universität wien

Diplomarbeit

Titel der Diplomarbeit

Conditional Logistic Regression and Odds Ratio Multifactor Dimensionality Reduction for the Analysis of Interactions of Environmental Risk Factors and the 5-HT_{2A} –1438 G/A Polymorphism in Anorexia Nervosa

Verfasser

DI (FH) Ingo Nader

angestrebter akademischer Grad

Magister der Naturwissenschaft (Mag.rer.nat.)

Wien, im Oktober 2008

Matrikelnummer:	9853097
Studienkennzahl lt. Studienblatt:	A 298
Studienrichtung lt. Studienblatt:	Psychologie
Betreuer:	Univ.-Ass. Mag. Dr. Karin Waldherr

Contents

Introduction 9

I. Theoretical Part 11

1. Eating Disorders 13

1.1. Anorexia Nervosa . . . 13

1.1.1. Clinical Features . . . 13

1.1.2. Epidemiology . . . 15

1.2. Bulimia Nervosa . . . 15

1.2.1. Clinical Features . . . 16

1.2.2. Epidemiology . . . 17

1.3. Other (Atypical) Eating Disorders . . . 17

1.3.1. Clinical Features . . . 18

1.3.2. Epidemiology . . . 18

1.3.3. Binge-Eating Disorder . . . 19

1.4. Risk Factors . . . 20

1.4.1. Genetic Risk Factors . . . 20

1.4.2. Personality Traits . . . 21

1.4.3. Environmental Risk Factors . . . 22

1.4.4. Specific Risk Factors for Subtypes of Eating Disorders . . . 24

2. Genes and Polymorphisms 27

2.1. What Genes Do . . . 27

2.1.1. The Genetic Code . . . 27

2.1.2. From Genes to Proteins . . . 28

2.2. Polymorphisms . . . 29

2.3. Gene and Environment Interplay . . . 31

2.3.1. Gene–Environment Interaction . . . 32

2.3.2. Gene–Environment Correlation . . . 32

2.3.3. Good Genes and Bad Genes . . . 33
2.4. Methods for Research of Genetic Influences . . . 34
2.4.1. Family Studies . . . 34
2.4.2. Twin Studies . . . 34
2.4.3. Adoption Studies . . . 35
2.4.4. Linkage Studies . . . 36
2.4.5. Association Studies . . . 36
2.4.6. Transmission Disequilibrium . . . 37
2.4.7. Gene–Environment Interaction (G × E) Studies . . . 37

3. Serotonin Regulation and Eating Disorders 39
3.1. Parts of the Serotonin Regulation . . . 39
3.2. Serotonin Regulation and Eating Behavior . . . 42
3.3. Serotonin Regulation and Anorexia Nervosa . . . 44
3.4. The 5-HT_{2A} Gene and Anorexia Nervosa . . . 49
3.4.1. The 5-HT_{2A} Gene and its Polymorphisms . . . 49
3.4.2. The 5-HT_{2A} –1438 G/A Polymorphism in Eating Behavior and Anorexia Nervosa . . . 51

4. Methods 57
4.1. Study Design . . . 57
4.2. Assessment of Risk Factors and other Variables . . . 57
4.3. Genotyping . . . 59
4.4. Statistical Analysis . . . 59
4.4.1. Logistic Regression . . . 59
4.4.2. Multiple Logistic Regression . . . 62
4.4.3. Conditional Logistic Regression . . . 63
4.4.4. Odds Ratio Multifactor Dimensionality Reduction . . . 73
4.5. Software for Statistical Analysis . . . 79

II. Empirical Part 81

5. Formulation of Questions and Hypotheses 83
5.1. Hardy-Weinberg Equilibrium . . . 83
5.2. Association of 5-HT_{2A} –1438 G/A Polymorphism with Anorexia Nervosa . . . 83
5.3. Gene–Environment Correlation . . . 84

5.4. Interaction of 5-HT_{2A} –1438 G/A Polymorphism and Environmental Risk Factors . . . 84
5.4.1. Models 1 and 2: Anorexia Nervosa and Gene–Environment Interaction 84
5.4.2. Application of Odds Ratio Multifactor Dimensionality Reduction to Identify Interactions . . . 85
5.4.3. Model 3: Body Mass Index and Gene–Environment Interaction . . . 85
5.4.4. Model 4: Age of Onset and Gene–Environment Interaction . . . 86

6. Results **87**
6.1. Description of Sample . . . 87
6.1.1. Recruitment of Subjects . . . 87
6.1.2. Sample Size . . . 87
6.1.3. Age . . . 88
6.1.4. Parents . . . 88
6.1.5. Eating Disorder Related Characteristics . . . 89
6.1.6. Exposure to Environmental Risk . . . 91
6.1.7. Allele and Genotype Frequencies . . . 91
6.2. Hardy-Weinberg Equilibrium . . . 92
6.3. Association of 5-HT_{2A} –1438 G/A Polymorphism with Anorexia Nervosa . 93
6.4. Gene–Environment Correlation . . . 94
6.5. Interaction of 5-HT_{2A} –1438 G/A Polymorphism and Environmental Risk Factors . . . 95
6.5.1. Anorexia Nervosa and Gene–Environment Interaction . . . 95
6.5.2. Body Mass Index and Gene–Environment Interaction . . . 101
6.5.3. Age of Onset and Gene–Environment Interaction . . . 103
6.6. Application of Odds Ratio Multifactor Dimensionality Reduction . . . 110
6.6.1. Recoding of Environmental Risk Factors . . . 110
6.6.2. Finding the Best Factor Models . . . 111

7. Interpretation and Discussion **121**
7.1. Association of 5-HT_{2A} –1438 G/A Polymorphism with Anorexia Nervosa . 121
7.2. Interaction of 5-HT_{2A} –1438 G/A Polymorphism and Environmental Risk Factors . . . 122
7.2.1. Anorexia Nervosa and Gene–Environment Interaction . . . 122
7.2.2. Body Mass Index and Gene–Environment Interaction . . . 125
7.2.3. Age of Onset and Gene–Environment Interaction . . . 126
7.3. Summary of Results . . . 128

7.3.1. Environmental Risk Effects 128
7.3.2. Effects of the 5-HT_{2A} –1438 G/A Polymorphism 128
7.3.3. Interactions of Genetic and Environmental Factors 129
7.4. The Odds Ratio Multifactor Dimensionality Reduction Method 129
7.5. Summary of Criticism and Suggestions for Future Research 130

III. Appendix **133**

A. R Code **135**
A.1. Libraries, Data Preperations, and Functions 135
A.2. Description of Sample 138
A.3. Hardy-Weinberg Equilibrium 146
A.4. Association tests 147
A.5. Gene–Environment Correlation 150
A.6. Conditional Logistic Regression Models 150
A.7. Linear Regression Models 153
A.8. OR MDR Analysis 159

List of Tables **169**

List of Figures **171**

Abstract **173**

Abstract (German) **175**

References **177**

Acknowledgements

First and foremost I want to express my gratitude to my supervisor, Univ.-Ass. Mag. Dr. Karin Waldherr, who has supported me throughout my thesis with her patience and knowledge. Furthermore I would like to thank Univ.-Prof. Dr. Andreas Karwautz and Mag. Gudrun Wagner for allowing me to be part of the project and for their confidence in my abilities to contribute to it.

I am also very grateful to Jim Malec for reading, correcting, and commenting the manuscript, and so helped me to improve style and readability of this paper.

I wish to thank all the people that encouraged my love for reading and studying. They are too many to name, but without them, I would never have had the chance to write this thesis.

I am thankful to my friends and colleagues Thomas Rusch, Mike Swazina, Stefan Dressler, and Jakob Pietschnig for their comments and inputs not only on statistical issues. Numerous conversations and discussions gave rise to a lot of interesting ideas that helped and inspired this paper.

I also wish to express my gratitude to Dr. Kurt Hornik, who encouraged my enthusiasm to work with R and introduced me to SWeave.

In my private life, I want to thank Margarete. For always pushing me to continue my work, for encouraging me to take the effort, for her patience, and for the time she sacrificed.

Lastly and most importantly, I wish to thank my parents, Christa and Hans Nader, for supporting me throughout all my life and my studies at University. To them I dedicate this thesis.

Introduction

In the last two decades, research on genetic influences on various psychiatric disorders has seen a virtual explosion. This is related to the identification of various polymorphisms in the human genome. The term "polymorphism" describes the fact that two or more variants of a gene (alleles) coexist in the same population. These variants can influence the phenotype, the observable characteristics of an organism. They may also influence the risk for diseases or disorders.

Newly arising molecular genetic studies (like linkage and association studies) are used to investigate possible influences of certain genetic variants (alleles) on disease susceptibility. In the field of anorexia nervosa, one polymorphism of interest is the 5-HT_{2A} –1438 G/A polymorphism. It is constituted by an alteration in the gene sequence coding the serotonin receptor 5-HT_{2A} (the base adenine is substituted by the base guanine at position –1438 in the promoter region of the DNA sequence). Various studies have found an association of this polymorphism with anorexia nervosa, but a considerable number of studies did not confirm this result.

This paper attempts to replicate the association of the 5-HT_{2A} –1438 G/A polymorphism with anorexia nervosa. Furthermore, as research on the association is contradictory, the possible interactions of the polymorphism with environmental risk factors will be explored. Moreover, the influence of the genetic main and interaction effects on life-time severity and age of onset will be investigated. For statistical analysis, conditional logistic regression and multiple linear regression are used. In addition to these "traditional" methods, an only recently developed data-mining algorithm will be applied: odds ratio multifactor dimensionality reduction. The results will be compared to those of the regression model.

Chapter one gives a short introduction on eating disorders. It includes the diagnostic criteria for anorexia nervosa, bulimia nervosa, and eating disorders not otherwise specified. Also, the risk factors for eating disorders in general and anorexia nervosa in particular are described.

In the second chapter, the function of genes and polymorphisms is explained, as well as possible study designs to investigate genetic influences on liability for diseases and mental disorders.

Chapter three is dedicated to giving an overview of the serotonin system with regard to anorexia nervosa. Influences of the serotonin regulation on eating behavior and eating disorders are reported. Furthermore, this chapter describes the role of the 5-HT_{2A} gene and the 5-HT_{2A} –1438 G/A polymorphism in eating disorders in general and anorexia nervosa in particular.

The fourth chapter describes the methods used for statistical analysis. It comprises an extensive description of the conditional logistic regression model as well as an introduction to odds ratio multifactor dimensionality reduction.

After this theoretical part, the hypotheses of this paper are listed in chapter five. This is followed by a description of the sample and the results of the statistical analysis in chapter six.

The last chapter, chapter seven, includes the interpretation of the results and the discussion and criticism concerning the study reported in this paper.

Part I.

Theoretical Part

1. Eating Disorders

Eating disorders have been known about for a long time, but only in the last thirty years have they been considered a psychiatric disorder and included in the diagnostic systems (Davison & Neale, 2002).

They are grouped in three diagnostic categories:

- anorexia nervosa
- bulimia nervosa
- atypical eating disorders

The borderline between those categories is not easy to draw, as the disorders have many features in common and patients frequently move between them (Fairburn & Harrison, 2003), particularly with regard to anorexia nervosa and bulimia nervosa. Nevertheless, there are distinct diagnostic criteria for these categories. They will be outlined in the following section.

1.1. Anorexia Nervosa

Anorexia nervosa is a psychiatric disorder that is mainly characterized by a strong fear of gaining weight. The literal translation of the expression however, meaning the loss of appetite due to emotional reasons, is somewhat misleading. People suffering from anorexia nervosa do not lose their appetite or their interest in eating. On the contrary, their focus is almost entirely dedicated towards issues related to eating. They often read cook books, or even prepare delicious meals for their family. But they are also involved in strictly controlling their food and calorie intake and in keeping a rigorous diet (Davison & Neale, 2002).

1.1.1. Clinical Features

The criteria for anorexia nervosa (according to DSM-IV-TR) are the following:

A. Refusal to maintain body weight at or above a minimally normal weight for age and height (e. g., weight loss leading to maintenance of body weight less than 85% of that expected; or failure to make expected weight gain during period of growth, leading to body weight less than 85% of that expected).

B. Intense fear of gaining weight or becoming fat, even though underweight.

C. Disturbance in the way in which one's body weight or shape is experienced, undue influence of body weight or shape on self-evaluation, or denial of the seriousness of the current low body weight.

D. In postmenarcheal females, amenorrhea, i. e., the absence of at least three consecutive menstrual cycles. (A woman is considered to have amenorrhea if her periods occur only following hormone, e. g., estrogen, administration.)

Additionally, the type of the disorder has to be specified:

1. Restricting Type: during the current episode of anorexia nervosa, the person has not regularly engaged in binge-eating or purging behavior (i. e., self-induced vomiting or the misuse of laxatives, diuretics, or enemas)

2. Binge-Eating/Purging Type: during the current episode of anorexia nervosa, the person has regularly engaged in binge-eating or purging behavior (i. e., self-induced vomiting or the misuse of laxatives, diuretics, or enemas)

The binge-eating/purging type seems to be more severe. Anorexic patients of this subtype have higher frequencies of other, comorbid disorders. They more often show comorbid depression, anxiety disorders, and substance abuse (Laessle, Wittchen, Fichter, & Pirke, 1988).

Anorexia nervosa is a disorder that affects the entire organism. Somatic manifestations and symptoms are secondary effects of starvation and include low blood pressure (hypotension), reduced resting heart rate (bradycardia), lower body temperature (hypothermia), and hypercarotenemia (a yellow-orange discoloration of the skin). The skin also becomes dry and fine body hair (lanugo hair) begins to grow. Also, due to starvation, loss of muscle mass and atrophy of the breasts can be observed. Osteopenia is another severe sign of the physical symptoms of anorexia nervosa. This reduction of the bone mineral density can lead to osteoporosis.

Dehydration and malnutrition can lead to headaches, lethargy, dizziness, and syncope (Gurenlian, 2002).

Malnutrition leads to a change in the hormone regulation and to an electrolyte imbalance. These substances are involved in the cerebral metabolism, consequently these changes are the cause of fatigue, adynamia, and heart rhythm disturbances. They can even lead to death.

In patients with anorexia nervosa, the size of the brain mass is reduced, the disorder frequently leads to neurological impairments (Davison & Neale, 2002).

Psychological manifestations include difficulties in concentration and decision making, depression, social withdrawal, and obsessiveness – especially with food (Gurenlian, 2002).

Patients with anorexia nervosa judge their self-worth in terms of weight and shape. They consider their low weight to be an accomplishment rather than a problem. Thus, they have limited motivation to change (Fairburn & Harrison, 2003).

1.1.2. Epidemiology

The age of onset of the disorder varies, though symptoms usually appear in the early or middle teenage years, often following a diet or a critical life event. The distribution of the age of onset has two peaks, at the age of 14 and 18 years (Gurenlian, 2002).

The life-time prevalence of anorexia nervosa is 2.0%, the point prevalence 0.3%. The restricting type is about three times more common than the binge-eating/purging type (Favaro, Ferrara, & Santonastaso, 2003).

Patients suffering from anorexia nervosa often have comorbid psychiatric disorders, the most common being depressive disorders, obsessive–compulsive disorders, anxiety disorders, alcohol abuse and personality disorders.

Affected female patients often have sexual disorders. A study on women at a mean age of 24 years showed that 20% did not yet have sexual intercourse, and that half of the sample did not experience an orgasm or sexual desire (Davison & Neale, 2002).

Among adolescents, dysthymia may be more strongly associated with eating disorders than depression (Zaider, Johnson, & Cockell, 2000).

1.2. Bulimia Nervosa

The origin of the term bulimia is the Greek word for ox hunger. It describes one of the main features that distinguishes bulimia nervosa from anorexia nervosa: repeated episodes of binge-eating, where large amounts of food are eaten in a relatively short time. These

episodes of eating are characterized by a loss of control – they only stop when the patient experiences an unpleasant feeling of fullness.

In many cases, these binge episodes are followed by compensatory purging behavior such as self-induced vomiting, abuse of laxatives, enemas, diuretics, caffeine, or other stimulants.

Bulimia nervosa is more difficult to detect than anorexia nervosa, because affected individuals do not show signs of illness and most are of normal weight (Gurenlian, 2002).

1.2.1. Clinical Features

The diagnostic criteria for bulimia nervosa (DSM-IV-TR) are the following:

A. Recurrent episodes of binge-eating. An episode of binge eating is characterized by both of the following:

1. Eating in a discrete period of time (e. g., within any 2-h period) an amount of food that is definitely larger than most people would eat in a similar period of time in similar circumstances; and,
2. A sense of lack of control over eating during the episode (e. g., a feeling that one cannot stop eating or control what or how much one is eating).

B. Recurrent inappropriate compensatory behavior to prevent weight gain, such as self-induced vomiting; misuse of laxatives, diuretics, or other medications; fasting; or excessive exercise.

C. The binge-eating and inappropriate compensatory behaviors both occur, on average, at least twice a week for 3 months.

D. Self-evaluation is unduly influenced by body shape and weight.

E. The disturbance does not occur exclusively during episodes of anorexia nervosa.

There are two subtypes:

1. Purging type: The person regularly engages in self-induced vomiting or the misuse of laxatives or diuretics.
2. Non-purging type: The person uses other inappropriate compensatory behaviors, such as fasting or excessive exercise, but does not regularly engage in self-induced vomiting or the misuse of laxatives or diuretics.

This distinction in two subtypes is new to DSM-IV and was not included in DSM-III-R. This ensures that the type of inappropriate compensatory behaviors is indicated with the diagnosis (Herzog & Delinski, 2001).

Patients of the non-purging type tend to have higher weight, less frequent eating attacks and fewer comorbid psychiatric disorders, although this is not supported by all studies – some findings suggest only minor differences between the two subtypes (Davison & Neale, 2002).

The binge-eating episodes are frequently kept secret and can be caused by stress and negative emotions. The ingested food is usually eaten quickly, and the chosen products are often sweets like ice cream or cake.

After the attack, feelings of disgust, unpleasantness, and fear of gaining weight are common. This triggers the purging behavior that is performed to revoke the calorie intake.

Like anorexic patients, people suffering from bulimia nervosa have a strong fear of gaining weight – their self-esteem is heavily dependent on their ability to maintain normal weight.

Bulimia nervosa leads to several somatic symptoms. The self-induced vomiting leads to potassium insufficiency, gastrointestinal lesions, and loss of adamantine due to corrosion from gastric acid. Use of laxatives can lead to diarrhea, to changes in the electrolyte homeostasis, and to irregularities in heart rhythm (Davison & Neale, 2002).

1.2.2. Epidemiology

The life-time prevalence of bulimia nervosa is about 4.6%, with a point prevalence of 1.8% (Favaro et al., 2003).

The greatest risk to develop bulimia nervosa is during the late teens and early twens, at a later age than for anorexia nervosa (Karwautz, 2004). 90% of the patients are female, some of them were overweight at the onset of the disorder and in many cases the eating attacks began during the course of a diet.

Frequent comorbid psychiatric disorders of bulimia nervosa are affective disorders (depression), personality disorders (borderline personality disorder, in particular), anxiety disorders, substance abuse and behavioral disorders. Suicide rates in bulimic patients are significantly higher than in the normal population (Davison & Neale, 2002).

1.3. Other (Atypical) Eating Disorders

The category “atypical eating disorders” or “eating disorders not otherwise specified” (EDNOS) is used to diagnose eating disorders of clinical severity that do not conform to the

diagnostic criteria for anorexia nervosa or bulimia nervosa. Most atypical eating disorders resemble anorexia nervosa and bulimia nervosa, and they may be as severe and long lasting (Fairburn & Harrison, 2003). People in this category may not meet all necessary criteria for anorexia nervosa or bulimia nervosa, e. g. they may still have a normal menstrual cycle or the frequency of purging may be too low for a diagnosis other than EDNOS.

1.3.1. Clinical Features

The diagnostic criteria for Eating Disorder Not Otherwise Specified (DSM-IV) are only given by examples:

1. For females, all of the criteria for anorexia nervosa are met except that the individual has regular menses.
2. All of the criteria for anorexia nervosa are met except that, despite significant weight loss, the individual's weight is in the normal range.
3. All of the criteria for bulimia nervosa are met except that the binge-eating and inappropriate compensatory mechanisms occur at a frequency of less than twice a week or for a duration of less than 3 months.
4. The regular use of inappropriate compensatory behavior by an individual of normal body weight after eating small amounts of food (e. g., self-induced vomiting after the consumption of two cookies).
5. Repeatedly chewing and spitting out, but not swallowing, large amounts of food.
6. Binge-eating disorder: recurrent episodes of binge-eating in the absence of the regular use of compensatory behaviors characteristic of bulimia nervosa.

This category represents a wide variety of clinical pictures, as it can range from persons who just fall short of full criteria to persons who have a qualitatively different disease pattern.

1.3.2. Epidemiology

A rather large proportion of eating-disordered patients falls in this category, with estimates ranging from 25% to 50%. Of the general population, 4% to 6% have an eating disorder not otherwise specified (EDNOS).

Some studies suggest that EDNOS is associated with later development of anorexia nervosa or bulimia nervosa (Herzog & Delinski, 2001).

1.3.3. Binge-Eating Disorder

Binge-Eating Disorder was introduced in the DSM-IV as a provisional diagnosis requiring further research. Thus, there are no diagnostic criteria, only research criteria. They are, according to DSM-IV, the following:

A. Recurrent episodes of binge-eating. An episode of binge eating is characterized by both of the following:

1. Eating, in a discrete period of time (e. g., within any 2-hour period), an amount of food that is definitely larger than most people would eat in a similar period of time under similar circumstances.
2. A sense of lack of control over eating during the episode (e. g., a feeling that one cannot stop eating or control what or how much one is eating).

B. The binge-eating episodes are associated with at least three of the following:

1. Eating much more rapidly than normal
2. Eating until feeling uncomfortably full
3. Eating large amounts of food when not feeling physically hungry
4. Eating alone because of being embarrassed by how much one is eating
5. Feeling disgusted with oneself, depressed, or feeling very guilty after overeating.

C. Marked distress regarding binge-eating.

D. The binge-eating occurs, on average, at least 2 days a week for 6 months.

E. The binge-eating is not associated with the regular use of inappropriate compensatory behaviors (e. g., purging, fasting, excessive exercise) and does not occur exclusively during the course of anorexia nervosa or bulimia nervosa.

Compared with obese patients who do not binge, obese patients who engage in binge-eating report significantly greater distress about eating and weight, as well as psychosocial impairment and psychopathology including depression. These patients also report lower self-esteem than non-binging obese subjects.

Although the introduction of the purging and non-purging subtype of bulimia nervosa with DSM-IV has clouded the distinction between binge-eating disorder and bulimia nervosa, it seems to be a distinct disorder (Herzog & Delinski, 2001).

Epidemiological research on binge-eating disorder is rare. A study on a representative Austrian sample (1000 female subjects) revealed a point prevalence rate of 3.3%, with

higher rates in overweight (4.1%) or obese (8.5%) women than in normal weight women (2.9%). This difference, however, was not statistically significant (Kinzl, Traweger, Trefalt, Mangweth, & Biebl, 1999).

The gender difference for the prevalence of binge-eating disorder does not seem to be as severe as in anorexia nervosa or bulimia nervosa. Several studies did not find any gender differences when comparing the rates of (white) women and men.

Studies have found very high rates of obesity and psychiatric comorbidity, especially mood and anxiety disorders (Striegel-Moore & Frank, 2003).

1.4. Risk Factors

In this section, the risk factors for eating disorders will be described. Anorexia nervosa, bulimia nervosa and atypical eating disorders share most of their risk factors, which is why they are not depicted separately.

Nonetheless, there are specific influences that are only relevant for either of these disorders. They will be mentioned at the end of this section, especially in regard to anorexia nervosa.

Generally, risk factors can be grouped in extrinsic (adverse experiences) and intrinsic (genetic vulnerability) factors (Karwautz et al., 2001). Another classification is to group them as

- genetic risk factors,
- personality traits and
- environmental risk factors (Karwautz, 2004).

1.4.1. Genetic Risk Factors

Genetic risk for eating disorder can be assessed in various ways, ranging from classical twin and family studies to only newly emerging molecular genetic studies. These methods will be discussed in more detail in section 2.4.

Twin studies showed a greater concordance rate of anorexia nervosa in monozygotic than in dizygotic twins. The concordance for monozygotic twins ranged from 30% to 65%, whereas for the dizygotic twins it only was between 0% and 9% (Karwautz, 2004). This suggests a relatively high importance of genetic factors, although no specific genetic risk factor (such as a vulnerability gene) can be identified by this approach.

For genetic traits, clustering within families should be observable. For anorexia nervosa, the prevalence rate in sisters of affected individuals ranges from 3% to 7%, and in female first-degree relatives from 4% to 10% (Karwautz, 2004).

Other studies also show familial clustering of eating disorders with other comorbid disorders. Higher prevalences of affective disorders and obsessive–compulsive disorders have been found in first-degree relatives of subjects suffering from anorexia nervosa or bulimia nervosa. Furthermore, a family history of an eating disorder, mood disorder, or alcohol and substance abuse seems to enhance the risk for anorexia nervosa or bulimia nervosa in adolescents (Gorwood, Kipman, & Foulon, 2003).

Molecular genetic studies indicated that genes involved in serotonergic regulation, especially the gene for the 5-HT_{2A} receptor, might be susceptibility factors for anorexia nervosa (Karwautz, 2004). The findings of these studies will be discussed in more detail in chapter 3, since they are of central interest to this paper.

1.4.2. Personality Traits

There are, amongst others, two very important personality traits that may be risk factors in anorexia nervosa: perfectionism and temperamental traits such as harm avoidance and novelty seeking.

Perfectionism is the strongest candidate for being related to anorexia nervosa. It has been found before onset, it enhances during the illness, and it is present after recovery (Karwautz, 2004). Persons with high levels of perfectionism are over-represented in first degree relatives of patients with an eating disorder (Gorwood et al., 2003), and family members of anorexic subjects also have higher rates of obsessive–compulsive personality disorders (Karwautz, 2004).

The trait has also been identified as a risk factor for bulimic women considering themselves as overweight.

Temperamental traits also seem to play an important role in eating disorders and anorexia nervosa. Studies have found higher harm avoidance, lower cooperativeness, and lower novelty seeking in subjects suffering from anorexia nervosa. All these differences were independent from body weight (Karwautz, 2004).

Personality traits are substantially influenced by genetic factors, and they may interact with environmental influences in anorexia nervosa. Moreover, they can be associated with specific neurotransmitter systems – novelty seeking is linked to dopaminergic functioning, and harm avoidance is linked to serotonergic functioning (Karwautz, 2004).

A third personality trait that might be associated with eating disorders is low self-esteem. Karwautz (2004) states that the risk of developing an eating disorder in children with low self-esteem is eight times higher than in those with high self-esteem. This might be explained by the fact that a negative self-evaluation enhances the drive to attain perfection in all areas of life.

Low self-esteem seems to be a potent risk factor for anorexia nervosa. High levels of shyness, loneliness and feelings of inferiority have been reported in adolescents preceding anorexia nervosa of the binge-purging type, but not in the restricting type (Karwautz, 2004).

1.4.3. Environmental Risk Factors

Familial risk and parental problems

Familial background and family environment are vital factors in the development of eating disorders. Both parental dieting and encouragement by parents to lose weight has significant impact on dieting concerns and behaviors in adolescent girls. Women with anorexia nervosa or bulimia nervosa report receiving more comments about weight or shape by family members during their childhood than controls (Wade, Gillespie, & Martin, 2007). Maternal restraint and disinhibition predicts higher body dissatisfaction and dieting behaviors even at the age of eight.

Families with an eating-disordered child share some common characteristics, but it is still unclear if these interaction patterns exist prior to the onset of the disorder. Some studies found that adverse parenting (especially low contact, high expectations and parental discord) constitutes a risk factor for eating disorders (e. g. Gorwood et al. (2003), Fairburn and Harrison (2003) or Wade et al. (2007)). Other longitudinal studies did not find any specific interaction pattern that proved to be predictive of anorexia nervosa or bulimia nervosa (Karwautz, 2004).

As is the case with most other psychiatric disorders, the presence of a psychiatric diagnose in the parents also increases the risk of developing an eating disorder. Family history of an eating disorder (and also of depression and obsessive–compulsive disorder) is a risk factor for developing an eating disorder.

In recent studies, genetic factors were found to explain a large part of this familial risk (Gorwood et al., 2003).

Life events

Stressful life events and psychosocial difficulties often precede anorexia nervosa and bulimia nervosa. These include negative life events concerning the patient's parents, interpersonal difficulties such as teasing and bullying at school or pudicity events. The latter seem to be especially important for anorexia nervosa – they occur in 20% of the cases.

However, such life events are experiences shared by the vast majority of people and cannot explain the specific development of an eating disorder. It has been proposed that the important factor is the *meaning* of the event: loss events are postulated to be related to depression, threat events to anxiety disorders. Pudicity events associated primarily with sexual disgust seem to be common in anorexia nervosa, which suggests that life events linked to the emotion of disgust and shame may be a trigger for eating disorders.

An alternative theory states that the response to an event is an important component in shaping the outcome. Eating-disordered patients were found to have a more helpless coping response: subjects suffering from anorexia nervosa applied more avoidance strategies, subjects suffering from bulimia nervosa ruminated more on their problems (Karwautz, 2004).

General cultural risk

The majority (90% to 96%) of subjects suffering from eating disorders is female. The greatest risk for developing anorexia nervosa is during the mid teens, whereas the risk for bulimia nervosa is highest in the late teens and early twens.

Race (white) and heterosexual orientation are also risk factors for eating disorders (Karwautz, 2004), as is living in a western society (Fairburn & Harrison, 2003).

However, social class is not correlated with the risk of developing an eating disorder, although adolescents from a higher social class were found to desire lower body weights, to diet more often, to binge more severely, and to exercise more for weight control (Karwautz, 2004).

Additionally, there is a higher risk for the development of an eating disorder, especially anorexia nervosa, in certain groups of people. These groups include dancers, runners, skaters, models, actors, gymnasts, wrestlers, and college sorority members, for whom thinness is emphasized and highly valued (Gurenlian, 2002).

Interpersonal Problems

Social support and peer group interaction play an important role as risk factors for eating disorders. Bullying and teasing of girls by peers when gaining weight during adolescence results in negative feelings towards body shape and further development of eating disorders.

Social support influences vulnerability for mental disorders in general. Anorexia nervosa patients and bulimia nervosa patients have deficits in their social network. Thus, insufficient or inadequate social support seems to be a predictor of bad outcome in eating disorders (Karwautz, 2004).

Traumatic Experiences

Sexual and physical abuse are non-specific risk factors for eating disorders, especially in connection with psychiatric comorbidity. Abuse is more strongly related with bulimia nervosa than with anorexia nervosa (Karwautz, 2004).

1.4.4. Specific Risk Factors for Subtypes of Eating Disorders

In summary, the risk factors for developing an eating disorder of any type are:

General factors:

- female sex
- adolescence and early adulthood
- living in a western society

Environmental risk factors:

- family dieting
- negative comments about weight or shape
- adverse parenting (questionable)
- family history of eating disorders, depression, or obsessive–compulsive disorders
- traumatic experiences (sexual and physical abuse)

Personality traits:

- perfectionism

- temperamental traits like harm avoidance, low novelty seeking, low self-esteem

Genetic risk factors:

- distinctive features of the serotonergic regulation
- other factors, like family history of eating disorders, may also be subsumed under genetic risk factors

As mentioned at the beginning of this section, most of the risk factors are shared for the different types of eating disorders, especially for anorexia nervosa and bulimia nervosa. Fairburn, Cooper, Doll, and Welch (1999) have proposed a classification into risk factors that increase the risk for a general psychiatric disorder and risk factors that increase the risk for a specific eating disorder.

Findings in this study suggest that there are some differences in the importance of the risk factors mentioned above for the different types of eating disorders.

Subjects suffering from anorexia nervosa have higher rates of parental eating disorders and of family dieting. Another rather specific risk factor for anorexia nervosa is the psychological trait perfectionism, which is especially common in anorexic subjects (Fairburn et al., 1999).

In addition, it has been found that high paternal expectations and paternal control is uniquely associated with the risk for developing anorexia nervosa. Maternal expectations or control did not play an important role for the distinction of the two types of eating disorders (Wade et al., 2007).

Bulimia nervosa patients, on the other hand, show higher rates of parental obesity during their childhood, they have an earlier menarche than anorexic patients, and there is a trend to having received more critical comments about eating, appearance, and weight from family members. Furthermore, the exposure to parental psychiatric disorders during the patient's childhood is higher for bulimia nervosa, especially for parental depression, alcohol abuse and drug abuse (Fairburn et al., 1999). This is in accordance with other studies (Fairburn & Harrison, 2003).

Another important risk factor that seems relevant for the distinction of bulimia nervosa from anorexia nervosa is premorbid repeated sexual and severe physical abuse, where bulimic patients show higher rates (Karwautz, 2004).

2. Genes and Polymorphisms

2.1. What Genes Do

Before starting to outline what genes do, it is important to make clear what genes do not do. Popular science writers often refer to genes "for" schizophrenia, intelligence, or depression. Needless to say, there are no genes "for" any of these traits or mental disorders. Genes will certainly play a contributory role in their development by influencing individual variations in liability for certain traits or disorders, but referring to a gene "for" some disorder can create the misleading impression that there is a direct genetic influence on these features.

There are, of course, some disorders that are linked to certain genes in a more direct way (meaning that the disorders are caused by a mutation of a single gene not depending on any environmental risk experience). Examples of these are cystic fibrosis, tuberous sclerosis, and Huntington's disease, all of which strictly follow Mendelian inheritance patterns and are therefore called Mendelian disorders. In more complex disorders like anorexia nervosa, vulnerability is a result of the interplay of many genes and environmental influences.

Even in Mendelian disorders however, genes do not actually cause the phenotype (the visible manifestation of the genetic liability). Rather, they are involved in the causation of specific biochemical reactions that serve to give rise to the phenotype. In the following, this process will be explained in detail.

2.1.1. The Genetic Code

From the mid-nineteenth century, it has been known that certain characteristics are passed on from generation to generation. This was discovered by the Austrian monk Gregor Mendel who studied patterns of inheritance in pea plants. He concluded that carriers of this information, the *genes*, exist in alternate forms, now called *alleles*.

The importance of this finding was not realized until much later, as Watson and Crick discovered in 1953 that the relevant genetic material, deoxyribonucleic acid (DNA), had a paired helix structure. This structure allows for replication of the genetic material, which is necessary for cell division (Rutter, 2006).

The carrier of the genetic information, the DNA, consists of combinations of four bases.

Theses bases are rings of carbon and hydrogen atoms and they constitute the *nucleotides* – adenine (A), cytosine (C), guanine (G), and thymine (T).

The nucleotides are organized in triplets – the so-called *codons*. The genetic information lies in the sequence of codons, like e. g. ATC CGA CTT ACC etc. They form the DNA, which is organized in a double helix. The two strands of DNA are reciprocal.

Although the DNA molecule is a very long chain, the information contained is not organized uninterruptedly. There are *exons* that code for polypeptides (and, in the long run, proteins), and the intervening *introns* that do not code for any specific genetic product.

2.1.2. From Genes to Proteins

The expression of genetic information in all cells is a one-way system. It begins with the DNA, which carries the inherited genetic information. The DNA is translated into a polypeptide, and ultimately into a protein, by processes known as transcription and translation.

Transcription

In the transcriptional phase, the DNA molecule is transformed into an mRNA (messenger ribonucleic acid) molecule. The mRNA differs from the DNA chemically, but the encoded genetic information is retained from original molecule. Not all genetic information in the DNA, however, is transcribed into mRNA. The *promoter region*, a section in the DNA that regulates the transcription process, is not transformed. The elements that influence the transcription process are called *transcription factors*, and they collectively constitute the *promoter*. Some of those factors are located distantly from the genes they are influencing (and therefore are called *trans-acting*), while some of them have a local function in the DNA duplex on which they reside (they are called *cis-acting*). Additionally, the promoter can contain so-called *enhancers* that enhance the transcriptional activity of specific genes, and *silencers* that inhibit that transcriptional activity. These transcriptional factors are not usually termed genes, but they are still made out of DNA sequences and will be inherited along with the rest of the DNA.

In the second transcription phase, after the removal of the promoter region, the introns are eliminated and the exons are connected – a process called *splicing*. This produces a continuous segment of end-to-end exons.

Many genes in humans show *alternative splicing*. This means that different exon combinations result from the same gene during RNA processing. The same gene may therefore give rise to multiple unique proteins with different effects.

Translation

During the process of translation, the mRNA segments are transformed into polypeptides and proteins. Each codon represents a specific amino acid. The information in the mRNA is processed sequentially, resulting in a chain of a large number of amino acids – the protein product coded in the specific gene (Brown, 1999).

The protein structure is highly varied and cannot easily be predicted from the amino acid sequence. In addition to the possible variations in the mRNA (as a result of alternative splicing), there are various post-translational modifications like the alteration of some amino acids and polypeptides.

The conversion of polypeptides into proteins involves the folding of proteins, which is crucial for them to unfold their effects. The precise mechanisms involved in the folding process are highly complex and still unclear. They are driven by genes, but they are also influenced by the environment of the cell.

These proteins coded in the DNA play a role in the liability to genetically influenced traits or disorders. In most cases however, it remains unclear how (and why) they lead to certain outcome behaviors. The ultimate behavioral outcome of gene action is a complicated process, and it is not directly determined by the inherited DNA in a single causal chain (Rutter, 2006).

2.2. Polymorphisms

The term "genetic polymorphism" (sometimes also called "morphism") describes the fact that genetic variants of a gene (alleles) coexist in temporary or permanent balance within a single interbreeding population in a single spatial region (Huxley, 1955).

Polymorphisms result from the evolutionary process, and they are extremely common. To be called a polymorphism, the allelic variant of a gene must have an occurrence of at least four percent; otherwise it is called a mutation. Polymorphisms serve to retain variety of form in a population living in a varied environment, and they ensure biodiversity by the process of genetic recombination (Sheppard, 1975).

Already, Mendel stated that a gene consists of a pair of alleles. One allele is inherited from each parent. Out of all existing polymorphisms for a specific gene, each individual possesses two allelic variants at the specific genetic locus that make up the subject's *genotype*. In this paper, the alleles of interest are the 5-HT_{2A} –1438 A and 5-HT_{2A} –1438 G alleles (explanation follows in section 3.4.1).

An individual can either possess two identical alleles (e. g. AA or GG) – the subject would then be called *homozygous* – or two different alleles (e. g. AG), where the subject would be said to be *heterozygous*.

A specific allele can be dominant or recessive. This specifies the effect of the allele on the *phenotype* (the observable characteristics of an organism). For *dominant* alleles, the phenotype is entirely determined by this allele. The other allele at that specific locus does not have any effect in the presence of the dominant co-allele – it remains silent and is called *recessive*. A recessive allele only determines the phenotype if the individual is homozygous for the recessive allele.

Another possibility (that was not described by Mendel, but later by Karl Correns) is *incomplete dominance*, where a heterozygous genotype creates an intermediate phenotype. In this case, both alleles influence the observable characteristics in a dosage-dependent manner.

Yet another way in which two alleles might interact is *co-dominance*. In co-dominance, neither of the alleles is recessive. If an individual is heterozygous, it expresses both phenotypes. An example for this would be the AB0 blood group system, where the alleles I^A and I^B are co-dominant. An I^AI^B genotype results in the individual having type AB blood as phenotype. The third allele, I^0, is recessive to the I^A and I^B alleles, so genotypes I^AI^A and I^AI^0 both cause type A blood as phenotype. This also exemplifies that there can be more than two alleles for a single genetic locus.

Mendel stated in his Law of Independent Assortment, also known as Inheritance Law, that the inheritance of one trait (in a specific genetic locus) does not affect the inheritance pattern of another trait (in another genetic locus). This is only true when the two genes reside on different chromosomes. When two genes reside on the same chromosome, they are inherited conjointly – a phenomenon called *genetic linkage*, in which the alleles are said to be in *linkage disequilibrium*.

There is, of course, an exception to this. During cell division (in the meiosis phase), chromosome pairs may intertwine and exchange fragments – a process called *crossing over*. This can cause two genes on the same chromosome to separate from each other. The two genes can then be inherited separately, causing the linkage disequilibrium to be incomplete. The further apart on a chromosome two genes reside, the higher the probability that they will separate and break the linkage (Brown, 1999).

Polymorphisms can be explored by the use of genetic markers. A genetic marker is a known DNA sequence that shows variation due to mutation or alteration. They can be used to study the relationship between a disease and its genetics.

There are different kinds of genetic markers, and the following list is far from being exhaustive (Spooner, Treuren, & Vicente, 2005):

- Restriction fragment length polymorphism (RFLP)
- Amplified fragment length polymorphism (AFLP)
- Random amplification of polymorphic DNA (RAPD)
- Variable number tandem repeat (VNTR)
- Short tandem repeat (STR)
- Microsatellite polymorphism
- Single nucleotide polymorphism (SNP)

The smallest unit of genetic variation is the single nucleotide polymorphism (SNP). It is a variation of the DNA sequence by only one nucleotide: a base in a certain genetic locus is substituted by a different one. SNPs reflect past mutations that were mostly (but not exclusively) unique events. Two individuals sharing a variant allele share a common evolutionary heritage.

SNPs are quite common in the human DNA. It has been estimated that when comparing human DNA sequences, an SNP can be found every 1000–2000 nucleotides. Hence, the whole human genome might contain 1.6 million to 3.2 million SNPs.

Most human variation that is influenced by genes can be traced to SNPs. They can influence the development of psychological traits or the susceptibility to certain diseases or disorders (Stoneking, 2001).

In this paper, a single nucleotide polymorphism in the serotonin system and its relevance for the development of anorexia nervosa will be investigated.

2.3. Gene and Environment Interplay

There is no direct connection between a particular genetic mutation and development of a mental disorder or disease. This is not the case for a single-gene disorder (where the mutation of one single gene is responsible for the outcome), nor is it the case in complex disorders. The latter are always a result of multiple genes and their interplay (including influences from non-protein-coding introns). Both types also depend on environmental factors.

2.3.1. Gene–Environment Interaction

Considering that both genetic and environmental factors are important for the development of a specific disorder, the question is raised whether genetic and environmental effects act independently. If so, the effect is said to be "additive", meaning that the combined effects are no more than the sum of both effects considered separately.

Recent research has shown that this is not the case for many disorders. It has been found that the effect of certain genes is not independent from environmental factors and vice versa, but that the effect of environmental hazards or risk factors is dependent on environmental influences. In other words, there is interaction between genetic and environmental factors (this is often written as G × E). The presence of certain environments may increase the genetic effects (or, viewed from a different perspective, the presence of a certain genetic configuration may increase the effects of a particular environmental factor). This would be termed a synergistic interaction. Otherwise, the effect of the one could be reduced by the other (Rutter, 2006).

The question whether there is a significant interaction between genes and environment can be assessed in many different ways (see section 2.4). One of them is to focus on the interaction between some identified and measured genetic allelic variation and some identified and measured environmental risk factor.

2.3.2. Gene–Environment Correlation

Another way in which genes can "interact" (although not in the statistical meaning of the word) with environment is termed *gene–environment correlation*. There are three types of gene–environment correlations – passive, active, and evocative.

Passive correlations mean that parents who pass "risky" genes on to their children also, on the whole, tend to create "risky" rearing environments for their children. Simply, the parental genes increase the likelihood that the children will experience more environmental risks. Genes thereby influence the children's environment and, indirectly, their vulnerability to a certain disease.

