinker, Hauke Bartsch, Norma Castro, Yoonho Chung, Anders M. Dale, Terry L. Jernigan, and Susan F. Tapert (July 2014). "Brain volume reductions in adolescent heavy drinkers". In: Developmental Cognitive Neuroscience 9, pp. 117–125.
- Squeglia, Lindsay M., Alecia Dager Schweinsburg, Carmen Pulido, and Susan F. Tapert (Oct. 2011). "Adolescent Binge Drinking Linked to Abnormal Spatial Working Memory Brain Activation: Differential Gender Effects". In: Alcoholism: Clinical and Experimental Research 35.10, pp. 1831-
- Squeglia, Lindsay M., Scott F. Sorg, Alecia Dager Schweinsburg, Reagan R. Wetherill, Carmen Pulido, and Susan F. Tapert (Apr. 2012). "Binge drinking differentially affects adolescent male and female brain morphometry". In: Psychopharmacology 220.3, pp. 529-539.
- Squeglia, Lindsay M., Susan F. Tapert, Edith V. Sullivan, Joanna Jacobus, M.J. Meloy, Torsten Rohlfing, and Adolf Pfefferbaum (June 2015). "Brain

- Development in Heavy-Drinking Adolescents". In: American Journal of Psychiatry 172.6, pp. 531-542.
- Stewart, Sherry H. and Heather Devine (Sept. 2000). "Relations between personality and drinking motives in young adults". In: Personality and *Individual Differences* 29.3, pp. 495–511.
- Sullivan, Edith V., Antony J. Harding, Roberta Pentney, Cynthia Dlugos, Peter R. Martin, Mitchell H. Parks, John E. Desmond, S. H. Annabel Chen, Michelle R. Pryor, Eve De Rosa, and Adolf Pfefferbaum (Feb. 2003). "Disruption of Frontocerebellar Circuitry and Function in Alcoholism". In: Alcoholism: Clinical & Experimental Research 27.2, pp. 301-309.
- Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the Lasso". In: Journal of the Royal Statistical Society. Series B (Methodological) 58.1, pp. 267-288.
- Tong, Yunxia, Qiang Chen, Thomas E Nichols, Roberta Rasetti, Joseph H Callicott, Karen F Berman, Daniel R Weinberger, and Venkata S Mattay (2016). "Seeking optimal region-of-interest (ROI) single-value summary measures for fMRI studies in imaging genetics". In: PloS one 11.3, e0151391.
- Trabert, W, T Betz, M Niewald, and G Huber (Aug. 1995). "Significant reversibility of alcoholic brain shrinkage within 3 weeks of abstinence." In: Acta psychiatrica Scandinavica 92.2, pp. 87–90.
- Tzourio-Mazoyer, N., B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot (Jan. 2002). "Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain". In: NeuroImage 15.1, pp. 273-289.
- VanderWeele, Tyler J and Ilya Shpitser (Feb. 2013). "On the definition of a confounder." In: Annals of statistics 41.1, pp. 196–220.
- Verbruggen, Frederick, Christopher D. Chambers, and Gordon D. Logan (Mar. 2013). "Fictitious Inhibitory Differences". In: Psychological Science 24.3, pp. 352-362.
- Verbruggen, Frederick and Gordon D. Logan (Nov. 2008). "Response inhibition in the stop-signal paradigm". In: Trends in Cognitive Sciences 12.11, pp. 418-424.
- Verbruggen, Frederick and Gordon D. Logan (May 2009). "Models of response inhibition in the stop-signal and stop-change paradigms". In: Neuroscience & Biobehavioral Reviews 33.5, pp. 647-661.