Active gene–environment correlations behave in a different way, although the genetic effects are still indirect. In the case of active correlation, the child's genetically influenced behavior has effects on the environmental situations that the child grows up in. A child that engages in aggressive reading, as an example, will have much more opportunity to learn from books than a child with different interests. The children's interests are, of course, influenced by their genes (as well as by their upbringing). Genetically influenced traits will lead to distinctive experiences.

Evocative gene–environment correlations work in a somewhat comparable fashion, the difference being that the effect is not on the selecting of environments, but rather on the responses induced in other people. Children vary in their ability to successfully socialize with others, thereby influencing their environment and their social experiences (humor and compassion, quarrel and rejection). This makes it difficult to separate genetic and environmental influences (Rutter, 2006).

2.3.3. Good Genes and Bad Genes

The identification of certain genes and their allelic variants as genetic risk factors raises the question whether there are certain genetic configurations that are undesirable or even bad. It might be assumed that genes could be subdivided into those that are good and those that are bad. This, however, is not the case.

Genes may be (partly) responsible for the development of traits or diseases. Considering traits, the first point that has to be made is that a specific gene only constitutes a small influence on a specific trait – the effect of the individual identified genes has proved to be very small. As an example, an identified susceptibility gene for the temperamental trait sensation seeking accounted for only about four percent of the variance on that trait.

The second point is even more crucial. The extension of the notion that genes can possibly be divided into good and bad implies that there are some human attributes or traits that are inherently good or bad in their effects. The temperamental feature of behavioral inhibition or emotional constraint or withdrawal in new or challenging situations can serve as an example. It constitutes a risk factor for anxiety disorders, but it is a protective factor against antisocial behavior. So, as traits cannot be divided into good and bad, neither can genes or their allelic variants.

Considering diseases, it might more easily be presumed that they are undesirable and therefore bad. This is certainly true for some (Huntington's disease, for example). But even in a single-gene disorder there might be an important mixture of good and bad effects. A well known example of this is provided by thalassemia, the condition giving rise to sickle-cell disease. Sickle-cell disease is a severe disorder with high mortality, and, obviously, it would be good to eliminate the suffering and death it brings about. However, the gene responsible for sickle-cell anaemia is incompletely recessive. Carriers of the allele responsible for sickle-cell disease also develop a small number of sickle red blood cells. These do not cause any symptoms, but they have a protective influence against malaria – an equally serious, equally deadly, disease.

Hence, the supposed subdivision of genes into good and bad is misleading and unhelpful.

The effect of genes always have to be seen in their interplay with environmental conditions (Rutter, 2006).

2.4. Methods for Research of Genetic Influences

There are several ways to study genetic influences on traits. Ramoz, Versini, and Gorwood (2007) adduce family and twin studies and linkage and association studies. Furthermore, the transmission of certain alleles from parents to their offspring can be analyzed with a specific study design (transmission disequilibrium test).

Another method is the gene–environment interaction study. Moffitt, Caspi, and Rutter (2005) add a description of a strategy to find these gene–environment interactions. A special case in this context is the use of case-control or case-sib designs (Witte, Gauderman, & Thomas, 1999).

2.4.1. Family Studies

Genetic risk for diseases or mental disorders runs in families, and this familial loading might reflect a genetic liability. This becomes even more plausible if the strength of the loading varies systematically with the degree of biological genetic relatedness – i. e. being highest in first-degree relatives, lower in second-degree relatives and even lower in third-degree relatives (Rutter, 2006). Based on this idea, family studies use the prevalence of mental disorders in family systems to estimate their heritability.

Family studies do not permit a clear separation of genetic and non-genetic influences, which is a major limitation. Differences in genetic relatedness are accompanied by differences in environmental risks, so the effects cannot be estimated separately.

Nevertheless, family data can be valuable in suggesting how many susceptibility genes are likely to be operative in a certain disorder. This can be accomplished through mathematical calculations. Since first-degree relatives share 50% of their genes it is possible to conclude that, if the risk is much less than 1 in 2, inheritance involves synergism between particular patterns of genes. A precise figure of genes involved cannot be calculated, however (Rutter, 2006).

2.4.2. Twin Studies

Twin studies are aimed at disentangling the variation that is caused by genes and environment. By comparison of monozygotic twins (who share 100% of their genetic configuration) and dizygotic twins (who share only 50% of their genes) it becomes possible to estimate

the percentage of the variance of a phenotype that can be attributed to genetic effects (heritability).

Twin studies allow for the parsing out of sources of familial aggregation, at least to some extent. Variance in susceptibility can be partitioned into additive genetic, shared environmental and unique environmental factors. Shared environmental factors reflect environmental influences to which both twins are exposed. They contribute to the similarity of twins. Unique environmental factors refer to environmental experiences to which only one member of a twin pair is exposed. This allows a more accurate calculation of heritability (Bulik, 2005).

Several studies have shown a family aggregation for anorexia nervosa and also bulimia nervosa. For anorexia nervosa, heritability estimations range from 60% to 88% (Ramoz et al., 2007).

There are various criticisms of the twin method. One of the more severe is that the heritability estimate may reflect factors other than genes. It has been shown that the application of heritability estimation on questionnaire data can lead to heritability rates greater than 100%, when the assumptions of the model are not met. Also, estimated heritability rates tend to overestimate the effect of genetic configuration (Schönemann, 1997).

As Schönemann (1997) points out, the biometric definition of heritability is limited to phenotypical *differences*. All genetically determined traits or features that have the same expression in all individuals are ignored. To illustrate this, he uses an example quoted from Hirsch (1981): Leggedness (the number of legs humans have) is clearly predetermined by genes. Under normal circumstances virtually all people are born with two legs. In wartimes however, people may lose one or both legs due to environmental influences. All *variability* in leggedness will be due to environment, not to genes. The heritability estimate for the trait leggedness will, therefore, be zero.

2.4.3. Adoption Studies

Adoption studies investigate the correlations between adopted children and either their biological parents or their adoptive parents (or both) on a certain trait or disorder. Adopted children are not reared by their biological parents, which allows for a more clear-cut separation of genetic and environmental influences.

In the field of anorexia nervosa, there are no adoption studies (Fairburn & Harrison, 2003).

2.4.4. Linkage Studies

Linkage studies have the goal to identify vulnerability genes for specific psychiatric disorders. In these studies, the whole human genome is screened, without an *a priori* hypothesis (Ramoz et al., 2007).

This type of study does not identify individual genes. Rather, it identifies particular segments of a chromosome. This is achieved by investigating the co-inheritance between a gene locus and the disorder (or trait) being studied in related couples (generally sib-pairs) that are both affected by the particular disorder. Statistics can determine if the extent to which two affected siblings share identical genetic alleles exceeds chance.

So-called polymorphic genetic markers are used to examine the transmission and segregation of genetic material. These are markers that vary across individuals in any population. In this manner, chromosomal regions that are linked to a specific disorder can be identified.

A limitation of linkage studies is that the identified chromosomal areas are quite large and therefore include a very large number of genes. So the task of finding the relevant susceptibility gene remains quite challenging (Rutter, 2006).

Furthermore, the effect of the susceptibility gene has to be quite strong to be detected, and the sample size therefore has to be large. This becomes, of course, more severe when the disorder examined is rare, as eating disorders are (Ramoz et al., 2007).

2.4.5. Association Studies

Just as with linkage studies, association studies are carried out to identify genes of vulnerability in eating disorders and other diseases, but they function on an entirely different principle.

Association studies compare the frequencies of specific alleles and genotypes between patients and controls in selected candidate genes. If the distribution of the alleles or genotypes varies between patients and healthy subjects, the gene is likely to be associated with the disorder. This approach is also known as *candidate gene approach* (Kwon & Goate, 2000).

Identifying variants of genes that are linked to a specific disorder helps the understanding of vulnerability for, or resilience against, that disorder. Another benefit could be the explanation for certain groups of patient's poor response to pharmacological treatment, if that treatment is linked to neurological pathways influenced by that gene (Ramoz et al., 2007).

The major advantage of this design is that it permits detection of susceptibility genes even when their effect is relatively small. This is especially important in the study of

multifactorial disorders (Rutter, 2006). Nevertheless, the samples used in this design have to be quite large (Hinney, Remschmidt, & Hebebrand, 2000).

One of the disadvantages of association studies is that association may arise as a result of population stratification. Population subgroups frequently differ in allele frequencies for reasons that do not have anything to do with the disorder of interest. So, if cases and controls are drawn from somewhat different populations, spurious associations may be detected and reported (Rutter, 2006).

Association studies carried out so far, both in the field of anorexia nervosa and in general, reveal occasionally significant but often unreplicated findings (Bulik, 2005).

2.4.6. Transmission Disequilibrium

The problem of population stratification is avoided by the use of the transmission disequilibrium test (TDT), developed by Spielman, McGinnis, and Ewens (1993). TDT uses family trios consisting of an index patient and both parents, requiring the parents to be heterozygous for the allele that is presumably associated with a certain disease or trait. The test evaluates the frequency with which the allele is transmitted to the affected offspring. If it is transmitted more frequently than expected by chance, this allele (or an allele in close proximity) is thought to predispose to the disease. If it is transmitted less frequently than expected by chance, it might be protective (Hinney et al., 2000).

The transmission disequilibrium test has relatively low statistical power and therefore needs larger sample sizes to detect significant effects (Klump & Gobrogge, 2005).

2.4.7. Gene–Environment Interaction (G × E) Studies

The research of genetic influences for eating disorders, and also other disorders and diseases, has yielded inconsistent and sometimes contradictory results. This leads to the (not very surprising) conclusion that not genes alone are responsible for vulnerability. Genes and environmental factors work together in a complex way. Therefore, in the last two decades, research has increasingly targeted the influences of both. These studies investigate the interaction of genes and environmental factors, and are therefore called gene–environment interaction studies – in short G × E studies (Thapar, Harold, Rice, Langley, & Michael, 2007).

The term G × E can describe different scenarios. In a methodological sense it refers to the situation where genetic factors influence individual sensitivity to certain environmental influences. This means that the presence of an interaction is deduced statistically. This

does not necessarily mean that this interaction is also present on a biological level, although this clearly is the hope.

The investigation of gene–environment interaction has become an important research area in psychiatry, psychology, and medicine (Thapar et al., 2007). It can help understanding and identification of both environmental and genetic risk factors and their interplay. Just like association studies, it can facilitate the understanding of biological causes of (not only) mental disorders and it can have benefits for the improvement of pharmacological treatment.

The study conducted in this paper is a gene–environment interaction study. The most common way of data analysis for G $\times$ E studies is the use of *logistic regression*. This method is outlined in sections 4.4.1 and 4.4.2.

There are various designs that can be used. One possibility is the use of a matched sample. For each affected subject (called the "case") an unaffected subject (called the "control") is selected. The matching can be done by age or other variables that are somehow related to the disorder of interest. A special case of a matched sample is the case-sibling design (Witte et al., 1999). For this design, the unaffected control group consists of siblings of the affected persons. This has the advantage that population stratification can be avoided and therefore cannot distort the findings.

Data analysis in this case can be done with *conditional logistic regression*. Some theoretical background on this is given in section 4.4.3.

A relatively new approach for both matched and unmatched samples is *odds ratio multifactor dimensionality reduction*. This method is introduced in section 4.4.4.

3. Serotonin Regulation and Eating Disorders

The topic of this paper is to investigate the role of a certain polymorphism in the 5-HT_{2A} receptor gene as a risk factor for anorexia nervosa. The 5-HT_{2A} receptor is part of the serotonin system, which is why this chapter will first outline the role of the serotonin regulation for eating behavior in general and anorexia nervosa in particular, followed by a closer look at the 5-HT_{2A} receptor gene and its polymorphisms. One of them, the 5-HT_{2A} –1438 G/A polymorphism, will be explained in more detail, and its role in anorexia nervosa will be elucidated.

3.1. Parts of the Serotonin Regulation

Serotonin (5-hydroxytryptamine, or 5-HT) is one of many important neurotransmitters in the central nervous system and other parts of the body. It was originally discovered as a vasoconstrictor substance in blood serum, a serum agent that affects vascular tone (that is also where the name "serotonin" comes from). Only later was serotonin's chemical composition and broad range of physiological roles elucidated. That is the reason why 5-HT is the preferred name in the pharmacological field (Serretti, Calati, Mandelli, & De Ronchi, 2006).

Serotonin is involved in many regulative circuits for different behaviors (perception of pain, regulation of sleep, mood and appetite being only a few examples), and it also plays a role in development and retention of various psychiatric disorders.

The serotonergic system is complex and consists of numerous ascending and descending nerve tracts. Origin of these tracts are the raphe nuclei (*nuclei raphes*), located in the brain stem. From there, various connections to different regions of the brain exist (e. g. to thalamus, hypothalamus, medial temporal lobe, neocortex, basal ganglia, cerebellum, and others).

The serotonin neurotransmitter belongs to the group of *monoamines*, along with adrenaline, noradrenaline, dopamine, and other neurotransmitter substances.

Starting point for the synthesis of serotonin is the amino acid L-tryptophan (L-TRP), which is derived from diet. In the first of the two steps, L-tryptophan is transformed into 5-hydroxy-L-tryptophan (5-HTP) by the enzyme tryptophan hydroxylase (TPH). The subsequent and final step in the synthesis is conducted by the enzyme 5-hydroxytryptophan-decarboxylase and leads to the final product, serotonin (5-HT). Its molecular formula is $C_{10}H_{12}N_2O$ (the chemical structure depicted in figure 3.1).

One major serotonin metabolite, a product of serotonin decomposition, is 5-hydroxyindoleacetic acid (5-HIAA). The transformation is performed with the help of the enzyme monoamine oxidase (MAO-A). After that, it can be transported back into the neuron that it was released from – a process called *reuptake*.

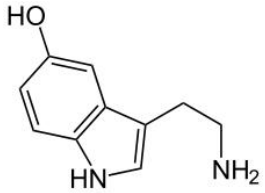

Figure 3.1.: *Chemical structure of serotonin (5-HT). (Source: http://de.wikipedia.org/wiki/Bild:Serotonin_-%285-HT%29.svg).*

Serotonin is stored in synaptic vesicles of the neurons until it is released into the synaptic cleft. The trigger for the release is an electric nerval stimulus, the action potential, that is received via the axon of the nerve cell. (For illustration, figure 3.2 shows a schematic drawing of a synapse.)

Once released, the function of the serotonin (and also of all other neurotransmitters) is to transmit the electric signal to the postsynaptic cell by a chemical pathway. This is achieved through the binding of the neurotransmitter, in this case serotonin, to a receptor in the membrane of the postsynaptic cell.

The postsynaptic receptor in the membrane of the cell to be affected has to recognize the transmitter (which is achieved by the specific form of the protein that the neurotransmitter consists of) and must consequently trigger a mechanism that enables signal transmission. This is done via the activation of ion channels.

Two types of receptors exist. In directly operating receptors, the ion channels are part of the same molecule that the receptor consists of (they are called *ionotropic receptors*). In indirectly operating receptors on the other hand, receptor (e. g. *metabotropic receptor*) and ion channel are two different molecules and have to communicate via *second messengers*.

The binding of a ligand (like serotonin) to a directly operating receptor causes the ion channels to open. The transmission is said to be conveyed by a *first messenger*. The

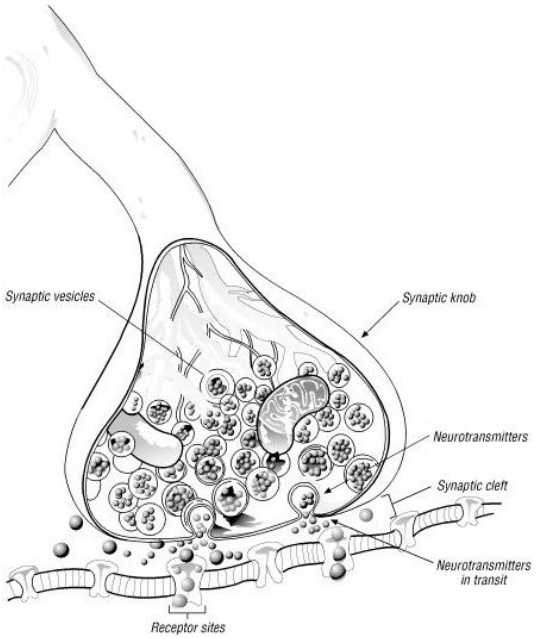

Figure 3.2.: *Synapse. (Source: http://www.patientcenters.com/autism/graphics/pdd0103.gif).*

opening of the ion channels allows ions of a certain type (like Na^+, K^+, Ca^{2+} or Cl^-) to penetrate the cellular membrane, which causes a depolarization in the target cell. The chemical signal is transferred back into an electric signal.

In indirectly operating receptors, the binding of the ligand causes the opening of the ion channel via a second messenger. In most cases, an additional protein (for instance, a G-protein that contains the amino acid guanin), is needed. The G-protein transforms inactive guanosine diphosphate (GDP) into active guanosine triphosphate (GTP). This stimulates enzymes like adenylate cyclase to produce (for example) cyclic adenosine monophosphate (cAMP) that acts as second messenger, opens the ion channels, and permits the signal transmission of the nerval impulse (Pritzel, Brand, & Markowitsch, 2003).

There is a great number of different serotonin receptors – at least 13 have been identified in humans. They are grouped into seven classes, named 5-HT_1 to 5-HT_7. They are located on the cell membrane of nerve cells and other cells, and they mediate the effect of serotonin and also of a broad range of pharmaceutical and hallucinogenic drugs. Except for the 5-HT_3 receptor, which is a ligand-gated ion channel, all other 5-HT receptors are G-protein coupled receptors that activate an intracellular second messenger cascade.

5-HT_1, 5-HT_5, 5-HT_6, and 5-HT_7 receptors are mainly located in the central nervous system (CNS), 5-HT_3 receptors are found in peripheral and central neurones, and 5-HT_2 and 5-HT_4 receptors are situated in both the central nervous system and peripherally.

The 5-HT_2 receptor group, which will be depicted in more detail, consists of three re-

ceptor subtypes that are found in the central nervous system and other parts of the body (Hoyer et al., 1994):

- **5-HT$_{2A}$ receptors** are located in vascular smooth muscles, platelets, lung, central nervous system, and in the gastrointestinal tract. This is the "classical" serotonin receptor, the first that was discovered in the 5-HT$_2$ receptor group – it was, therefore, also named the 5-HT$_2$ receptor for a long time, until the different receptors in this group were discovered. It is also the focus of this paper, so it will be discussed below.
- **5-HT$_{2B}$ receptors** are situated peripherally, as far as research has shown up to now.
- **5-HT$_{2C}$ receptors** are only found in the central nervous system, with a high density appearing in the choroid plexus (*plexus choroideus*), the area in the ventricles where the cerebrospinal fluid is produced.

The 5-HT$_{2A}$ receptor consists of 471 amino acids. The DNA coding for the receptor is located on locus 13q14–q21, that is on the long arm of chromosome 13, in the range of band 14 to band 21 (Ferrari et al., 2007).

It is found in various peripheral tissues and in many areas of the cortex – in the claustrum (a region that is connected to the visual cortex), some components of the limbic system (particularly the olfactory nuclei), and parts of the basal ganglia. The functions mediated by the receptor include contractile responses in many vascular smooth muscle preparations, contractile response in bronchial, uterine, and urinary smooth muscle, as well as platelet aggregation and increased capillary permeability. Central nervous effects include some behavioral effects and neuronal depolarization. Most of these were found in animal studies (Hoyer et al., 1994). Functions of the 5-HT$_{2A}$ receptor and the serotonin system generally related to eating will be discussed below, in section 3.2.

3.2. Serotonin Regulation and Eating Behavior

Serotonin regulation plays an important part in eating and eating behavior. It has been shown that spontaneous food intake and the termination of food intake correlates with intrahypothalamic release of serotonin. In the course of ingestion, a phaseal release of serotonin leads to satiation and consequently to the cessation of ingestion. Through this mechanism, the amount of consumed food is regulated.

The related area in the brain of the phaseal release of serotonin and therefore of appetite regulation is the suprachiasmatic nucleus (*nucleus suprachiasmaticus*) located in

the paraventricular and ventromedial hypothalamus – a structure that is also responsible for controlling endogenous circadian rhythms. Experiments with serotonin agonists and antagonists show that this region plays an important role in the regulation of the eating behavior. This part of the brain has an exceptionally high density of 5-HT_{2C} and 5-HT_{1B} receptors.

Furthermore, it has been shown in animal experiments that an increased availability of serotonin agonists (like dexfenfluramine, a pharmaceutical product that is used to treat obesity) in extracellular space of the brain leads to a reduction of the amount of food intake and also occasionally to a reduction of body weight. Inversely, a reduction of serotonin availability in cerebral extracellular space leads to an augmentation of food intake and, as the case may be, body weight (Baumgarten & Grozdanovic, 1996).

In a multicenter study conducted in nine European countries, the use of dexfenfluramine caused an increased weight loss for six months in obese patients, with the lower body weight maintained for a further six months (Guy-Grand et al., 1989). This suggests that the serotonergic system has an effect on the appetite system and on energy intake.

Other studies showed that pharmacological enhancement of 5-HT neurotransmission in both animals and humans generally leads to increased satiety. An exception to this is the 5-HT_{1A} receptor, who's stimulation, in animals, leads to the induction of feeding. This suggests that there are opposing components within the serotonin system (Brewerton, 1995).

Regarding the composition of food consumed, it becomes evident that the amount of carbohydrates in particular is influenced by the serotonergic system. This seems to be mediated especially by 5-HT_{2C} and 5-HT_{1B} receptors, as has been shown in studies with mutated mice that did not express the 5-HT_{2C} receptor: the mice showed a significant increase of body mass (13%) when they had unlimited access to food sources. This also suggests the pathogenic role of selective 5-HT receptor alteration in the development of eating disorders in humans. A compromised homeostasis of satiety could lead to uncontrolled and excessive food intake, potentially resulting in obesity (Baumgarten & Grozdanovic, 1996).

In addition, other studies show a reduction of fat intake. This has first been shown in animal studies that provided different diets which the animals could choose from. Serotonin admission in these experiments gave rise to a selective reduction of both carbohydrate and fat intake. Studies of the same kind were also performed on humans. Subjects that received dexfenfluramine reduced the intake of high carbohydrate food and, because of the nutrient composition of the food, also the consumption of fat.

Furthermore, in humans serotonergic drugs lead not only to a reduction of meal size, but also to a reduction of snacking (the elimination of some snacks from the eating repertoire) (Blundell, Lawton, & Halford, 1995).

Although the development of eating disorders of the type bulimia nervosa cannot, like any complex, multifactorial disorder, be explained by serotonergic dysregulation alone, it seems to play a significant role. Several studies show that severity of symptoms, especially regarding the frequency of binge eating attacks, correlates with the concentration of 5-HIAA (5-hydroxyindoleacetic acid) in bulimic patients. High frequency of eating attacks involved low levels of 5-HIAA in liquor (Baumgarten & Grozdanovic, 1996). Binging and vomiting has been shown to decrease serotonin synthesis in other studies (Kaye, Gwirtsman, & George, 1989), and a significant correlation of high frequencies of binging and low concentrations of serotonin in cerebrospinal fluid has been found (Jimerson, Lesem, Kaye, & Brewerton, 1992). A preliminary trial study showed a decrease in the frequency of bulimic symptoms in severely symptomatic patients who were treated with ondanserton, a 5-HT_3 receptor antagonist (Faris et al., 2000).

Concerning the symptoms of bulimia nervosa in general, insulin resistance (which might be present in both anorexia nervosa and bulimia nervosa) impairs the body's ability to produce serotonin from L-tryptophan (Brewerton, 1995). Furthermore, the symptom of compulsive exercising may be related to changes in serotonin metabolism. It has been shown that the selective serotonin reuptake inhibitor (SSRI) fluoxetine leads to a reduction in compulsive exercising (Patrick, 2002).

Another cardinal symptom of bulimia nervosa is the loss of control during the binge eating attacks. The serotonin metabolism is related to controlling punishment-suppressed behaviors. Serotonin enhancement leads to suppression of these "unwanted" behaviors, whereas serotonin attenuation leads to the release of those behaviors. Theoretically, a functional serotonin deficit in bulimic patients could contribute to the loss of control of typical "punishment-suppressed" bulimic behaviors, such as bingeing, vomiting, purging or other forms of socially unacceptable behaviors (Brewerton, 1995).

3.3. Serotonin Regulation and Anorexia Nervosa

Serotonin regulation has been shown to be a significant factor in eating disorders (Patrick, 2002). The function of serotonin regulation in relation to eating and satiety has already been discussed in section 3.2.

In anorexic patients, low levels of serotonin have been reported (Patrick, 2002). These

disturbances could be a consequence of dietary abnormalities, or they could be a premorbid trait. This is why subjects who have recovered from anorexia nervosa are also studied – persisting disturbances suggest that these are existent already before the onset of the disorder.

Several studies have shown alterations in the serotonin activity in the ill state, and also that these disturbances persist after recovery (Kaye et al., 2005; Kaye, Gendall, & Michael, 1998).

When underweight, anorexia nervosa patients also show a significant reduction in concentration of the serotonin metabolite 5-hydroxyindoleacetic acid (5-HIAA) in the cerebrospinal fluid (CSF) compared to healthy controls. These findings suggest a reduced serotonergic activity (which could also be secondary to a reduced availability of tryptophan due to reduced food intake) (Kaye et al., 2005).

Another study by Kaye, Gwirtsman, George, and Ebert (1991) reports elevated concentrations of 5-HIAA in cerebrospinal fluid after long-term recovery from anorexia nervosa, a finding that, at first sight, appears contradictory. It might be explained by the fact that dieting also lowers plasma tryptophan levels in healthy women. In anorexic patients, resumption of normal eating may unveil intrinsic abnormalities in the serotonergic system that might be responsible for core behavioral or temperamental risks (Kaye et al., 1998).

Some studies use platelets as a model for the serotonergic synapse in the brain. The platelet 5-HT_{2A} receptor has characteristics similar to those reported for the 5-HT_{2A} receptor in the human brain. In fact, the nucleotide sequence of the human platelet 5-HT_{2A} receptor is identical to that of the human frontal cortex 5-HT_{2A} receptor (except for one nucleotide, which does not alter the amino acid structure).

It could be shown that patients suffering from eating disorders (both anorexia nervosa and bulimia nervosa) show altered binding characteristics for the platelet 5-HT_{2A} receptor. Certain binding characteristics were elevated, indicating an enhanced receptor binding in anorexia and bulimia nervosa. This indicates that the 5-HT_{2A} receptor is in some way related to body weight and body weight regulation (Spigset, Andersen, Hägg, & Mjöndal, 1999).

Brain imaging studies using serotonin ligands offer the potential for understanding previously inaccessible brain serotonin neurotransmitter function. Results showed reduced 5-HT_{2A} receptor activity in cortical regions for subjects that are ill with, or have recovered from, anorexia nervosa (Kaye et al., 2005).

Another hint that the serotonin regulation is an important factor in the development of

anorexia nervosa is the fact that it is known to modulate several hormones which have been reported to be dysfunctional in anorexic patients. These include corticotropin releasing hormone (CRH) and related hormones, gonadotropins, arginine vasopressin (AVP) and prolactin (PRL) (Brewerton & Jimerson, 1996).

Symptoms of Anorexia Nervosa

Serotonin regulation is also involved with a number of symptoms and comorbidities typical for anorexia nervosa.

Eating disorders have a considerable relation to affective disorders, particularly depression, which also involve a reduction in serotonin function. Anorexic patients also have high frequencies of depressed mood and lifetime histories of major depressive disorders. It is possible that the impaired synaptic transmission in serotonin pathways may be a shared cause for both eating disorders and depression (Jimerson, Lesemb, Kaye, Heggd, & Brewerton, 1990).

Anxiety is another symptom commonly reported by eating disordered patients. They also show high frequencies of obsessional symptoms and higher rates of obsessive–compulsive disorder (OCD) and other anxiety disorders than would be expected from normal controls. Serotonin plays an important role in the modulation of anxiety. Both animal and human studies suggest that anxiety is generally increased with enhanced serotonin function. An exception to this is the 5-HT_{1A} receptor, as it seems to have an anxiolytic effect. The stimulation of this receptor also leads to the induction of feeding in animals, as was noted in section 3.2, which could link eating with anxiety reduction (Brewerton, 1995).

Another important trait that is typical of anorexic patients (and that is linked to anxiety) is harm avoidance. Subjects suffering from anorexia nervosa typically show high scores for this personality trait (Brewerton, Hand, & Bishop Jr., 1993), and harm avoidance is linked to the serotonin system: high scores on the harm avoidance dimension are related to an increased serotonergic activity (Hansenne & Ansseau, 1999; Melke et al., 2003).

Eating disordered patients commonly have a distorted view of their actual body shape (Kaye et al., 1998). They perceive themselves as chubby even when at the brink of starvation. Interestingly, psychedelic drugs such as lysergic acid diethylamide (LSD) – which structurally resembles the serotonin neurotransmitter – are known to produce mood-congruent alterations in the perception of self and body, including changes in body size.

Serotonin receptor dysfunction could therefore be involved in changes in body image and self perception in anorexia nervosa patients (Brewerton, 1995).

It should also be noted that the serotonin antagonist cyproheptadine decreased the "fear of becoming fat" in anorexic patients compared to a placebo group in a study by Goldberg et al. (1980).

It is a known fact that the majority of patients with anorexia nervosa are female. This asymmetry is often explained by social or perhaps hormonal factors. Another possible explanation could be differences in central serotonin regulation between the sexes. This issue has first been addressed in studies on rats, where female rats showed higher brain serotonin and 5-HIAA levels, a greater response to certain serotonin agonists, and a higher activity of tryptophan hydroxylase. Moreover, the serotonin system of female rats appears to be less adaptive to stress than that of male rats (Brewerton, 1995).

Some of these differences also have been found in humans. Like in the animal studies, female subjects show higher levels of 5-HIAA in the cerebrospinal fluid (CSF) and also higher responses (increase in serum prolactin level) to injection of L-tryptophan in comparison to male subjects (Heninger, Charney, & Sternberg, 1984). In addition, these prolactin responses to injection of L-tryptophan are markedly increased after three weeks of dieting in women, but not in men – a finding that suggests that dieting alters the serotonin function more in women than in men (Goodwin, Fairburn, & Cowen, 1987).

Medication Studies

The role of serotonin regulation in relation to anorexia nervosa can also be investigated with the help of medication studies. The reaction of individuals suffering from anorexia nervosa and the respective reaction of controls to serotonin agonists and antagonists can give insight into the function of serotonin in the disorder.

In contrast to bulimia nervosa, the role of medical treatment in anorexia nervosa is quite limited, especially in low-weight patients (Kaye et al., 2005).

Although selective serotonin reuptake inhibitors (SSRIs) seemed to be quite promising for the treatment of eating disorders, no consistent effect of weight gain could be shown (Ramoz et al., 2007).

The serotonin agonist cyproheptadine increases the rate of weight gain in anorexic patients, but the effect is only marginal (Halmi, Eckert, LaDu, & Cohen, 1986).

Given the high rate of relapse for eating disorders, it is only reasonable that a considerable number of studies focuses on this topic. It has been found in a double-blind, placebo-

controlled study that the SSRI fluoxetine, when given after weight restoration, significantly reduces the rate of relapse in anorexia nervosa patients and also facilitates weight gain.

The fact that selective serotonin reuptake inhibitors are not effective in malnourished and underweight anorexia nervosa patients might be explained by the reduced availability of tryptophan, a precursor of serotonin that is gathered from diet. SSRIs are dependent on neuronal release of serotonin, because they only impair the reuptake of the released neurotransmitter. Malnourished patients with anorexia nervosa show reduced 5-HIAA in cerebrospinal fluid, which suggests reduced synaptic serotonin availability.

As weight restoration normalizes nutrition, 5-HIAA concentration becomes elevated in anorexic patients. These changes in nutrition and serotonin activity could be the underlying cause for patients responding to fluoxetine after weight gain (Kaye et al., 1998).

Polymorphisms in the Serotonin System

There are various polymorphisms in the serotonergic system. For some of them, the research of their importance for eating disorders has only begun very recently. As this is not the main topic of this paper, this section does not attempt to give a comprehensive or exhaustive review of ongoing research in this area. Instead, only a few examples of genetic studies shall be given as an introduction to the field of research that this study is embedded in.

Genetic polymorphisms that have been studied in relation to anorexia nervosa are (amongst others) the following (Klump & Gobrogge, 2005):

- The Pro-279-Leu polymorphism 5-HT_7 receptor gene,
- the Phe-124-Cys polymorphism in the 5-$HT_{1D\beta}$ receptor gene,
- various other polymorphisms in the 5-HT_{1D} receptor gene,
- the Cys-23-Ser polymorphism in the 5-HT_{2C} receptor gene,
- the serotonin transporter (5-HTT) gene linked polymorphism (5-HTTLPR; long and short alleles),
- the 1095 T/C polymorphism in the tryptophan hydroxylase (TPH) gene,
- the –1438 G/A polymorphism in the promoter region of the 5-HT_{2A} receptor gene.

Most of the studies conducted to research the influence of these genes in their role related to anorexia nervosa reported negative findings. No associations between allele frequency and anorexia nervosa have been found for the 5-HT$_7$, 5-HT$_{1D\beta}$, 5-HTT, or the TPH gene.

It bears noting that there are different ways to assess the influence of genetic polymorphisms on psychiatric (or other) disorders. The association studies mentioned here are only one possibility. Other possibilities, such as gene–environment interaction studies (G × E studies), are explained in section 2.4.

For the 5-HT$_{2C}$ receptor gene, initial studies draw a more promising picture. Two of four studies have found an increased frequency of the ser23 allele for individuals suffering from anorexia nervosa in relation to controls. That could suggest an influence of this polymorphism in the development of this disorder (Klump & Gobrogge, 2005).

The most promising results for genes in the serotonin system have been found for the 5-HT$_{2A}$ receptor gene. These will be outlined in section 3.4.2, after a short introduction to the genetics of the 5-HT$_{2A}$ receptor.

3.4. The 5-HT$_{2A}$ Gene and Anorexia Nervosa

3.4.1. The 5-HT$_{2A}$ Gene and its Polymorphisms

The 5-HT$_{2A}$ Gene

The 5-hydroxytryptamine type 2A receptor gene (HTR2A) was one of the first human serotonin receptor genes to be cloned. It is located on the long arm of chromosome 13, in the bands 14 to 21 (locus 13q14–21). The gene location is depicted in figure 3.3. The gene includes three exons and two introns, spanning 20 kilobases (kb) (Parsons, D'Souza, Arranz, Kerwin, & Makoff, 2004).

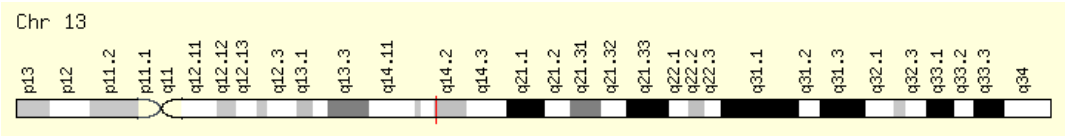

Figure 3.3.: *Chromosome 13. (Source: http://www.genecards.org/cgi-bin/carddisp.pl?id=5293&id_type=-hgnc&search=5293).*

Polymorphisms in the 5-HT$_{2A}$ Gene

There are six polymorphisms related to the 5-HT$_{2A}$ gene that have been targets of research to present (Jönsson et al., 2001):

- Thr-25-Asn polymorphism (amino acid substitution)
- His-452-Tyr polymorphism (amino acid substitution)
- Ala-447-Val polymorphism (amino acid substitution)
- 102 T/C polymorphism (nucleotide exchange)
- 516 C/T polymorphism (nucleotide exchange)
- –1438 G/A polymorphism (nucleotide exchange)

The three amino acid substitutions, Thr-25-Asn, His-452-Tyr, and Ala-447-Val, have not yielded any significant results in the field of eating disorders, nor have the two nucleotide exchange polymorphisms, 102 T/C and 516 C/T. The latter two polymorphisms are so-called silent mutations, which do not lead to an alteration of the amino acid sequence in protein synthesis (Hinney et al., 2000).

In contrast to these polymorphisms, the –1438 G/A polymorphism is not a mutation of the 5-HT_{2A} gene itself – it is located in the promoter region of that gene. Numerous studies have been performed to evaluate the role of this polymorphism in eating disorders and other mental diseases. Important results will be depicted in section 3.4.2.

The 5-HT_{2A} –1438 G/A Polymorphism

Detection of the –1438 G/A polymorphism was first published by Spurlock et al. (1998). They screened the promoter region of the 5-HT_{2A} gene and detected an A to G polymorphism at base position –1438. The base adenine was substituted by the base guanine. They also found that this newly detected polymorphism was in complete linkage disequilibrium with the 102 T/C polymorphism. All subjects in the study with the –1438 A allele were found to have the 102 T allele, while those with the –1438 G allele were found to have the 102 C allele.

Functionality of the 5-HT_{2A} –1438 G/A Polymorphism

The initial study of Spurlock et al. (1998) did not find any significant differences in basal activity or promoter activity of the –1438 A and G alleles. The polymorphism does not alter the expression or structure of the 5-HT_{2A} receptor protein.

Nacmias et al. (1999) first suggested that the polymorphism might modify the transcription of the gene. Parsons et al. (2004) investigated the functionality of the –1438 G/A polymorphism in greater detail. They found that promoter activity was significantly

higher in the presence of the –1438 A allele when a specific enhancer (SV40) was present. They concluded that the –1438 G/A polymorphism does enhance promoter activity, under certain conditions.

To date, this issue is still in discussion and could not be clarified satisfactorily.

3.4.2. The 5-HT$_{2A}$ –1438 G/A Polymorphism in Eating Behavior and Anorexia Nervosa

In the following, the results of different studies concerning the 5-HT$_{2A}$ –1438 G/A polymorphism in relation to eating behavior and anorexia nervosa will be reported. It must always be kept in mind that in complex diseases like anorexia nervosa, a single mutation or polymorphism can only confer a small effect to predisposition to the disease. The risk of individuals carrying a specific allele will only be slightly elevated (Hinney et al., 2000).

The 5-HT$_{2A}$ –1438 G/A Polymorphism in Eating Behavior in General

The –1438 A allele is related to food and energy intake. This has been shown in a study of overweight adults, were the A allele was associated with lower energy intake. This was due to the lower intakes of all main nutrients. The effects of the A allele were not related to differences in body mass index or biological variables, and it is important to mention that the 5-HT$_{2A}$ –1438 G/A polymorphism is not associated with obesity (Aubert, Betoulle, Herbeth, Siest, & Fumeron, 2000).

These results have also been found in children and adolescents that were not overweight. The –1438 A allele was significantly associated with lower energy intake. This was mainly due to lower ingestion of total, monounsaturated, and saturated fat (both measured in g/d and as percentage of daily energy levels) (Herbeth et al., 2005).

The 5-HT$_{2A}$ –1438 G/A Polymorphism in Anorexia Nervosa

The first study to report an association of the 5-HT$_{2A}$ –1438 G/A polymorphism and anorexia nervosa was performed by Collier et al. (1997). They performed a study on 81 affected patients of British origin and 226 healthy white controls and found an increased frequency of the –1438 A allele in anorexic patients. They also reported a significantly increased frequency of the AA genotype and concluded that the 5-HT$_{2A}$ gene might be a susceptibility factor for anorexia nervosa. The reported frequencies in anorexia nervosa patients were 0.51 for the –1438 A allele and 0.31 for the AA genotype (controls: 0.41, 0.15).

Two subsequent studies (Hinney, Ziegler, Nothen, Remschmidt, & Hebebrand, 1997;

Campbell, Sundaramurthy, Markham, & Pieri, 1998) failed to replicate this association. It has to be noted that the first of these two studies did not use a control group, and the study by Campbell et al. observed a very high frequency of the AA genotype in the control group.

Sorbi et al. (1998) performed another study, this time replicating the results of the initial study. They used a sample of 77 patients with anorexia nervosa and 107 controls, both of Italian origin. They found an elevated frequency of the –1438 A allele and the AA genotype in anorexia nervosa patients, the reported frequencies being 0.565 for the A allele and 0.299 for the AA genotype. In their study, they also analyzed the association of the –1438 G/A polymorphism with different subgroups of anorexia nervosa patients. They found that both A allele frequency and AA genotype frequency were increased in particular in subjects with anorexia nervosa of the restricting subtype (A allele frequency 0.662, AA genotype frequency 0.419) and suggested that restricting and purging subtypes of anorexia nervosa have different involvement of the 5-HT_{2A} –1438 G/A polymorphism.

More studies were performed, some supporting an association of the –1438 G/A polymorphism with anorexia nervosa, others not supporting it. A meta-analysis of nine studies (four in favor of the association, five not supporting this hypothesis) found a significantly increased frequency of the –1438 A allele in patients with anorexia nervosa (reported A allele frequency 0.468 in patients, 0.436 in controls; AA genotype frequency not reported). All of the studies mentioned above were included in the meta-analysis, and it was stated that the inter-study heterogeneity was high. Also, the study reported a large variation of the –1438 A allele frequency in controls (0.36 to 0.54) and admitted that the diversity of inclusion criteria and the differences in clinical characteristics of the samples (comorbidity, age of onset, ...) could contaminate the impact of the 5-HT_{2A} –1438 G/A polymorphism (Gorwood et al., 2003). Furthermore, this study only investigated the association of the –1438 G/A polymorphism with the whole anorexia nervosa group, not distinguishing between restricting and purging subtype.

One of the studies (Kipman et al., 2002) included in this meta-analysis also investigated the transmission of the –1438 A allele from parents to their offspring (via transmission disequilibrium test). In a sample of French origin, they did not find an excess of transmission of the A allele, nor did they find any association of the A allele with anorexia nervosa. Furthermore, the allele frequencies between the restricting and purging subtype did not significantly differ from each other. The study also did not succeed in finding an association of the A allele with lifetime severity of anorexia nervosa (assessed by worst lifetime Body Mass Index).

However, the results of the study indicated a relation of the A allele to a later age of onset.

Patients with the A allele had an older age of onset (mean = 15.28) than patients without this allele (mean = 14.01). Although this association was rather weak, the conclusion of the study was that the 5-HT$_{2A}$ –1438 A allele might act as a modifying rather than global susceptibility factor, delaying the onset rather than properly increasing the risk for the disorder.

A study by Gorwood et al. (2002) also used the transmission disequilibrium test in a relatively big sample of 316 trios collected in six European centers. As mentioned in section 2.4.6, this method is not vulnerable to report false positive findings that result from stratification bias. Like the study by Kipman et al. (2002) (that used the same test), no association of the –1438 A allele or AA genotype could be found in this sample. Furthermore, no differences in age of onset and minimum adult body mass index were found in *post hoc* analyses when comparing patients with an A allele to patients that are homozygous for the G allele.

In more recent studies, the association could be replicated at least in part. Ricca et al. (2004) performed a study on a sample of 148 eating disordered patients (anorexia nervosa and bulimia nervosa) and 89 controls recruited in Italy. They found that patients with some kind of eating disorder had higher frequencies of both A allele and AA genotype (A allele frequency in total eating disorder sample 0.524, AA genotype frequency in this group 0.216; controls: 0.358, 0.101). The study also reported significant differences in the frequency of the –1438 A allele and the AA genotype for the anorexia nervosa restricting subgroup in comparison to controls, A allele frequency being 0.554, AA genotype frequency 0.297. No significant differences were found when the anorexia nervosa purging subgroup was compared to controls.

Rybakowski et al. (2006) analyzed the association of the 5-HT$_{2A}$ –1438 G/A polymorphism with anorexia nervosa in a Polish sample of 132 female adolescent patients suffering from anorexia nervosa and 93 healthy controls. They could not find a significant effect, but they reported a statistical trend towards the association of the A allele with anorexia nervosa. The A allele frequency in this sample was 0.649, the AA genotype frequency 0.42 in anorexia nervosa patients, 0.567 and 0.292 in healthy controls. The A allele and AA genotype frequencies in the restricting subgroup were slightly higher (A allele 0.661, AA genotype 0.448), but no separate statistical test was performed to assess the significance of these frequencies in comparison to the control subjects.

The 5-HT_{2A} –1438 G/A Polymorphism in Comorbidities and Traits related to Anorexia Nervosa

A different and less restrictive approach of investigating the possible relation of the 5-HT_{2A} –1438 G/A polymorphism to anorexia nervosa is to take a look at personality traits and comorbidities involved in this disorder. Since certain traits or comorbidities are quite frequent, it can be assumed that the susceptibility genes do not convey a higher susceptibility to anorexia nervosa in particular, but rather that they are responsible for certain character attributes which might be risk factors for disorders that share some of their features.

This was first stated in a study by Enoch et al. (1998). They adduced that anorexia nervosa and obsessive–compulsive disorder are heritable and often comorbid disorders that share several personality traits, including harm avoidance, perfectionism, and obsessionality. The study replicated the association of the 5-HT_{2A} –1438 A allele and AA genotype with anorexia nervosa and extended it to obsessive compulsive disorder, which was (only) associated with a higher frequency of the A allele, not the AA genotype in this study. As a conclusion the study suggests the possibility that the 5-HT_{2A} –1438 G/A promoter polymorphism may contribute to a behavioral trait that is common to both anorexia nervosa and obsessive–compulsive disorder, like one of the traits mentioned above.

This was replicated in a study by Walitza et al. (2002), who not only found an increased frequency of the –1438 A allele in patients suffering from obsessive–compulsive disorder, but also an elevated AA genotype frequency. They also adduced traits that frequently occur in both anorexia nervosa and obsessive–compulsive disorder as a possible explanation for the association of the –1438 G/A polymorphism with the two disorders, explicitly mentioning behavioral inhibition, high harm avoidance and obsessive concerns with exactness and perfectionism.

A study by Ricca et al. (2004) investigated eating disorder pathology by the means of a standardized interview, the Eating Disorder Examination (EDE). Results showed that the pathology differed for the three genotypes (AA, AG, and GG). Subjects with the –1438 AA genotype showed higher weight concern (EDE 3 subscale) compared to subjects with GA and GG genotype, and they showed higher shape concern (EDE 4 subscale) compared to subjects with GG genotype. The overall severity of eating disorder psychopathology (EDE total score) was also found to be higher in subjects with AA genotype compared to subjects with the GG genotype. These results were independent of the diagnostic group (anorexia nervosa restricting and purging type, and bulimia nervosa).

Personality traits related to anorexia nervosa were assessed directly in a study by Rybakowski

et al. (2006). This study found that the –1438 AA genotype is associated with lower reward dependence (compared to the GG genotype). It is stated that subjects who score low in this dimension are characterized as emotionally withdrawn, socially aloof and not sentimental, and previous research has shown that low reward dependence is characteristic of restrictive-type anorexia nervosa subjects.

The study also found the –1438 AG genotype to be associated with higher harm avoidance (compared to the AA genotype).

The findings presented in this chapter indicate a possible role of the 5-HT_{2A} receptor and the –1438 G/A polymorphism in anorexia nervosa, although the results of the studies are somewhat conflicting. It has to be noted that, at present, there are no gene–environment interaction studies for the 5-HT_{2A} gene or the –1438 G/A polymorphism and anorexia nervosa.

4. Methods

This chapter will outline the instruments for data acquisition and for statistical analysis. Furthermore, the genotyping procedure will be delineated.

4.1. Study Design

A case–control design was used (matched 1–1 study) in this study, comparing sister pairs that are discordant for anorexia nervosa – the anorexia nervosa patient being the case, her sister serving as the control.

4.2. Assessment of Risk Factors and other Variables

The environmental risk factors were assessed with the Oxford Risk Factor Interview for Eating Disorders (ORFI). It was developed and first used by Fairburn, Welch, Doll, Davies, and O'Connor (1997).

The Oxford Risk Factor Interview for Eating Disorders is a semi-structured investigator-based interview designed to examine the specific risk factors associated with an eating disorder. It consists of 37 items. The items are coded dichotomously by the investigator – they can either be considered a risk factor (1) or not (0). The questions can be classified in five environmental risk domains:

1. Parental problems: This subscale includes low parental contact, separation from parents, arguments with parents, parental criticism, parental high expectations, parental overinvolvement or underinvolvement, parental minimal affection, parental control, family conflict avoidance, and family social situation.

2. Disruptive events: This domain includes parental death, death of a sibling or another close person, parental divorce, parental chronic illness, frequent house moves, critical life events in the year before onset, sexual and physical abuse.

3. Parental psychiatric disorder: This part of the interview asks for distress by parental mood disorder, parental anxiety disorder, and parental substance-related disorder

(all before the onset of the eating disorder).

4. Interpersonal problems: This subscale includes questions about teasing (not related to shape, weight, eating, or appearance), bullying, not having close friends, not having male friends, missing preparations for menarche, and teasing about breast or breast development.

5. Family dieting environment: This domain covers dieting with a family member before onset of anorexia nervosa, repeated critical comments about shape, weight, or eating (by family or others), parental eating disorder or obesity, family fitness influences, teasing about shape, weight, or appearance, and eating disorder among acquaintances.

The interview starts with establishing a time line with index age (the age of onset of the eating disorder). This ensures that the variable of interest preceded the outcome (Kazdin, Kraemer, Kessler, Kupfer, & Offord, 1997).

It has to be noted that for this study, it has not been assessed whether these risk factors were present in the environment of the interrogated subject. Rather, the questions were phrased such that the *subjective distress* caused by a certain environmental factor was recorded (e.g. "Were you *distressed by* low parental contact?"). The interviewer judged whether the event or circumstance in question was distressing for the patient. This ensured that it was not the presence of a risk factor that was assessed, but rather the subjective distress that that risk factor caused. (This is especially important because the study consisted of sib-pairs. It is, of course, highly probable that the present risk factors for the siblings are almost identical. The *subjective impact* of these risk factors, on the other hand, can be different for the two siblings.)

In addition to the environmental risk factors described above, the following variables were recorded:

- age of patient and control
- age of onset
- life-time diagnosis (either anorexia nervosa restricting subtype or binge-eating/purging subtype)
- Body Mass Index (current, lowest and highest)
- current employment status

- highest education level and age (at the time of birth of patient) of mother and father
- age and sex of siblings

4.3. Genotyping

Blood samples and cheek cell swabs were collected from patients and their sisters. The genomic DNA was prepared according to procedures described by Freeman et al. (1997).

The sequence of the forward primer Pro2F is 59-CTA GCC ACC CTG AGC CTA TG-39; the sequence of the reverse primer Pro2R is 59-TTG TGC AGA TTC CCA TTA AGG-39. The polymerase chain reaction (PCR) program consisted of 30 cycles of 95°C for 30 seconds, 59°C for 30 seconds, and 72°C for 30 seconds. The resulting PCR product was then digested with *MspI* restriction endonuclease. After digestion, the products were separated on agarose gel by electrophoresis. An A at position –1438 leads to an uncut fragment of 200 base pairs (bp), a G at position –1438 leads to two fragments of length 121 bp and 79 bp (Masellis et al., 1998).

4.4. Statistical Analysis

For statistical analysis, regression models have been used. Conditional Regression has become a standard method to predict dichotomous outcomes (like disease status: anorexia nervosa vs. no eating disorder). As this study uses a matched sample, the method of choice is conditional logistic regression.

To facilitate the understanding of the model, this part starts with an explanation of the standard logistic regression analysis, followed by a closer look at conditional logistic regression model in section 4.4.3.

As an alternative, a data-mining algorithm was applied. This method, the odds ratio multifactor dimensionality reduction, was designed to identify interactions that predict a dichotomous outcome. It is described in section 4.4.4.

Multiple linear regression models are also used in the analysis. They serve to identify genetic and environmental effects on continuous outcome variables (like body mass index and age of onset). Since this method is relatively well known, it will not be described here.

4.4.1. Logistic Regression

To start with, the method will be outlined with only a single independent variable. Subsequently, the model of the multiple logistic regression will be explained.

The Logistic Regression Model

It is the intent of the logistic regression model to explain a single dichotomous outcome variable y dependent on a single independent variable x. In this context, the independent variable is called the predictor.

The logistic model is defined by

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{4.1}$$

with

- $\pi(x)$ being the expected value of the outcome at x, in mathematical terms $E(Y|x)$
- β_0 being the intercept of the model and
- β_1 being the slope parameter of predictor x.

The application of the so-called *logit-transformation* results in

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x \tag{4.2}$$

which is called the *logit*. It has several desirable properties of the linear regression model. The logit $g(x)$ is linear in its parameters, it can be continuous and can have a range from $-\infty$ to $+\infty$.

However, there are differences to the linear regression model. In the logistic regression model, an observation of the outcome variable may be expressed as

$$y = E(Y|x) + \epsilon$$

or

$$y = \pi(x) + \epsilon$$

with

- ϵ being the error (the observation's deviation from the conditional mean).

Opposed to the linear regression model, ϵ is not normally distributed, because it can only assume two possible values. For $y = 1$, the error results to $\epsilon = 1 - \pi(x)$, for $y = 0$ it is $\epsilon = -\pi(x)$. So the distribution of ϵ is a binomial distribution with the variance of $\pi(x)(1-\pi(x))$.

Maximum Likelihood Estimation of Coefficients

In the linear regression model, the method of least squares is often used for estimating the unknown parameters of the model. The values for β_0 and β_1 are chosen so that the sum of squared deviations of the observed values of Y from the predicted values is a minimum.

For the linear model, those estimates have several desirable properties. This is not the case in the logistic model. For this reason, the more general method of *maximum likelihood* is chosen. To apply this method, a likelihood function has to be constructed. This function expresses the probability of the observed data as a function of the unknown model parameters (Hosmer & Lemeshow, 2000). The intention of the maximum likelihood method is to find those model parameters that maximize the likelihood. The result is a set of parameters that ensure that the model describes the observed data most closely.

If the outcome variable Y is coded as 0 or 1, the expression for $\pi(x)$ in equation (4.1) yields the conditional probability that Y is equal to 1, given x and an arbitrary value of $\boldsymbol{\beta} = (\beta_0, \beta_1)$. This will be denoted as $P(Y = 1|x)$. As a consequence, the expression $1 - \pi(x)$ gives the conditional probability that Y equals 0 given x, $P(y = 0|x)$.

The contribution to the likelihood function for the pair (x_i, y_i) can be expressed as

$$\pi(x_i)^{y_i} \left(1 - \pi(x_i)\right)^{1-y_i}. \tag{4.3}$$

The observations of the i subjects are assumed to be independent, so the likelihood function for the whole dataset is the product of the likelihoods of all subjects ($i = 1 \ldots n$),

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left(1 - \pi(x_i)\right)^{1-y_i} \tag{4.4}$$

with

$\boldsymbol{\beta}$ as the vector of parameters, (β_0, β_1).

We now need to find those values of $\boldsymbol{\beta}$ that maximize the value of the likelihood function in (4.4). This is accomplished by differentiating the equation with respect to the unknown parameters, β_0 and β_1. As this is mathematically complex for a term consisting of products, we work with the logarithm of equation (4.4). This is called the *log likelihood*:

$$L(\boldsymbol{\beta}) = \ln l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \ln(\pi(x_i)) + (1 - y_i) \ln\left(1 - \pi(x_i)\right)\right) \tag{4.5}$$

After differentiation of (4.5) with respect to β_0 and β_1, we set the resulting equations to zero to obtain the values of β_0 and β_1 that yield the greatest likelihood. Those equations are called the *likelihood equations*:

$$\sum (y_i - \pi(x_i)) = 0 \tag{4.6}$$

and

$$\sum x_i (y_i - \pi(x_i)) = 0 \quad . \tag{4.7}$$

These equations have to be solved by means of iterative methods, since their solutions are nonlinear in β_0 and β_1.

The resulting values of equations (4.6) and (4.7) are the *maximum likelihood estimates*, denoted as $\hat{\boldsymbol{\beta}}$.

4.4.2. Multiple Logistic Regression

The Multiple Logistic Regression Model

In the case of more than one predictor, $\boldsymbol{x}$ is a vector with p elements, one element per predictor. The conditional probability of the outcome is $P(Y = 1|\boldsymbol{x} = \pi(\boldsymbol{x}))$.

The logit of the multiple regression model is therefore

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p \tag{4.8}$$

and the regression model itself is defined by

$$\pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{1 + e^{g(\boldsymbol{x})}} \quad . \tag{4.9}$$

Equation (4.9) is the model specification for continuous independent variables. If the model includes predictors on an ordinal scale, dummy variables have to be used. For a predictor x_j with k_j levels, $k_j - 1$ dummy variables are required. They are denoted as D_{jl} with $l = 1 \ldots k_j - 1$ (Hosmer & Lemeshow, 2000).

The logit for p predictors, with the j^{th} being discrete (ordinal), would be written as

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{l=1}^{k_j - 1} \beta_{jl} D_{jl} + \cdots + \beta_p x_p \quad . \tag{4.10}$$

In the following section, the summation and double subscripting for design variables will be omitted when possible to improve readability.

Maximum Likelihood Estimation

For the estimation of the parameters of the model, we have a sample of n independent observations $(\boldsymbol{x}_i, y_i)$ with $i = 1 \ldots n$. By means of the maximum likelihood method, the unknown coefficient vector $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$ should be estimated. The likelihood equation is the same as for the univariate logistic regression model

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i} \left(1 - \pi(\boldsymbol{x}_i)\right)^{1-y_i} \tag{4.11}$$

with $\pi(\boldsymbol{x})$ defined as in equation (4.9). Analogous to the univariate model, the logarithm of the likelihood function will be differentiated with respect to the $p+1$ coefficients, resulting in $p + 1$ likelihood equations:

$$\sum \left(y_i - \pi(\boldsymbol{x}_i)\right) = 0$$

and

$$\sum x_i \left(y_i - \pi(\boldsymbol{x}_i)\right) = 0$$

for $j = 1 \ldots p$. These equations have to be solved iteratively, usually using special software. The solution of these equations are the estimated coefficients $\hat{\boldsymbol{\beta}}$.

4.4.3. Conditional Logistic Regression

The use of the logistic regression model is not limited to simple random samples. It can also be used in cohort studies and in case-control studies, when the likelihood function is adjusted accordingly.

Case-Control Studies

Case-control studies are studies where the sampling is performed conditional on the outcome variable. In fact, two samples are drawn; one out of a population of cases, the other out of a population of controls.

By means of a logistic regression model, adjusted odds-ratios can be obtained from the estimated slope coefficients. The fact that the odds ratio is invariant under different study designs (cohort or case-control) has first been noted by Cornfield (1951). The mathematical base for this was demonstrated later, by Farewell (1979) and Prentice and Pyke (1979).

Unlike in conventional logistic regression, the binary outcome variable in case-control studies is fixed by stratification. All the cases have outcome $y = 1$, while all the controls

have outcome $y = 0$. The dependent variables in this case are the covariates $\boldsymbol{x}$. So the aim of the logistic regression is not to model the probability of the outcome, but to model the probability of the covariates $\boldsymbol{x}$ given a predefined outcome. The likelihood function is the product of the stratum-specific likelihood functions, which depend on

- the probability that the subject was selected for the sample, and
- the probability distribution of the covariates (Hosmer & Lemeshow, 2000).

In order to construct the likelihood function, we introduce a variable that specifies whether a subject of the population was selected for the sample ($s = 1$) or not ($s = 0$).

The full likelihood for a sample of n_1 cases (outcome $y = 1$) and n_0 controls (outcome $y = 0$) is

$$\prod_{i=1}^{n_1} P\left(\boldsymbol{x}_i | y_i = 1, s_i = 1\right) \prod_{i=1}^{n_0} P\left(\boldsymbol{x}_i | y_i = 0, s_i = 1\right) \quad . \tag{4.12}$$

The key in developing a likelihood function for case-control studies is the application of Bayes theorem. For an individual term in the likelihood function shown in equation (4.12), the application of the theorem results in

$$P(\boldsymbol{x}|y, s = 1) = \frac{P(y|\boldsymbol{x}, s = 1)P(\boldsymbol{x}|s = 1)}{P(y|s = 1)} \quad . \tag{4.13}$$

Bayes theorem is then used on the first term in the numerator in equation (4.13), which yields for $y = 1$

$$P(y = 1|\boldsymbol{x}, s = 1) = \frac{P(y = 1|\boldsymbol{x})P(s = 1|\boldsymbol{x}, y = 1)}{P(y = 0|\boldsymbol{x})P(s = 1|\boldsymbol{x}, y = 0) + P(y = 1|\boldsymbol{x})P(s = 1|\boldsymbol{x}, y = 1)} \quad . \tag{4.14}$$

An important assumption is that the selection of cases and controls is independent of the covariate values. Let τ_1 be the probability for a case to be selected for the sample and τ_0 the probability of selection of the control:

$$\tau_1 = P(s = 1|y = 1, \boldsymbol{x}) = P(s = 1|y = 1)$$

$$\tau_0 = P(s = 1|y = 0, \boldsymbol{x}) = P(s = 1|y = 0)$$

These probabilities, and also the logistic regression model, $\pi(\boldsymbol{x})$, are now substituted into equation (4.14), leading to

$$P(y=1|\boldsymbol{x},s=1)=\frac{\tau_1\pi(\boldsymbol{x})}{\tau_0(1-\pi(\boldsymbol{x}))+\tau_1\pi(\boldsymbol{x})} \quad . \tag{4.15}$$

If both the numerator and denominator of the expression on the right side of equation (4.15) are divided by $\tau_0(1-\pi(\boldsymbol{x}))$, the result is a logistic regression model with intercept $\beta_0^*=\ln(\tau_1/\tau_0)+\beta_0$.

To simplify notation we introduce

$$\pi^*(\boldsymbol{x})=P(y=1|\boldsymbol{x},s=1)=\frac{\tau_1\pi(\boldsymbol{x})}{\tau_0(1-\pi(\boldsymbol{x}))+\tau_1\pi(\boldsymbol{x})} \quad .$$

Because of the assumption that sampling is carried out independent of covariate values, it results that $P(\boldsymbol{x}|s=1)=P(\boldsymbol{x})$, with $P(\boldsymbol{x})$ denoting the probability distribution of the covariates (Hosmer & Lemeshow, 2000). The general term in the likelihood of equation (4.13) then becomes, for $y=1$

$$P(\boldsymbol{x}|y=1,s=1)=\frac{\pi^*(\boldsymbol{x})P(\boldsymbol{x})}{P(y=1|s=1)} \tag{4.16}$$

and for $y=0$

$$P(\boldsymbol{x}|y=0,s=1)=\frac{(1-\pi^*(\boldsymbol{x}))P(\boldsymbol{x})}{P(y=0|s=1)} \quad . \tag{4.17}$$

We can now express the likelihood for cases and controls with

$$L^*(\boldsymbol{\beta})=\prod_{i=1}^{n}\pi^*(\boldsymbol{x}_i)^{y_i}(1-\pi^*(\boldsymbol{x}_i))^{1-y_i} \tag{4.18}$$

The likelihood function of the full sample shown in equation (4.12) now becomes

$$L^*(\boldsymbol{\beta})\cdot\prod_{i=1}^{n}\left(\frac{P(\boldsymbol{x}_i)}{P(y_i|s_i=1)}\right) \quad . \tag{4.19}$$

As can easily be seen, the first term in equation (4.19) is the same as for a non-stratified sample, with the outcome of interest being modeled as the dependent variable. Assuming that the probability distribution of the covariates $\boldsymbol{x}$, $P(\boldsymbol{x})$, does not contain information about the coefficients in the logistic regression model, maximizing the full likelihood is only restricted by $P(y_i=1|s_i=1)=n_1/n$ and $P(y_i=0|s_i=1)=n_0/n$. The likelihood equations obtained by differentiating, with respect to the parameter β_0^*, satisfy this condition. As a consequence, the analysis of data from case-control studies via logistic regression can be performed in the same way as for conventional studies.

The maximum likelihood estimators have the same properties as in conventional samples:

they are asymptotically normally distributed, with their covariance matrix obtained from the inverse of the information matrix (Hosmer & Lemeshow, 2000).

The Matched Study

The matched study is a special case of a stratified case-control study. Subjects are stratified on variables that are thought to be associated with the outcome (e. g. sex or age). Within each stratum, samples of cases (outcome $y = 1$) and controls (outcome $y = 0$) are selected.

Generally, it is not necessary for the number of cases and controls within each stratum to be constant, but a typical design is the 1–m matched study, in which for each case there are m controls (Hosmer & Lemeshow, 2000).

Another typical design is the 1–1 matched design. Here, the subjects are grouped in pairs, with one subject being the case and the other being the control. One way to conduct this stratification is to select pairs of genetically related subjects, with one sibling being the case ($y = 1$) and the other being the control ($y = 0$). In this case, there are only two subjects per stratum, the case and the control. This design is used in this paper. It is explained in greater detail in section 4.4.3.

Assuming that we would have n case-control pairs (and therefore n strata) and p covariates, the number of parameters to be estimated by the maximum likelihood method would be $n + p$, consisting of the constant term, p slope coefficients for the covariates and $n - 1$ coefficients for the stratum-specific design variables. The sample size would be $2n$, so the number of parameters to estimate would increase with the sample size. This would lead to biased maximum likelihood estimates (Breslow & Day, 1980).

A solution for this problem is to consider the $n - 1$ stratum-specific parameters as nuisance parameters that need not be estimated. In this case it is possible to use methods of conditional inference to create a maximum likelihood function who's estimators of the slope coefficients in the logistic regression model are consistent and asymptotically normally distributed.

The derivation of the conditional likelihood function is illustrated in section 4.4.3.

Stratum-Specific Logistic Regression Model

The stratum-specific logistic regression model for K strata with n_{1k} cases and n_{0k} controls in stratum k ($k = 1 \dots K$) is

$$\pi_k(\boldsymbol{x}) = \frac{e^{\alpha_k + \boldsymbol{\beta}'\boldsymbol{x}}}{1 + e^{\alpha_k + \boldsymbol{\beta}'\boldsymbol{x}}} \tag{4.20}$$

with

α_k as contribution to the logit of all terms constant within the k^{th} stratum (e. g. the matching or stratification variables) and

$\boldsymbol{\beta}$ as vector of the p slope coefficients, $\boldsymbol{\beta'} = (\beta_1, \beta_2, \ldots, \beta_p)$ without an intercept.

The interpretation of the p slope coefficients is the same as in any other logistic regression model. Each coefficient represents the change in log-odds for a one-unit-increase in the covariate holding all other covariates constant in every stratum (Hosmer & Lemeshow, 2000).

Conditional Likelihood

For the 1–m matched case study, the conditional likelihood for the k^{th} stratum is dependent on the total number of subjects in the stratum $n_k = n_{1k} + n_{0k}$ and the total number of cases in the stratum n_{1k}. The question is how many possibilities there are to assign n_{1k} persons as cases out of the n_k subjects in the stratum. The number of possible assignments c_k is given by

$$c_k = \binom{n_k}{n_{1k}} = \frac{n_k!}{n_{1k}!(n_k - n_{1k})!} \quad . \tag{4.21}$$

To specify the stratum-specific likelihood, the subjects 1 to n_{1k} are the cases ($y = 1$) and the subjects $n_{1k} + 1$ to n_k are the controls. The stratum-specific likelihood function can then be written as

$$l_k(\boldsymbol{\beta}) = \frac{\prod_{i=1}^{n_{1k}} P(\boldsymbol{x}_i|y_i = 1) \prod_{i=n_{1k}+1}^{n_k} P(\boldsymbol{x}_i|y_i = 0)}{\sum_{j=1}^{c_k} \left(\prod_{i_j=1}^{n_{1k}} P(\boldsymbol{x}_{ji_j}|y_{i_j} = 1) \prod_{i_j=n_{1k}+1}^{n_k} P(\boldsymbol{x}_{ji_j}|y_{i_j} = 0) \right)} \quad . \tag{4.22}$$

Here, the index i is used for the observed data, and the index i_j is used for the j^{th} possibility out of the c_k assignments.

The full conditional likelihood is given by

$$l(\boldsymbol{\beta}) = \prod_{k=1}^{K} l_k(\boldsymbol{\beta}) \tag{4.23}$$

which is the product over the stratum-specific likelihoods. This is legitimate due to their independence.

Application of Bayes' theorem to each $P(\boldsymbol{x}|y)$ term in (4.22) yields

$$l_k(\boldsymbol{\beta}) = \frac{\prod_{i=1}^{n_{1k}} e^{\boldsymbol{\beta}'\boldsymbol{x}_i}}{\sum_{j=1}^{c_k} \prod_{i_j=1}^{n_{1k}} e^{\boldsymbol{\beta}'\boldsymbol{x}_{ji_j}}} \quad . \tag{4.24}$$

The 1–1 Matched Study

For the 1–1 matched study with only two subjects in each stratum, the stratum-specific likelihood can be simplified to the expression

$$l_k(\boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}_{1k}}}{e^{\boldsymbol{\beta}'\boldsymbol{x}_{1k}} + e^{\boldsymbol{\beta}'\boldsymbol{x}_{0k}}} \tag{4.25}$$

with

- $\boldsymbol{x}_{1k}$ as the data vector for the case ($y = 1$) in the k^{th} stratum/pair and
- $\boldsymbol{x}_{0k}$ as the data vector for the controls ($y = 0$) in the k^{th} stratum/pair.

As described in section 4.4.3, equation (4.25) does not model the probability of the outcome. It estimates the probability of the covariate values given a certain outcome.

Equation (4.25) describes the likelihood that the subject that is identified as the case is in fact the case (given specific values of $\boldsymbol{\beta}$, $\boldsymbol{x}_{1k}$ and $\boldsymbol{x}_{0k}$) under the following assumptions (Hosmer & Lemeshow, 2000):

1. that there are two subjects: one subject as the case, the other subject as the control and

2. that the logistic regression model in equation (4.20) is the correct model.

To facilitate understanding of this fact, we assume as an example a model with one dichotomous covariate and a slope parameter $\beta = 0.8$. The observed covariate data are $x_{1k} = 1$ for the case and $x_{0k} = 0$ for the control. Equation (4.25) results to a value of

$$l_k(\beta = 0.8) = \frac{e^{0.8 \cdot 1}}{e^{0.8 \cdot 1} + e^{0.8 \cdot 0}} = 0.690 \quad .$$

This means that the probability is 0.69 that a subject with covariate value $x = 1$ is the case, compared to a subject with the covariate value $x = 0$.

Otherwise, if the value of the covariate for the case is $x_{1k} = 0$ and that of the control is $x_{0k} = 1$, the result would be

$$l_k(\beta = 0.8) = \frac{e^{0.8 \cdot 0}}{e^{0.8 \cdot 0} + e^{0.8 \cdot 1}} = 0.310$$

So the probability is 0.31 that a subject with $x = 0$ is the case, compared to a subject with $x = 1$.

For the values of the covariates of case and control being identical ($\boldsymbol{x}_{1k} = \boldsymbol{x}_{0k}$), equation (4.25) results to $l_k(\boldsymbol{\beta}) = 0.5$ for any value of $\boldsymbol{\beta}$. Therefore, if case and control have the same values in their covariates, the model considers it equally likely for that subject to be case or control. Thus, any case-control pair with the same value for any covariate is uninformative for the estimation of that covariate's coefficient.

This means that, in practice, an estimator may only be based on a small fraction of the total number of possible pairs, namely on the discordant pairs. This can be examined by forming 2×2 tables cross-classifying case vs. control for all dichotomous covariates (Hosmer & Lemeshow, 2000).

Effects of Ignoring the Matching

Breslow and Day (1980) state that it was common practice in earlier days (when special software was not easily available) to ignore the matching of the data and perform a standard, unconditional regression analysis. In most cases, this does not alter the estimates of the relative risk. However, there are certain circumstances under which the differences between the matched and unmatched analysis become more severe.

The two conditions under which the matching can be ignored with only slight alterations of the estimates, according to Breslow and Day (1980), are that the stratification variables are either

1. conditionally independent of disease status given the risk factors or
2. conditionally independent of the risk factors given disease status.

Nevertheless, in an unmatched analysis with data collected in a matched design, the results tend to be biased in direction of conservatism.

Other authors, for example Hosmer and Lemeshow (2000), also advise not to ignore the matching.

Relation of Conditional and Unconditional Logistic Regression

Hosmer and Lemeshow (2000) state that it is also possible to use the method of standard logistic regression to fit a conditional logistic regression model. To do this, the data has to be reorganized. A new data vector $\boldsymbol{x}^*$ has to be computed by subtracting the covariate values of the control from the covariate values of the case, $\boldsymbol{x}^* = \boldsymbol{x}_{1k} - \boldsymbol{x}_{0k}$. The new

sample size is the number of the case-control pairs (the number of strata in the conditional analysis).

For the fitted standard logistic regression model, the intercept has to be excluded from the model specification, and the outcome is set to 1 for all subjects. This yields identical results as an unconditional logistic regression analysis on the original data.

It has to be noted that not all software packages are able to allow a constant outcome for all cases.

Model Interpretation

In the linear regression model, the interpretation of the slope coefficients (the estimated $\hat{\beta}$) is relatively straightforward. The slope coefficients represent the change in the dependent variable when the independent variable changes by one unit (when all other independent variables are held constant). The mathematical reason for this is that there is no link function involved (or, expressed more correctly, the link function is the identity function $y = y$).

In the logistic regression model, the link function is the logit transformation $g(\boldsymbol{x}) = \log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \boldsymbol{\beta}'\boldsymbol{x}$. The slope coefficient therefore represents the change in the logit corresponding to a change of one unit in the independent variable (all other independent variables held constant). The interpretation depends on placing meaning on the difference between two logits.

Slope Coefficients and Odds Ratio. A measure to make interpretation easier is the *odds ratio*. The *odds* for the outcome to be present amongst individuals with a covariate value of $x = 1$ are defined as relation of the probability of the outcome to be present to the probability of the outcome not to be present:

$$\frac{\pi(x=1)}{1-\pi(x=1)} \tag{4.26}$$

Vice versa, the odds for the outcome being present among individuals with the covariate value $x = 0$ are defined by:

$$\frac{\pi(x=0)}{1-\pi(x=0)} \tag{4.27}$$

So the odds ratio (OR) is given by the equation

$$OR = \frac{\frac{\pi(x=1)}{1-\pi(x=1)}}{\frac{\pi(x=0)}{1-\pi(x=0)}} = \exp(\beta) \tag{4.28}$$

An odds ratio expresses how much more likely it is for the outcome to be present among those with covariate values of $x = 1$ than among those with $x = 0$. An odds ratio of two would mean that the risk for a certain outcome of a subject with covariate $x = 1$ is twice as high as the risk of a subject with $x = 0$ in the study population. An odds ratio lower than one indicates that a covariate value of $x = 1$ reduces the risk, while an odds ratio of one indicates that the covariate does not have an effect. For this reason, 95% confidence intervalls for the odds ratios are specified.

For dichotomous variables, the odds ratio indicates the change of risk in relation to the reference category (usually $x = 0$). In the case of polytomous variables with $k > 2$ categories, $k-1$ slope coefficients have to be estimated. The odds ratios of these coefficients also indicate the change of risk in relation to the reference category (if the proper coding for the design variables is used).

The adequacy of this interpretation is entirely dependent on the adequacy of the model assumptions, namely linearity and constant slope. Departure from this makes interpretation inappropriate.

Interpretation of Interaction. One of these departures is called interaction. The relationship of independent variable and logit is still linear, but the slopes differ for certain groups. When interaction is present, the association between risk factor and outcome variable depends in some way on the level of the covariate. The effect of the risk factor cannot be specified without also specifying the level of the covariate (Hosmer & Lemeshow, 2000).

Assessment of Fit

There are various ways to assess the fit of a (conditional) logistic regression model. The following is a brief outline of the methods used in this paper.

Likelihood Ratio Test (LR-Test). The likelihood ratio test (LR-Test, LRT) is a statistical test that can be used to compare the goodness-of-fit of two nested models. It compares a relatively complex model to a simpler model, and judges which of those models better fits the particular dataset. The test provides an objective criterion for selecting among possible models. It approximately follows a χ^2 distribution, with the number of parameters (degrees of freedom, df) being the number of additional parameters in the more complex model. The test statistic is defined by

$$LR = -2 \cdot \log \frac{L_0}{L_1} = -2 \cdot (\log L_0 - \log L_1) \tag{4.29}$$

with

L_0 being the maximum likelihood of the simple model and
L_1 being the maximum likelihood of the more complex model.

In most cases (including this paper), the test compares a particular model with the null-model that only contains an intercept parameter. In other words, it is a test of the omnibus null hypothesis that all parameters (all estimated slope coefficients β) are zero (Huelsenbeck & Crandall, 1997).

Akaike Information Criterion (AIC). The Akaike information criterion (Akaike, 1974) is a measure of the goodness of fit of an estimated statistical model. It is not a test in the sense of hypothesis testing, but serves for model selection. Given several competing models, the one with the lowest AIC is the one fitting the data best. The AIC "punishes" more complex models because it takes the number of parameters into account. It is defined by

$$AIC = -2 \cdot \log L + 2k \tag{4.30}$$

with

L as the maximum likelihood of the model in question and
k as the number of parameters of the model.

Variance explained (R^2). The dependent variable y of the regression model shows a certain variance. The fraction of variance that is explained by the regression model is indicated by R^2:

$$R^2_{SS} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{\pi}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \tag{4.31}$$

with

y_i observed value for subject i and
$\hat{\pi}_i$ predicted value of the regression model for subject i.

As the variance explained automatically increases with the number of parameters in the model, this figure can be adjusted (adjusted R^2). This measure of the amount of variance explained is based on the deviance of the expected values from the observed values – it is a sum-of-squares measure (therefore the index).

Another measure of explained variance is based on the model likelihoods. When applied to linear models, it is identical to the standard multiple R^2, but it can also be applied to generalized linear models like the conditional logistic regression model. It is defined by

$$R_L^2 = 1 - \left(\frac{L_0}{L_1}\right)^{\frac{2}{n}} \tag{4.32}$$

with

- L_0 being the maximum likelihood of the null model,
- L_1 being the maximum likelihood of the model in question, and
- n being the number of observations (sample size).

As this measure cannot attain a value of one, the maximum value can be calculated by

$$R_{L,(\max)}^2 = 1 - L_0^{\frac{2}{n}} \tag{4.33}$$

and the likelihood-based measure can be scaled more adequately (Mittlböck & Schemper, 1996).

Hosmer and Lemeshow (2000) adduce that for logistic regression models, the values for R^2 are generally very low in comparison to linear regression models. Thus, they do not recommend publishing this figure with results from fitted logistic regression models. But they also state that it may be useful in the model building stage.

4.4.4. Odds Ratio Multifactor Dimensionality Reduction

Multifactor dimensionality reduction (MDR) is the implementation of a data-mining strategy developed by Ritchie et al. (2001), inspired by the combinatorial-partitioning method (Nelson, Kardia, Ferrell, & Sing, 2001). The MDR approach is an alternative to logistic regression that is suited to find high-order interactions in relatively small data sets. It was designed to find gene–environment (G × E) or gene–gene (G × G) interactions. Logistic regression is less practical when dealing with high-dimensional data, because in the estimation of coefficients for high-order interactions, the contingency table cells may contain no or few observations. This leads to large coefficient estimates and high standard errors (Hosmer & Lemeshow, 2000).

The MDR approach is a data-reduction method for the exploratory analysis of quantitative traits. It pools variables into high-risk and low-risk groups. The predictors are thereby reduced from n dimensions to one dimension. The method is model free (as no particular genetic model is assumed) and nonparametric (as no parameters are estimated). It is also applicable to discordant sib-pair studies (Ritchie et al., 2001).

The MDR method was first used to detect high-order genetic interaction in sporadic breast cancer (Ritchie et al., 2001). A freeware program has been developed by Hahn, Ritchie, and Moore (2003).

The method was later enhanced to include the estimation of odds ratios by Chung, Lee, Elston, and Park (2007). The method is called odds ratio multifactor dimensionality reduction (OR MDR).

MDR Algorithm

The MDR algorithm consists of six steps (Coffey et al., 2004) that will be outlined in the following (see also figure 4.1). It can be used to find interaction patterns of polytomous predictors (that have two or three categories) when predicting a dichotomous outcome.

1. The procedure starts with the subdivision of the data into 10 equal subsets. The so-called *training set* consists of 9/10 of the matched pairs, while the independent *holdout set* contains the remaining 1/10 of the pairs. The model is first developed in the training set, while the holdout set is used for cross-validation.

2. In the second step, a set of n factors is selected. In fact, only the number n of how many factors will be included in the model is specified.

3. Then, the n factors are represented in n-dimensional space. As an example, two polymorphisms with three genotypes would each be represented as nine (3^2) genotype combinations.

4. In step 4, the ratio of cases to controls is computed in each multifactor cell. Each cell is labeled as "high-risk" if the ratio of cases to controls exceeds a certain threshold (e. g. ratio ≥ 1). Otherwise, the cell is labeled as "low-risk". This reduces the n-dimensional multifactor classes into a one-dimensional model with two multifactor classes: high-risk and low-risk.

5. Steps 2–4 are repeated for all other possible n factor combinations. The model that has the fewest misclassified individuals in the training set is chosen as best n factor model.

6. This best n factor model is then used to predict the disease status for the remaining 1/10 of the data, the holdout set. All empty cells in either the training or holdout set are ignored since there is nothing to predict. In this step, the *prediction error* is computed. It is the ratio of incorrect predictions in the holdout set.

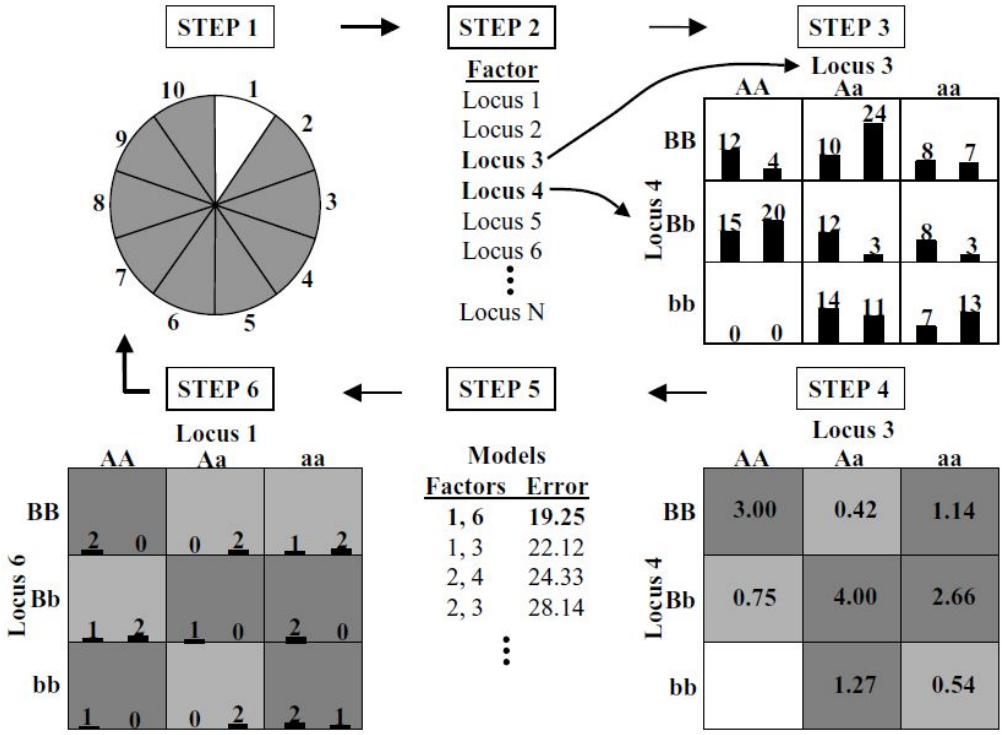

Figure 4.1.: *MDR algorithm. (Source: Coffey et al., 2004)*

This process is repeated for each possible 9:1 split of the data, resulting in a 10-fold *cross-validation*. To protect against chance divisions of the data, this 10-fold-cross validation is repeated 10 times, the data being randomized before each trial.

The reported statistics for the final model are the average prediction error of the 100 estimates and the *cross-validation consistency*, which is the number of times a particular set of n factors is identified across the 10-fold cross-validation (ranging from 0 to 10).

As the number n of factors has to be chosen, it is advisable to perform the procedure for a range of values, which results in one model for each value of n (e. g. one best two-factor model, one best three-factor model, one best four-factor model).

It can be assessed whether the prediction error is lower than expected by chance. In a balanced sample, the prediction error expected by chance would be 0.5, as the number of cases equals the number of controls (Coffey et al., 2004).

The MDR algorithm can be enhanced by the integration of machine learning methods for classification (Moore et al., 2006).

Advantages and Disadvantages of the MDR Approach

The primary advantage of the MDR approach is that it facilitates the simultaneous detection of multiple genetic and environmental loci associated with a discrete clinical endpoint. This is accomplished by the pooling into high-risk and low-risk groups, reducing the dimensionality of the data set to one.

Furthermore, the method is nonparametric. In parametric methods like the generalized linear model, the number of possible interaction terms grows exponentially with each additional main effect included in the model. This can result in having too many independent variables in relation to the number of observed outcome events. Fitting a full model with all interaction terms and then using backward elimination to derive a more parsimonious model would not be possible in this case. The MDR method avoids this problem.

A third advantage is that MDR does not assume a specific genetic model – no mode of inheritance has to be specified. This is especially important for diseases where the mode of inheritance is unknown and likely very complex.

A disadvantage of the MDR approach is that it can be computationally intensive when more polymorphisms or environmental risk factors are involved.

In the present implementation, the MDR method is only applicable to case-control studies that are balanced (that have the same number of cases and controls). This limitation may be patched by results of future research.

Furthermore, MDR models can be difficult to interpret. There might not be an obvious trend in the distribution of high-risk and low-risk groupings across the multidimensional genotype space. The lack of an obvious trend might be a sign of epistasis (the interaction of multiple genes to produce a certain phenotype; e. g. a disease or mental disorder) (Ritchie et al., 2001).

Another disadvantage in model interpretation may become evident when MDR is used in the presence of main effects or known covariates, where it becomes much harder to disentangle the final model. For example, if an MDR analysis suggests an optimal model that contains four factors, it is not clear whether this model represents a four-way interaction, two separate two-way interactions, or two main effects and a two-way interaction, etc. (Coffey et al., 2004).

Perhaps the most severe limitation is that the MDR approach is vulnerable to false positive and negative errors when the ratio of cases to controls in a combination of genotypes (or environmental factors, or both) is similar to that of the entire data, or when the number of cases and controls are very small in a combination. If the number of subjects in a cell is small, the ratio is not robust against small changes in the number of cases or controls

– small changes in the frequencies can result in a change of the classification of that cell, thereby altering the resulting model (Chung et al., 2007).