- Verbruggen, Frederick, Gordon D. Logan, and Michaël A. Stevens (May 2008). "STOP-IT: Windows executable software for the stop-signal paradigm". In: Behavior Research Methods 40.2, pp. 479–483.
- Walesiak, Marek (1999). "Distance measure for ordinal data". In: Argumenta Oeconomica 2.8, pp. 167-173.
- Ward, L. Charles (July 1990). "Prediction of verbal, performance, and full scale IQs from seven subtests of the wais-r". In: Journal of Clinical Psychology 46.4, pp. 436-440.
- Weafer, Jessica and Mark T. Fillmore (Nov. 2012). "Comparison of alcohol impairment of behavioral and attentional inhibition". In: Drug and Alcohol Dependence 126.1-2, pp. 176-182.
- Wechsler, David (1949). Wechsler Intelligence Scale for Children. San Antonio, TX, US: Psychological Corporation.
- Wechsler, David (2003). Wechsler intelligence scale for children-WISC-IV. Psychological Corporation.
- Whelan, Robert, Patricia J Conrod, Jean-Baptiste Poline, Anbarasu Lourdusamy, Tobias Banaschewski, Gareth J Barker, Mark A Bellgrove, Christian Büchel, Mark Byrne, Tarrant D R Cummins, Mira Fauth-Bühler, Herta Flor, Jürgen Gallinat, Andreas Heinz, Bernd Ittermann, Karl Mann, Jean-Luc Martinot, Edmund C Lalor, Mark Lathrop, Eva Loth, Frauke Nees, Tomas Paus, Marcella Rietschel, Michael N Smolka, Rainer Spanagel, David N Stephens, Maren Struve, Benjamin Thyreau, Sabine Vollstaedt-Klein, Trevor W Robbins, Gunter Schumann, and Hugh Garavan (June 2012). "Adolescent impulsivity phenotypes characterized by distinct brain networks". In: Nature Neuroscience 15.6, pp. 920–925.
- Wilsnack, Richard W., Nancy D. Vogeltanz, Sharon C. Wilsnack, and T. Robert Harris (Feb. 2000). "Gender differences in alcohol consumption and adverse drinking consequences: cross-cultural patterns". In: Addiction 95.2, pp. 251-265.
- Wilsnack, Richard W., Sharon C. Wilsnack, Arlinda F. Kristjanson, Nancy D. Vogeltanz-Holm, and Gerhard Gmel (Sept. 2009). "Gender and alcohol consumption: patterns from the multinational GENACIS project". In: Addiction 104.9, pp. 1487–1500.
- Windle, Michael, Eun Young Mun, and Rebecca C Windle (May 2005). "Adolescentto-young adulthood heavy drinking trajectories and their prospective predictors." In: Journal of studies on alcohol 66.3, pp. 313-22.
- Woicik, Patricia A., Sherry H. Stewart, Robert O. Pihl, and Patricia J. Conrod (2009). "The substance use risk profile scale: A scale measuring traits linked to reinforcement-specific substance use profiles". In: Addictive Behaviors 34.12, pp. 1042-1055.

- Yarkoni, Tal and Jacob Westfall (Nov. 2017). "Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning". In: Perspectives on Psychological Science 12.6, pp. 1100-1122.
- Zelazo, Philip David and Stephanie M. Carlson (June 2012). "Hot and Cool Executive Function in Childhood and Adolescence: Development and Plasticity". In: *Child Development Perspectives* 6.4, n/a-n/a.
- Zhang, Shichao (Nov. 2012). "Nearest neighbor selection for iteratively kNN imputation". In: Journal of Systems and Software 85.11, pp. 2541-2552.
- Zhang, Xiaochu, Betty Jo Salmeron, Thomas J. Ross, Hong Gu, Xiujuan Geng, Yihong Yang, and Elliot A. Stein (Jan. 2011). "Anatomical differences and network characteristics underlying smoking cue reactivity". In: NeuroImage 54.1, pp. 131–141.
- Zilberman, Noam, Michal Lavidor, Gal Yadid, and Yuri Rassovsky (Mar. 2019). "Qualitative review and quantitative effect size meta-analyses in brain regions identified by cue-reactivity addiction studies." In: Neuropsychology 33.3, pp. 319-334.