This limitation is remedied with the introduction of estimated odds ratios into the MDR method, resulting in the odds ratio multifactor dimensionality reduction method (OR MDR).

OR MDR Algorithm

The odds ratio multifactor dimensionality reduction method (OR MDR) is based on the MDR method. Developed by Chung et al. (2007), its major advantage is that it allows estimation of odds ratios as quantitative measure of risk for each combination of genotypes – the genotype combinations can be ordered from highest to lowest risk. Furthermore, confidence intervals can be estimated with this method, allowing a comparison of disease risk between different genotype combinations.

In an OR MDR analysis, the first step is to perform a conventional MDR analysis on the data. This results in a best model (the combination of genetic and/or environmental factors that has the lowest prediction error). For this model, the odds ratios for each combination of genotypes (or environmental factors) are computed as a quantitative measure of disease risk (Chung et al., 2007).

The whole process is illustrated in figure 4.2, where the MDR analysis is performed in stage 1 and the odds ratios are computed in stage 2.

To illustrate the estimation of the odds ratios, two polymorphisms (P_1, P_2) each with three genotypes are assumed to be selected as the best model by the MDR analysis. The two polymorphisms and the binary variable distinguishing cases and controls yield a $3\times3\times2$ contingency table for N subjects.

The odds of the disease for a given genotype combination ($P_1 = i$, $P_2 = j$) are

$$\frac{P\left(\text{Disease}|P_1 = i, P_2 = j\right)}{P\left(\text{Normal}|P_1 = i, P_2 = j\right)} = \frac{P\left(P_1 = i, P_2 = j|\text{Disease}\right)}{P\left(P_1 = i, P_2 = j|\text{Normal}\right)} \cdot \frac{P\left(\text{Disease}\right)}{P\left(\text{Normal}\right)} \tag{4.34}$$

Then, the odds for the genotypes (i, j), θ_{ij}, is given as follows:

$$\theta_{ij} = \frac{P\left(P_1 = i, P_2 = j|\text{Disease}\right)}{P\left(P_1 = i, P_2 = j|\text{Normal}\right)} = \frac{P\left(\text{Disease}|P_1 = i, P_2 = j\right)}{P\left(\text{Normal}|P_1 = i, P_2 = j\right)} \cdot \frac{P\left(\text{Normal}\right)}{P\left(\text{Disease}\right)} \tag{4.35}$$

The right-hand side can be interpreted as the odds of the disease for the genotypes (i, j) divided by the odds of the disease for all data (disregarding the genotype information). From the data, θ_{ij} is estimated as follows:

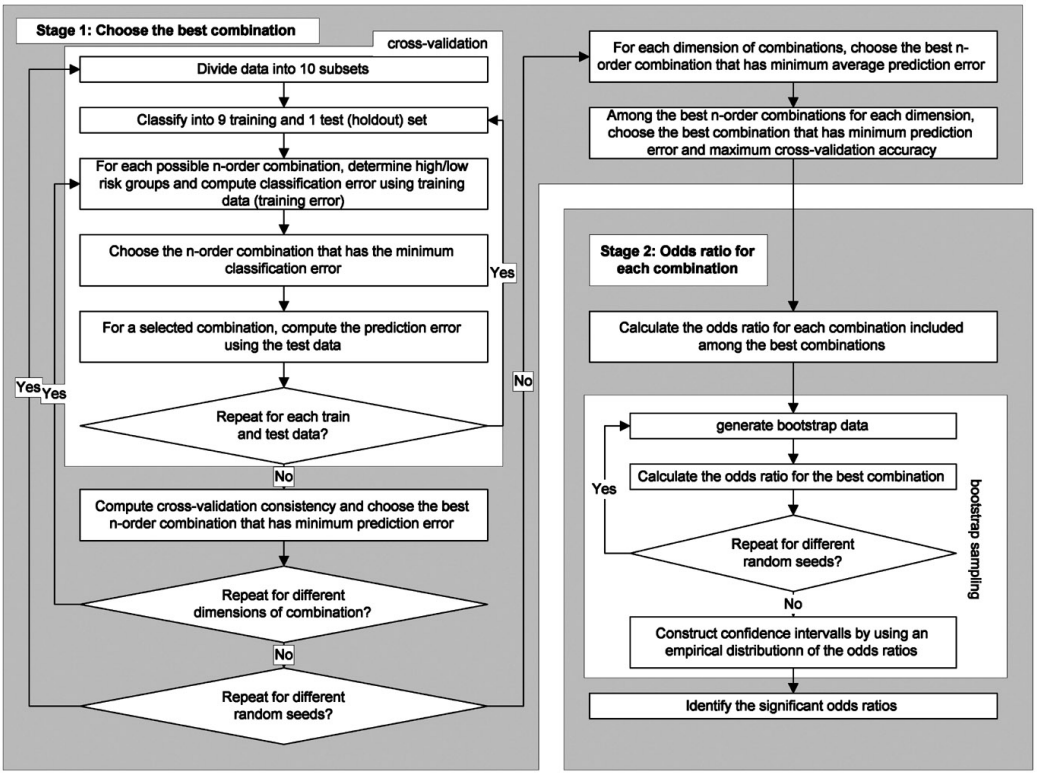

Figure 4.2.: *OR MDR algorithm. (Source: Chung et al. (2007)*

$$\theta_{ij} = \frac{\frac{n_{ij0}}{n_{..0}}}{\frac{n_{ij1}}{n_{..1}}} \tag{4.36}$$

with

n_{ij0} as the number of controls with genotypes i and j

n_{ij1} as the number of cases with genotypes i and j

$n_{..0}$ as the total number of controls

$n_{..1}$ as the total number of cases

This odds ratio is slightly different from the ordinary odds ratio estimator. If it is equal to one, the odds of the genotype is equal to the odds of the case for the entire data (the risk associated with the disease for the given genotype combination is the same as the overall risk in the whole sample), so there is no association with the disease. Odds ratios greater than one indicate a positive association, odds ratios of less than one a negative (protective) association between genotype (or environmental) combination and the disease.

This modified odds ratio ensures that, for a chosen threshold of one, the binary classification of the OR MDR method is the same as that of the MDR method (Chung et al., 2007).

The OR MDR method also provides a measure of accuracy by deriving a cell-specific confidence interval for θ_{ij}. This can be used to compare the odds ratios for specific combinations of genotypes (or environmental risks).

Two types of confidence intervals are provided by the method. One type is the conventional asymptotic confidence interval for the ratio of two success probabilities derived from two independent binomial distributions. The other type is a bootstrap confidence interval that is especially useful when sample sizes are small (Chung et al., 2007).

4.5. Software for Statistical Analysis

All statistical analyses were performed with the software package R, version 2.7.0 (R Development Core Team, 2008). For the conditional logistic regression analysis, the *survival* and *MASS* packages (Venables & Ripley, 2001) were used. For the MDR and OR MDR analysis, the *ORMDR* package was used.

It must be noted that there is another MDR package available in R, namely the *Rmdr* package. This package seems to yield incorrect results, as the results were contradictory to both the freeware program by Hahn et al. (2003) and the *ORMDR* package.

The code for the analyses is found in the appendix.

Part II.

Empirical Part

5. Formulation of Questions and Hypotheses

This chapter will detail the questions and hypotheses that are investigated in this paper.

5.1. Hardy-Weinberg Equilibrium

The Hardy-Weinberg principle (Hardy, 1908) states that the genotype frequencies in a population remain constant, or are in equilibrium from generation to generation, unless there are specific disturbing influences. These disturbing influences include non-random mating, mutations, selection, limited population size, random genetic drift, and gene flow.

If the Hardy-Weinberg principle applies, the (expected) genotype frequencies can be calculated as a function of the allele frequencies:

$$\begin{aligned} f(AA) &= p^2 \\ f(AG) &= 2pq \\ f(GG) &= q^2 \end{aligned}$$

with

- p being the frequency of the A allele and
- q being the frequency of the G allele.

These expected genotype frequencies will be compared to the empirical (observed) genotype frequencies in the sample by means of a χ^2-test. It is assumed that the empirical frequencies do not deviate from the expected ones – in other words the genotype frequencies are assumed to be in Hardy-Weinberg equilibrium.

5.2. Association of 5-HT_{2A} –1438 G/A Polymorphism with Anorexia Nervosa

Although this is not the paper's main question, it will analyze whether the association of the –1438 A allele with anorexia nervosa can be replicated in the collected sample. This will

be done for the whole anorexia nervosa group (both restricting and binge-eating/purging type) and for the restricting group alone (as this group has been associated with an elevated frequency of the –1438 a allele, e.g. by Sorbi et al. (1998); Ricca et al. (2004)). The statistical analysis used for this will once again be the χ^2-test. According to some of the association studies that revealed a significant effect of the 5-HT_{2A} –1438 A allele and AA genotype (e.g. Collier et al., 1997; Sorbi et al., 1998; Gorwood et al., 2003; Ricca et al., 2004), the expectation is that both in the whole sample and in the restricting subgroup, the frequencies of the –1438 A allele and of the AA genotype are higher than in the control group.

5.3. Gene–Environment Correlation

In order to evaluate possible genetic effects or interactions of environmental and genetic factors, it is necessary to assess whether the exposure to the environmental risk factors is associated with the genetic factor. If this is the case, the measure of environmental exposure is contaminated with genetic variation, compromising the interpretation of the results (Caspi & Mofftt, 2006). It is assumed that there is no gene–environment correlation. This is assessed by means of an analysis of variance (ANOVA) and the corresponding effect size η^2.

5.4. Interaction of 5-HT_{2A} –1438 G/A Polymorphism and Environmental Risk Factors

Research shows that the 5-HT_{2A} –1438 G/A polymorphism might be associated with anorexia nervosa. Since the results are conflicting, it is assumed that there are interactions of environmental risk factors (assessed via Oxford Risk Factor Interview) and the 5-HT_{2A} –1438 G/A polymorphism. The possible interaction effects will be evaluated using regression models.

5.4.1. Models 1 and 2: Anorexia Nervosa and Gene–Environment Interaction

In the first model, the predicted variable will be the diagnosis (anorexia nervosa vs. no eating disorder). A conditional logistic regression model will be used for the analysis. The predictors in the model are the 5-HT_{2A} –1438 genotype, the total exposure to environmental risk factors, and their interaction. From this initial model, unnecessary predictors are excluded in a stepwise backward selection process based on the Akaike information criterion

– the terms that lower the AIC the most are excluded first, until the model only contains significant or relevant predictors.

In the second model, the environmental risk factors will be investigated in more detail. The Oxford Risk Factor Interview distinguishes five environmental risk domains. So the initial model includes:

- the main effects
  - of the 5-HT$_{2A}$ –1438 G/A polymorphism and
  - of all five environmental subdomains ($ed1$: parental problems; $ed2$: disruptive events; $ed3$: parental psychiatric disorder; $ed4$: interpersonal problems; $ed5$: family dieting behavior) and
- the interaction effects of the 5-HT$_{2A}$ –1438 G/A polymorphism with all five environmental subdomains:
  - 5-HT$_{2A}$ –1438 G/A Polymorphism × parental problems ($ed1$)
  - 5-HT$_{2A}$ –1438 G/A Polymorphism × disruptive events ($ed2$)
  - 5-HT$_{2A}$ –1438 G/A Polymorphism × parental psychiatric disorder ($ed3$)
  - 5-HT$_{2A}$ –1438 G/A Polymorphism × interpersonal problems ($ed4$)
  - 5-HT$_{2A}$ –1438 G/A Polymorphism × family dieting behavior ($ed5$)

As in model 1, unnecessary predictors are excluded in a stepwise backward selection process based on the AIC.

5.4.2. Application of Odds Ratio Multifactor Dimensionality Reduction to Identify Interactions

In addition to the conditional logistic regression model that evaluates possible interactions of the 5-HT$_{2A}$ –1438 G/A polymorphism with the five environmental risk domains, a new (exploratory) method of analysis is employed: odds ratio multifactor dimensionality reduction (Chung et al., 2007). This method was designed to find gene–environment (and gene–gene) interactions.

5.4.3. Model 3: Body Mass Index and Gene–Environment Interaction

Research suggests that the possible impact of the 5-HT$_{2A}$ –1438 G/A polymorphism might not be directly linked to the diagnosis of anorexia nervosa. A possible measure of the severity of the disorder is the lowest body mass index during the illness (Hebebrand et al.,

1996). To evaluate a possible genetic influence, a multiple linear regression model will be used to predict the lowest recorded body mass index in the anorexic subjects (the healthy controls are not used for the analysis). Studies to present do not reveal a genetic influence of the 5-HT_{2A} –1438 G/A polymorphism on life-time severity assessed by lowest body mass index (e.g. Kipman et al., 2002; Gorwood et al., 2002). Therefore, the regression model not only includes genetic main effects but also interactions with environmental risk factors.

Model 3 includes the main effects of the 5-HT_{2A} –1438 G/A polymorphism and all five environmental subdomains, as well as all their interactions. Again, a stepwise backward selection process is performed.

Since the body mass index might be lower for the restricting subgroup, the influence of the anorexia nervosa subtype is investigated by the introduction of an according covariate into the resulting model.

5.4.4. Model 4: Age of Onset and Gene–Environment Interaction

Literature states that the 5-HT_{2A} –1438 G/A polymorphism might act as a modifying factor in anorexia nervosa. The –1438 A allele was found to delay the onset in a study by Kipman et al. (2002), while one other study did not report any differences in the age of onset related to the 5-HT_{2A} –1438 G/A polymorphism (Gorwood et al., 2002). The question whether the effect of the 5-HT_{2A} –1438 G/A polymorphism can be clarified in this sample, as well as the additional issue of potential interactions with environmental risk factors, will be answered by means of a multiple linear regression model. As above, the model (model 4) includes genetic and five environmental main effects and their interactions, the dependent variable being the age of onset. A stepwise backward selection process will determine the model with the best fit.

To evaluate a possible influence of the subtype (restricting vs. binge-eating/purging), an according covariate is added to the model.

6. Results

6.1. Description of Sample

6.1.1. Recruitment of Subjects

The study conducted in this paper is a multicenter study. Participants were recruited in three centers: the King's College Institute of Psychiatry, Eating Disorder Department (London), the Hospital Princeps d'Espanya, Department of Psychiatry (Barcelona), and the University Clinic of of Neuropsychiatry of Childhood and Adolescence, Eating Disorders Unit (Vienna).

The protocol for the study was approved by the ethic committees of the three centers. Patients and their sisters gave written informed consent for all procedures before inclusion in the study.

The data were collected during a period of four years, between 2000 and 2003.

6.1.2. Sample Size

For the study, 128 sister pairs were interviewed and genotyped. The genotyping procedure could be performed successfully for only 89 sister pairs. Therefore the sample consists of 89 women suffering from anorexia nervosa (cases), and their 89 healthy sisters (controls).

Table 6.1 gives the number of case-control pairs with respect to the center in which the data were collected and to the anorexia nervosa subtype.

	AN-R	AN-BP	AN (total)
London	9	6	15
Vienna	25	29	54
Barcelona	6	14	20
Total	40	49	89

Table 6.1.: *Sample size per center and diagnose*

6.1.3. Age

The mean age of the patients is 24.18 years (standard deviation 7.75, ranging from 14 years to 62 years), the mean age of the sisters is 24.64 years (standard deviation 8.37, ranging from 14 years to 56 years). Accordingly, the mean age difference between the affected and the healthy sister is -0.49 years, with a standard deviation of 3.95. The patients are slightly younger.

The distribution of the age difference is depicted in figure 6.1.

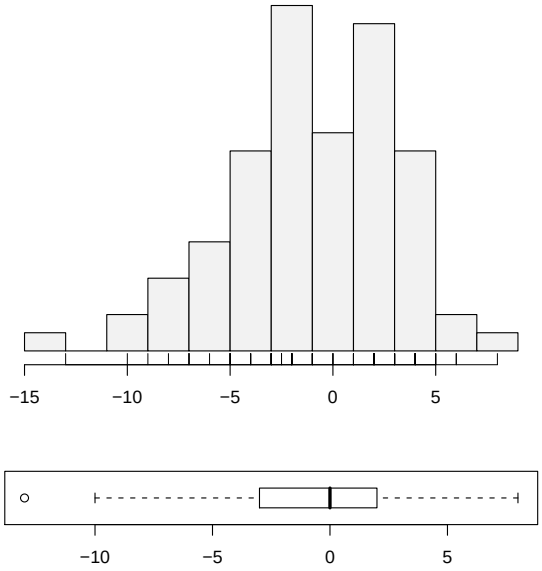

Figure 6.1.: *Distribution of age difference*

6.1.4. Parents

The education levels of the subject's parents are given in table 6.2. For both mothers and fathers, secondary degree has the highest frequency.

	primary school	professional degree	secondary degree	university degree
Mother	0.292	0.270	0.303	0.135
Father	0.227	0.250	0.273	0.250

Table 6.2.: *Highest education levels of parents*

The age at birth of the affected sister also has been recorded for the parents. For the mothers this variable averages 27.59 years (standard deviation 5.15), while for the fathers, the mean age at birth of the future anorexic daughter is 30.86 years (standard deviation

6.02).

6.1.5. Eating Disorder Related Characteristics

The mean age of onset in the total anorexia nervosa group (all patients) is 16.3 years (standard deviation 4.36). The mean age of onset for the restricting subtype (15.88 years) is lower than in the binge-eating/purging subtype (16.63 years). The difference is significant (Mann-Whitney test, $W = 7832$, $p < 0.001$), and the distribution of this variable is visualized in figure 6.2.

The mean duration of illness in the total anorexia nervosa group is 4.97 years (standard deviation 5.32). For the restricting subgroup, the duration of illness averages 3.46 years, while for the binge-eating/purging subgroups it is 6.17 years. This difference is also significant (Mann-Whitney test, $W = 6320.5$, $p < 0.001$), and a boxplot for the variable duration of illness is depicted in figure 6.2.

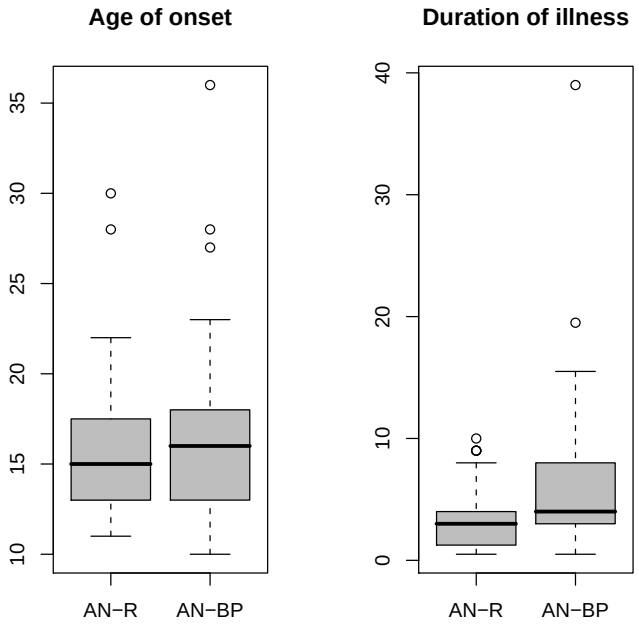

Figure 6.2.: *Boxplots of age of onset (left) and duration of illness (right)*

Concerning the Body Mass Index (BMI), three measures have been recorded for both the patients and their healthy sisters: lowest, current, and highest BMI.

The current BMI for the patients is 18.22 on the average (standard deviation 1.94), whereas for the healthy controls it is at a somewhat higher level of 21.85 (standard deviation

3.24). Concerning the subtypes, the current BMI is somewhat lower for the restricting subgroup (17.85), compared to the binge-eating/purging subgroup (18.52).

The difference in the current body mass index was assessed with a Kruskal-Wallis test to compare the three groups, indicating that the difference is significant ($\chi^2_{(2)} = 63.28$, $p < 0.0001$). Furthermore, the significance of the difference of the restricting and binge-eating/purging subgroup was judged with an exact Mann-Whitney U-test. The two subgroups do not differ significantly in their current body mass index ($W = 794.5$, $p = 0.1778$).

The lowest BMI (that can be considered a measure of severity of the eating disorder, especially in the subgroup of restricting patients) is 14.57 for the affected and 19.76 for the healthy sisters.

The numbers are given in table 6.3, with their standard deviations in parentheses. For visualization, figure 6.3 shows boxplots of lowest, current and highest BMI for each group separately.

The difference of lowest body mass index in a three-group comparison (healthy sisters, restricting, and binge-eating/purging subgroup) is significant (Kruskal-Wallis test; $\chi^2_{(2)} = 111.87$, $p < 0.0001$). Moreover, the difference between the patients in the two anorexia nervosa subgroups is also significant (exact Wilcoxon test; $W = 676.5$, $p = 0.029$)

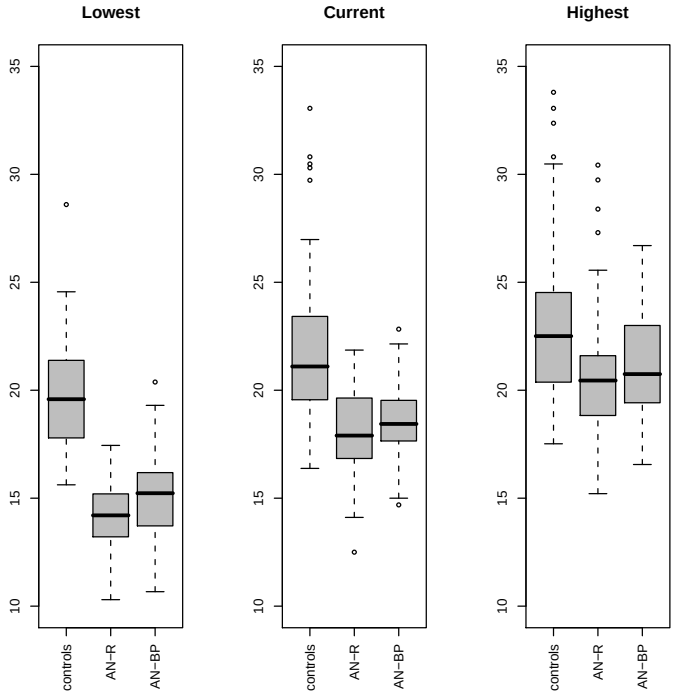

Figure 6.3.: *Boxplots of BMI*

	AN-R	AN-BP	AN (total)	controls
lowest BMI	14.08 (1.59)	14.96 (1.91)	14.57 (1.82)	19.76 (2.5)
current BMI	17.85 (2.12)	18.52 (1.74)	18.22 (1.94)	21.85 (3.24)
highest BMI	20.99 (3.4)	21.26 (2.4)	21.14 (2.88)	23.09 (3.69)

Table 6.3.: *Body Mass Index*

6.1.6. Exposure to Environmental Risk

The exposure to environmental risk factors (distressing events or circumstances) was assessed by the Oxford Risk Factor Interview (see section 4.2). The environmental risk factors are grouped into five domains: parental problems ($ed1$), disruptive events ($ed2$), parental psychiatric disorder ($ed3$), interpersonal problems ($ed4$), and family dieting environment ($ed5$). The mean numbers of circumstances or events that the subjects felt distressed by are listed in table 6.4, along with the corresponding standard deviations in parentheses.

	AN-R	AN-BP	AN (total)	controls
Parental problems ($ed1$)	2.38 (2.14)	2.92 (1.86)	2.67 (2)	1.92 (1.91)
Disruptive events ($ed2$)	1.12 (1.14)	1.59 (1.37)	1.38 (1.28)	0.88 (0.96)
Parental psych. disorder ($ed3$)	0.25 (0.63)	0.45 (0.77)	0.36 (0.71)	0.26 (0.53)
Interpersonal problems ($ed4$)	1.1 (1.1)	0.92 (0.95)	1 (1.02)	0.62 (0.86)
Family dieting env. ($ed5$)	1.85 (1.23)	1.96 (1.43)	1.91 (1.34)	1.39 (1.39)
Total (edt)	6.7 (3.18)	7.84 (3.54)	7.33 (3.41)	5.07 (3.53)

Table 6.4.: *Environmental risk exposure*

Furthermore, the distributions of the five environmental domains (and the total environmental exposure (edt), which is the sum of the five subdomains) are visualized in figure 6.4. For each subdomain, the boxplot on the left corresponds to the healthy sisters, the box in the middle to the anorexic patients of restricting subtype, and the plot on the right to the patients of binge-eating/purging subtype.

As further description of the sample, the correlations between the five environmental subdomains are given in table 6.5. The highest correlation is observed between subdomains one and four, parental problems and interpersonal problems.

6.1.7. Allele and Genotype Frequencies

The allele frequencies of the 5-HT_{2A} –1438 G/A polymorphism are shown in table 6.6, the genotype distribution is shown in table 6.7.

The frequency of the –1438 A allele is highest in the subgroup of restricting patients, and lowest in controls. The subgroup of binge-eating/purging anorexic patients shows an

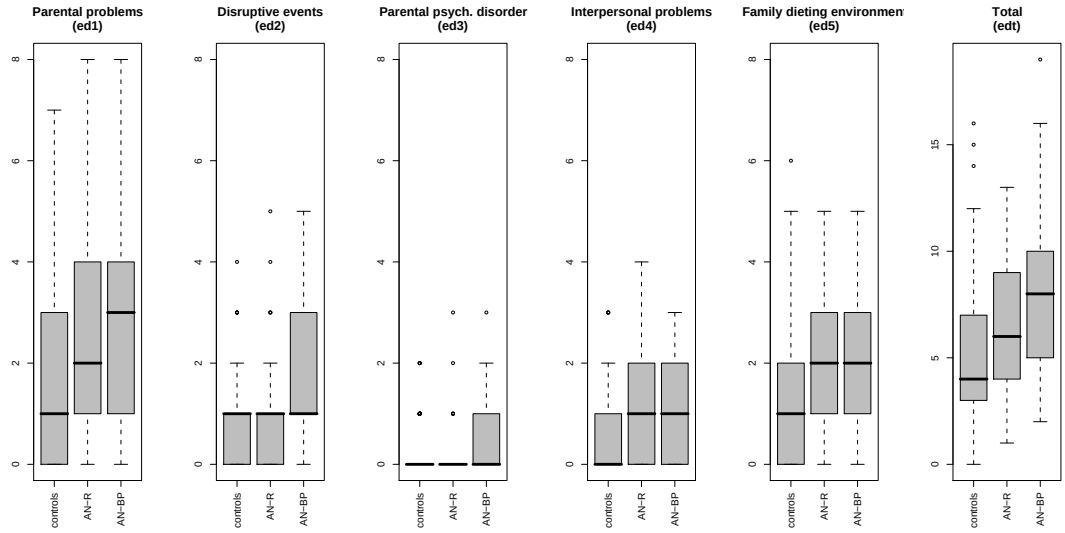

Figure 6.4.: *Boxplots of environmental risk exposure*

	ed1	ed2	ed3	ed4	ed5
ed1	1.0000	0.2139	0.2563	0.3232	0.1426
ed2	0.2139	1.0000	0.2161	-0.0436	0.1933
ed3	0.2563	0.2161	1.0000	0.0795	0.0723
ed4	0.3232	-0.0436	0.0795	1.0000	0.0219
ed5	0.1426	0.1933	0.0723	0.0219	1.0000

Table 6.5.: *Environmental domain correlations*

intermediate frequency. In total, the whole group of anorexic patients still shows a higher frequency of the A allele, compared to controls.

The same can be said for the –1438 AA genotype.

	AN-R	AN-BP	AN (total)	controls
A allele frequency	0.475	0.449	0.461	0.421
G allele frequency	0.525	0.551	0.539	0.579

Table 6.6.: *Allele frequencies*

6.2. Hardy-Weinberg Equilibrium

If the Hardy-Weinberg principle (Hardy, 1908) applies, the (expected) genotype frequencies are a function of the allele frequencies in the sample (see section 5.1). The observed and expected genotype frequencies are shown in table 6.8.

The χ^2-test performed to compare observed and expected frequencies was non-significant

	AN-R	AN-BP	AN (total)	controls
GG genotype frequency	0.250	0.286	0.270	0.270
AG genotype frequency	0.550	0.531	0.539	0.618
AA genotype frequency	0.200	0.184	0.191	0.112

Table 6.7.: *Genotype frequencies*

	observed	expected
GG genotype	48.00	55.62
AG genotype	103.00	87.76
AA genotype	27.00	34.62

Table 6.8.: *Observed and expected genotype frequencies*

($\chi^2_{(2)} = 2.72$, $p = 0.2567$). The genotype frequencies in the sample are in Hardy-Weinberg equilibrium.

6.3. Association of 5-HT$_{2A}$ –1438 G/A Polymorphism with Anorexia Nervosa

In order to assess a possible association of the 5-HT$_{2A}$ –1438 G/A polymorphism and anorexia nervosa, a comparison of absolute frequencies of alleles and genotypes between anorexic patients and controls was performed (using χ^2 tests). The absolute and relative frequencies of the –1438 alleles are shown in table 6.9, while the frequencies of the –1438 genotypes are shown in table 6.10.

	AN (total)	AN-R	controls
A allele frequency	82 (0.461)	38 (0.475)	75 (0.421)
G allele frequency	96 (0.539)	42 (0.525)	103 (0.579)

Table 6.9.: *Allele frequencies for test of association*

The statistical tests were performed for the anorexia nervosa total group and for the restricting subgroup, both allele-wise and genotype-wise.

The tests did not yield any significant association of the 5-HT$_{2A}$ –1438 G/A polymorphism with the diagnosis of anorexia nervosa. The results are listed in table 6.11, and they show that neither the distribution of alleles nor the distribution of genotypes have an association with the diagnosis, both for the total anorexia nervosa group and for the restricting subtype subgroup.

	AN (total)	AN-R	controls
GG genotype	24 (0.27)	10 (0.25)	24 (0.27)
AG genotype	48 (0.539)	22 (0.55)	55 (0.618)
AA genotype	17 (0.191)	8 (0.2)	10 (0.112)

Table 6.10.: *Genotype frequencies for test of association*

	χ^2	*df*	p value
allele-wise, AN (total)	0.41	1	0.52187
allele-wise, AN-R	0.45	1	0.5043
genotype-wise, AN (total)	2.29	2	0.31814
genotype-wise, AN-R	1.77	2	0.41205

Table 6.11.: *Association tests*

6.4. Gene–Environment Correlation

In order to investigate a possible gene–environment correlation, it has to be computed whether the exposure to the five environmental domains differs for the three genotypes. This was done using an analysis of variance (ANOVA) for each environmental domain separately.

	effect size η^2	p value
Parental Problems ($ed1$)	0.002146	0.828663
Disruptive Events ($ed2$)	0.003747	0.720047
Parental Psychiatric Disorder ($ed3$)	0.009797	0.422536
Interpersonal Problems ($ed4$)	0.031079	0.063127
Family dieting environment ($ed5$)	0.006787	0.551092

Table 6.12.: *Gene–environment correlation*

The results (p values and effect sizes η^2) are shown in table 6.12. None of the three genotypes did show a significantly different exposure to the environmental factors assessed in this study – no gene–environment correlation could be detected in the sample.

Concerning the assumptions of the ANOVA it must be noted that the assumption of normality of the errors (residuals) was not met. The Shapiro-Wilk test revealed deviations from normality ($p < 0.0001$ for all five subdomains). The nonparametric alternative, the Kruskal-Wallis test, requires the number of ties in the data to be small (Fischer, 1996), and the ANOVA seems quite robust to deviations from normality if homogeneity of variance is given (Ferguson & Takane, 2005). Therefore it can be assumed that this violation will not have any consequence here. The assumption of equal variances was met (Levene's test, $p \geq 0.362$ for all five subdomains).

6.5. Interaction of 5-HT$_{2A}$ –1438 G/A Polymorphism and Environmental Risk Factors

6.5.1. Anorexia Nervosa and Gene–Environment Interaction

The primary intent of this paper is to investigate the role of the 5-HT$_{2A}$ –1438 G/A polymorphism and its interaction with environmental risk factors in relation to anorexia nervosa. For this purpose, a regression model was used. The outcome of interest, which is the diagnosis (anorexia nervosa or no eating disorder), is dichotomous. Because of the study design (1–1 matched sample), the data analysis was performed by means of a conditional logistic regression model.

The sample consists of 89 sister pairs, so the sample size for the following analyses (using conditional logistic regression) is $N = 178$.

Model 1: Anorexia Nervosa and one Total Environmental Factor

The first regression model used to examine genetic and environmental effects, as well as a possible interaction between the two factors, is shown in equation (6.1). It is the full (saturated) model – the starting point of the analysis. It shall be called model 1a.

$$g_{an}(\boldsymbol{x}) = \beta_{genotype} \cdot x_{genotype} + \beta_{edt} \cdot x_{edt} + \beta_{genotype \times edt} \cdot x_{genotype} \cdot x_{edt} \tag{6.1}$$

with

- $g_{an}(\boldsymbol{x})$ being the logit of the probability $P(Y = 1)$ (the probability of showing the outcome of interest, namely a diagnosis of anorexia nervosa) dependent on the values of the covariates,
- β being the slope coefficient and
- x being the covariate value.

The model uses one global, continuous environmental factor (*edt*, environmental domain (total)). It is constituted by the number of questions of the Oxford Risk Factor Interview that have been answered with "yes", so it is the number of risky environmental events or circumstances the subject felt distressed by (see section 4.2 for further description). It can (theoretically) range from 0 (not distressed by any environmental risk) to 36 (distressed by all assessed environmental risk events). The maximum value in the sample is 19.

The *genotype* variable has three levels, representing the three possible 5-HT$_{2A}$ –1438 genotypes: AA (two –1438 A alleles), AG (one A allele and one G allele), and GG (two G alleles). For model estimation, dummy variables must be used (see section 4.4.2), so, in

fact, the model does not only contain one variable for the genotype: for the three levels, two dummy variables must be used (and therefore, two slope coefficients must be estimated: $\beta_{genotypeGA}$ and $\beta_{genotypeAA}$). The third level, genotype GG, is the reference level.

The model does not contain an intercept, as the intercepts would have to be estimated for each stratum (sister pair) separately. For this analysis, the intercepts are not of interest and therefore regarded as nuisance parameters (they are not estimated by the model).

The parameters (slope coefficients) of model 1a are shown in table 6.13.

	$\hat{\beta}$	OR	SE($\hat{\beta}$)	$\text{OR}_{\text{lower .95}}$	$\text{OR}_{\text{upper .95}}$	p value	
genAG	-0.4645	0.628	0.976	0.09280	4.26	0.630	
genAA	-1.2784	0.278	1.743	0.00915	8.47	0.460	
edt	0.2988	1.348	0.141	1.02212	1.78	0.034	*
genAG:edt	-0.0143	0.986	0.159	0.72164	1.35	0.930	
genAA:edt	0.4060	1.501	0.308	0.81997	2.75	0.190	

Table 6.13.: *Model 1a (coefficients)*

Model 1a explains 17.69% of the total variance and 35.39% of the variance that can maximally be explained ($R^2_L = 0.1769$, $R^2_{L,(max)} = 0.5$). It differs significantly from the null-model (likelihood ratio test, $\chi^2_{(5)} = 34.66$, $p < 0.0001$) and has an AIC of 98.72.

From this model, the non-significant terms were excluded stepwise (backward stepwise selection), based on the Akaike information criterion (AIC). The terms that lowered the AIC the most were excluded first, until the model only contained significant or relevant terms (for interaction effects, the main effects were of course kept in the model). So, the first term excluded was the term for the interaction. This yielded the following model (model 1b), shown in equation (6.2).

$$g_{an}(\boldsymbol{x}) = \beta_{genotype} \cdot x_{genotype} + \beta_{edt} \cdot x_{edt} \tag{6.2}$$

Model 1b only includes the main effects of the 5-HT_{2A} –1438 G/A polymorphism and environmental risk. The likelihood ratio test to compare models 1a and 1b was not significant ($\chi^2_{(2)} = 3.71$, $p = 0.15671$), so the fit of model 1b is equally good while using less parameters. The estimated coefficients are shown in table 6.14.

Model 1b explains 15.96% of the total variance and 31.93% of the variance that can possibly be explained ($R^2_L = 0.1596$, $R^2_{L,(max)} = 0.5$). It differs significantly from the null-model (likelihood ratio test, $\chi^2_{(3)} = 30.96$, $p < 0.0001$) and has an AIC of 98.42.

As can be ascertained from table 6.14, no genetic main effect could be found. The odds ratios of the coefficients for the different genotypes did not differ significantly from one

	$\hat{\beta}$	OR	SE($\hat{\beta}$)	$OR_{lower\ .95}$	$OR_{upper\ .95}$	p value	
genAG	-0.569	0.566	0.5521	0.192	1.67	0.300000	
genAA	0.961	2.615	0.8323	0.512	13.36	0.250000	
edt	0.308	1.360	0.0718	1.182	1.57	0.000018	*

Table 6.14.: *Model 1b (coefficients)*

(this is also indicated by the confidence interval that includes the value one). It must be noted that the odds ratios for the effects of the genotype are rather high for genotype AA and rather low for genotype AG, but they are not significant. This may be explained by the fact that the sample size is rather small – only 27 persons had genotype AA.

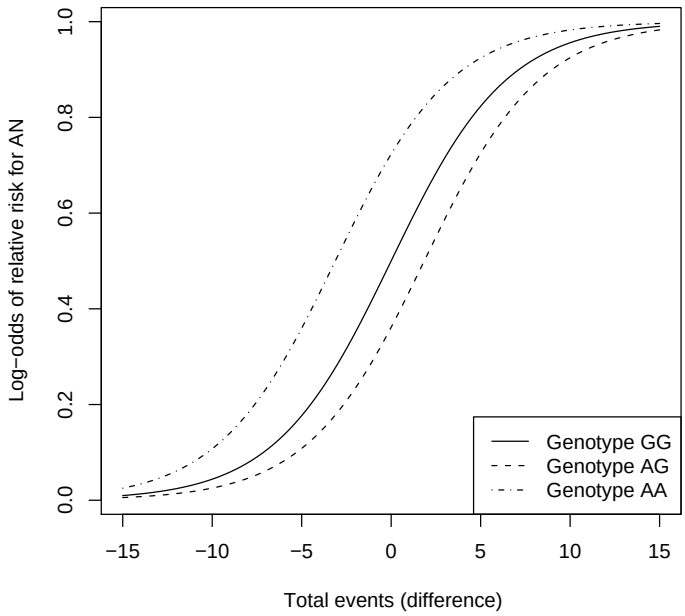

Figure 6.5.: *Model 1b*

The odds ratio for the variable *edt* shows that the environmental influences that were assessed with the Oxford Risk Factor Interview constitute a risk factor. The more events or circumstances the subjects are distressed by, the higher the risk of being the case (of having developed an eating disorder of type anorexia nervosa). Each event increases the relative risk by the factor of 1.360.

The model is illustrated in figure 6.5. Although the genetic main effect is non-significant, the relative log-odds of risk for anorexia nervosa is shown separately for the three genotypes.

The model building process was not continued for model 1, since the hypotheses in this paper focus on genetic main effect and genetic interaction, neither of which could be found with model 1.

Model 2: Anorexia Nervosa and Five Environmental Subdomains

Looking at the environmental risk factors in more detail, we find that they can be divided into five subdomains. These subdomains (that already have been listed in section 4.2) are:

- Environmental domain 1 ($ed1$): parental problems
- Environmental domain 2 ($ed2$): disruptive events
- Environmental domain 3 ($ed3$): parental psychiatric disorder
- Environmental domain 4 ($ed4$): interpersonal problems
- Environmental domain 5 ($ed5$): family dieting environment

The conditional logistic regression analysis started with the full (saturated) model, model 2a, that is shown in equation (6.3).

$$\begin{aligned} g_{an}(\boldsymbol{x}) = \quad & \beta_{genotype} \cdot x_{genotype} + \beta_{ed1} \cdot x_{ed1} + \beta_{ed2} \cdot x_{ed2} + \\ & \beta_{ed3} \cdot x_{ed3} + \beta_{ed4} \cdot x_{ed4} + \beta_{ed5} \cdot x_{ed5} + \\ & \beta_{genotype \times ed1} \cdot x_{genotype} \cdot x_{ed1} + \beta_{genotype \times ed2} \cdot x_{genotype} \cdot x_{ed2} + \\ & \beta_{genotype \times ed3} \cdot x_{genotype} \cdot x_{ed3} + \beta_{genotype \times ed4} \cdot x_{genotype} \cdot x_{ed4} + \\ & \beta_{genotype \times ed5} \cdot x_{genotype} \cdot x_{ed5} \end{aligned} \tag{6.3}$$

It includes a genetic main effect, all five environmental subdomains, and five interaction terms. The estimated coefficients of model 2a are listed in table 6.15.

Model 2a explains 23.64% of the total variance and 47.27% of the variance that can possibly be explained ($R^2_L = 0.2364$, $R^2_{L,(max)} = 0.5$). It differs significantly from the null-model (likelihood ratio test, $\chi^2_{(17)} = 48$, $p < 0.0001$) and has an AIC of 109.38.

From this model, in which only the influence of environmental domain 5 (family dieting environment) is significant, the non-significant terms are once again removed in a stepwise backward selection process (based on the AIC).

The resulting model is called model 2b, shown in equation (6.4).

	$\hat{\beta}$	OR	SE($\hat{\beta}$)	OR$_{\text{lower .95}}$	OR$_{\text{upper .95}}$	p value	
genAG	-0.09750	0.9071	1.218	0.083392	9.87	0.940	
genAA	-2.32086	0.0982	3.023	0.000262	36.76	0.440	
ed1	0.06580	1.0680	0.225	0.687066	1.66	0.770	
ed2	0.75090	2.1189	0.521	0.763005	5.88	0.150	
ed3	0.00448	1.0045	0.812	0.204489	4.93	1.000	
ed4	0.57354	1.7745	0.432	0.761467	4.14	0.180	
ed5	0.85471	2.3507	0.433	1.005379	5.50	0.049	*
genAG:ed1	0.18257	1.2003	0.316	0.646488	2.23	0.560	
genAA:ed1	0.91298	2.4917	0.642	0.708283	8.77	0.150	
genAG:ed2	0.19472	1.2150	0.557	0.407523	3.62	0.730	
genAA:ed2	-1.34469	0.2606	1.145	0.027609	2.46	0.240	
genAG:ed3	0.03071	1.0312	1.197	0.098736	10.77	0.980	
genAA:ed3	4.69006	108.8602	3.142	0.230134	51494.01	0.140	
genAG:ed4	-0.30597	0.7364	0.519	0.266514	2.03	0.560	
genAA:ed4	3.64370	38.2329	2.653	0.211059	6925.81	0.170	
genAG:ed5	-0.78432	0.4564	0.468	0.182448	1.14	0.094	
genAA:ed5	0.48202	1.6193	0.962	0.245544	10.68	0.620	

Table 6.15.: *Model 2a (coefficients)*

$$
\begin{aligned}
g_{an}(\boldsymbol{x}) = \quad & \beta_{genotype} \cdot x_{genotype} + \\
& \beta_{ed2} \cdot x_{ed2} + \beta_{ed4} \cdot x_{ed4} + \beta_{ed5} \cdot x_{ed5} + \\
& \beta_{genotype \times ed5} \cdot x_{genotype} \cdot x_{ed5}
\end{aligned} \tag{6.4}
$$

	$\hat{\beta}$	OR	SE($\hat{\beta}$)	OR$_{\text{lower .95}}$	OR$_{\text{upper .95}}$	p value	
genAG	0.286	1.331	0.779	0.289	6.13	0.7100	
genAA	1.517	4.559	1.020	0.618	33.64	0.1400	
ed2	0.731	2.078	0.251	1.270	3.40	0.0036	*
ed4	0.569	1.766	0.206	1.179	2.65	0.0058	*
ed5	0.912	2.488	0.396	1.146	5.40	0.0210	*
genAG:ed5	-0.762	0.467	0.417	0.206	1.06	0.0680	
genAA:ed5	-0.476	0.621	0.529	0.220	1.75	0.3700	

Table 6.16.: *Model 2b (coefficients)*

The estimated coefficients of model 2b are shown in table 6.16. The model contains three significant parameters, namely environmental domain 2 (disruptive events), environmental domain 4 (interpersonal problems), and environmental domain 5 (family dieting environment). Comparison of models 2a and 2b through a likelihood ratio test showed that model

2b fits the data equally well while being more parsimonious ($\chi^2_{(10)} = 11.93$, $p = 0.28972$).

Model 2b explains 18.34% of the total variance and 36.69% of the variance that can possibly be explained ($R^2_L = 0.1834$, $R^2_{L,(max)} = 0.5$). It differs significantly from the null-model (likelihood ratio test, $\chi^2_{(7)} = 36.07$, $p < 0.0001$) and has an AIC of 101.31.

The model building process for model 2 was discontinued at this point, because it could be seen that the issues of interest, the genetic main effect and its interaction with environmental risks, were not significant. Nevertheless it must be noted that the interaction effect of the AG genotype and environmental domain 5 (family dieting environment) is borderline significant.

Table 6.16 shows that a possible genetic main effect of the 5-HT_{2A} –1438 G/A polymorphism could not be found in the data used for this analysis. The odds ratios for the different genotypes do not significantly differ from one, indicating that they do not convey any increased risk for development of anorexia nervosa. The same can be said for the interaction terms of genotype AG and environmental subdomain 5 (family dieting behavior), although it must be noted that this term is borderline significant.

As for the environmental influences, the presence of disruptive events ($ed2$), interpersonal problems ($ed4$), or family dieting environment ($ed5$) increases the risk for anorexia nervosa.

The odds ratios for these factors all are significantly greater than one. This means that the risk for anorexia nervosa increases with the number of events in the specific subdomain. The biggest increase in risk is for the family dieting subdomain ($ed5$), the odds ratio being 2.488. This means that with each additional event or circumstance in the family dieting environment that a subject feels distressed by, the risk for anorexia nervosa increases by this factor.

Being distressed by a disruptive event ($ed2$) increases the risk by a factor of 2.078, while each additional event in the interpersonal problem subdomain increases the risk by 1.766. Whether these coefficients differ from each other has not been assessed.

Both models (model 1b with a total environmental factor and model 2b with the environmental influences split up in five subdomains) indicate that there are no genetic influences when the outcome of interest is the diagnosis (anorexia nervosa vs. no eating disorder). But the interaction term of environmental domain 5 (family dieting environment) and the 5-HT_{2A} –1438 AG genotype was borderline significant. Disregarding the non-significance, this indicates that subjects with the AG genotype are less severely influenced by distressing events in the family dieting environment than persons with genotype AA. It must be noted that only 27 persons in the sample have genotype AA. This results in a high standard error for the genetic main effects (table 6.16), and may impair parameter estimation.

6.5.2. Body Mass Index and Gene–Environment Interaction

To investigate the role of the 5-HT$_{2A}$ –1438 G/A polymorphism in other factors related to anorexia nervosa, a multiple linear regression on the lowest body mass index (BMI) was computed. This measure can serve as indicator of the severity of the disorder (Hebebrand et al., 1996).

The data set for this regression analysis contained only the subjects that were diagnosed with anorexia nervosa (the cases), so the sample size is $N = 89$.

Model 3: Linear Regression Model for Body Mass Index

The initial model (model 3a) included the genetic and environmental main effects as well as all possible interactions. Additionally, it includes the subtype of anorexia nervosa (as the body mass index is assumed to be lower for the patients of restricting subtype) as well as the age of onset (as the body mass index is generally lower in younger subjects (Hebebrand, Wehmeier, & Remschmidt, 2000)). It is shown in equation (6.5).

$$\begin{aligned}\hat{y}_{bmi}(\boldsymbol{x}) = \quad & \beta_0 + \beta_{genotype} \cdot x_{genotype} + \beta_{ed1} \cdot x_{ed1} + \beta_{ed2} \cdot x_{ed2} + \\ & \beta_{ed3} \cdot x_{ed3} + \beta_{ed4} \cdot x_{ed4} + \beta_{ed5} \cdot x_{ed5} + \\ & \beta_{genotype \times ed1} \cdot x_{genotype} \cdot x_{ed1} + \beta_{genotype \times ed2} \cdot x_{genotype} \cdot x_{ed2} + \\ & \beta_{genotype \times ed3} \cdot x_{genotype} \cdot x_{ed3} + \beta_{genotype \times ed4} \cdot x_{genotype} \cdot x_{ed4} + \\ & \beta_{genotype \times ed5} \cdot x_{genotype} \cdot x_{ed5} + \beta_{subtype} \cdot x_{subtype} + \beta_{onset} \cdot x_{onset} \end{aligned} \tag{6.5}$$

From this model, the non-significant terms were excluded stepwise (backward stepwise selection), based on the Akaike information criterion (AIC). The terms that lowered the AIC the most were excluded first, until the model only contained significant or relevant terms (for interaction effects, the main effects were, of course, kept in the model). This yielded model 3b, shown in equation (6.6).

$$\hat{y}_{bmi}(\boldsymbol{x}) = \quad \beta_0 + \beta_{ed3} \cdot x_{ed3} + \beta_{subtype} \cdot x_{subtype} + \beta_{onset} \cdot x_{onset} \tag{6.6}$$

The estimated coefficients for model 3b are shown in table 6.17. Comparison of models 3a and 3b through a likelihood ratio test showed that model 3b fits the data equally well while being more parsimonious ($\chi^2_{(16)} = 20.53$, $p = 0.19721$).

In model 3b, there is no significant genetic main or interaction effect. The terms have

	$\hat{\beta}$	Std. Error	t value	p value	
(Intercept)	11.634	0.8910	13.06	<0.0001	*
ed3	-0.485	0.2645	-1.83	0.07050	
onset	0.106	0.0431	2.45	0.01639	*
subtype	0.893	0.3750	2.38	0.01957	*

Table 6.17.: *Model 3b, BMI (Coefficients)*

been excluded by the stepwise selection procedure. This shows that the 5-HT_{2A} –1438 G/A polymorphism does not influence the severity of anorexia nervosa.

The environmental subdomain parental psychiatric disorder (*ed*3) is borderline significant (the body mass index would be lower for individuals that are more distressed by parental disorders, indicating a higher severity of the illness). The covariates for age of onset and subtype are both significant. The reference level for the subtype is the restricting subgroup, so the positive coefficient can be interpreted as a higher body mass index in the binge-eating/purging subgroup of patients (as expected). The coefficient for age of onset is also positive, indicating that the body mass index rises with later onset of the disease (as presumed).

It must be noted that the global measures of model 3b are not very satisfying. It only explains a small part of the total variance ($R^2 = 0.1411$ and $R^2_{adj} = 0.1101$). The global F-test of the null hypothesis that all coefficients are zero is significant ($F_{(3,83)} = 4.55$, $p = 0.0053$). The AIC of the model is 346.87.

Test of Model Assumptions

The model assumptions for the linear regression model are linearity of relationship between dependent and independent variables, homoscedasticity (constant variance) of the errors, independence of the errors (no serial correlation), and normality of the error distribution. Furthermore, correlation between independent variables (multicollinearity) inflates the estimated standard errors of the coefficients, so this should also be checked for.

Linearity was tested graphically. Neither the plot observed vs. predicted values nor the plot of residuals vs. predicted values revealed any major deviations from linearity. Of course, the plots also show the poor fit of the model. This is especially true for the plot of observed vs. predicted values – in a model with better fit, the points in the scatter plot should be close to a diagonal line (as observed and predicted value should be in immediate vicinity). The plots are shown in figure 6.6.

The Harrison-McCabe test was not significant and therefore did not reveal any sign of heteroscedasticity ($HMC = 0.403$, $p = 0.105$), the Shapiro-Wilk test for normality of the

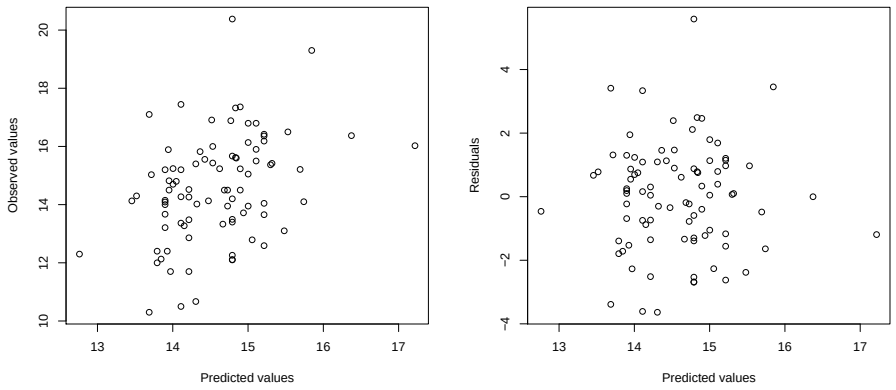

Figure 6.6.: *Model 3b: Observed vs. predicted (left), residuals vs. predicted (right)*

residuals also was non-significant ($p = 0.32$), and the variance inflating factors (Fox & Monette, 1992) all were below the critical value of 4 ($GVIF <= 1.049$). The Durbin-Watson statistic was non-significant, showing no sign of autocorrelation ($DW = 1.732$, $p = 0.206$).

6.5.3. Age of Onset and Gene–Environment Interaction

Another potential impact of the 5-HT$_{2A}$ –1438 G/A polymorphism is its possible role in the moderation of the age of onset. In a study by Kipman et al. (2002) the –1438 A allele was found to delay the onset. In this paper that issue was assessed through a linear regression model.

Model 4: Linear Regression Model for Age of Onset

The initial model, model 4a, included genetic and environmental main effects and all possible interactions. It is shown in equation (6.7).

$$\begin{aligned}\hat{y}_{onset}(\boldsymbol{x}) = \quad & \beta_0 + \beta_{genotype} \cdot x_{genotype} + \beta_{ed1} \cdot x_{ed1} + \beta_{ed2} \cdot x_{ed2} + \\ & \beta_{ed3} \cdot x_{ed3} + \beta_{ed4} \cdot x_{ed4} + \beta_{ed5} \cdot x_{ed5} + \\ & \beta_{genotype \times ed1} \cdot x_{genotype} \cdot x_{ed1} + \beta_{genotype \times ed2} \cdot x_{genotype} \cdot x_{ed2} + \\ & \beta_{genotype \times ed3} \cdot x_{genotype} \cdot x_{ed3} + \beta_{genotype \times ed4} \cdot x_{genotype} \cdot x_{ed4} + \\ & \beta_{genotype \times ed5} \cdot x_{genotype} \cdot x_{ed5} \end{aligned} \tag{6.7}$$

In a stepwise backward selection process, predictors were removed (based on the AIC)

until the model only contained significant or relevant terms. This leads to model 4b, shown in equation (6.8).

$$\begin{aligned}\hat{y}_{onset}(\boldsymbol{x}) = \quad & \beta_0 + \beta_{genotype} \cdot x_{genotype} + \beta_{ed2} \cdot x_{ed2} + \\ & \beta_{ed5} \cdot x_{ed5} + \beta_{genotype \times ed2} \cdot x_{genotype} \cdot x_{ed2} \end{aligned} \tag{6.8}$$

	$\hat{\beta}$	Std. Error	t value	p value	
(Intercept)	15.388	1.274	12.076	<0.0001	*
genAG	0.729	1.444	0.505	0.61519	
genAA	1.180	1.839	0.641	0.52302	
ed2	2.522	0.694	3.635	0.00048	*
ed5	-0.706	0.330	-2.137	0.03560	*
genAG:ed2	-1.529	0.825	-1.853	0.06752	
genAA:ed2	-2.152	1.131	-1.902	0.06067	

Table 6.18.: *Model 4b, age of onset (coefficients)*

When model 4b is compared to model 4a through a likelihood ratio test it can be shown that the more parsimonious model 4b fits the data equally well ($\chi^2_{(11)} = 10.81, p = 0.45973$).

The model contains two significant (environmental) main effects, namely for the domains disruptive events ($ed2$) and family dieting environment ($ed5$). The presence of disruptive events seems to delay the onset of anorexia nervosa, whereas the presence of family dieting environment seems to lead to an earlier onset of the disease.

The model does not contain a significant main effect of the 5-HT_{2A} –1438 G/A polymorphism. The standard errors for the genetic main effects are relatively large (due to small sample size). The coefficient for the AA genotype main effect is positive, which would (in case of significance) indicate that a later onset is related to more disruptive events for subjects with that genotype.

The interaction coefficients in model 4b are only borderline significant. If they were significant (i. e. in a larger sample), they would indicate that both the –1438 AA and AG genotype seem to lead to an earlier onset only in the presence of disruptive events. If this interaction is excluded from the model, the coefficients for the genetic main effects change little, and remain non-significant (results not shown).

Model 4b explains 14.92% of the total variance ($R^2 = 0.2079$ and $R^2_{adj} = 0.1492$), the global F-test of the null hypothesis that all coefficients are zero is significant ($F_{(6,81)} = 3.54$, $p = 0.0036$). The AIC of the model is 503.34.

Inclusion of a covariate for the subtype of the eating disorder (restricting vs. binge-eating/purging subtype) did not change the model. The subtype covariate was not significant, so the type of eating disorder does not seem to influence the effects of the other covariates in the model predicting the age of onset. The likelihood ratio test to compare model 4b and the model including the subtype covariate (model 4c) shows that their fit does not differ significantly, indicating that the inclusion of the subtype covariate does not have any benefit ($\chi^2_{(1)} = 0.02$, $p = 0.89407$).

For the sake of completeness, model 4c is shown in equation (6.9). Its estimated coefficients are listed in table 6.19.

$$\begin{aligned} \hat{y}_{onset}(\boldsymbol{x}) = \quad & \beta_0 + \beta_{genotype} \cdot x_{genotype} + \beta_{ed2} \cdot x_{ed2} + \\ & \beta_{ed5} \cdot x_{ed5} + \beta_{genotype \times ed2} \cdot x_{genotype} \cdot x_{ed2} + \\ & \beta_{subtype} \cdot x_{subtype} \end{aligned} \tag{6.9}$$

	$\hat{\beta}$	Std. Error	t value	p value	
(Intercept)	15.216	1.868	8.147	<0.0001	*
genAG	0.741	1.456	0.509	0.61213	
genAA	1.194	1.854	0.644	0.52127	
ed2	2.512	0.702	3.579	0.00058	*
ed5	-0.706	0.332	-2.124	0.03676	*
subtype	0.113	0.893	0.127	0.89928	
genAG:ed2	-1.530	0.830	-1.842	0.06913	
genAA:ed2	-2.156	1.138	-1.894	0.06188	

Table 6.19.: *Model 4c, age of onset (coefficients)*

Model 4c would explain 13.88% of the total variance ($R^2 = 0.2081$ and $R^2_{adj} = 0.1388$), the global F-test of the null hypothesis that all coefficients are zero would be significant ($F_{(7,80)} = 3$, $p = 0.0075$). The AIC of the model would be 505.32.

Test of Model Assumptions and Comments on Model 4b

Although the global summary statistics of model 4b mentioned above are better than for model 3c, it may not be ideal, as the model could be influenced by outliers. This becomes evident when exploring the assumption of linearity with the plot of observed vs. predicted values and residuals vs. predicted values (figure 6.7).

These plots reveal that there are some individuals with an untypically late age of onset, which the model does not predict well. What is slightly more crucial is that the model

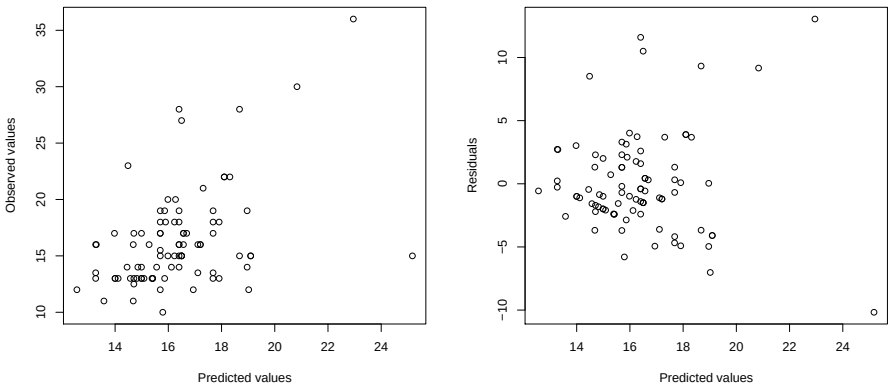

Figure 6.7.: *Model 4b: Observed vs. predicted (left), residuals vs. predicted (right)*

itself is influenced to a great extent by those outliers, in particular by the value of subject 111, as is seen in the plot of residuals vs. leverage, where Cook's distance is also charted (figure 6.8).

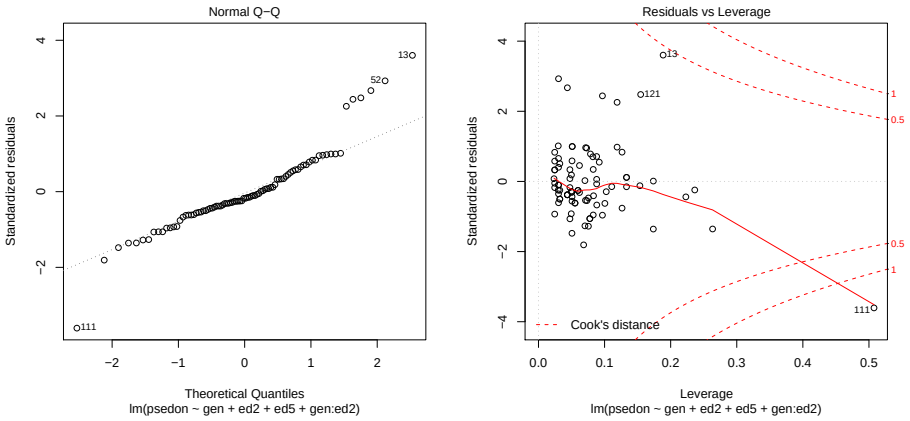

Figure 6.8.: *Model 4b: QQ-plot (left), residuals vs. leverage, Cook's distance (right)*

The values of the predictors of subject 111 are shown in table 6.20, along with the corresponding means of the sample used for the regression model (that is, including subject 111). This subject has a very high exposure to distressing disruptive events ($ed2$) and family dieting environment ($ed5$). Only very few subjects in the sample show an equally extreme exposure, which explains this subject's severe influence on the model. The model would predict a rather late age of onset of 25.2 years for subject 111, but the observed age of onset is 15.0 years.

This also explains why the residuals deviated significantly from normality, as the Shapiro-Wilk test revealed ($p < 0.0001$). The other assumptions of the linear regression model were

	gen	ed1	ed2	ed3	ed4	ed5	Age of Onset	Predicted
Subject 111	GG	3.00	5.00	0.000	0.00	4.00	15.0	25.2
Mean	NA	2.62	1.35	0.364	1.01	1.91	16.3	NA

Table 6.20.: *Predictors of subject 111 and sample means*

met for model 4b. The Harrison-McCabe test did not detect heteroscedasticity ($HMC = 0.496$, $p = 0.486$), the Durbin-Watson statistic was non-significant indicating that there were no auto-correlations ($DW = 1.798$, $p = 0.311$), and the variance inflating factors (Fox & Monette, 1992) were below the critical value of 4 for each predictor ($GVIF <= 2.028$).

Model 4d: Exclusion of Outlier

As the one subject mentioned above influenced the model severely and had a very extreme exposure to environmental risks, the model building process was repeated, this time based on a dataset where subject 111 was excluded. As true above, starting point was the model including all main and interaction effects, model 4a (equation (6.7)). From this model, terms were excluded based on the AIC. This process revealed almost the same model as the analysis based on the full dataset, model 4b. The resulting model is labeled model 4d.

$$\begin{aligned} \hat{y}_{onset}(\boldsymbol{x}) = \quad & \beta_0 + \beta_{genotype} \cdot x_{genotype} + \beta_{ed2} \cdot x_{ed2} + \\ & \beta_{genotype \times ed2} \cdot x_{genotype} \cdot x_{ed2} \end{aligned} \tag{6.10}$$

The only difference to model 4b is that the main effect of environmental domain 5, family dieting environment, is no longer significant and was therefore removed from the model. The coefficients of model 4d are listed in table 6.21.

	$\hat{\beta}$	Std. Error	t value	p value	
(Intercept)	12.30	1.136	10.82	<0.0001	*
genAG	2.52	1.415	1.78	0.07853	
genAA	3.56	1.752	2.03	0.04537	*
ed2	4.94	0.888	5.57	<0.0001	*
genAG:ed2	-4.02	0.982	-4.09	0.00010	*
genAA:ed2	-4.93	1.215	-4.05	0.00011	*

Table 6.21.: *Model 4d, age of onset (coefficients)*

The results show that the influence of disruptive events ($ed2$) is significant, just as in model 4b. A later onset of anorexia nervosa seems to be related with being distressed by

more disruptive events, as the model based on the sample data suggests. Furthermore, there is a significant genetic main effect. Subjects with the 5-HT_{2A} –1438 AA genotype seem to have a later age of onset, delayed by about 3.5 years opposed to the GG genotype. Having the AG genotype also seems to delay the onset by 2.5 years, although this effect is not significant.

However, model 4d contains a significant interaction term. The delay for AA and AG genotypes (genetic main effect) would be true only in absence of any disruptive events. When events of this domain are present, the genetic and environmental main effects are modified by the interaction. The environmental main effect states that more disruptive events and a later onset are related. This is only true, however, for subjects with genotypes AA and AG. Patients with genotype GG do not show any correlation of onset and the number of disruptive events.

To make interpretation a little clearer, this will be exemplified with a numeric example. The predicted age of onset for a subject with the GG genotype not being distressed by any disruptive event is 12.3 years. For an individual with the AA genotype and no disruptive event, the predicted age of onset is 15.86 years, about 3.5 years later. Due to the main effect of disruptive events ($ed2$), the predicted age of onset rises with the number of events that an individual is distressed by. But, as there is an interaction, this is only true for individuals with the GG genotype. The predicted onset with one distressing event for a subject with the GG genotype is 17.24 years. But for an individual with the AA genotype, the predicted age is 15.88 years – so the onset is (almost) *not* delayed for subjects of the AA (and AG) genotype by the distressing event (the main effect of the disruptive event is compensated by the interaction of disruptive events and the AA (and AG) genotype). This is also visualized in figure 6.9. The 5-HT_{2A} –1438 G/A polymorphism seems to be a modifying factor in the onset of anorexia nervosa.

Model 4d explains 26.77% of the total variance ($R^2 = 0.3103$ and $R^2_{adj} = 0.2677$), the global F-test of the null hypothesis that all coefficients are zero is significant ($F_{(5,81)} = 7.29$, $p < 0.0001$). The AIC of the model is 484.66.

Inclusion of the subtype (restricting vs. binge-eating/purging) of the eating disorder did not change the model. The subtype coefficient was not significant and the other coefficients remained almost unchanged. This is also indicated by the likelihood ratio test comparing this new model (model 4e) to model 4d. It is non-significant, showing that there is no difference in the fit of these two models. ($\chi^2_{(1)} = 0.06$, $p = 0.81048$).

For the sake of completeness, model 4e is shown in equation (6.11) and its coefficients are listed in table 6.22.

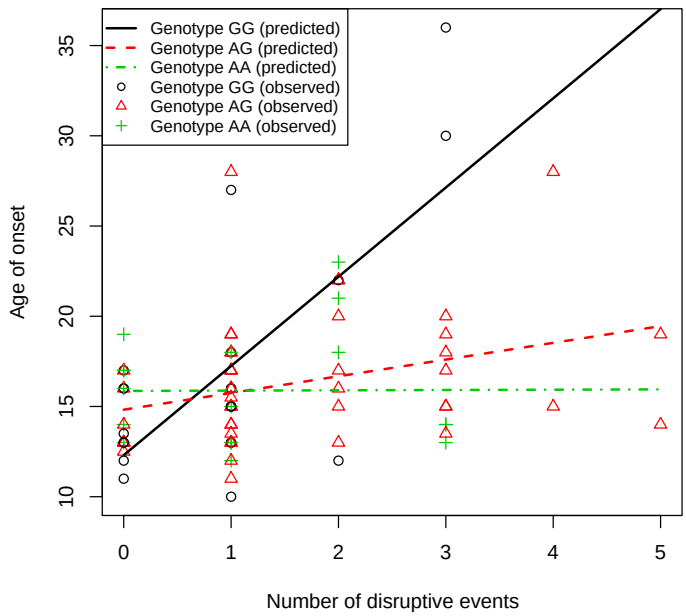

Figure 6.9.: *Model 4d: Observed and predicted values*

$$\hat{y}_{onset}(\boldsymbol{x}) = \beta_0 + \beta_{genotype} \cdot x_{genotype} + \beta_{ed2} \cdot x_{ed2} + \beta_{genotype \times ed2} \cdot x_{genotype} \cdot x_{ed2} + \beta_{subtype} \cdot x_{subtype} \tag{6.11}$$

	$\hat{\beta}$	Std. Error	t value	p value	
(Intercept)	12.008	1.712	7.01	<0.0001	*
genAG	2.545	1.427	1.78	0.07834	
genAA	3.588	1.766	2.03	0.04553	*
ed2	4.931	0.895	5.51	<0.0001	*
subtype	0.192	0.833	0.23	0.81868	
genAG:ed2	-4.022	0.988	-4.07	0.00011	*
genAA:ed2	-4.937	1.223	-4.04	0.00012	*

Table 6.22.: *Model 4e, age of onset (coefficients)*

Model 4e would explain 25.91% of the total variance ($R^2 = 0.3108$ and $R^2_{adj} = 0.2591$), the global F-test of the null hypothesis that all coefficients are zero would be significant ($F_{(6,80)} = 6.01$, $p < 0.0001$). The AIC of the model would be 486.61.

Test of Model Assumptions for Model 4d

The graphical test of linearity is not depicted here, because, obviously, it looks very similar to model 4b (as only one subject has been excluded). Still, the residuals deviated significantly from normality, as the Shapiro-Wilk test showed ($p = 0.0015$). The Harrison-McCabe test did not show any sign of heteroscedasticity ($HMC = 0.555$, $p = 0.761$); the Durbin-Watson statistic was non-significant, indicating that there were no auto-correlations ($DW = 1.624$, $p = 0.072$), and the variance inflating factors (Fox & Monette, 1992) were below the critical value of 4 for each predictor ($GVIF <= 2.645$).

6.6. Application of Odds Ratio Multifactor Dimensionality Reduction

As an alternative to conditional logistic regression, the possible gene–environment interactions of the 5-HT_{2A} –1438 G/A polymorphism and the five environmental subdomains, with respect to the diagnosis (anorexia nervosa vs. no eating disorder), were analyzed with odds ratio multifactor dimensionality reduction (OR MDR).

6.6.1. Recoding of Environmental Risk Factors

In order to render the application of the odds ratio multifactor dimensionality reduction (OR MDR) method possible, the continuous environmental risk factors must be recoded into dichotomous or polytomous variables with three levels.

The levels (categories) of the resulting variables (named $ed1r$ to $ed5r$) have been determined with the help of a conditional logistic regression model that has been estimated for each environmental domain. In each model, the environmental domain in question is the only predictor for the disease status (eating disorder vs. no eating disorder). But the model does not use the environmental domain as a linear predictor – instead it treats it as if it were a variable of nominal scale. As a consequence, a number of dummy variables are created and their coefficients are estimated by the regression model. The odds ratio of each coefficient is the increase in risk relative to the reference category (which is "not feeling distressed by any environmental risk event of the environmental domain in question", e. g. $ed1 = 0$).

This reference category (e. g. $ed1 = 0$) was also the reference category for the new variable (e. g. $ed1r = 0$). The next higher levels of the environmental domain that had similar odds ratios were combined to category 1 of the new variable (e. g. $ed1r = 1$, which will be referred to as "moderate exposure"), while all other environmental risk exposures

with even higher odds ratios were combined to category 2 (e. g. $ed1r = 2$, referred to as "high exposure").

For environmental domain 1 (parental problems), environmental exposures of 1, 2, or 3 events were recoded to category 1 ($ed1r = 1$).

For environmental domains 2 and 5 (disruptive events and family dieting environment), having experienced 1 or 2 distressing events in that domain was recoded to $ed2r = 1$ and $ed5r = 1$, respectively.

For the remaining domains 3 and 4 (parental psychiatric disorder and interpersonal problems), exposure to 1 distressing event was recoded into level 1 for the new variable ($ed3r = 1$, $ed4r = 1$). Having experienced more than 1 event was coded as category 2.

6.6.2. Finding the Best Factor Models

The OR MDR method allows to specify the number of factors that are combined to predict the outcome of interest, in this case the diagnosis. As this number is not known in advance, a range of values was used, and the resulting models are listed. In the following, the model calculations for two, three, and four factors are described.

In each application of the OR MDR method, the result is one best model. This best model is the model with the lowest prediction error and the highest cross-validation consistency (the number of times that the model has been identified as the best model in the 10 cross-validation trials, so the number can range from 0 to 10). Additionally, the OR MDR method returns a table of all models that were found in the 10 cross-validations. In the best case scenario, the best model is found in all 10 trials. If this is not the case, all other identified models are also reported.

Like in most data mining algorithms, chance findings are possible. For this reason, the OR MDR analysis is performed not only once, but 10000 times (in this paper). The reported results consist of two tables. The first table is the table of best models: It specifies how many times a specific model has been identified as the best model. In this table, the mean cross-validation consistency is also reported for the model(s).

The second table contains the results of the 10-fold cross-validations of the 10000 OR MDR applications. Like the first table it contains the absolute and relative frequencies with which the model was identified as the best model (in the 10-fold cross-validations), and it reports the mean training error (percentage of wrong classifications in the training set) and the mean test error (prediction error: percentage of wrong classifications in the holdout set) for the models.

Two Factor Model

The two factor model revealed the combination of the environmental factors disruptive events ($ed2r$) and interpersonal problems ($ed4r$) as best model to predict disease status. In all of the 10000 trials, this model resulted as the best model. The model has a cross-validation consistency of 9.4, meaning that is was selected 9.4 times in the 10-fold cross-validation, on the average (table 6.23). Considering the 100000 cross-validations, this model was selected 94.26% of the times (table 6.24). So the results for the two factor case are unequivocal.

The model does not contain an effect of the 5-HT_{2A} –1438 G/A polymorphism.

	Frequency	Percentage	CV-consistency
ed2r × ed4r	10000	1	9.4262

Table 6.23.: *Two factor model (best models)*

	Frequency	Percentage	Avg.train.er	Avg.test.er
ed1r × ed3r	2	0.00002	0.3375	0.7778
ed1r × ed4r	140	0.00140	0.3344	0.6093
ed1r × ed5r	4103	0.04103	0.3320	0.4975
ed2r × ed4r	94262	0.94262	0.3193	0.3294
ed2r × ed5r	27	0.00027	0.3329	0.6577
ed4r × ed5r	1466	0.01466	0.3236	0.5235

Table 6.24.: *Two factor model (cross validation models)*

For an easier interpretation, the odds ratios of the factor combinations are listed below. For some of the combinations the risk ratio cannot be estimated, because the combination does not contain enough subjects. The highest risk (odds ratio) is observed in the group of subjects exposed to high frequencies of disruptive events ($ed2r = 2$) and to no interpersonal problems ($ed4r = 0$). In this combination the asymptotic 95% confidence interval does not contain the value of one, indicating that this combination has a significant effect, increasing the risk of anorexia nervosa by the factor of 3.667, relative to the overall ratio of cases to controls.

The only other odds ratio that exceeds one is the combination of moderate exposure to both disruptive events and interpersonal problems ($ed2r = 1$, $ed4r = 1$). This combination also increases the risk of being affected by the disorder (by a factor of 2.111). Not being distressed by any events in those two categories has an odds ratio of 0.4, the confidence interval not including one. So this combination has a lower risk of developing anorexia nervosa. This means that the ratio of cases to controls in this combination is lower than

the ratio of the whole sample.

When the ranking of the odds ratios is considered (table 6.25), it can be seen that the risk is elevated whenever one of the environmental exposures is high ($ed2r = 2$ and/or $ed4r = 2$). The odds ratio of the group that shows high exposure in both groups cannot be estimated, as this combination only contains cases and no controls (which is also a conclusive fact).

Furthermore, when a subject shows moderate exposure in both domains, the risk is also elevated. In the author's view, this may constitute an interaction.

On the whole, the prediction error (average test error) is 0.3294, meaning that 32.94% of the subjects in the holdout set are classified incorrectly. This number is notably below the classification error that would be expected by chance, where 50% of the subjects would be classified incorrectly.

	ed2r	ed4r	Cases	Cont.	Cat.	OR	lower .95	upper .95
1	0	1	4	11	Low	0.364	0.201	1.099
2	0	0	10	25	Low	0.4	0.27	0.783
3	1	0	14	24	Low	0.583	0.391	1.052
4	1	2	15	11	High	1.364	0.755	2.803
5	2	1	5	3	High	1.667	0.538	6.765
6	1	1	19	9	High	2.111	1.098	4.41
7	0	2	9	3	High	3	0.968	10.717
8	2	0	11	3	High	3.667	1.183	12.7
9	2	2	2	0	High	Inf	NaN	Inf

Table 6.25.: *Two factor model (odds ratios)*

Three Factor Model

For the three factor model, the results are not as conclusive as for the two factor model. Here the 10000 applications of the (OR) MDR method revealed two models, both of them resulting with almost equal frequency (table 6.26). Interestingly, the model that resulted more often ($ed1r \times ed2r \times ed4r$) has the lower average cross-validation consistency (4.4 times out of 10). The other model ($ed2r \times ed4r \times ed5r$) has a cross-validation frequency of 4.9. This second model also has a lower prediction error and occurs more often in the 10-fold cross validations (table 6.27).

Analogous to the two factor model, no genetic influence of the 5-HT_{2A} –1438 G/A polymorphism is reported by any three factor model.

The odds ratios will be estimated for both models. The first model ($ed1r \times ed2r \times ed4r$) includes the environmental domains parental problems, disruptive events, and interpersonal

	Frequency	Percentage	CV-consistency
ed1r × ed2r × ed4r	5084	0.5084	4.4079
ed1r × ed2r × ed5r	562	0.0562	4.0801
ed2r × ed4r × ed5r	4354	0.4354	4.9118

Table 6.26.: *Three factor model (best models)*

	Frequency	Percentage	Avg.train.er	Avg.test.er
ed1r × ed2r × ed4r	36385	0.36385	0.2724	0.3996
ed1r × ed2r × ed5r	21333	0.21333	0.2717	0.4415
ed1r × ed3r × ed5r	7	0.00007	0.2768	0.7143
ed1r × ed4r × ed5r	1700	0.01700	0.2730	0.5428
ed2r × ed3r × ed4r	6	0.00006	0.2719	0.6019
ed2r × ed4r × ed5r	39556	0.39556	0.2676	0.4130
gen1 × ed2r × ed4r	1013	0.01013	0.2751	0.5206

Table 6.27.: *Three factor model (cross validation models)*

problems. The results of the OR MDR estimation are shown in table 6.28. The frequencies in the cells are rather low, leading to large confidence intervals. Only three intervals do not contain the value of one (indicating a significant odds ratio). The first is for the combination of no exposure to any environmental risk factors, where the odds ratio is below one (in this category, the ratio of cases to controls is lower than in the whole sample). The other confidence intervals are higher than one.

The combinations of environmental risk factors that have an odds ratio significantly greater than one are: high exposure to environmental domains 1 and 4 and moderate exposure to domain 2, and moderate exposure to all three of those factors.

A three-way interaction is hard to judge, and the results have to interpreted critically, as the number of subjects per combination is low. The results might not be very stable.

The prediction error of this model is 44.15%, only somewhat lower than would be expected from chance.

The second model ($ed2r \times ed4r \times ed5r$) includes the environmental domains disruptive events, interpersonal problems, and family dieting environment. Table 6.29 reveals that the frequencies in all combinations are relatively low, resulting in large confidence intervals. Some of the cells do not contain any observations at all, so no odds ratio estimations are possible.

Only two intervals do not include the value of one. One of them is for low exposure to environmental risk (no events in the domains disruptive events and interpersonal problems, moderate exposure in the family dieting environment), where the odds ratio is lower than

	ed1r	ed2r	ed4r	Cases	Cont.	Cat.	OR	lower .95	upper .95
1	0	0	1	0	2	Low	0	NaN	NaN
2	0	0	2	0	1	Low	0	NaN	NaN
3	0	0	0	2	12	Low	0.167	0.095	0.723
4	2	1	0	1	4	Low	0.25	0.094	2.193
5	2	0	1	1	4	Low	0.25	0.094	2.193
6	1	0	0	5	11	Low	0.455	0.252	1.255
7	2	2	1	1	2	Low	0.5	0.125	5.416
8	1	0	1	3	5	Low	0.6	0.25	2.435
9	1	1	0	8	13	Low	0.615	0.357	1.412
10	1	1	2	5	8	Low	0.625	0.313	1.837
11	0	1	1	2	3	Low	0.667	0.215	3.894
12	0	1	0	5	7	Low	0.714	0.341	2.166
13	0	2	0	1	1	High	1	0.141	15.741
14	0	1	2	1	1	High	1	0.141	15.741
15	2	0	0	3	2	High	1.5	0.375	8.762
16	2	0	2	5	2	High	2.5	0.625	12.549
17	1	1	1	14	5	High	2.8	1.165	7.445
18	2	1	1	3	1	High	3	0.423	28.295
19	1	2	1	4	1	High	4	0.563	35.088
20	2	1	2	9	2	High	4.5	1.125	20.244
21	1	2	0	5	1	High	5	0.704	41.943
22	2	2	0	5	1	High	5	0.704	41.943
23	1	0	2	4	0	High	Inf	NaN	Inf
24	2	2	2	2	0	High	Inf	NaN	Inf
25	0	2	1	0	0		NaN	NaN	NaN
26	0	2	2	0	0		NaN	NaN	NaN
27	1	2	2	0	0		NaN	NaN	NaN

Table 6.28.: *Three factor model A (odds ratios)*

one, indicating a lower cases to controls ratio than in the whole sample. The other one is for the combination of moderate exposure in all three domains, where the odds ratio is significantly higher than one (higher case–control ratio than in the whole sample).

In the author's point of view, no noticeable pattern that would account for an interaction can be found in the table of odds ratios. Generally viewed, the risk for having developed anorexia nervosa (odds ratio) rises with the exposure to environmental risk.

The prediction error of the model $ed2r \times ed4r \times ed5r$ is 41.3%, slightly lower than expected by chance.

	ed2r	ed4r	ed5r	Cases	Cont.	Cat.	OR	lower .95	upper .95
1	0	1	0	0	4	Low	0	NaN	NaN
2	0	1	2	0	3	Low	0	NaN	NaN
3	0	0	0	2	9	Low	0.222	0.116	1
4	1	0	0	2	8	Low	0.25	0.125	1.145
5	0	0	1	4	12	Low	0.333	0.189	0.994
6	1	0	1	6	14	Low	0.429	0.254	1.065
7	1	1	0	2	4	Low	0.5	0.188	2.661
8	2	1	1	2	3	Low	0.667	0.215	3.894
9	1	2	0	4	4	High	1	0.375	3.875
10	0	1	1	4	4	High	1	0.375	3.875
11	0	2	1	2	2	High	1	0.25	6.944
12	0	0	2	4	4	High	1	0.375	3.875
13	1	2	1	6	4	High	1.5	0.563	5.135
14	1	2	2	5	3	High	1.667	0.538	6.765
15	2	0	2	5	2	High	2.5	0.625	12.549
16	1	0	2	6	2	High	3	0.75	14.465
17	1	1	1	13	4	High	3.25	1.22	9.584
18	2	0	1	4	1	High	4	0.563	35.088
19	1	1	2	4	1	High	4	0.563	35.088
20	0	2	2	4	1	High	4	0.563	35.088
21	2	0	0	2	0	High	Inf	NaN	Inf
22	2	1	0	2	0	High	Inf	NaN	Inf
23	0	2	0	3	0	High	Inf	NaN	Inf
24	2	2	1	1	0	High	Inf	NaN	Inf
25	2	1	2	1	0	High	Inf	NaN	Inf
26	2	2	2	1	0	High	Inf	NaN	Inf
27	2	2	0	0	0		NaN	NaN	NaN

Table 6.29.: *Three factor model B (odds ratios)*

Four Factor Model

In the case of four factors, the 10000 (OR) MDR trials lead to a unique solution, as tables 6.30 and 6.31 show. The model contains only environmental risk domains, namely parental problems ($ed1r$), disruptive events ($ed2r$), interpersonal problems ($ed4r$), and family dieting environment ($ed5r$). The cross-validation consistency is 9.2 (a relatively high value). The prediction error is 37%, considerably lower than expected by chance.

Again, the model does not contain an effect of the 5-HT_{2A} –1438 G/A polymorphism. As the frequencies in the cells are obviously lower than in the three factor model (because of the additional variable in the model), the odds ratios are not listed for the model. Their confidence intervals are quite large.

	Frequency	Percentage	CV-consistency
ed1r × ed2r × ed4r × ed5r	10000	1	9.2234

Table 6.30.: *Four factor model (best models)*

	Frequency	Percentage	Avg.train.er	Avg.test.er
ed1r × ed2r × ed3r × ed4r	23	0.00023	0.2178	0.5734
ed1r × ed2r × ed3r × ed5r	5520	0.05520	0.2080	0.4628
ed1r × ed2r × ed4r × ed5r	92234	0.92234	0.2003	0.3703
gen1 × ed1r × ed2r × ed4r	6	0.00006	0.2219	0.6481
gen1 × ed1r × ed2r × ed5r	1559	0.01559	0.2112	0.5364
gen1 × ed1r × ed4r × ed5r	655	0.00655	0.2125	0.5454
gen1 × ed2r × ed4r × ed5r	3	0.00003	0.2188	0.6296

Table 6.31.: *Four factor model (cross validation models)*

Conclusion

The two factor model seems to describe the data best, as it has the lowest prediction error. As mentioned in section 4.4.4, it cannot be judged whether the included variables (disruptive events and interpersonal problems) act as two separate main effects or if they constitute an interaction effect.

The three and four factor models also have a prediction error that is lower than expected by chance (50%), so the detected effects seem valuable for a prediction of disease status.

The results of the odds ratio multifactor dimensionality reduction (OR MDR) method can be interpreted with the resulting odds ratios, but there are two disadvantages here (an in-depth discussion will follow in section 7.4). First, the two, three, and four factor models are reported one at a time, with the odds ratios only listed for each model separately – thereby making a combined interpretation of all findings not easily possible.

Second, the predictors in the (OR) MDR analysis are only of nominal scale. This makes interpretation difficult.

Another possible way to evaluate the results of the (odds ratio) multifactor dimensionality reduction is to use a conditional logistic regression model. The model included the interaction terms the (OR) MDR method revealed, along with all the necessary main effects and subordinate interaction terms. The model equation is shown in equation (6.12). It does not use the recoded versions of the environmental risk variables, as this is not necessary for the regression model. The estimated coefficients are listed in table 6.32.

$$
\begin{aligned}
g_{an}(\boldsymbol{x}) = \quad & \beta_{ed1} \cdot x_{ed1} + \beta_{ed2} \cdot x_{ed2} + \beta_{ed4} \cdot x_{ed4} + \beta_{ed5} \cdot x_{ed5} + \\
& \beta_{ed2 \times ed4} \cdot x_{ed2} \cdot x_{ed4} + \\
& \beta_{ed2 \times ed5} \cdot x_{ed2} \cdot x_{ed5} + \\
& \beta_{ed4 \times ed4} \cdot x_{ed4} \cdot x_{ed5} + \\
& \beta_{ed1 \times ed2 \times ed4} \cdot x_{ed1} \cdot x_{ed2} \cdot x_{ed4} + \\
& \beta_{ed1 \times ed2 \times ed5} \cdot x_{ed1} \cdot x_{ed2} \cdot x_{ed5} + \\
& \beta_{ed1 \times ed4 \times ed5} \cdot x_{ed1} \cdot x_{ed4} \cdot x_{ed5} + \\
& \beta_{ed2 \times ed4 \times ed5} \cdot x_{ed2} \cdot x_{ed4} \cdot x_{ed5} + \\
& \beta_{ed1 \times ed2 \times ed4 \times ed5} \cdot x_{ed1} \cdot x_{ed2} \cdot x_{ed4} \cdot x_{ed5}
\end{aligned} \tag{6.12}
$$

	$\hat{\beta}$	OR	SE($\hat{\beta}$)	$\text{OR}_{\text{lower .95}}$	$\text{OR}_{\text{upper .95}}$	p value	
ed1	0.0637	1.066	0.185	0.7416	1.532	0.730	
ed2	0.8465	2.332	0.490	0.8921	6.094	0.084	
ed4	1.0227	2.781	0.546	0.9545	8.100	0.061	
ed5	0.4429	1.557	0.294	0.8745	2.773	0.130	
ed2:ed4	-1.3342	0.263	0.743	0.0614	1.130	0.073	
ed2:ed5	-0.0865	0.917	0.258	0.5526	1.522	0.740	
ed4:ed5	-0.2916	0.747	0.432	0.3205	1.742	0.500	
ed1:ed2:ed4	0.4142	1.513	0.196	1.0311	2.220	0.034	*
ed1:ed2:ed5	0.0198	1.020	0.055	0.9162	1.136	0.720	
ed1:ed4:ed5	0.0368	1.038	0.106	0.8426	1.278	0.730	
ed2:ed4:ed5	0.4600	1.584	0.398	0.7260	3.456	0.250	
ed1:ed2:ed4:ed5	-0.1545	0.857	0.096	0.7096	1.035	0.110	

Table 6.32.: *Conditional logistic regression model of MDR results*

The estimated coefficients show that none of the main effects remain significant. Environmental domains disruptive events ($ed2$) and interpersonal problems ($ed4$) are borderline significant, with both odds ratios higher than two. If interpreted, each distressing event in the domains disruptive events ($ed2$) and interpersonal problems ($ed4$) would increase the risk of anorexia nervosa by the factor of two, approximately.

The two-way interaction found by the two factor model of disruptive events and interpersonal problems ($ed2 \times ed4$) is also only borderline significant. Interestingly, the odds ratio is below one. This means that when distressing events by one of the two domains are experienced, additional events of the other domain do not seem to be as severe as when occurring alone (if the fact that the coefficients are not significant was ignored). The other

two-way interactions are clearly non-significant. They have to remain in the model so that the higher-order interactions can be estimated correctly.

Of the two three-way interactions that were revealed by the three factor models, only one is significant ($ed1 \times ed2 \times ed4$). When parental problems ($ed1$), disruptive events ($ed2$), and interpersonal problems ($ed4$) occur together, the risk for anorexia nervosa increases more dramatically as the (borderline significant) main effects of these domains would suggest.

The coefficient of the four-way interaction is not significant.

The model explains 19.65% of the total variance and 39.31% of the variance that can possibly be explained ($R^2_L = 0.1965$, $R^2_{L,(max)} = 0.5$). It differs significantly from the null-model (likelihood ratio test, $\chi^2_{(12)} = 38.95$, $p = 0.00011$) and has an AIC of 108.43.

Neither of the (OR) MDR models identified a genetic effect. Also, the domain parental psychiatric disorder ($ed3$) did not have a significant impact on the risk of developing anorexia nervosa.

The application of the (OR) MDR method revealed some interaction effects that would not have been found with the regression model alone, as a full model including all two-way, three-way, and four-way interactions would have too many parameters to use backward stepwise selection to identify them.

On the other hand, the results of the (OR) MDR analysis are not easy to interpret, as it cannot be judged whether the models are based only on an interaction or if they include also main effects. Furthermore, the odds ratio estimations have quite large confidence intervals, as the sample size is relatively low. Another disadvantage is that the predictors used in the MDR analysis are only categorial.

Combining both methods is quite beneficial, as possible main and interaction effects of higher order can be evaluated in a regression model without having to use a backward stepwise selection process for model building.

7. Interpretation and Discussion

To the author's knowledge, this is the first study of possible interactions of the 5-HT_{2A} –1438 G/A polymorphism with environmental risk factors in anorexia nervosa. This section will attempt to embed the reported results in the current state of research.

7.1. Association of 5-HT_{2A} –1438 G/A Polymorphism with Anorexia Nervosa

In a number of studies of the role of the 5-HT_{2A} –1438 G/A polymorphism in anorexia nervosa, the 5-HT_{2A} –1438 A allele and/or the AA genotype has been found to be associated with the disorder (Collier et al., 1997; Sorbi et al., 1998; Nacmias et al., 1999; Ricca et al., 2002, 2004). In other studies, only a statistical trend towards association was found (Rybakowski et al., 2006), while some did not find any association (Hinney et al., 1997; Campbell et al., 1998; Kipman et al., 2002). A meta-analysis including nine studies (four in favor of the association, five not supporting this hypothesis) found an increased frequency of the –1438 A allele in anorexia nervosa (Gorwood et al., 2003).

In this sample, the association of the 5-HT_{2A} –1438 A allele with anorexia nervosa could not be replicated. The analysis showed that the frequency of the A allele was not significantly higher in the group of anorexic subjects (total sample) compared to controls. Some earlier studies reported elevated frequencies only in the subgroup of patients with anorexia nervosa restricting subtype (Sorbi et al., 1998; Nacmias et al., 1999; Ricca et al., 2002, 2004), so the analysis was also carried out for this subgroup separately. Again, the association of the A allele with anorexia nervosa could not be replicated.

The same results have been found when the frequencies of the genotypes were compared. The AA genotype was not significantly associated with anorexia nervosa (total sample), nor did the results support an association of the AA genotype in the restricting subgroup.

Interpreting this result, it always has to be kept in mind that the effect of a single gene can only be a small contribution in the susceptibility for a complex disorder like anorexia nervosa. So non-replications in this field are quite frequent (Hinney et al., 2000).

Another fact that must be considered is that the frequency of alleles and genotypes varies in different populations (Hinney et al., 2000). This is the reason why in other studies ethnic admixture is a possible explanation for a significant association (Witte et al., 1999). The study in hand used a sample of discordant sister pairs, so the possibility of ethnic admixture is one point of criticism that does not apply here. Therefore, the non-replication of the association of the 5-HT_{2A} –1438 G/A polymorphism and anorexia nervosa (total and restricting subgroup) in this sample may indicate a higher probability that the positive findings were false-positives. This interpretation is supported by the fact that other studies that used the transmission disequilibrium test (which is not affected by ethnic admixture) also could not replicate an association of the 5-HT_{2A} –1438 G/A polymorphism (e.g. Kipman et al., 2002; Gorwood et al., 2002).

On the other hand, a few points of criticism apply to the test of association in this study. Using (only) the diagnostic criteria for measuring the outcome may not be sufficient to allow meaningful associations. Clinical heterogeneity might contaminate the results (Ricca et al., 2004). It might be more beneficial to use a continuous psychopathological measure to identify an association.

7.2. Interaction of 5-HT_{2A} –1438 G/A Polymorphism and Environmental Risk Factors

7.2.1. Anorexia Nervosa and Gene–Environment Interaction

The possible genetic and environmental effects and their potential interactions, with respect to the diagnosis of anorexia nervosa, have been assessed with two conditional logistic regression models (model 1b and model 2b) and with the odds ratio multifactor dimensionality reduction method.

Conditional Logistic Regression Model Results

Neither of the two regression models found a genetic main effect of the 5-HT_{2A} –1438 G/A polymorphism on the predicted outcome (anorexia nervosa vs. no eating disorder).

Model 1b. Model 1b used a global measure of environmental risk and found, as expected, a significant influence of environmental risk factors. The more distressed by environmental events an individual feels, the higher the risk of having developed an eating disorder of type anorexia nervosa.

No significant interaction effects of the genetic and environmental factors could be found with model 1b.

Model 2b. Model 2b separated the environmental risk factor into five subdomains. This model found significant environmental main effects of three domains: disruptive events, interpersonal problems, and family dieting environment. Distressing events in those domains increase the risk of having developed anorexia nervosa.

The domains parental problems and parental psychiatric disorder did not have a significant impact on the risk for the disorder.

Concordant to model 1b, no significant interaction effect of environmental risk factors and the 5-HT$_{2A}$ –1438 G/A polymorphism could be detected with model 2b. But it must be noted that the interaction effect of family dieting environment and 5-HT$_{2A}$ –1438 AG genotype was borderline significant. This non-significance is possibly due to a lack of statistical power because of insufficient sample size. Neglecting the non-significance, the odds ratio of the interaction coefficient would show that subjects with the –1438 AG genotype were less negatively influenced by distressing family dieting environment.

Inclusion of a covariate for the anorexia nervosa subtype (restricting vs. binge-eating/purging) was not possible in model 1b or 2b. The maximum likelihood estimations did not converge in the model with the additional covariate, resulting in extremely high standard errors. The reason for this may be insufficient sample size (not enough discordant pairs in the included covariates).

Neither was it possible to perform a separate conditional logistic regression analysis in the subsample of anorexic subjects of restricting subtype and their healthy sisters (for the same reasons).

Still, the models explain a rather large part of the explainable variance (31.93% for model 1b and 36.69% for model 2b), so the model fit seems sufficient.

Criticism of Regression Models. One possible point of criticism related to the regression analysis is the assumption of linearity of the covariates for the environmental subdomains. In the framework of the regression model, it is postulated that each additional environmental event or circumstance that a subject is distressed by increases the risk in a linear fashion. The increase in risk is presumed to be the same whether the environmental risk changes from no distressing event to one event, or from four events to five events. This is relatively unlikely, so further research may benefit from the use of different models (or a mathematical transformation of the covariates).

Results of the Odds Ratio Multifactor Dimensionality Reduction Analysis

The odds ratio multifactor dimensionality reduction analysis did not reveal any interaction of environmental risk factors with the 5-HT_{2A} –1438 G/A polymorphism. Instead, it revealed four possible interaction effects among the environmental risks.

In the (OR) MDR analysis, the number of predictors that are considered must be specified. As this number is not known in advance, the analysis has been performed three times (with two, three, and four predictors).

The two factor model revealed a possible interaction of disruptive events and interpersonal problems.

The three factor model indicated that there are two possible three-way interactions; disruptive events, interpersonal problems, family dieting; and parental problems, disruptive events, interpersonal problems.

The four factor model suggested an interaction between parental problems, disruptive events, interpersonal problems, and family dieting environment.

Due to methodological issues (that will be discussed in section 7.4), the interpretation of these possible interaction effects is not an easy task. This is why a conditional logistic regression analysis was performed. The model included the presumed interactions and all necessary lower-order effects.

The results of the conditional logistic regression analysis showed that the two-way interaction of disruptive events and interpersonal problems was borderline significant. The odds ratio for the coefficient was below one, indicating that the combined occurrence of distressing events in these two domains is not as severe as suggested by their main effects.

Of the suggested three-way interactions, only the coefficient of parental problems, disruptive events, and interpersonal problems was significant. The odds ratio was greater than one. This means that when events in two of these domains are present, an additional occurrence of a distressing event in the third domain has a more severe influence on the risk for anorexia nervosa than the main effects would suggest.

The four-way interaction of parental problems, disruptive events, interpersonal problems, and family dieting environment was not significant.

Of the main effects (that must be included in the model to estimate the coefficients of the interactions correctly), none were significant. The effects of disruptive events and interpersonal problems were only borderline significant, although their odds ratios were quite high (the relative risk was increased by the factor of more than two).

The regression model explains 39.31% of the explainable variance. This is about as much as the regression models that have been deduced by a backward stepwise selection

process. Nevertheless, the (OR) MDR method revealed interactions among environmental risk factors that would not have been found with the regression models alone (partly because they were not searched for, and partly because the number of parameters would have been to high for coefficient estimation).

Criticism of Analysis

As previously noted, it may be more promising to use a continuous measure instead of the clinical diagnosis, as the clinical diagnosis may be contaminated by heterogeneity, thereby clouding possible genetic effects on specific subpopulations that are united into a single diagnostic category.

It may also be beneficial to perform an analogous analysis on the restricting subgroup only. This would require a larger sample be collected. Also, in a larger sample, the interaction effect of family dieting environment and the 5-HT$_{2A}$ –1438 G/A polymorphism may prove significant.

It has been previously discussed that there is a possibility the 5-HT$_{2A}$ –1438 G/A polymorphism is not related to anorexia nervosa in a direct way. The effects of the polymorphism may be contributory for an underlying personality trait that moderates the risk of development of the disorder. Herbeth et al. (2005) mentioned perfectionism and obsessionality, and Rybakowski et al. (2006) adduced high harm avoidance and low reward dependence. Future research should address this possibility.

7.2.2. Body Mass Index and Gene–Environment Interaction

One of the hypotheses in this paper is that the 5-HT$_{2A}$ –1438 G/A polymorphism and its interaction with environmental risk factors may be a moderator for the severity of anorexia nervosa. Other studies (e. g. Kipman et al., 2002; Gorwood et al., 2003) did not find an association of life-time severity (also assessed by lowest body mass index) and the 5-HT$_{2A}$ –1438 G/A polymorphism. In this paper, the question was investigated by means of a linear regression model (model 3b) that predicted the lowest body mass index, which can be used as an indicator of the severity of the disease (Hebebrand et al., 1996). The model was based on the subjects suffering from anorexia nervosa – the healthy controls were not used for the analysis.

Main Effects. In the regression model (model 3b), no significant genetic main effects were found. The analysis in this study supports the findings mentioned above – the 5-HT$_{2A}$ –1438 G/A polymorphism does not seem to have an influence on the severity of anorexia

nervosa.

Concerning the environmental risk factors, only the domain of parental psychiatric disorders might have an influence on severity. It must be noted that the effect is only borderline significant, which could be attributed to the small sample size.

The subtype of the disorder (restricting or binge-eating/purging type) and the age of onset seem to be predictive of the severity of the disorder, as has been expected. Patients of the restricting subtype tend to have a lower body mass index, as well as patients with an earlier onset of the disease (as the body mass index is generally lower in younger people).

The analysis supports previous findings that the 5-HT_{2A} –1438 G/A polymorphism does not influence the severity of anorexia nervosa, as no main or interaction effect could be found in the regression model.

Criticism of Regression Model. A major point of criticism is that the analysis could not be carried out on a data set including only patients of the restricting subtype. This was hindered by the insufficient sample size. As the body mass index is generally lower for the restricting subtype, a separate analysis for the restricting subgroup (or on a bigger overall sample) would be desirable.

Another point of criticism that applies is the assumption of linearity of the predictors in the environmental subdomains (just as in models 1b and 2b). It may be more plausible that the environmental risk factors do not act linearly on the severity of the disorder. Future research should consider this.

7.2.3. Age of Onset and Gene–Environment Interaction

One study indicated that the 5-HT_{2A} –1438 G/A polymorphism might be related to the age of onset of anorexia nervosa. It was reported that the A allele was associated with a later age of onset (Kipman et al., 2002).

To investigate this matter, a linear regression model was used, predicting the age of onset in the sample of eating disordered patients. As the initial analysis was severely influenced by an outlier in the sample, the regression analysis was performed twice – based on the complete data set (model 4b) and based on a data set where that subject was excluded (model 4d).

Main Effects. The initial model including the outlier (model 4b) did not find a genetic main effect of the 5-HT_{2A} –1438 G/A polymorphism, but two environmental risk factors showed a significant influence on the age of onset. A later onset is related to having experi-

enced more distressing disruptive events, whereas events in the family dieting environment domain seem to lead to an earlier onset of the disorder.

Interactions. The model did not find any significant interactions of the 5-HT$_{2A}$ –1438 G/A polymorphism with environmental risk factors, but the coefficients for the interaction of AG and AA genotype with disruptive events were borderline significant (both interactions would lead to an earlier onset when significant).

The model explains 14.92% of the total variance, but there were deviations from normality of the residuals, so not all model assumptions were met.

Furthermore, as was mentioned above, the model was severely influenced by an outlier. The subject in question displayed very high scores for the environmental domains disruptive events and family dieting environment.

Exclusion of Outlier. So the analysis was repeated, excluding the outlier. The resulting model (model 4d) found a significant main effect of disruptive events. The later the onset of anorexia nervosa, the more the individual feels distressed by disruptive events. The main effect of the family dieting environment domain was no longer significant in model 4d and was therefore excluded. The model found a significant main effect of the 5-HT$_{2A}$ –1438 AA genotype, delaying the onset. This is in accordance with the results of Kipman et al. (2002).

Furthermore, the interaction coefficients of AG and AA genotype and disruptive events (that were only borderline significant in model 4b) were now significant. For subjects with AG and AA genotypes, the age of onset is (almost) not influenced by distressing disruptive events (or vice versa).

Influence of Subtype. Inclusion of subtype did not change the models, so the subtype of the eating disorder (restricting vs. binge-eating/purging) does not seem to affect the age of onset or the other coefficients in the model.

Model 4d explains 26.77% of total variance, indicating a good fit for a regression model. But like in model 4b, the assumption of normality of the error was not met.

Criticism of the Regression Model. One major point of criticism is that the model does not predict well when the observed age of onset is high. This is especially severe for the subject that was excluded. Additionally, that subject also displayed very high environmental risk exposure. Due to scarce data with high environmental risk scores, the model was highly influenced by this outlier. This leads to another point of criticism: The

models do not predict well for (untypically) high environmental risk exposure. Of course, this is only natural as unduly extrapolation is always a problem in regression models. This is also key in the criticism of the assumption of linearity of the predictors (that has already been phrased for the preceding models in this paper, and also applies to this model).

7.3. Summary of Results

7.3.1. Environmental Risk Effects

Environmental risk factors play an important role in the field of anorexia nervosa. This was once again confirmed by the analysis in this paper.

Distressing events in the domains disruptive events, interpersonal problems, and family dieting environment seem to induce an increased risk for a diagnosis of anorexia nervosa. Furthermore, these environmental risk domains might interact with each other (as the odds ratio multifactor dimensionality reduction revealed). Disruptive events and interpersonal problems occurring conjointly might not be fully additive in their effects, as the risk does not increase as much as predicted by the main effects (although only borderline significant). On the other hand, when events of these two domains co-occur with an event of the parental problem domain, the risk for anorexia nervosa increases more than anticipated.

Life-time severity (assessed by lowest body mass index) seems to be influenced by the environmental domain of parental psychiatric disorder. More events in this domain seem to lead to a lower body mass index, indicating a higher severity.

7.3.2. Effects of the 5-HT_{2A} –1438 G/A Polymorphism

The analysis in this paper did not show a main effect of the 5-HT_{2A} –1438 G/A polymorphism, concerning the risk for anorexia nervosa (the outcome of interest being the diagnosis). This was shown with means of an association test, a conditional logistic regression model, and an odds ratio multifactor dimensionality reduction analysis.

The 5-HT_{2A} –1438 G/A polymorphism does not seem to have an effect on the life-time severity of anorexia nervosa. A linear regression model did not reveal any genetic effects in the prediction of lowest body mass index.

The AA and AG genotype seem to delay the onset of anorexia nervosa. This was shown in a linear regression analysis, after exclusion of an outlier. (The analysis also revealed interaction effects, see section 7.3.3.)

7.3.3. Interactions of Genetic and Environmental Factors

Although no main effect of the 5-HT_{2A} –1438 G/A polymorphism could be found when predicting the diagnosis (anorexia nervosa vs. no eating disorder), the analysis revealed an interaction that was borderline significant. The AG genotype seems to act in a protective way in the presence of distressing events in the family dieting domain. When distressed by those events (like family fitness influences, parental eating disorders or obesity, dieting with other family members before onset, or critical comments about shape, weight, or eating by family and others), the risk for subjects with AG genotype is not as high as for subjects with other genotypes.

Concerning the age of onset, the analysis revealed a significant interaction of the 5-HT_{2A} –1438 G/A polymorphism and environmental risk factors (after exclusion of an outlier). For subjects with AA and AG genotypes, the presence of disruptive events does (almost) not influence the age of onset of anorexia nervosa, compared to subjects with GG genotype that experience distress in this environmental domain (for them, a later onset correlates with more distressing events).

This might be surprising, as the AA genotype has been found to delay the onset as a main effect. But viewed in a wider context, this may not be as unrealistic as it seems. If a certain genetic configuration would only have negative effects, it would most likely be very rare. Individuals with that unfavorable genotype would be disadvantaged (from a evolutionary standpoint), and therefore the frequency of that genotype or the respective allele could be expected to be very low. As this is not the case for the –1438 A allele or AA genotype, the fact that this genotype can also act protectively might not be as startling as initially thought.

7.4. The Odds Ratio Multifactor Dimensionality Reduction Method

The odds ratio multifactor dimensionality reduction method certainly has some advantages (which have already been mentioned in section 4.4.4). It is non-parametric and model-free (no genetic model is assumed). Most of all, it facilitates the simultaneous detection of high-order interactions among genetic and environmental factors.

A major point of criticism of the odds ratio multifactor dimensionality reduction method is the fact that the results cannot be easily interpreted. There are three facets that contribute to this problem. First, the model is not easy to disentangle. A reported four-way

interaction might consist of two separate two-way interactions, a two-way interaction and two separate main effects, or some other combination. With the results of the OR MDR method alone, this cannot be judged satisfactorily.

Second, the odds ratios reported by the OR MDR method are unorthodox. They do not represent an increase or decrease in risk relative to a specific reference category, but only an increased ratio of cases to controls (in a certain combination of risk factors) in relation to the case–control ratio in the whole sample. This may not be a severe problem for finding gene–gene interactions, but in case of interactions that only include environmental risk factors it is nonetheless awkward. Speaking of a "reduced" risk in a combination of zero exposure to two risk factors seems, in this context, inappropriate.

The third issue that complicates interpretation is the fact that the models for two, three and four factors (and all other computed models) are only reported separately. This makes a combined interpretation almost impossible. It can easily be judged which model is the better one by interpreting the prediction error, but how the findings of the models interact may not be judged by the means of the OR MDR method.

A criticism that does not only apply to OR MDR but also to conditional logistic regression analysis is that false positives and false negatives findings are possible when sample size is small. This is especially grave when only MDR is used. With the OR MDR method, the size of the odds ratio confidence intervals can reveal how assured the effect of a combination of risk factors is, but in order to obtain small confidence intervals, the sample size seems to have to be relatively big (especially when the number of predictors is high).

It has to be noted that the OR MDR method is a data-mining strategy. This means that it is predominantly used for exploratory data analysis. For this purpose it seems highly valuable, because it can identify (possible) interaction effects more easily than a regression model.

For interpretation, it has several limitations. In the author's opinion it is best to use a conditional logistic regression model that includes the interaction effects revealed by the MDR method to facilitate the understanding of the reported results.

7.5. Summary of Criticism and Suggestions for Future Research

Like any other empiric study, the analysis in this paper is subject to limitations and criticism. Some of the aspects have already been mentioned above. This section will attempt to give a summary, and will also suggest some improvements for future research.

To begin at a very basic level, the assessment of environmental risk factors is of major

importance for the results of the analysis and their interpretation. Although the Oxford Risk Factor Interview (ORFI) is a well-recognized instrument in the field of risk factor research in eating disorders, it seems questionable whether the impact of the various events in the subdomains is of equal importance. This could be assessed by means of a Rasch model.

Furthermore it is well-known that a (forced) dichotomous response (yes/no) yields a number of problems. This "forced choice" does not allow any nuances in the response, which might lead to the measure being biased by the phenomenon of reactance. This criticism could be diminished by the fact that the ORFI is conducted by a trained researcher and that this researcher judged whether the event or circumstance was distressing for the individual (which might also be criticized). Still, it might be beneficial to assess the influence of environmental risk on a finer scale (e. g. a Likert scale).

At a more technical level, criticism might be aroused by the assumptions of the regression models. Here, the environmental predictors are postulated to act in a linear fashion. Each event in an environmental domain is presumed to increase the risk by the same amount. This assumption is, clearly, unrealistic. It still can serve as a first approximation, but other models may be advantageous.

Moreover, exposure to a large number of distressing environmental risk events is not especially common. This means both that the model is highly influenced by subjects that show high exposure, and that the model does not predict well for high exposures.

The scores for the environmental risk factors could be recoded, as has been done for the odds ratio multifactor dimensionality reduction. Of course, such recoding is always more or less arbitrary and can also be criticized.

Future studies should also take into account that the effect of the 5-HT_{2A} –1438 G/A polymorphism might only be related to the subgroup of patients with restricting-type anorexia nervosa. This is especially important when predicting the disease status (diagnosis). The sample in this paper was too small to conduct a separate analysis on this subset of patients. Generally it has to be said that the results in this study are quite speculative, as the sample size is relatively small.

The regression model for predicting the age of onset revealed an outlier that influenced the model severely. In this paper, the outlier was excluded. It would be beneficial to repeat the analysis in a different sample.

It was already mentioned in this paper that the use of a categorial outcome measure (disease status: diagnosis of anorexia nervosa vs. no eating disorder) is not favorable. Heterogeneity may cloud some genetic effects. For future research it could be advantageous

to use a continuous measure of outcome.

Another approach of future research might be to investigate underlying personality traits (like perfectionism, obsessionality, harm avoidance, or reward dependence). It is highly probable that genetic variations do not influence disease status in a direct fashion. These personality traits (that are risk factors for anorexia nervosa) might be a link in the causal chain.

One last point of criticism applies to every empiric study. The reported significant effects might by false positives, resulting from type I errors. In the field of molecular genetic research, false positives (and also false negatives) are extremely common, because the effects of one single genetic polymorphism are usually very small. It is the task of future research to replicate the findings in this paper, or to dismiss them as false positives.

Part III.

Appendix

A. R Code

A.1. Libraries, Data Preperations, and Functions

```
#####################################################
### chunk number 1: preamble
#####################################################

library(foreign)
library(survival)      # clogit
library(MASS)          # stepAIC
library(xtable)            # LaTeX-Tabellen
library(UsingR)
library(car)   # recode
library(lmtest)
library(exactRankTests)


#workdir <- "E:/Uni-Psycho/Diplomarbeit/Projekt 5ht2a Esstoerung/Auswertung/"
                # Memory Stick Uni
workdir <- "G:/Uni-Psycho/Diplomarbeit/Projekt 5ht2a Esstoerung/Auswertung/"
                # Memory Stick zu Hause
#workdir <- "D:/_new/Uni-Psycho/Diplomarbeit/Projekt 5ht2a Esstoerung/Auswertung/"
           # zu Hause

origdatfile <- paste(workdir, "5ht2a-data-v006-q.csv", sep = "")     # Daten noch im
    "abhängigen" Format, zur Kontrolle der Variablen
filename <- paste(workdir, "5ht2a-data-v007-q.csv", sep = "")        # Daten für log.
    Reg

origdat <- read.csv2(origdatfile)
names(origdat) <- tolower(names(origdat))
dat <- read.csv2(filename)
names(dat) <- tolower(names(dat))

## ===========================================================================
## Erstellen der dichotomisierten Variablen für Umweltfaktoren
## ===========================================================================
  # evtl. schon bei der Datenaufbereitung?? und im File speichern?

envdom1d <- as.integer(dat$envdom1 > 0)
envdom2d <- as.integer(dat$envdom2 > 0)
```

```
envdom3d <- as.integer(dat$envdom3 > 0)
envdom4d <- as.integer(dat$envdom4 > 0)
envdom5d <- as.integer(dat$envdom5 > 0)
envdomtd <- as.integer(dat$envdomt > 0)

envdom1dm <- as.integer(dat$envdom1 > median(dat$envdom1))
envdom2dm <- as.integer(dat$envdom2 > median(dat$envdom2))
envdom3dm <- as.integer(dat$envdom3 > median(dat$envdom3))
envdom4dm <- as.integer(dat$envdom4 > median(dat$envdom4))
envdom5dm <- as.integer(dat$envdom5 > median(dat$envdom5))
envdomtdm <- as.integer(dat$envdomt > median(dat$envdomt))

dat <- data.frame(dat, envdom1d, envdom2d, envdom3d, envdom4d, envdom5d, envdomtd,
    envdom1dm, envdom2dm, envdom3dm, envdom4dm, envdom5dm, envdomtdm)
rm(envdom1d, envdom2d, envdom3d, envdom4d, envdom5d, envdomtd, envdom1dm, envdom2dm,
     envdom3dm, envdom4dm, envdom5dm, envdomtdm)

nomisdat <- subset(dat, dat$gencompl)    #89
orignomisdat <- subset(origdat, origdat$gencompl)

ed1 <- nomisdat$envdom1
ed2 <- nomisdat$envdom2
ed3 <- nomisdat$envdom3
ed4 <- nomisdat$envdom4
ed5 <- nomisdat$envdom5
edt <- nomisdat$envdomt

gen <- factor(nomisdat$ht2a1438)
levels(gen) <- c("GG", "AG", "AA")
nomisdat <- data.frame(nomisdat, gen, ed1, ed2, ed3, ed4, ed5, edt)

strpar <- function(a, b, rda=2, rdb=2) {
  a <- round(a, rda)
  b <- round(b, rdb)
  s1 <- as.character(a)
  s2 <- as.character(b)
  return(paste(a, " (", b, ")", sep=""))
}

strchitab <- function(chitestobj, cr=2, pr=5) {
  c1 <- as.character(round(chitestobj$statistic, cr))
  c2 <- as.character(round(chitestobj$parameter, 0))
  c3 <- as.character(round(chitestobj$p.value, pr))
  return(c(c1,c2,c3))
}

recode <-
function (var, recodes, as.factor.result, levels)
{
    recode.list <- rev(strsplit(recodes, ";")[[1]])
    is.fac <- is.factor(var)
```

```
    if (missing(as.factor.result))
        as.factor.result <- is.fac
    if (is.fac)
        var <- as.character(var)
    result <- var
    if (is.numeric(var)) {
        lo <- min(var, na.rm = TRUE)
        hi <- max(var, na.rm = TRUE)
    }
    for (term in recode.list) {
        if (0 < length(grep(":", term))) {
            range <- strsplit(strsplit(term, "=")[[1]][1], ":")
            low <- eval(parse(text = range[[1]][1]))
            high <- eval(parse(text = range[[1]][2]))
            target <- eval(parse(text = strsplit(term, "=")[[1]][2]))
            result[(var >= low) & (var <= high)] <- target
        }
        else if (0 < length(grep("else", term))) {
            target <- eval(parse(text = strsplit(term, "=")[[1]][2]))
            result[1:length(var)] <- target
        }
        else {
            set <- eval(parse(text = strsplit(term, "=")[[1]][1]))
            target <- eval(parse(text = strsplit(term, "=")[[1]][2]))
            for (val in set) {
                if (is.na(val))
                  result[is.na(var)] <- target
                else result[var == val] <- target
            }
        }
    }
    if (as.factor.result) {
        result <- if (!missing(levels))
            factor(result, levels = levels)
        else as.factor(result)
    }
    else if (!is.numeric(result)) {
        result.valid <- na.omit(result)
        save.warn <- options(warn=2)
        res <- try(if (length(result.valid) == length(grep("[0-9]",
result.valid)))
            as.numeric(result), silent=TRUE)
        result <- if (class(res) == "try-error" || is.null(res)) result
else res
        options(save.warn)
    }
    result
}

AIC.self <- function(mod) return(-2*mod$loglik[2] + 2*length(mod$coefficients))
```

```
eta <- function(aovobj) {
  SS <- anova(aovobj)$"Sum Sq"
  return(SS[1]/sum(SS)) #SS factor N/SS Total
}
condregtab <- function(coxpho) {
  taba <- summary(coxpho)$coef
  tabb <- summary(coxpho)$conf.int
  sig <- as.integer(taba[,5] < 0.05)
  sig <- recode(sig, "0='';1='*'")
  #tab <- cbind(taba[,c(1,2,3)], tabb[,c(3,4)], taba[,5])
  tabtmp <- cbind(taba[,c(1,2,3)], tabb[,c(3,4)], taba[,5])
  tab <- cbind(
    format(tabtmp[,1],digits=3, scientific=FALSE), # coef
    format(tabtmp[,2],digits=3, scientific=FALSE), # exp(coef)
    format(tabtmp[,3],digits=3, scientific=FALSE), # se(coef)
    format(tabtmp[,4],digits=3, scientific=FALSE), # lower .95 CI of exp(coef)
    format(tabtmp[,5],digits=3, scientific=FALSE), # upper .95 CI of exp(coef)
    format(tabtmp[,6],digits=5, scientific=FALSE)) # p-value
  tab <- cbind(tab, sig)
  rownames(tab) <- rownames(taba)
  colnames(tab) <- c("$\\hat \\beta$", "OR", "SE($\\hat \\beta$)", "OR\\subscr{lower
      .95}", "OR\\subscr{upper .95}", "$p$ value", " ")
  return(tab)
}

linregtab <- function(lmo) {
  tmptab <- summary(lmo)$coefficients
  sig <- as.integer(tmptab[,4] < 0.05)
  sig <- recode(sig, "0='';1='*'")
  tab <- cbind(
  format(tmptab[,1],digits=3),
  format(tmptab[,2],digits=3),
  format(tmptab[,3],digits=3),
  format(tmptab[,4],digits=10, scientific=FALSE)
  )
  tab[,4] <- substr(tab[,4], 0, 7)
  tab[,4] <- recode(tab[,4], "'0.00000'='<0.0001'")
  tab <- cbind(tab, sig)
  colnames(tab) <- c("$\\hat \\beta$", "Std. Error", "$t$ value", "$p$ value", " ")
  return(tab)
}

lrtest.self <- function(simp, comp) {
  chisq <- abs(-2*(simp$loglik[2] - comp$loglik[2]))
  dfd <- length(comp$coefficients) - length(simp$coefficients)
  lrt.p <- pchisq(chisq, abs(dfd), lower.tail=FALSE)
  return(c(chisq, dfd, lrt.p))
}
```

A.2. Description of Sample

```
#####################################################
### chunk number 2: samplesize
#####################################################
n.cases.all <- length(dat$uniqid[dat$outcome==1])
n.cont.all <-  length(dat$uniqid[dat$outcome==0])

n.cases <- length(nomisdat$uniqid[nomisdat$outcome==1])
n.cont <-  length(nomisdat$uniqid[nomisdat$outcome==0])

n.cases.london <- length(nomisdat$uniqid[(nomisdat$outcome==1) & (nomisdat$centid
    ==1)])
n.cases.vienna <- length(nomisdat$uniqid[(nomisdat$outcome==1) & (nomisdat$centid
    ==2)])
n.cases.spain <- length(nomisdat$uniqid[(nomisdat$outcome==1) & (nomisdat$centid==3)
    ])

n.anr <- length(nomisdat$uniqid[nomisdat$psdiag2==1])
n.anbp <- length(nomisdat$uniqid[nomisdat$psdiag2==2])

n.anr.london <- length(nomisdat$uniqid[(nomisdat$psdiag2==1) & (nomisdat$centid==1)
    ])
n.anr.vienna <- length(nomisdat$uniqid[(nomisdat$psdiag2==1) & (nomisdat$centid==2)
    ])
n.anr.spain <- length(nomisdat$uniqid[(nomisdat$psdiag2==1) & (nomisdat$centid==3)])

n.anbp.london <- length(nomisdat$uniqid[(nomisdat$psdiag2==2) & (nomisdat$centid==1)
    ])
n.anbp.vienna <- length(nomisdat$uniqid[(nomisdat$psdiag2==2) & (nomisdat$centid==2)
    ])
n.anbp.spain <- length(nomisdat$uniqid[(nomisdat$psdiag2==2) & (nomisdat$centid==3)
    ])



#####################################################
### chunk number 3: samplesize.tab
#####################################################
tab <- rbind(
  c(n.anr.london, n.anbp.london, n.cases.london),
  c(n.anr.vienna, n.anbp.vienna, n.cases.vienna),
  c(n.anr.spain, n.anbp.spain, n.cases.spain),
  c(n.anr, n.anbp, n.cases))

rownames(tab) <- c("London", "Vienna", "Barcelona", "Total")
colnames(tab) <- c("AN-R", "AN-BP", "AN (total)")

print(xtable(tab, label="tab.n", caption="Sample size per center and diagnose")) #
    align = ; sanitize.text.function = function(x) x
```

```
#################################################
### chunk number 4: age
#################################################
age.cases.mean <- mean(nomisdat$p1age[nomisdat$outcome==1], na.rm=TRUE)
age.cont.mean <- mean(nomisdat$p1age[nomisdat$outcome==0], na.rm=TRUE)

age.cases.min <- min(nomisdat$p1age[nomisdat$outcome==1], na.rm=TRUE)
age.cases.max <- max(nomisdat$p1age[nomisdat$outcome==1], na.rm=TRUE)
age.cont.min <- min(nomisdat$p1age[nomisdat$outcome==0], na.rm=TRUE)
age.cont.max <- max(nomisdat$p1age[nomisdat$outcome==0], na.rm=TRUE)

age.cases.sd <- sd(nomisdat$p1age[nomisdat$outcome==1], na.rm=TRUE)
age.cont.sd <- sd(nomisdat$p1age[nomisdat$outcome==0], na.rm=TRUE)

age.diff.mean <- mean(orignomisdat$p1age - orignomisdat$p1ages, na.rm=TRUE)
age.diff.sd <- sd(orignomisdat$p1age - orignomisdat$p1ages, na.rm=TRUE)


#################################################
### chunk number 5: figage
#################################################
simple.hist.and.boxplot(orignomisdat$p1age - orignomisdat$p1ages,
  main=NULL,
  breaks=seq(from=-15, to=10, by=2))


#################################################
### chunk number 6: edu
#################################################
momedu <- factor(orignomisdat$a33) # mother's highest education
levels(momedu) <- c("primary school", "professional degree", "secondary degree", "
    university degree")
dadedu <- factor(orignomisdat$a30)
levels(dadedu) <- c("primary school", "professional degree", "secondary degree", "
    university degree")

momage.mean <- mean(origdat$a11, na.rm=TRUE)
momage.sd <- sd(origdat$a11, na.rm=TRUE)
dadage.mean <- mean(origdat$a12, na.rm=TRUE)
dadage.sd <- sd(origdat$a12, na.rm=TRUE)

tab <- rbind(prop.table(table(momedu)), prop.table(table(dadedu)))
rownames(tab) <- c("Mother", "Father")

print(xtable(tab, label="tab:edu", caption="Highest education levels of parents",
    digits = 3))


#################################################
### chunk number 7: edrelatedaon
#################################################
```

```
# Age of Onset
aon.cases.mean <- mean(nomisdat$psedon[nomisdat$outcome==1], na.rm=TRUE)
aon.cases.sd <- sd(nomisdat$psedon[nomisdat$outcome==1], na.rm=TRUE)

aon.anr.mean <- mean(nomisdat$psedon[nomisdat$psdiag2==1], na.rm=TRUE)
aon.anr.sd <- sd(nomisdat$psedon[nomisdat$psdiag2==1], na.rm=TRUE)
aon.anr.min <- min(nomisdat$psedon[nomisdat$psdiag2==1], na.rm=TRUE)
aon.anr.max <- max(nomisdat$psedon[nomisdat$psdiag2==1], na.rm=TRUE)

aon.anbp.mean <- mean(nomisdat$psedon[nomisdat$psdiag2==2], na.rm=TRUE)
aon.anbp.sd <- sd(nomisdat$psedon[nomisdat$psdiag2==2], na.rm=TRUE)
aon.anbp.min <- min(nomisdat$psedon[nomisdat$psdiag2==2], na.rm=TRUE)
aon.anbp.max <- max(nomisdat$psedon[nomisdat$psdiag2==2], na.rm=TRUE)

tmp <- subset(nomisdat, psdiag2 != 0)
psedon.stat <- wilcox.test(tmp$psedon, tmp$psdiag2)$statistic
psedon.p <- wilcox.test(tmp$psedon, tmp$psdiag2)$p.value


#####################################################
### chunk number 8: edrelateddur
#####################################################
# Duration of illness
dur.cases.mean <- mean(nomisdat$dur[nomisdat$outcome==1], na.rm=TRUE)
dur.cases.sd <- sd(nomisdat$dur[nomisdat$outcome==1], na.rm=TRUE)

dur.anr.mean <- mean(nomisdat$dur[nomisdat$psdiag2==1], na.rm=TRUE)
dur.anr.sd <- sd(nomisdat$dur[nomisdat$psdiag2==1], na.rm=TRUE)

dur.anbp.mean <- mean(nomisdat$dur[nomisdat$psdiag2==2], na.rm=TRUE)
dur.anbp.sd <- sd(nomisdat$dur[nomisdat$psdiag2==2], na.rm=TRUE)

tmp <- subset(nomisdat, psdiag2 != 0)
dur.stat <- wilcox.test(tmp$dur, tmp$psdiag2)$statistic
dur.p <- wilcox.test(tmp$dur, tmp$psdiag2)$p.value


#####################################################
### chunk number 9: figduraon
#####################################################
#Age of Onset + Duration of Illness - Boxplots
par(mfrow=c(1,2))
boxplot(nomisdat$psedon[nomisdat$outcome==1] ~ nomisdat$psdiag2[nomisdat$outcome
    ==1], col="gray", names=c("AN-R", "AN-BP"), main="Age of onset")
boxplot(nomisdat$dur[nomisdat$outcome==1] ~ nomisdat$psdiag2[nomisdat$outcome==1],
    col="gray", names=c("AN-R", "AN-BP"), main="Duration of illness")


#####################################################
### chunk number 10: edrelatedbmi
#####################################################
```

```
#Body Mass Index (BMI) - current
bmi.cur.cases.mean <- mean(nomisdat$p1cbmi[nomisdat$outcome==1], na.rm=TRUE)
bmi.cur.cases.sd <- sd(nomisdat$p1cbmi[nomisdat$outcome==1], na.rm=TRUE)

bmi.cur.anr.mean <- mean(nomisdat$p1cbmi[nomisdat$psdiag2==1], na.rm=TRUE)
bmi.cur.anr.sd <- sd(nomisdat$p1cbmi[nomisdat$psdiag2==1], na.rm=TRUE)

bmi.cur.anbp.mean <- mean(nomisdat$p1cbmi[nomisdat$psdiag2==2], na.rm=TRUE)
bmi.cur.anbp.sd <- sd(nomisdat$p1cbmi[nomisdat$psdiag2==2], na.rm=TRUE)

bmi.cur.cont.mean <- mean(nomisdat$p1cbmi[nomisdat$outcome==0], na.rm=TRUE)
bmi.cur.cont.sd <- sd(nomisdat$p1cbmi[nomisdat$outcome==0], na.rm=TRUE)

#Body Mass Index (BMI) - lowest
bmi.low.cases.mean <- mean(nomisdat$p2lbmi[nomisdat$outcome==1], na.rm=TRUE)
bmi.low.cases.sd <- sd(nomisdat$p2lbmi[nomisdat$outcome==1], na.rm=TRUE)

bmi.low.anr.mean <- mean(nomisdat$p2lbmi[nomisdat$psdiag2==1], na.rm=TRUE)
bmi.low.anr.sd <- sd(nomisdat$p2lbmi[nomisdat$psdiag2==1], na.rm=TRUE)

bmi.low.anbp.mean <- mean(nomisdat$p2lbmi[nomisdat$psdiag2==2], na.rm=TRUE)
bmi.low.anbp.sd <- sd(nomisdat$p2lbmi[nomisdat$psdiag2==2], na.rm=TRUE)

bmi.low.cont.mean <- mean(nomisdat$p2lbmi[nomisdat$outcome==0], na.rm=TRUE)
bmi.low.cont.sd <- sd(nomisdat$p2lbmi[nomisdat$outcome==0], na.rm=TRUE)

#Body Mass Index (BMI) - highest
bmi.high.cases.mean <- mean(nomisdat$p5hbmi[nomisdat$outcome==1], na.rm=TRUE)
bmi.high.cases.sd <- sd(nomisdat$p5hbmi[nomisdat$outcome==1], na.rm=TRUE)

bmi.high.anr.mean <- mean(nomisdat$p5hbmi[nomisdat$psdiag2==1], na.rm=TRUE)
bmi.high.anr.sd <- sd(nomisdat$p5hbmi[nomisdat$psdiag2==1], na.rm=TRUE)

bmi.high.anbp.mean <- mean(nomisdat$p5hbmi[nomisdat$psdiag2==2], na.rm=TRUE)
bmi.high.anbp.sd <- sd(nomisdat$p5hbmi[nomisdat$psdiag2==2], na.rm=TRUE)

bmi.high.cont.mean <- mean(nomisdat$p5hbmi[nomisdat$outcome==0], na.rm=TRUE)
bmi.high.cont.sd <- sd(nomisdat$p5hbmi[nomisdat$outcome==0], na.rm=TRUE)

kw <- kruskal.test(nomisdat$p1cbmi ~ factor(nomisdat$psdiag2))
kw.cbmi.stat <- kw$statistic
kw.cbmi.df <- kw$parameter
kw.cbmi.p <- kw$p.value

tmp <- subset(nomisdat, psdiag2 != 0)
wilcox <- wilcox.exact(tmp$p1cbmi ~ factor(tmp$psdiag2))
wilcox.cbmi.stat <- wilcox$statistic
wilcox.cbmi.p <- wilcox$p.value

kw <- kruskal.test(nomisdat$p2lbmi ~ factor(nomisdat$psdiag2))
kw.lbmi.stat <- kw$statistic
```

```
kw.lbmi.df <- kw$parameter
kw.lbmi.p <- kw$p.value

wilcox <- wilcox.exact(tmp$p2lbmi ~ factor(tmp$psdiag2))
wilcox.lbmi.stat <- wilcox$statistic
wilcox.lbmi.p <- wilcox$p.value


#####################################################
### chunk number 11: figbmi
#####################################################
par(mfrow=c(1,3))
boxplot(nomisdat$p2lbmi ~ nomisdat$psdiag2, col="gray", main="Lowest", names=c("
    controls", "AN-R", "AN-BP"), las=3, ylim=c(10,35))
boxplot(nomisdat$p1cbmi ~ nomisdat$psdiag2, col="gray", main="Current", names=c("
    controls", "AN-R", "AN-BP"), las=3, ylim=c(10,35))
boxplot(nomisdat$p5hbmi ~ nomisdat$psdiag2, col="gray", main="Highest", names=c("
    controls", "AN-R", "AN-BP"), las=3, ylim=c(10,35))


#####################################################
### chunk number 12: bmitab
#####################################################
tab <- rbind(
c(strpar(bmi.low.anr.mean, bmi.low.anr.sd), strpar(bmi.low.anbp.mean, bmi.low.anbp.
    sd), strpar(bmi.low.cases.mean, bmi.low.cases.sd), strpar(bmi.low.cont.mean, bmi
    .low.cont.sd)),
c(strpar(bmi.cur.anr.mean, bmi.cur.anr.sd), strpar(bmi.cur.anbp.mean, bmi.cur.anbp.
    sd), strpar(bmi.cur.cases.mean, bmi.cur.cases.sd), strpar(bmi.cur.cont.mean, bmi
    .cur.cont.sd)),
c(strpar(bmi.high.anr.mean, bmi.high.anr.sd), strpar(bmi.high.anbp.mean, bmi.high.
    anbp.sd), strpar(bmi.high.cases.mean, bmi.high.cases.sd), strpar(bmi.high.cont.
    mean, bmi.high.cont.sd)))

rownames(tab) <- c("lowest BMI", "current BMI", "highest BMI")
colnames(tab) <- c("AN-R", "AN-BP", "AN (total)", "controls")

print(xtable(tab, label="tab:bmi", caption="Body Mass Index"))  #align = ; sanitize.
    text.function = function(x) x


#####################################################
### chunk number 13: edexptab
#####################################################
edmean <- rbind(
unlist(lapply(split(nomisdat$ed1, nomisdat$psdiag2), mean, na.rm=TRUE)),
unlist(lapply(split(nomisdat$ed2, nomisdat$psdiag2), mean, na.rm=TRUE)),
unlist(lapply(split(nomisdat$ed3, nomisdat$psdiag2), mean, na.rm=TRUE)),
unlist(lapply(split(nomisdat$ed4, nomisdat$psdiag2), mean, na.rm=TRUE)),
unlist(lapply(split(nomisdat$ed5, nomisdat$psdiag2), mean, na.rm=TRUE)),
unlist(lapply(split(nomisdat$edt, nomisdat$psdiag2), mean, na.rm=TRUE))
```

```
)

edsd <- rbind(
unlist(lapply(split(nomisdat$ed1, nomisdat$psdiag2), sd, na.rm=TRUE)),
unlist(lapply(split(nomisdat$ed2, nomisdat$psdiag2), sd, na.rm=TRUE)),
unlist(lapply(split(nomisdat$ed3, nomisdat$psdiag2), sd, na.rm=TRUE)),
unlist(lapply(split(nomisdat$ed4, nomisdat$psdiag2), sd, na.rm=TRUE)),
unlist(lapply(split(nomisdat$ed5, nomisdat$psdiag2), sd, na.rm=TRUE)),
unlist(lapply(split(nomisdat$edt, nomisdat$psdiag2), sd, na.rm=TRUE))
)

edmean.tot <- c(
mean(nomisdat$ed1[nomisdat$outcome==1], na.rm=TRUE),
mean(nomisdat$ed2[nomisdat$outcome==1], na.rm=TRUE),
mean(nomisdat$ed3[nomisdat$outcome==1], na.rm=TRUE),
mean(nomisdat$ed4[nomisdat$outcome==1], na.rm=TRUE),
mean(nomisdat$ed5[nomisdat$outcome==1], na.rm=TRUE),
mean(nomisdat$edt[nomisdat$outcome==1], na.rm=TRUE)
)

edsd.tot <- c(
sd(nomisdat$ed1[nomisdat$outcome==1], na.rm=TRUE),
sd(nomisdat$ed2[nomisdat$outcome==1], na.rm=TRUE),
sd(nomisdat$ed3[nomisdat$outcome==1], na.rm=TRUE),
sd(nomisdat$ed4[nomisdat$outcome==1], na.rm=TRUE),
sd(nomisdat$ed5[nomisdat$outcome==1], na.rm=TRUE),
sd(nomisdat$edt[nomisdat$outcome==1], na.rm=TRUE)
)

edtab <- cbind(
strpar(edmean[,2], edsd[,2]),
strpar(edmean[,3], edsd[,3]),
strpar(edmean.tot, edsd.tot),
strpar(edmean[,1], edsd[,1])
)

colnames(edtab) <- c("AN-R", "AN-BP", "AN (total)", "controls")
rownames(edtab) <- c("Parental problems ($ed1$)", "Disruptive events ($ed2$)", "
    Parental psych. disorder ($ed3$)", "Interpersonal problems ($ed4$)", "Family
    dieting env. ($ed5$)", "Total ($edt$)")

print(xtable(edtab, label="tab.edexp", caption="Environmental risk exposure"), table
    .placement="!ht", sanitize.text.function = function(x) x)


#####################################################
### chunk number 14: figedexp1
#####################################################
par(mfrow=c(1,3))
boxplot(nomisdat$ed1 ~ nomisdat$psdiag2, col="gray", main="Parental problems\n(ed1)"
    , names=c("controls", "AN-R", "AN-BP"), las=3, ylim=c(0,8))
```

```
boxplot(nomisdat$ed2 ~ nomisdat$psdiag2, col="gray", main="Disruptive events\n(ed2)"
    , names=c("controls", "AN-R", "AN-BP"), las=3, ylim=c(0,8))
boxplot(nomisdat$ed3 ~ nomisdat$psdiag2, col="gray", main="Parental psych. disorder\
    n(ed3)", names=c("controls", "AN-R", "AN-BP"), las=3, ylim=c(0,8))


#####################################################
### chunk number 15: figedexp2
#####################################################
par(mfrow=c(1,3))
boxplot(nomisdat$ed4 ~ nomisdat$psdiag2, col="gray", main="Interpersonal problems\n(
    ed4)", names=c("controls", "AN-R", "AN-BP"), las=3, ylim=c(0,8))
boxplot(nomisdat$ed5 ~ nomisdat$psdiag2, col="gray", main="Family dieting
    environment\n(ed5)", names=c("controls", "AN-R", "AN-BP"), las=3, ylim=c(0,8))
boxplot(nomisdat$edt ~ nomisdat$psdiag2, col="gray", main="Total\n(edt)", names=c("
    controls", "AN-R", "AN-BP"), las=3)


#####################################################
### chunk number 16: edcortab
#####################################################
edcor <- format(cor(nomisdat[72:76]), digits=3, scientific=FALSE)
print(xtable(edcor, label="tab.edcor", caption="Environmental domain correlations",
    align="rrrrrr"))  #align = ; sanitize.text.function = function(x) x


#####################################################
### chunk number 17: allelefreqtab
#####################################################
# cases and controls separately
tmp <- nomisdat$ht2a[nomisdat$outcome==0]
absfreq.g <- sum(tmp == 11, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)
absfreq.a <- sum(tmp == 22, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)
cont.relfreq.a <- absfreq.a / (absfreq.a+absfreq.g)   # 0.4213483
cont.relfreq.g <- absfreq.g / (absfreq.a+absfreq.g)

# AN (total)
tmp <- nomisdat$ht2a[nomisdat$outcome==1]
absfreq.g <- sum(tmp == 11, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)
absfreq.a <- sum(tmp == 22, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)
case.relfreq.a <- absfreq.a / (absfreq.a+absfreq.g) # case.relfreq.a
case.relfreq.g <- absfreq.g / (absfreq.a+absfreq.g)

# AN-R
tmp <- nomisdat$ht2a[nomisdat$psdiag2==1]
absfreq.g <- sum(tmp == 11, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)
absfreq.a <- sum(tmp == 22, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)
anr.relfreq.a <- absfreq.a / (absfreq.a+absfreq.g) # case.relfreq.a
anr.relfreq.g <- absfreq.g / (absfreq.a+absfreq.g)

# AN-BP
```

```
tmp <- nomisdat$ht2a[nomisdat$psdiag2==2]
absfreq.g <- sum(tmp == 11, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)
absfreq.a <- sum(tmp == 22, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)
anbp.relfreq.a <- absfreq.a / (absfreq.a+absfreq.g) # case.relfreq.a
anbp.relfreq.g <- absfreq.g / (absfreq.a+absfreq.g)

tab <- rbind(
c(anr.relfreq.a, anbp.relfreq.a, case.relfreq.a, cont.relfreq.a),
c(anr.relfreq.g, anbp.relfreq.g, case.relfreq.g, cont.relfreq.g))

rownames(tab) <- c("A allele frequency", "G allele frequency")
colnames(tab) <- c("AN-R", "AN-BP", "AN (total)", "controls")

print(xtable(tab, label="tab:allelefreq", caption="Allele frequencies", digits = 3))
      #align = ; sanitize.text.function = function(x) x

# Genotype frequencies

tab <- rbind(
prop.table(table(nomisdat$ht2a1438[nomisdat$psdiag2==1])), # AN-R
prop.table(table(nomisdat$ht2a1438[nomisdat$psdiag2==2])), # AN-BP
prop.table(table(nomisdat$ht2a1438[nomisdat$outcome==1])), # AN (total)
prop.table(table(nomisdat$ht2a1438[nomisdat$outcome==0]))) # controls
tab <- t(tab)
rownames(tab) <- c("GG genotype frequency", "AG genotype frequency", "AA genotype
    frequency")
colnames(tab) <- c("AN-R", "AN-BP", "AN (total)", "controls")

print(xtable(tab, label="tab:genotypefreq", caption="Genotype frequencies", digits =
     3))  #align = ; sanitize.text.function = function(x) x
```

A.3. Hardy-Weinberg Equilibrium

```
#####################################################
### chunk number 18: hw
#####################################################
# 1=G (cut), 2=A (uncut)

absfreq.g <- sum(nomisdat$ht2a == 11, na.rm=TRUE)*2 + sum(nomisdat$ht2a == 12, na.rm
    =TRUE)
absfreq.a <- sum(nomisdat$ht2a == 22, na.rm=TRUE)*2 + sum(nomisdat$ht2a == 12, na.rm
    =TRUE)

relfreq.a <- absfreq.a / (absfreq.a+absfreq.g)      # 0.4522613
relfreq.g <- absfreq.g / (absfreq.a+absfreq.g)      # 0.5477387

prop.table(table(nomisdat$ht2a))
hw.gg <- relfreq.g^2
hw.aa <- relfreq.a^2
```

```
hw.ga <- 2 * relfreq.a * relfreq.g

data.absfreq <- table(nomisdat$ht2a)
hw.absfreq <- c(hw.gg, hw.ga, hw.aa)*length(nomisdat$ht2a)

tab <- matrix(c(data.absfreq, hw.absfreq), nrow=3)
rownames(tab) <- c("GG genotype", "AG genotype", "AA genotype")
colnames(tab) <- c("observed", "expected")

chsq <- chisq.test(tab)
chsq.val <- chsq$statistic
chsq.df <- chsq$parameter
chsq.p <- chsq$p.value


#######################################################
### chunk number 19: hwtab
#######################################################
print(xtable(tab, label="tab:hardy-weinberg", caption="Observed and expected
    genotype frequencies"))  #align = ; sanitize.text.function = function(x) x
```

A.4. Association tests

```
#######################################################
### chunk number 20: tabassoc
#######################################################
## ===================================================================
## Association of A allele and AN
## ===================================================================
# 1=G (cut), 2=A (uncut)

# AN (total) vs. controls

# controls
tmp <- nomisdat$ht2a[nomisdat$outcome==0]
cont.absfreq.g <- sum(tmp == 11, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)
cont.absfreq.a <- sum(tmp == 22, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)

tmp <- nomisdat$ht2a[nomisdat$outcome==1]
cases.absfreq.g <- sum(tmp == 11, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)
cases.absfreq.a <- sum(tmp == 22, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)

tab <- matrix(c(cases.absfreq.a, cases.absfreq.g, cont.absfreq.a, cont.absfreq.g),
    nrow=2)
rownames(tab) <- c("A allele frequency", "G allele frequency")
colnames(tab) <- c("AN (total)", "controls")

chisq.allele.cases <- chisq.test(tab)  # [?]

proptab <- prop.table(tab, 2)
```

```
prtab.cases <- matrix(c(strpar(tab[1,1], proptab[1,1], 0, 3), strpar(tab[1,2],
    proptab[1,2], 0, 3), strpar(tab[2,1], proptab[2,1], 0, 3), strpar(tab[2,2],
    proptab[2,2], 0, 3)), nrow=2, byrow=TRUE)
rownames(prtab.cases) <- c("A allele frequency", "G allele frequency")
colnames(prtab.cases) <- c("AN (total)", "controls")
#print(xtable(prtab.cases, label="tab:alleleassoc", caption="Allele frequencies, AN
    (total) vs. controls")) #align = ; sanitize.text.function = function(x) x

# AN-R vs. controls
tmp <- nomisdat$ht2a[nomisdat$outcome==0]
cont.absfreq.g <- sum(tmp == 11, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)
cont.absfreq.a <- sum(tmp == 22, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)

tmp <- nomisdat$ht2a[nomisdat$psdiag2==1] # an-r
anr.absfreq.g <- sum(tmp == 11, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)
anr.absfreq.a <- sum(tmp == 22, na.rm=TRUE)*2 + sum(tmp == 12, na.rm=TRUE)

tab <- matrix(c(anr.absfreq.a, anr.absfreq.g, cont.absfreq.a, cont.absfreq.g), nrow
    =2)
rownames(tab) <- c("A allele frequency", "G allele frequency")
colnames(tab) <- c("AN-R", "controls")

chisq.allele.anr <- chisq.test(tab) # [?]

proptab <- prop.table(tab, 2)

prtab.anr <- matrix(c(strpar(tab[1,1], proptab[1,1], 0, 3), strpar(tab[1,2], proptab
    [1,2], 0, 3), strpar(tab[2,1], proptab[2,1], 0, 3), strpar(tab[2,2], proptab
    [2,2], 0, 3)), nrow=2, byrow=TRUE)
rownames(prtab.anr) <- c("A allele frequency", "G allele frequency")
colnames(prtab.anr) <- c("AN-R", "controls")

prtab.ges <- cbind(prtab.cases[,1], prtab.anr)
colnames(prtab.ges) <- c("AN (total)", "AN-R", "controls")

print(xtable(prtab.ges, label="tab:alleleassoc", caption="Allele frequencies for
    test of association")) #align = ; sanitize.text.function = function(x) x

## ============================================================================
## Association of AA genotype and AN
## ============================================================================

# AN (total) vs. controls
tab <- t(rbind(
  table(nomisdat$ht2a1438[nomisdat$outcome==1]),
  table(nomisdat$ht2a1438[nomisdat$outcome==0])))

colnames(tab) <- c("AN (total)", "controls")
rownames(tab) <- c("GG genotype", "AG genotype", "AA genotype")
```

```
chisq.genotype.cases <- chisq.test(tab)  # [?]

# AN-R vs. controls
tab <- t(rbind(
  table(nomisdat$ht2a1438[nomisdat$psdiag2==1]),
  table(nomisdat$ht2a1438[nomisdat$outcome==0])))

colnames(tab) <- c("AN-R", "controls")
rownames(tab) <- c("GG genotype", "AG genotype", "AA genotype")

chisq.genotype.anr <- chisq.test(tab)  # [?]

# table for paper
tab <- t(rbind(
  table(nomisdat$ht2a1438[nomisdat$outcome==1]),
  table(nomisdat$ht2a1438[nomisdat$psdiag2==1]),
  table(nomisdat$ht2a1438[nomisdat$outcome==0])))
colnames(tab) <- c("AN (total)", "AN-R", "controls")
rownames(tab) <- c("GG genotype", "AG genotype", "AA genotype")

proptab <- prop.table(tab, 2)
prtab <- matrix(c(
  strpar(tab[1,1], proptab[1,1], 0, 3),
  strpar(tab[1,2], proptab[1,2], 0, 3),
  strpar(tab[1,3], proptab[1,3], 0, 3),
  strpar(tab[2,1], proptab[2,1], 0, 3),
  strpar(tab[2,2], proptab[2,2], 0, 3),
  strpar(tab[2,3], proptab[2,3], 0, 3),
  strpar(tab[3,1], proptab[3,1], 0, 3),
  strpar(tab[3,2], proptab[3,2], 0, 3),
  strpar(tab[3,3], proptab[3,3], 0, 3)), nrow=3, byrow=TRUE)

colnames(prtab) <- colnames(tab)
rownames(prtab) <- rownames(tab)
print(xtable(prtab, label="tab:genotypeassoc", caption="Genotype frequencies for
    test of association"))  #align = ; sanitize.text.function = function(x) x




#####################################################
### chunk number 21: tabassoc2
#####################################################
chisqtab <- rbind(
  strchitab(chisq.allele.cases),
  strchitab(chisq.allele.anr),
  strchitab(chisq.genotype.cases),
  strchitab(chisq.genotype.anr))

colnames(chisqtab) <- c("$\\chi^2$", "$df$", "$p$ value")
rownames(chisqtab) <- c("allele-wise, AN (total)", "allele-wise, AN-R", "genotype-
```

```
    wise, AN (total)", "genotype-wise, AN-R")

print(xtable(chisqtab, label="tab:assoc", caption="Association tests", align="rrrr")
    ,sanitize.text.function = function(x) x)
```

A.5. Gene–Environment Correlation

```
#####################################################
### chunk number 22: rgetab
#####################################################
rge1 <- aov(nomisdat$envdom1 ~ gen)
rge2 <- aov(nomisdat$envdom2 ~ gen)
rge3 <- aov(nomisdat$envdom3 ~ gen)
rge4 <- aov(nomisdat$envdom4 ~ gen)
rge5 <- aov(nomisdat$envdom5 ~ gen)
rgetab <- cbind(
  format(c(eta(rge1), eta(rge2), eta(rge3), eta(rge4), eta(rge5)), digits=4),
  format(c(anova(rge1)$"Pr(>F)"[1],anova(rge2)$"Pr(>F)"[1],anova(rge3)$"Pr(>F)"[1],
      anova(rge4)$"Pr(>F)"[1],anova(rge5)$"Pr(>F)"[1]), digits=5))
colnames(rgetab) <- c("effect size $\\eta^2$", "$p$ value")
rownames(rgetab) <- c("Parental Problems ($ed1$)", "Disruptive Events ($ed2$)", "
    Parental Psychiatric Disorder ($ed3$)", "Interpersonal Problems ($ed4$)", "
    Family dieting environment ($ed5$)")
print(xtable(rgetab, label="tab.rge", caption="Gene--environment correlation", align
    = "rrr"), table.placement="!ht", sanitize.text.function = function(x) x)

levene.min <- min(
  levene.test(rge1)$"Pr(>F)"[1],
  levene.test(rge2)$"Pr(>F)"[1],
  levene.test(rge3)$"Pr(>F)"[1],
  levene.test(rge4)$"Pr(>F)"[1],
  levene.test(rge5)$"Pr(>F)"[1])

shapiro.range <- range(
  shapiro.test(rge1$residuals)$p.value,
  shapiro.test(rge2$residuals)$p.value,
  shapiro.test(rge3$residuals)$p.value,
  shapiro.test(rge4$residuals)$p.value,
  shapiro.test(rge5$residuals)$p.value)
```

A.6. Conditional Logistic Regression Models

```
#####################################################
### chunk number 23: clr.sample.size
#####################################################
n.pairs <- length(nomisdat$outcome[nomisdat$outcome==1])
n.clr <- length(nomisdat$outcome)
```

```
edtmax <- max(nomisdat$edt, na.rm=TRUE)


#####################################################
### chunk number 24: model1a
#####################################################
model1a <- clogit(outcome ~ gen*edt + strata(uniqid), data=nomisdat, method="exact")
#model1a

model1a.rsq <- summary(model1a)$rsq[1]
model1a.rsqmax <- summary(model1a)$rsq[2]
model1a.rsqrel <- model1a.rsq / model1a.rsqmax
model1a.AIC <- AIC.self(model1a)
model1a.LRT.chisq <- summary(model1a)$logtest[1]
model1a.LRT.df <- summary(model1a)$logtest[2]
model1a.LRT.p <- summary(model1a)$logtest[3]

tab <- condregtab(model1a)
print(xtable(tab, label="tab.model1a", caption="Model 1a (coefficients)", align = "
    rrrrrrrl"), table.placement="!ht", sanitize.text.function = function(x) x)


#####################################################
### chunk number 25: lrtab
#####################################################
model1b <- stepAIC(model1a, trace=FALSE)
lrt.1ab <- lrtest.self(model1b, model1a)
lrt.1ab.chisq <- lrt.1ab[1]
lrt.1ab.df <- lrt.1ab[2]
lrt.1ab.p <- lrt.1ab[3]


#####################################################
### chunk number 26: model1b
#####################################################
#model1b$anova
#summary(model1b)
#AIC.self(model1b)
  # Modell enthält keine WW, nur HE gen und HE umwelt

model1b.rsq <- summary(model1b)$rsq[1]
model1b.rsqmax <- summary(model1b)$rsq[2]
model1b.rsqrel <- model1b.rsq / model1b.rsqmax
model1b.AIC <- AIC.self(model1b)
model1b.LRT.chisq <- summary(model1b)$logtest[1]
model1b.LRT.df <- summary(model1b)$logtest[2]
model1b.LRT.p <- summary(model1b)$logtest[3]

tab <- condregtab(model1b)
print(xtable(tab, label="tab.model1b", caption="Model 1b (coefficients)", align = "
    rrrrrrrl"), table.placement="!ht", sanitize.text.function = function(x) x)
```

```
##################################################
### chunk number 27: figmodel1b
##################################################
model1b.pre <- function(genAG, genAA, edt,co) {
  g <- genAG*co$coefficients[1]+genAA*co$coefficients[2]+edt*co$coefficients[3]
  return(exp(g)/(1+exp(g)))
}
curve(model1b.pre(0,0,x, model1b), from=-15, to=15, ylab="Log-odds of relative risk
    for AN", xlab="Total events (difference)") #?
curve(model1b.pre(1,0,x, model1b), from=-15, to=15, add=TRUE, lty="dashed")
curve(model1b.pre(0,1,x, model1b), from=-15, to=15, add=TRUE, lty="dotdash")
legend("bottomright", legend=c("Genotype GG", "Genotype AG", "Genotype AA"), lty=c("
    solid", "dashed", "dotdash")) #$


##################################################
### chunk number 28: model2a
##################################################
model2a <- clogit(outcome ~ gen*ed1 + gen*ed2 + gen*ed3 + gen*ed4 + gen*ed5 + strata
    (uniqid), data=nomisdat, method="exact")
#model2a

model2a.rsq <- summary(model2a)$rsq[1]
model2a.rsqmax <- summary(model2a)$rsq[2]
model2a.rsqrel <- model2a.rsq / model2a.rsqmax
model2a.AIC <- AIC.self(model2a)
model2a.LRT.chisq <- summary(model2a)$logtest[1]
model2a.LRT.df <- summary(model2a)$logtest[2]
model2a.LRT.p <- summary(model2a)$logtest[3]

model2a.tab <- condregtab(model2a)
print(xtable(model2a.tab, label="tab.model2a", caption="Model 2a (coefficients)",
    align = "rrrrrrrl"), table.placement="!ht", sanitize.text.function = function(x)
     x)


##################################################
### chunk number 29: model2b
##################################################
model2b <- clogit(outcome ~ gen + ed2+ed4+ed5+gen:ed5+ strata(uniqid), data=nomisdat
    , method="exact")

model2b.rsq <- summary(model2b)$rsq[1]
model2b.rsqmax <- summary(model2b)$rsq[2]
model2b.rsqrel <- model2b.rsq / model2b.rsqmax
model2b.AIC <- AIC.self(model2b)
model2b.LRT.chisq <- summary(model2b)$logtest[1]
model2b.LRT.df <- summary(model2b)$logtest[2]
```

```
model2b.LRT.p <- summary(model2b)$logtest[3]

model2b.tab <- condregtab(model2b)
print(xtable(model2b.tab, label="tab.model2b", caption="Model 2b (coefficients)",
    align = "rrrrrrrl"), table.placement="!ht", sanitize.text.function = function(x)
    x)

lrt.2ab <- lrtest.self(model2b, model2a)
lrt.2ab.chisq <- lrt.2ab[1]
lrt.2ab.df <- lrt.2ab[2]
lrt.2ab.p <- lrt.2ab[3]
```

A.7. Linear Regression Models

```
#####################################################
### chunk number 30: linregn
#####################################################
n.lin <- length(nomisdat$ht2a1438[nomisdat$outcome==1]) #$


#####################################################
### chunk number 31: model3b
#####################################################
subtype <- nomisdat$psdiag2
levels(subtype) <- c("AN-R", "AN-BP")
onset <- nomisdat$psedon

model3a <- lm(p2lbmi ~ gen*ed1 + gen*ed2+gen*ed3 + gen*ed4+gen*ed5 + onset + subtype
    , data=nomisdat, subset=(outcome==1))
#summary(model3a)
model3a.step <- stepAIC(model3a, trace=FALSE)
#summary(model3a.step)
#AIC(model3a.step)

model3b <- lm(p2lbmi ~ ed3+onset+subtype, data=nomisdat, subset=(outcome==1))

model3b.rsq <- summary(model3b)$r.squared
model3b.adjrsq <- summary(model3b)$adj.r.squared
model3b.f <- summary(model3b)$fstatistic[1]
model3b.fdf1 <- summary(model3b)$fstatistic[2]
model3b.fdf2 <- summary(model3b)$fstatistic[3]
model3b.fp <- pf(summary(model3b)$fstatistic[1],summary(model3b)$fstatistic[2],
    summary(model3b)$fstatistic[3], lower.tail=FALSE)
model3b.AIC <- AIC(model3b)

model3b.tab <- linregtab(model3b)
print(xtable(model3b.tab, label="tab.model3b", caption="Model 3b, BMI (Coefficients)
    ", align = "rrrrrl"), table.placement="!ht", sanitize.text.function = function(x
    ) x)
```

```
lrt.3ab <- lrtest(model3b, model3a)
lrt.3ab.chisq <- lrt.3ab$Chisq[2]
lrt.3ab.df <- lrt.3ab$Df[2]
lrt.3ab.p <- lrt.3ab$"Pr(>Chisq)"[2]



##################################################
### chunk number 32: figmodel3blinearity1
##################################################
plot(model3b$fitted.values,model3b$fitted.values+model3b$residuals, xlab="Predicted
    values", ylab="Observed values")  #linearity (1): observed vs.\ predicted values
    -- diagnonal line


##################################################
### chunk number 33: figmodel3blinearity2
##################################################
plot(model3b$fitted.values, model3b$residuals, xlab="Predicted values", ylab="
    Residuals")  #linearity (2): residuals vs. predicted -- horizontal line #$


##################################################
### chunk number 34: model3bass
##################################################
model3b.hmc <- hmctest(model3b, plot = FALSE)  #homoscedasticity (harrison-mccabe-
    test)
  model3b.hmc.s <- model3b.hmc$statistic
  model3b.hmc.p <- model3b.hmc$p.value
model3b.dw <- dwtest(model3b, alternative = "two.sided")  #autocorrelation        (
    knapp signifikant!)
  model3b.dw.s <-model3b.dw$statistic
  model3b.dw.p <- model3b.dw$p.value
model3b.shap <- shapiro.test(model3b$residuals)  #nv der residuen
  model3b.shap.s <- model3b.shap$statistic
  model3b.shap.p <- model3b.shap$p.value
model3b.maxgvif <- max(vif(model3b))  #multicollinearity; gvif^(1/2df) <= 4 [?] fox
    (1992)


##################################################
### chunk number 35: model4b
##################################################
tmpdat <- subset(nomisdat, (nomisdat$outcome==1) & !is.na(nomisdat$psedon))

model4a <- lm(psedon ~ gen*ed1 + gen*ed2+gen*ed3 + gen*ed4+gen*ed5, data=tmpdat)
model4b <- lm(psedon ~ gen+ed2+ed5+gen:ed2, data=tmpdat)

model4b.rsq <- summary(model4b)$r.squared
model4b.adjrsq <- summary(model4b)$adj.r.squared
```

```
model4b.f <- summary(model4b)$fstatistic[1]
model4b.fdf1 <- summary(model4b)$fstatistic[2]
model4b.fdf2 <- summary(model4b)$fstatistic[3]
model4b.fp <- pf(summary(model4b)$fstatistic[1],summary(model4b)$fstatistic[2],
    summary(model4b)$fstatistic[3], lower.tail=FALSE)
model4b.AIC <- AIC(model4b)

model4b.tab <- linregtab(model4b)
print(xtable(model4b.tab, label="tab.model4b", caption="Model 4b, age of onset (
    coefficients)", align = "rrrrrl"), table.placement="!ht", sanitize.text.function
    = function(x) x)

lrt.4ab <- lrtest(model4b, model4a)
lrt.4ab.chisq <- lrt.4ab$Chisq[2]
lrt.4ab.df <- lrt.4ab$Df[2]
lrt.4ab.p <- lrt.4ab$"Pr(>Chisq)"[2]


#####################################################
### chunk number 36: lrtest4bc
#####################################################
subtype <- nomisdat$psdiag2
levels(subtype) <- c("AN-R", "AN-BP")
model4c <- lm(psedon ~ gen+ed2+ed5+gen:ed2+subtype, data=nomisdat, subset=(outcome
    ==1))

lrt.4bc <- lrtest(model4b, model4c)
lrt.4bc.chisq <- lrt.4bc$Chisq[2]
lrt.4bc.df <- lrt.4bc$Df[2]
lrt.4bc.p <- lrt.4bc$"Pr(>Chisq)"[2]


#####################################################
### chunk number 37: model4c
#####################################################

model4c.rsq <- summary(model4c)$r.squared
model4c.adjrsq <- summary(model4c)$adj.r.squared
model4c.f <- summary(model4c)$fstatistic[1]
model4c.fdf1 <- summary(model4c)$fstatistic[2]
model4c.fdf2 <- summary(model4c)$fstatistic[3]
model4c.fp <- pf(summary(model4c)$fstatistic[1],summary(model4c)$fstatistic[2],
    summary(model4c)$fstatistic[3], lower.tail=FALSE)
model4c.AIC <- AIC(model4c)

model4c.tab <- linregtab(model4c)
print(xtable(model4c.tab, label="tab.model4c", caption="Model 4c, age of onset (
    coefficients)", align = "rrrrrl"), table.placement="!ht", sanitize.text.function
    = function(x) x)
```

```
##################################################
### chunk number 38: figmodel4blinearity1
##################################################
plot(model4b$fitted.values,model4b$fitted.values+model4b$residuals, xlab="Predicted
    values", ylab="Observed values")  #linearity (1): observed vs. predicted values
    -- diagnonal line


##################################################
### chunk number 39: figmodel4blinearity2
##################################################
plot(model4b$fitted.values, model4b$residuals, xlab="Predicted values", ylab="
    Residuals")  #linearity (2): residuals vs. predicted -- horizontal line$


##################################################
### chunk number 40: figmodel4blinearity3
##################################################
plot(model4b, which=2)


##################################################
### chunk number 41: figmodel4blinearity4
##################################################
plot(model4b, which=5)


##################################################
### chunk number 42: tabs111val
##################################################
  s111 <- rbind(tmpdat[75,c(71:76, 6)], mean(tmpdat[,c(71:76, 6)]))
  s111 <- cbind(s111, c(  model4b$fitted.values[75], NA))
  colnames(s111) <- c("gen", "ed1", "ed2", "ed3", "ed4", "ed5", "Age of Onset", "
      Predicted")
  rownames(s111) <- c("Subject 111", "Mean")
  s111 <- format(s111, digits=3, scientific=FALSE)


##################################################
### chunk number 43: tabs111
##################################################
  print(xtable(s111, label="tab.s111", caption="Predictors of subject 111 and sample
      means", align = "rrrrrrrrr"), table.placement="!ht", sanitize.text.function =
      function(x) x)

model4b.hmc <- hmctest(model4b, plot = FALSE)  #homoscedasticity (harrison-mccabe-
    test)
  model4b.hmc.s <- model4b.hmc$statistic
  model4b.hmc.p <- model4b.hmc$p.value
model4b.dw <- dwtest(model4b, alternative = "two.sided")  #autocorrelation        (
    knapp signifikant!)
```

```
  model4b.dw.s <-model4b.dw$statistic
  model4b.dw.p <- model4b.dw$p.value
model4b.shap <- shapiro.test(model4b$residuals)  #nv der residuen
  model4b.shap.s <- model4b.shap$statistic
  model4b.shap.p <- model4b.shap$p.value
model4b.maxgvif <- max(vif(model4b)[,3])  #multicollinearity; gvif^(1/2df) <= 4 [?]
    fox (1992)



#######################################################
### chunk number 44: model4d
#######################################################
  noout<- tmpdat[-75,]
  model4d <- lm(psedon ~ gen+ed2+gen:ed2, data=noout)

    model4d.rsq <- summary(model4d)$r.squared
    model4d.adjrsq <- summary(model4d)$adj.r.squared
    model4d.f <- summary(model4d)$fstatistic[1]
    model4d.fdf1 <- summary(model4d)$fstatistic[2]
    model4d.fdf2 <- summary(model4d)$fstatistic[3]
    model4d.fp <- pf(summary(model4d)$fstatistic[1],summary(model4d)$fstatistic[2],
        summary(model4d)$fstatistic[3], lower.tail=FALSE)
    model4d.AIC <- AIC(model4d)

    model4d.tab <- linregtab(model4d)
    print(xtable(model4d.tab, label="tab.model4d", caption="Model 4d, age of onset (
        coefficients)", align = "rrrrrl"), table.placement="!ht", sanitize.text.
        function = function(x) x)


#######################################################
### chunk number 45: model4dpredict
#######################################################
  pre <- noout[3,]
  pre$ed2<-0; pre$gen<-factor("GG")
  preGG0 <- predict(model4d, pre) #, interval="prediction")
  pre$ed2<-1; pre$gen<-factor("GG")
  preGG1 <- predict(model4d, pre) #, interval="prediction")
  pre$ed2<-0; pre$gen<-factor("AA")
  preAA0 <- predict(model4d, pre) #, interval="prediction")
  pre$ed2<-1; pre$gen<-factor("AA")
  preAA1 <- predict(model4d, pre) #, interval="prediction")


#######################################################
### chunk number 46: figmodel4cinteract
#######################################################
  hlp <- function(model, gen, ed2) {
    pre<-noout[3,]
    pre$ed2 <- ed2
```

```
    pre$gen <- gen
    return(predict(model, pre))
  }
  ed2.p <- c(0,5,0,5,0,5)
  gen.p <- c("GG","GG","AG","AG","AA","AA")
  y <- NULL
  for (i in 1:6) y <- c(y, hlp(model4d, factor(gen.p[i]), ed2.p[i]))

  plot(noout$ed2, noout$psedon, col=as.integer(noout$gen), pch=as.integer(noout$gen)
      , xlab="Number of disruptive events", ylab="Age of onset")
  lines(ed2.p[1:2],y[1:2], lwd=2, col=which(levels(noout$gen)=="GG"), lty=which(
      levels(noout$gen)=="GG"))
  lines(ed2.p[3:4],y[3:4], lwd=2, col=which(levels(noout$gen)=="AG"), lty=which(
      levels(noout$gen)=="AG"))
  lines(ed2.p[5:6],y[5:6], lwd=2, col=which(levels(noout$gen)=="AA"), lty=which(
      levels(noout$gen)=="AA")+1)
  legend("topleft", legend=c("Genotype GG (predicted)", "Genotype AG (predicted)", "
      Genotype AA (predicted)", "Genotype GG (observed)", "Genotype AG (observed)",
      "Genotype AA (observed)"), col=c(1,2,3,1,2,3), pch=c(NA,NA,NA,1,2,3), lty=c
      (1,2,4,0,0,0), lwd=c(2,2,2,1,1,1), cex=0.8)


#####################################################
### chunk number 47: model4delrt
#####################################################
  subtype <- noout$psdiag2
  model4e <- lm(psedon ~ gen+ed2+gen:ed2+subtype, data=noout)
    lrt.4de <- lrtest(model4d, model4e)
    lrt.4de.chisq <- lrt.4de$Chisq[2]
    lrt.4de.df <- lrt.4de$Df[2]
    lrt.4de.p <- lrt.4de$"Pr(>Chisq)"[2]


#####################################################
### chunk number 48: model4e
#####################################################
    model4e.rsq <- summary(model4e)$r.squared
    model4e.adjrsq <- summary(model4e)$adj.r.squared
    model4e.f <- summary(model4e)$fstatistic[1]
    model4e.fdf1 <- summary(model4e)$fstatistic[2]
    model4e.fdf2 <- summary(model4e)$fstatistic[3]
    model4e.fp <- pf(summary(model4e)$fstatistic[1],summary(model4e)$fstatistic[2],
        summary(model4e)$fstatistic[3], lower.tail=FALSE)
    model4e.AIC <- AIC(model4e)

    model4e.tab <- linregtab(model4e)
    print(xtable(model4e.tab, label="tab.model4e", caption="Model 4e, age of onset (
        coefficients)", align = "rrrrrl"), table.placement="!ht", sanitize.text.
        function = function(x) x)
```

```
#####################################################
### chunk number 49: model4dass
#####################################################
model4d.hmc <- hmctest(model4d, plot = FALSE)  #homoscedasticity (harrison-mccabe-
    test)
  model4d.hmc.s <- model4d.hmc$statistic
  model4d.hmc.p <- model4d.hmc$p.value
model4d.dw <- dwtest(model4d, alternative = "two.sided")  #autocorrelation        (
    knapp signifikant!)
  model4d.dw.s <-model4d.dw$statistic
  model4d.dw.p <- model4d.dw$p.value
model4d.shap <- shapiro.test(model4d$residuals)  #nv der residuen
  model4d.shap.s <- model4d.shap$statistic
  model4d.shap.p <- model4d.shap$p.value
model4d.maxgvif <- max(vif(model4d)[,3])  #multicollinearity; gvif^(1/2df) <= 4 [?]
    fox (1992)
```

A.8. OR MDR Analysis

```
#####################################################
### chunk number 1: prepare
#####################################################
library(foreign)
library(survival)      # clogit
library(MASS)          # stepAIC
library(car)           # vif()   and recode()
#library(Rmdr)
library(ORMDR)
library(xtable)


#workdir <- "E:/Uni-Psycho/Diplomarbeit/Projekt 5ht2a Esstoerung/Auswertung/"
                # Memory Stick Uni
workdir <- "G:/Uni-Psycho/Diplomarbeit/Projekt 5ht2a Esstoerung/Auswertung/"
                # Memory Stick zu Hause
#workdir <- "D:/_new/Uni-Psycho/Diplomarbeit/Projekt 5ht2a Esstoerung/Auswertung/"
          # zu Hause

origdatfile <- paste(workdir, "5ht2a-data-v006-q.csv", sep = "")     # Daten noch im
    "abhängigen" Format, zur Kontrolle der Variablen
filename <- paste(workdir, "5ht2a-data-v007-q.csv", sep = "")        # Daten für log.
    Reg

origdat <- read.csv2(origdatfile)
names(origdat) <- tolower(names(origdat))
dat <- read.csv2(filename)
names(dat) <- tolower(names(dat))

## ===================================================================
## Erstellen der dichotomisierten Variablen für Umweltfaktoren
```

```
## ============================================================================
  # evtl. schon bei der Datenaufbereitung?? und im File speichern?

envdom1d <- as.integer(dat$envdom1 > 0)
envdom2d <- as.integer(dat$envdom2 > 0)
envdom3d <- as.integer(dat$envdom3 > 0)
envdom4d <- as.integer(dat$envdom4 > 0)
envdom5d <- as.integer(dat$envdom5 > 0)
envdomtd <- as.integer(dat$envdomt > 0)

envdom1dm <- as.integer(dat$envdom1 > median(dat$envdom1))
envdom2dm <- as.integer(dat$envdom2 > median(dat$envdom2))
envdom3dm <- as.integer(dat$envdom3 > median(dat$envdom3))
envdom4dm <- as.integer(dat$envdom4 > median(dat$envdom4))
envdom5dm <- as.integer(dat$envdom5 > median(dat$envdom5))
envdomtdm <- as.integer(dat$envdomt > median(dat$envdomt))

dat <- data.frame(dat, envdom1d, envdom2d, envdom3d, envdom4d, envdom5d, envdomtd,
    envdom1dm, envdom2dm, envdom3dm, envdom4dm, envdom5dm, envdomtdm)
rm(envdom1d, envdom2d, envdom3d, envdom4d, envdom5d, envdomtd, envdom1dm, envdom2dm,
     envdom3dm, envdom4dm, envdom5dm, envdomtdm)

nomisdat <- subset(dat, dat$gencompl)   #89
orignomisdat <- subset(origdat, origdat$gencompl)

ed1 <- nomisdat$envdom1
ed2 <- nomisdat$envdom2
ed3 <- nomisdat$envdom3
ed4 <- nomisdat$envdom4
ed5 <- nomisdat$envdom5
edt <- nomisdat$envdomt

gen <- factor(nomisdat$ht2a1438)
levels(gen) <- c("GG", "AG", "AA")
nomisdat <- data.frame(nomisdat, gen, ed1, ed2, ed3, ed4, ed5, edt)


ed1r <- recode(nomisdat$envdom1, "0=0;c(1,2,3)=1;else=2")
ed2r <- recode(nomisdat$envdom2, "0=0;c(1,2)=1;else=2")
ed3r <- recode(nomisdat$envdom3, "0=0;1=1;else=2")
ed4r <- recode(nomisdat$envdom4, "0=0;1=1;else=2")
ed5r <- recode(nomisdat$envdom5, "0=0;c(1,2)=1;else=2")

## ============================================================================
## MDR: Daten zusammenstellen
## ============================================================================

gen1 <- recode(nomisdat$ht2a1438, "11=0;12=1;22=2")
#gen2 <- recode(nomisdat$httlpr, "11=0;12=1;22=2")
attach(nomisdat)
  ormdrdat1 <- data.frame("Class"=outcome, gen1, ed1r, ed2r, ed3r, ed4r, ed5r)
```

```
  #ormdrdat2 <- data.frame("Class"=outcome, gen1, gen2, ed1r, ed2r, ed3r, ed4r, ed5r
      )
  #ormdrdat1.d <- data.frame("Class"=outcome, gen1, envdom1d, envdom2d, envdom3d,
      envdom4d, envdom5d)
  #ormdrdat1.dm <- data.frame("Class"=outcome, gen1, envdom1dm, envdom2dm, envdom3dm
      , envdom4dm, envdom5dm)
detach(nomisdat)

mdrsummary <- function(mdro) {
  varn <- names(mdro$data)
  cat("Variable names: ", varn, "\n")
  cat("Combinations: \n")
  print(mdro$min.comb)
  cat("Training Error:\n")
  print(mdro$train.erate)
  cat("Test Error:\n")
  print(mdro$test.erate)
  cat("Best combi: ", mdro$best.combi, "(", varn[as.integer(mdro$best.combi)], ")\n"
      )
  # mean training error and test error
  # number of (successful) cross-validations
}
mult.mdr <- function(count, dat, combi, colresp, cs, cv.fold, randomize) {
  mdrr <- mdr.c(dat, combi=combi, colresp=colresp, cs=cs, cv.fold=cv.fold, randomize
      =randomize)
  varn <- names(mdrr$data)
  tmp <- paste(varn[as.integer(mdrr$best.combi)], collapse=",")
  for (i in 2:count) {
    mdrr <- mdr.c(dat, combi=combi, colresp=colresp, cs=cs, cv.fold=cv.fold,
        randomize=randomize)
    tmp <- c(tmp, paste(varn[as.integer(mdrr$best.combi)], collapse=","))
  }
  return(table(tmp))
}
mult.mdr.cv <- function(count, dat, combi, colresp, cs, cv.fold, randomize) {
  mdrr <- mdr.c(dat, combi=combi, colresp=colresp, cs=cs, cv.fold=cv.fold, randomize
      =randomize)
  varn <- names(mdrr$data)
  train.er <- NULL
  test.er <- NULL
  besttmp <- NULL
  cvtmp <- NULL
  bestcvc <- NULL
  for (i in 1:count) {
    mdrr <- mdr.c(dat, combi=combi, colresp=colresp, cs=cs, cv.fold=cv.fold,
        randomize=randomize)
    train.er <- c(train.er, mdrr$train.erate)
    test.er <- c(test.er, mdrr$test.erate)
    thisbest <- paste(varn[as.integer(mdrr$best.combi)], collapse=" x ")
    besttmp <- c(besttmp, thisbest)
      cvthistmp <- c(paste(varn[mdrr$min.comb[,1]], collapse=" x "),
```

```
      paste(varn[mdrr$min.comb[,2]], collapse=" x "),
      paste(varn[mdrr$min.comb[,3]], collapse=" x "),
      paste(varn[mdrr$min.comb[,4]], collapse=" x "),
      paste(varn[mdrr$min.comb[,5]], collapse=" x "),
      paste(varn[mdrr$min.comb[,6]], collapse=" x "),
      paste(varn[mdrr$min.comb[,7]], collapse=" x "),
      paste(varn[mdrr$min.comb[,8]], collapse=" x "),
      paste(varn[mdrr$min.comb[,9]], collapse=" x "),
      paste(varn[mdrr$min.comb[,10]], collapse=" x "))
    cvtmp <- c(cvtmp, cvthistmp)
    bestcvc <- c(bestcvc, sum(cvthistmp==thisbest))
    #cvc <- c(cvc, sum(cvthistmp==cvthistmp[1]),
    #              sum(cvthistmp==cvthistmp[2]),
    #              sum(cvthistmp==cvthistmp[3]),
    #              sum(cvthistmp==cvthistmp[4]),
    #              sum(cvthistmp==cvthistmp[5]),
    #              sum(cvthistmp==cvthistmp[6]),
    #              sum(cvthistmp==cvthistmp[7]),
    #              sum(cvthistmp==cvthistmp[8]),
    #              sum(cvthistmp==cvthistmp[9]),
    #              sum(cvthistmp==cvthistmp[10]))
    #for(i in 1:cv.fold) cvtmp <- c(cvtmp, paste(varn[mdrr$min.comb[,i]], collapse="
        x "))
  }
  #return(list(besttmp,bestcvc))
  retcv <- data.frame(model=cvtmp, train.er, test.er)
  retcvtab <- t(t(table(retcv$model)))
  traintmp <- NULL
  testtmp <- NULL
  cvctmp <- NULL
  for (i in 1:length(retcvtab)) {
    traintmp <- c(traintmp, mean(retcv$train.er[retcv$model==rownames(retcvtab)[i]])
        )
    testtmp <- c(testtmp, mean(retcv$test.er[retcv$model==rownames(retcvtab)[i]]))
    #cvctmp <- c(cvctmp, mean(retcv$cvc[retcv$model==rownames(retcvtab)[i]]))
  }
  retcvtab <- cbind(retcvtab, prop.table(retcvtab), traintmp, testtmp)
  colnames(retcvtab) <- c("Frequency", "Percentage", "Avg.train.er", "Avg.test.er")

  retbest <- data.frame(model=besttmp, bestcvc)
  retbesttab <- t(t(table(retbest$model)))
  for (i in (1:length(retbesttab)))
    cvctmp <- c(cvctmp,   mean(retbest$bestcvc[retbest$model==rownames(retbesttab)[i
        ]]))

  retbesttab <- cbind(retbesttab, prop.table(retbesttab), cvctmp)
  colnames(retbesttab) <- c("Frequency", "Percentage", "CV-consistency")
  return(list("best"=retbesttab, "crossval"=retcvtab))
}

texformat <- function(tab) {
```

```
  tabneu <- tab
  tabneu$best[,1] <- as.character(tab$best[,1])# , digits=0, scientific=FALSE)
  tabneu$best[,2] <- format(tab$best[,2], digits=4, scientific=FALSE)
  tabneu$best[,3] <- format(tab$best[,3], digits=5, scientific=FALSE)

  tabneu$crossval[,1] <- as.character(tab$crossval[,1]) #, digits=0, scientific=
      FALSE)
  tabneu$crossval[,2] <- format(tab$crossval[,2], digits=3, scientific=FALSE)
  tabneu$crossval[,3] <- format(tab$crossval[,3], digits=4, scientific=FALSE)
  tabneu$crossval[,4] <- format(tab$crossval[,4], digits=4, scientific=FALSE)

  #rownames(tabneu$best) <-  gsub("$", "$", rownames(tabneu$best))
  #rownames(tabneu$best) <-  gsub("^", "$", rownames(tabneu$best))
  rownames(tabneu$best) <-  gsub("x", "$\\\\times$", rownames(tabneu$best))
  #rownames(tabneu$crossval) <-  gsub("$", "$", rownames(tabneu$crossval))
  #rownames(tabneu$crossval) <-  gsub("^", "$", rownames(tabneu$crossval))
  rownames(tabneu$crossval) <-  gsub("x", "$\\\\times$", rownames(tabneu$crossval))
    #$ zu beginn und ende, $edrx$, x ersetzen durch \\times

  return(tabneu)
}

AIC.self <- function(mod) return(-2*mod$loglik[2] + 2*length(mod$coefficients))
eta <- function(aovobj) {
  SS <- anova(aovobj)$"Sum Sq"
  return(SS[1]/sum(SS)) #SS factor N/SS Total
}

recode <-
function (var, recodes, as.factor.result, levels)
{
    recode.list <- rev(strsplit(recodes, ";")[[1]])
    is.fac <- is.factor(var)
    if (missing(as.factor.result))
        as.factor.result <- is.fac
    if (is.fac)
        var <- as.character(var)
    result <- var
    if (is.numeric(var)) {
        lo <- min(var, na.rm = TRUE)
        hi <- max(var, na.rm = TRUE)
    }
    for (term in recode.list) {
        if (0 < length(grep(":", term))) {
            range <- strsplit(strsplit(term, "=")[[1]][1], ":")
            low <- eval(parse(text = range[[1]][1]))
            high <- eval(parse(text = range[[1]][2]))
            target <- eval(parse(text = strsplit(term, "=")[[1]][2]))
            result[(var >= low) & (var <= high)] <- target
        }
        else if (0 < length(grep("else", term))) {
```

```
            target <- eval(parse(text = strsplit(term, "=")[[1]][2]))
            result[1:length(var)] <- target
        }
        else {
            set <- eval(parse(text = strsplit(term, "=")[[1]][1]))
            target <- eval(parse(text = strsplit(term, "=")[[1]][2]))
            for (val in set) {
                if (is.na(val))
                  result[is.na(var)] <- target
                else result[var == val] <- target
            }
        }
    }
    if (as.factor.result) {
        result <- if (!missing(levels))
            factor(result, levels = levels)
        else as.factor(result)
    }
    else if (!is.numeric(result)) {
        result.valid <- na.omit(result)
        save.warn <- options(warn=2)
        res <- try(if (length(result.valid) == length(grep("[0-9]",
result.valid)))
            as.numeric(result), silent=TRUE)
        result <- if (class(res) == "try-error" || is.null(res)) result
else res
        options(save.warn)
    }
    result
}

condregtab <- function(coxpho) {
  taba <- summary(coxpho)$coef
  tabb <- summary(coxpho)$conf.int
  sig <- as.integer(taba[,5] < 0.05)
  sig <- recode(sig, "0='';1='*'")
  #tab <- cbind(taba[,c(1,2,3)], tabb[,c(3,4)], taba[,5])
  tabtmp <- cbind(taba[,c(1,2,3)], tabb[,c(3,4)], taba[,5])
  tab <- cbind(
    format(tabtmp[,1],digits=3, scientific=FALSE), # coef
    format(tabtmp[,2],digits=3, scientific=FALSE), # exp(coef)
    format(tabtmp[,3],digits=2, scientific=FALSE), # se(coef)
    format(tabtmp[,4],digits=3, scientific=FALSE), # lower .95 CI of exp(coef)
    format(tabtmp[,5],digits=4, scientific=FALSE), # upper .95 CI of exp(coef)
    format(tabtmp[,6],digits=5, scientific=FALSE)) # p-value
  tab <- cbind(tab, sig)
  rownames(tab) <- rownames(taba)
  colnames(tab) <- c("$\\hat \\beta$", "OR", "SE($\\hat \\beta$)", "OR\\subscr{lower
      .95}", "OR\\subscr{upper .95}", "$p$ value", " ")
  return(tab)
}
```

```
lrtest.self <- function(simp, comp) {
  chisq <- abs(-2*(simp$loglik[2] - comp$loglik[2]))
  dfd <- length(comp$coefficients) - length(simp$coefficients)
  lrt.p <- pchisq(chisq, abs(dfd), lower.tail=FALSE)
  return(c(chisq, dfd, lrt.p))
}
ormdrtab <- function(datset, bestcomb) {
  tmp <- ormdr(datset,bestcombi=bestcomb,cs=1,colresp=1,CI.Asy=TRUE,CI.Boot=FALSE,B
      =1000)
  tmp <- data.frame(tmp, stringsAsFactors=FALSE)
  tmp <- with(tmp, tmp[order(Odds.ratio, decreasing=FALSE),])
  tmp <- tmp[-(which(names(tmp)=="Rank"))]  # Rank-Spalte weg

  unchangedpos <- which(names(tmp)=="Hi.Low")
  tab1 <- as.matrix(tmp[1:unchangedpos])
  tab2 <- as.matrix(format(tmp[,(unchangedpos+1):length(tmp)], digits=3))

  #tab2 <- NULL
  #for (i in (unchangedpos+1):length(tmp)) {
  #  tab2 <- cbind(tab2, format(tmp[,i], digits=3)) # , scientific=FALSE))
  #}
  tab <- cbind(tab1, tab2)
  rownames(tab) <- rep(" ", length(tab[,1]))       # different character? [?] [...] [
      here]
  rownames(tab)[1] <- "  "
  rownames(tab)[3] <- "  "
  colnames(tab)[(unchangedpos-2):(unchangedpos+3)] <- c("Cases", "Cont.", "Cat.", "
      OR", "lower .95", "upper .95")
  return(tab)
}

tab1 <- mult.mdr.cv(10000, ormdrdat1, combi=1, colresp=1, cs=1, cv.fold=10,
    randomize=TRUE)
tab2 <- mult.mdr.cv(10000, ormdrdat1, combi=2, colresp=1, cs=1, cv.fold=10,
    randomize=TRUE)
tab3 <- mult.mdr.cv(10000, ormdrdat1, combi=3, colresp=1, cs=1, cv.fold=10,
    randomize=TRUE)
tab4 <- mult.mdr.cv(10000, ormdrdat1, combi=4, colresp=1, cs=1, cv.fold=10,
    randomize=TRUE)
tab5 <- mult.mdr.cv(10000, ormdrdat1, combi=5, colresp=1, cs=1, cv.fold=10,
    randomize=TRUE)


#save(tab1, file=paste(workdir, "/tab1.Robj", sep=""))
#save(tab2, file=paste(workdir, "/tab2.Robj", sep=""))
#save(tab3, file=paste(workdir, "/tab3.Robj", sep=""))
#save(tab4, file=paste(workdir, "/tab4.Robj", sep=""))
#save(tab5, file=paste(workdir, "/tab5.Robj", sep=""))

#load(paste(workdir, "/tab1.Robj", sep=""))
```

```
#load(paste(workdir, "/tab2.Robj", sep=""))
#load(paste(workdir, "/tab3.Robj", sep=""))
#load(paste(workdir, "/tab4.Robj", sep=""))
#load(paste(workdir, "/tab5.Robj", sep=""))

tab1tex <- texformat(tab1)
tab2tex <- texformat(tab2)
tab3tex <- texformat(tab3)
tab4tex <- texformat(tab4)
tab5tex <- texformat(tab5)


###################################################
### chunk number 2: 2fac
###################################################
print(xtable(tab2tex$best, label="tab.2fbest", caption="Two factor model (best
    models)", align = "rrrr"), table.placement="!ht", sanitize.text.function =
    function(x) x)
print(xtable(tab2tex$crossval, label="tab.2fcv", caption="Two factor model (cross
    validation models)", align = "rrrrr"), table.placement="!ht", sanitize.text.
    function = function(x) x)


###################################################
### chunk number 3: ormdr2
###################################################
x <- ormdrtab(ormdrdat1,bestcomb=c(4, 6))
print(xtable(x, label="tab.2for", caption="Two factor model (odds ratios)", align =
    "rrrrrrrrr"), table.placement="!ht", sanitize.text.function = function(x) x)
#x <- data.frame(ormdr(ormdrdat1,bestcombi=c(4, 6),cs=1,colresp=1,CI.Asy=TRUE,CI.
    Boot=FALSE,B=1000))
#ormdr2sort <- with(x, x[order(Odds.ratio, decreasing=FALSE),])
#print(ormdr2sort)


###################################################
### chunk number 4: 3fac
###################################################
print(xtable(tab3tex$best, label="tab.3fbest", caption="Three factor model (best
    models)", align = "rrrr"), table.placement="!ht", sanitize.text.function =
    function(x) x)
print(xtable(tab3tex$crossval, label="tab.3fcv", caption="Three factor model (cross
    validation models)", align = "rrrrr"), table.placement="!ht", sanitize.text.
    function = function(x) x)


###################################################
### chunk number 5: ormdr3a
###################################################
x <- ormdrtab(ormdrdat1,bestcomb=c(3, 4, 6))
print(xtable(x, label="tab.3forA", caption="Three factor model A (odds ratios)",
```

```
    align = "rrrrrrrrrr"), table.placement="!ht", sanitize.text.function = function(
    x) x)


#####################################################
### chunk number 6: ormdr3b
#####################################################
x <- ormdrtab(ormdrdat1,bestcomb=c(4, 6, 7))
print(xtable(x, label="tab.3forB", caption="Three factor model B (odds ratios)",
    align = "rrrrrrrrrr"), table.placement="!ht", sanitize.text.function = function(
    x) x)

#x <- data.frame(ormdr(ormdrdat1,bestcombi=c(4, 6, 7),cs=1,colresp=1,CI.Asy=TRUE,CI.
    Boot=FALSE,B=1000))
#ormdr3sort <- with(x, x[order(Odds.ratio, decreasing=FALSE),])
#print(ormdr3sort)


#####################################################
### chunk number 7: 4fac
#####################################################
print(xtable(tab4tex$best, label="tab.4fbest", caption="Four factor model (best
    models)", align = "rrrr"), table.placement="!ht", sanitize.text.function =
    function(x) x)
print(xtable(tab4tex$crossval, label="tab.4fcv", caption="Four factor model (cross
    validation models)", align = "rrrrr"), table.placement="!ht", sanitize.text.
    function = function(x) x)


#####################################################
### chunk number 8: ormdr4
#####################################################
#x <- ormdrtab(ormdrdat1,bestcomb=c(3, 4, 6, 7))
#print(xtable(x, label="tab.4for", caption="Four factor model (odds ratios)", align
    = "rrrrrrrrrrr"), table.placement="!ht", sanitize.text.function = function(x) x)
x <- data.frame(ormdr(ormdrdat1,bestcombi=c(3, 4, 6, 7),cs=1,colresp=1,CI.Asy=TRUE,
    CI.Boot=FALSE,B=1000))
ormdr4sort <- with(x, x[order(Odds.ratio, decreasing=FALSE),])
print(ormdr4sort)


#####################################################
### chunk number 9: mdrtest
#####################################################
mdrtest <- clogit(outcome ~ ed1+ed2+ed4+ed5 + ed2:ed4 + ed2:ed5 + ed4:ed5 + ed1:ed2:
    ed4 + ed1:ed2:ed5 + ed1:ed4:ed5 + ed2:ed4:ed5 + ed1:ed2:ed4:ed5 + strata(uniqid)
    , data=nomisdat, method="exact")

mdrtest.rsq <- summary(mdrtest)$rsq[1]
mdrtest.rsqmax <- summary(mdrtest)$rsq[2]
mdrtest.rsqrel <- mdrtest.rsq / mdrtest.rsqmax
```

```
mdrtest.AIC <- AIC.self(mdrtest)
mdrtest.LRT.chisq <- summary(mdrtest)$logtest[1]
mdrtest.LRT.df <- summary(mdrtest)$logtest[2]
mdrtest.LRT.p <- summary(mdrtest)$logtest[3]

mdrtest.tab <- condregtab(mdrtest)
print(xtable(mdrtest.tab, label="tab.mdrtest", caption="Conditional logistic
    regression model of MDR results", align = "rrrrrrrl"), table.placement="!ht",
    sanitize.text.function = function(x) x)
```

List of Tables

6.1. Sample size per center and diagnose . . . 87
6.2. Highest education levels of parents . . . 88
6.3. Body Mass Index . . . 91
6.4. Environmental risk exposure . . . 91
6.5. Environmental domain correlations . . . 92
6.6. Allele frequencies . . . 92
6.7. Genotype frequencies . . . 93
6.8. Observed and expected genotype frequencies . . . 93
6.9. Allele frequencies for test of association . . . 93
6.10. Genotype frequencies for test of association . . . 94
6.11. Association tests . . . 94
6.12. Gene–environment correlation . . . 94
6.13. Model 1a (coefficients) . . . 96
6.14. Model 1b (coefficients) . . . 97
6.15. Model 2a (coefficients) . . . 99
6.16. Model 2b (coefficients) . . . 99
6.17. Model 3b, BMI (Coefficients) . . . 102
6.18. Model 4b, age of onset (coefficients) . . . 104
6.19. Model 4c, age of onset (coefficients) . . . 105
6.20. Predictors of subject 111 and sample means . . . 107
6.21. Model 4d, age of onset (coefficients) . . . 107
6.22. Model 4e, age of onset (coefficients) . . . 109
6.23. Two factor model (best models) . . . 112
6.24. Two factor model (cross validation models) . . . 112
6.25. Two factor model (odds ratios) . . . 113
6.26. Three factor model (best models) . . . 114
6.27. Three factor model (cross validation models) . . . 114
6.28. Three factor model A (odds ratios) . . . 115
6.29. Three factor model B (odds ratios) . . . 116

6.30. Four factor model (best models) 117
6.31. Four factor model (cross validation models) 117
6.32. Conditional logistic regression model of MDR results 118

List of Figures

3.1. Chemical structure of serotonin (5-HT). (Source: http://de.wikipedia.org/-wiki/Bild:Serotonin_%285-HT%29.svg). 40
3.2. Synapse. (Source: http://www.patientcenters.com/autism/graphics/pdd-0103.gif). 41
3.3. Chromosome 13. (Source: http://www.genecards.org/cgi-bin/carddisp.pl-?id=5293&id_type=hgnc&search=5293). 49

4.1. MDR algorithm. (Source: Coffey et al., 2004) 75
4.2. OR MDR algorithm. (Source: Chung et al. (2007) 78

6.1. Distribution of age difference 88
6.2. Boxplots of age of onset (left) and duration of illness (right) 89
6.3. Boxplots of BMI 90
6.4. Boxplots of environmental risk exposure 92
6.5. Model 1b 97
6.6. Model 3b: Observed vs. predicted (left), residuals vs. predicted (right) . . . 103
6.7. Model 4b: Observed vs. predicted (left), residuals vs. predicted (right) . . . 106
6.8. Model 4b: QQ-plot (left), residuals vs. leverage, Cook's distance (right) . . 106
6.9. Model 4d: Observed and predicted values 109

Abstract

The serotonin system is known to play a role in eating behavior and eating disorders. Recently, a number of polymorphisms have been identified in the serotonin system, one of them being the –1438 G/A polymorphism in the gene for the 5-HT_{2A} receptor (HTR2A). The A allele and the AA genotype have been associated with the diagnosis of anorexia nervosa in some studies. Other studies did not confirm the association.

In this paper, the association could not be replicated in a 1–1 matched sib-pair sample collected in three European centers. The sample consisted of 89 subjects suffering from anorexia nervosa and their 89 healthy sisters as controls.

The inconsistent findings on the association of the 5-HT_{2A} –1438 G/A polymorphism suggest that not this polymorphism alone is responsible for the development of anorexia nervosa. As a new approach, the interactions of the 5-HT_{2A} –1438 G/A polymorphism and five environmental risk domains (assessed by the Oxford Risk Factor Interview) are evaluated by the use of conditional logistic regression and odds ratio multifactor dimensionality reduction. The latter method is a model-free, nonparametric data-mining strategy that was specifically developed for identifying gene–environment and gene–gene interactions in relatively small samples.

The conditional logistic regression analysis did not find any significant genetic main effect that altered the risk of being diagnosed with anorexia nervosa. Concerning the environmental risk factors, the domains of disruptive events, interpersonal problems, and family dieting environment significantly increased the risk for anorexia nervosa. For the presumed interaction effects, only borderline significance could be shown. The AG genotype might act protectively in presence of distressing events in the family dieting environment domain.

The odds ratio multifactor dimensionality reduction analysis did not support that interaction. But it revealed interactions between environmental domains disruptive events, interpersonal problems, and parental problems.

The two methods came to somewhat similar conclusions concerning the genetic influence of the 5-HT_{2A} –1438 G/A polymorphism on the risk for anorexia nervosa.

It was also suggested that the 5-HT_{2A} –1438 G/A polymorphism might be a modifying

factor in the age of onset. This issue was investigated with a linear regression model. After exclusion of an outlier, the model found a significant genetic main effect. The AA genotype seems to delay the onset of anorexia nervosa. Being distressed by disruptive events seems to constitute an environmental risk that also seems to delay the onset. Furthermore, a significant interaction effect of the AA and AG genotype with the environmental domain disruptive events was found: Both genotypes seem to lead to an earlier onset of anorexia nervosa in presence of distressing disruptive events, compared to subjects with GG genotype.

The paper also investigated possible genetic and environmental main and interaction effects with respect to the life-time severity of anorexia nervosa (assessed by the lowest body mass index, BMI). The analysis did not reveal any significant genetic main or interaction effects.

In summary, the study found no genetic main or interaction effect when the outcome of interest was the diagnosis (only a borderline significant interaction effect between the AG genotype and family dieting environment, which was not confirmed by the OR MDR analysis). Neither were there any genetic effects influencing life-time severity (assessed by lowest BMI). But there were genetic main and interaction effects in the prediction of age of onset.

Zusammenfassung

Es ist bekannt, dass das Serotonin-System mit Essverhalten und Essstörungen in Zusammenhang steht. Im Zuge jüngerer genetischer Forschung wurden mehrere sogenannte Polymorphismen im Serotonin-System identifiziert. Einer davon ist der 5-HT_{2A} –1438 G/A Polymorphismus im Gen für den 5-HT_{2A} Rezeptor (HTR2A). Das A Allel und der AA Genotyp dieses Polymorphismus wurden in einigen Studien mit Anorexia Nervosa in Zusammenhang gebracht. Andere Studien wiederum konnten diese Assoziation nicht replizieren.

Die Assoziation konnte auch in dieser Studie nicht repliziert werden. Die verwendete Stichprobe bestand aus Patientinnen dreier europäischer Zentren, die an Anorexia Nervosa litten, und deren gesunden Schwestern als Kontrollgruppe.

Die uneinheitlichen Ergebnisse der Assoziationsstudien legen nahe, dass womöglich auch Umweltfaktoren eine wichtige Rolle spielen. In der vorliegenden Arbeit wurden als neuer Ansatz neben den Einflüssen des 5-HT_{2A} –1438 G/A Polymorphismus und Umweltrisikofaktoren (erhoben mittels des Oxford Risk Factor Interviews) auch deren Wechselwirkungen untersucht. Die Auswertung erfolgte mittels bedingter logistischer Regression als auch mit der Odds Ratio Multifactor Dimensionality Reduction Methode. Letzteres ist ein relativ neues, modell- und parameterfreies Verfahren aus dem Data Mining Bereich, das speziell für die Identifikation von Gen–Gen- und Gen–Umwelt-Wechselwirkungen entwickelt wurde.

Mittels bedingter logistischer Regression konnte kein genetischer Haupteffekt des 5-HT_{2A} –1438 G/A Polymorphismus gefunden werden, der das Risiko einer Anorexia Nervosa Diagnose erhöhen würde. Von den fünf untersuchten Umweltrisikofaktoren gingen drei mit einem erhöhten Risiko einher (*disruptive events*, *interpersonal problems* und *family dieting environment*). Bezüglich der vermuteten Wechselwirkungen konnte auch lediglich eine knapp nicht-signifikante Wechselwirkung identifiziert werden. Personen mit dem Genotyp AG scheinen bei einem vorhandenen Umweltrisiko in der Domäne *family dieting environment* eine (relativ zu dem Genotyp GG) leicht verminderte Wahrscheinlichkeit für eine Diagnose von Anorexia Nervosa aufzuweisen.

Die Odds Ratio Multifactor Dimensionality Reduction Methode stützt diese mögliche Wechselwirkung nicht. Sie identifizierte allerdings Interaktionen zwischen verschiedenen

Umweltfaktoren (Domänen *disruptive events*, *interpersonal problems* und *parental problems*).

Eine mögliche moderierende Wirkung des 5-HT_{2A} –1438 G/A Polymorphismus in Bezug auf den Erkrankungsbeginn (*age of onset*) wurde schon in früheren Artikeln diskutiert. In dieser Arbeit wurde dies mittels multipler linearer Regression bestätigt. Nach dem Ausschluss eines Ausreißers ergab sich ein genetischer Haupteffekt: Personen mit Genotyp AA zeigten einen verspäteten Beginn der Essstörung (relativ zu Personen vom Genotyp GG). Zusätzlich identifizierte die Analyse Wechselwirkungen. Während bei Personen mit Genotyp AA mehr Ereignisse aus der Domäne *disruptive events* mit einem späteren Beginn der Erkrankung zusammenhängen, scheint diese Beziehung für Personen mit den Genotypen AG und GG nicht zu gelten.

Weiters wurde in der vorliegenden Arbeit der Einfluss von genetischen und Umwelteffekten (und deren Wechselwirkungen) auf den Schweregrad der Erkrankung untersucht (erfasst durch den niedrigsten Body Mass Index). Es konnten keine genetischen Haupteffekte oder Wechselwirkungen gefunden werden.

Zusammenfassend kann gesagt werden, dass in der Studie keine signifikanten genetischen Haupteffekte oder Wechselwirkungen gefunden wurden, die das Risiko einer Diagnose von Anorexia Nervosa beeinflussen würden (es wurde lediglich eine knapp nicht-signfikante Wechselwirkung zwischen dem AG Gentotyp und *family dieting environment* gefunden, die aber durch die Odds Ratio Multifactor Dimensionality Reduction Methode nicht bestätigt wurde). Ebensowenig wurden genetische Haupteffekte oder Wechselwirkungen in Bezug auf den Schweregrad der Essstörung gefunden. Der 5-HT_{2A} –1438 G/A Polymorphismus dürfte aber Einfluss auf den Krankheitsbeginn haben (sowohl als Haupteffekt als auch in Form einer Wechselwirkung).

References

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

Aubert, R., Betoulle, D., Herbeth, B., Siest, G., & Fumeron, F. (2000). 5-HT2A receptor gene polymorphism is associated with food and alcohol intake in obese people. *International Journal of Obesity*, *24*(7), 920–924.

Baumgarten, H. G., & Grozdanovic, Z. (1996). Serotonin – essverhalten und essstörungen. *TW Neurologie Psychiatrie*, *10*(9), 660–665.

Blundell, J. E., Lawton, C. L., & Halford, J. C. G. (1995). Serotonin, eating behavior, and fat intake. *Obesity Research*, *3*(4), 471s–476s.

Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research* (Vols. Volume 1 – The Analysis of Case-Control Studies). Lyon: International Agency for Research on Cancer.

Brewerton, T. D. (1995). Toward a unified theory of serotonin dysregulation in eating and related disorders. *Psychoneuroendocrinology*, *20*(6), 561–590.

Brewerton, T. D., Hand, L. D., & Bishop Jr., E. R. (1993). The tridimensional personality questionnaire in eating disorder patients. *International Journal of Eating Disorders*, *14*(2), 213–218.

Brewerton, T. D., & Jimerson, D. C. (1996). Studies of serotonin function in anorexia nervosa. *Psychiatry Research*, *62*, 31–42.

Brown, T. A. (1999). *Moderne Genetik* (2nd ed.). Heidelberg, Berlin: Spektrum, Akademischer Verlag.

Bulik, C. M. (2005). Exploring the gene-environment nexus in eating disorders. *Journal of Psychiatry and Neuroscience*, *30*(5), 335–339.

Campbell, D., Sundaramurthy, D., Markham, A., & Pieri, L. (1998). Lack of association between 5-HT2A gene promoter polymorphism and susceptibility to anorexia nervosa. *Lancet*, *351*(9101), 499.

Caspi, A., & Moffitt, T. (2006). Gene-environment interactions in psychiatry: joining forces with neuroscience. *Nature Reviews: Neuroscience*, *7*, 583–590.

Chung, Y., Lee, S., Elston, R., & Park, T. (2007). Odds ratio based multifactor-

dimensionality reduction method for detecting gene-gene interactions. *Bioinformatics*, *23*(1), 71–76.

Coffey, C., Hebert, P., Ritchie, M., Krumholz, H., Gaziano, J., Ridker, P., et al. (2004). An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation. *Bioinformatics*, *5*, 49.

Collier, D. A., Arranz, M. J., Li, T., Mupita, D., Brown, N., & Treasure, J. (1997). Association between 5-HT2A gene promoter polymorphism and anorexia nervosa. *Lancet*, *350*, 412–413.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *Journal Of The National Cancer Institute*, *11*(6), 1269–1275.

Davison, G. C., & Neale, J. M. (2002). *Klinische Psychologie* (Eight Edition ed.). Weinheim: Beltz, Psychologie Verlags Union.

Enoch, M., Kaye, W., Rotondo, A., Greenberg, B., Murphy, D., & Goldman, D. (1998). 5-HT2A promoter polymorphism –1438G/A, anorexia nervosa, and obsessive-compulsive disorder. *Lancet*, *351*(9118), 1785–1786.

Fairburn, C. G., Cooper, Z., Doll, H. A., & Welch, S. L. (1999). Risk factors for anorexia nervosa: Three integrated case-control comparisons. *Archives of General Psychiatry*, *56*(5), 468–476.

Fairburn, C. G., & Harrison, P. J. (2003). Eating disorders. *Lancet*, *361*, 407–416.

Fairburn, C. G., Welch, S., Doll, H., Davies, B., & O'Connor, M. (1997). Risk factors for bulimia nervosa. a community-based case-control study. *Archives of general psychiatry*, *54*(6), 509–517.

Farewell, V. T. (1979). Some results on the estimation of logistic models based on retrospective data. *Biometrika*, *66*(1), 27–32.

Faris, P. L., Kim, S. W., Mellwer, W. H., Goodale, R. L., Oakman, S. A., & Hofbauer, R. D. (2000). Effect of decreasing afferent vagal activity with ondansetron on symptoms of bulimia nervosa: A randomised double-blind trial. *Lancet*, *355*(9206), 792–799.

Favaro, A., Ferrara, S., & Santonastaso, P. (2003). The spectrum of eating disorders in young women: A prevalence study in a general population sample. *Psychosomatic Medicine*, *65*, 701–708.

Ferguson, G. A., & Takane, Y. (2005). *Statistical analysis in psychology and education* (6th ed.). Montréal, Quebec: McGraw-Hill Ryerson Limited.

Ferrari, F., Bortoluzzi, S., Coppe, A., Sirota, A., Safran, M., Shmoish, M., et al. (2007). Novel definition files for human genechips based on geneannot. *BMC Bioinformatics*,

8(446).

Fischer, G. H. (1996). *Statistische Auswertung Psychologischer Experimente II. Skriptum zur Vorlesung.* Wien: Universität Wien.

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, *87*(417), 178–183.

Freeman, B., Powell, J., Ball, D., Hill, L., Craig, I., & Plomin, R. (1997). Dna by mail: an inexpensive and noninvasive method for collecting dna samples from widely dispersed populations. *Behavior Genetics*, *27*(3), 251–257.

Goldberg, S. C., Eckert, E. D., Halmi, K. A., Casper, R. C., Davis, J. M., & Roper, M. (1980). Effects of cyproheptadine on symptoms and attitudes in anorexia nervosa. *Archives of General Psychiatry*, *37*(9), 1083.

Goodwin, G. M., Fairburn, C. G., & Cowen, P. J. (1987). Dieting changes serotonergic function in women, not men: implications for the aetiology of anorexia nervosa? *Psychological Medicine*, *17*(4), 839–842.

Gorwood, P., Ades, J., Bellodi, L., Cellini, E., Collier, D. A., Di Bella, D., et al. (2002). The 5-HT2A -1438G/A polymorphism in anorexia nervosa: a combined analysis of 316 trios from six european centres. *Molecular Psychiatry*, *7*(1), 90–94.

Gorwood, P., Kipman, A., & Foulon, C. (2003). The human genetics of anorexia nervosa. *European Journal of Pharmacology*, *480*(1), 163–170.

Gurenlian, J. R. (2002). Eating disorders. *Journal of Dental Hygiene*, *76*(3), 219–234.

Guy-Grand, B., Crepaldi, G., Lefevre, P., Apfelbaum, M., Gries, A., & Turner, P. (1989). International trial of long-term dexfenfluramine in obesity. *Lancet*, 1142–1145.

Hahn, L., Ritchie, M., & Moore, J. (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *BMC Bioinformatics*, *19*(3), 376–382.

Halmi, K., Eckert, E., LaDu, T., & Cohen, J. (1986). Anorexia nervosa. treatment efficacy of cyproheptadine and amitriptyline. *Archives Of General Psychiatry*, *43*(2), 177–181.

Hansenne, M., & Ansseau, M. (1999). Harm avoidance and serotonin. *Biological Psychology*, *51*(1), 77–81.

Hardy, G. (1908). Mendelian proportions in a mixed population. *Science*, *28*, 49–50.

Hebebrand, J., Himmelmann, G. W., Wewetzer, C., Gutenbrunner, C., Heseker, H., Schäfer, H., et al. (1996). Body weight in acute anorexia nervosa and at follow-up assessed with percentiles for the body mass index : Implications of a low body weight at referral. *International Journal of Eating Disorders*, *19*(4), 347–357.

Hebebrand, J., Wehmeier, P., & Remschmidt, H. (2000). Weight criteria for diagnosis of

anorexia nervosa. *American Journal of Psychiatry*, *157*(6), 1024.

Heninger, G., Charney, D., & Sternberg, D. (1984). Serotonergic function in depression. prolactin response to intravenous tryptophan in depressed patients and healthy subjects. *Archives Of General Psychiatry*, *41*(4), 398–402.

Herbeth, B., Aubry, E., Fumeron, F., Aubert, R., Cailotto, F., Siest, G., et al. (2005). Polymorphism of the 5-HT2A receptor gene and food intakes in children and adolescents: the stanislas family study. *Am J Clin Nutr*, *82*(2), 467–470.

Herzog, D. B., & Delinski, S. S. (2001). Eating disorders. Innovative directions in research and practice. In R. H. Striegel-Moore & L. Smolak (Eds.), (chap. Classification of Eating Disorders). Washington, DC: American Psychological Association.

Hinney, A., Remschmidt, H., & Hebebrand, J. (2000). Candidate gene polymorphisms in eating disorders. *European Journal of Pharmacology*, *410*, 147–459.

Hinney, A., Ziegler, A., Nothen, M., Remschmidt, H., & Hebebrand, J. (1997). 5-HT2A receptor gene polymorphism, anorexia nervosa, and obesity. *Lancet*, *350*, 1324–1325.

Hirsch, J. (1981). To 'unfrock the charlatans'. *Sage Race Relations Abstracts*, *6*, 1–67.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York, NY: Wiley.

Hoyer, D., Clarke, D. E., Fozard, J. R., Hartig, P. R., Martin, G. R., Mylecharane, E. J., et al. (1994). International union of pharmacology classification of receptors for 5-hydroxytryptamine (serotonin). *Pharmacological Reviews*, *46*(2), 157–203.

Huelsenbeck, J. P., & Crandall, K. A. (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics*, *28*, 437–466.

Huxley, J. (1955). Morphism and evolution. *Heredity*, *9*, 1–52.

Jimerson, D. C., Lesem, M. D., Kaye, W. H., & Brewerton, T. D. (1992). Low serotonin and dopamine metabolite concentrations in cerebrospinal fluid from bulimic patients with frequent binge episodes. *Archives Of General Psychiatry*, *49*, 132–138.

Jimerson, D. C., Lesemb, M. D., Kaye, W. H., Heggd, A. P., & Brewerton, T. D. (1990). Eating disorders and depression: Is there a serotonin connection? *Biological Psychiatry*, *28*(5), 443–454.

Jönsson, E., Nöthen, M., Gustavsson, J., Berggård, C., Bunzel, R., Forslund, K., et al. (2001). No association between serotonin 2a receptor gene variants and personality traits. *Psychiatric Genetics*, *11*(1), 11–17.

Karwautz, A. (2004). Anorexia nervosia (anorexia nervosa: From its origins to treatment). In L. R. Moreno (Ed.), (chap. Factores de riesgo y de proteccion para la anorexia nerviosa (Risk and Protection Factors for Anorexia Nervosa)). Barcelona: Librerias Yenny.

Karwautz, A., Rabe-Hesketh, S., Hu, X., Zhao, J., Sham, P., A., C. D., et al. (2001).

Individual-specific risk factors for anorexia nervosa: A pilot study using a discordant sister-pair design. *Psychological Medicine*, *31*, 317–329.

Kaye, W. H., Frank, G. K., Bailer, U. F., Henry, S. E., Meltzer, C. C., Price, J. C., et al. (2005, May). Serotonin alterations in anorexia and bulimia nervosa: new insights from imaging studies. *Physiology & Behavior*, *85*(1), 73–81.

Kaye, W. H., Gendall, K., & Michael, S. (1998). Sertononin neuronal function and selective serotonin reuptake inhibitor treatment in aorexia dn bulimia nervosa. *Biological Psychiatry*, *44*, 825–838.

Kaye, W. H., Gwirtsman, H. E., & George, D. T. (1989). The effect of bingeing and vomiting on hormonal secretion. *Biological Psychiatry*, *25*, 768–780.

Kaye, W. H., Gwirtsman, H. E., George, D. T., & Ebert, H. (1991). Altered serotonin activity in anorexia nervosa after long-term weight recovery. *Archives Of General Psychiatry*, *48*, 556–562.

Kazdin, A., Kraemer, H., Kessler, R., Kupfer, D., & Offord, D. (1997). Contributions of risk-factor research to developmental psychopathology. *Clinical Psychology Review*, *17*(4), 375–406.

Kinzl, J. F., Traweger, C., Trefalt, E., Mangweth, B., & Biebl, W. (1999). Binge eating disorder in females: A population-based investigation. *International Journal of Eating Disorders*, *25*(3), 287–292.

Kipman, A., Bruins-Slot, L., Boni, C., Hanoun, N., Ades, J., Blot, P., et al. (2002). 5-HT2A gene promoter polymorphism as a modifying rather than a vulnerability factor in anorexia nervosa. *European Psychiatry*, *17*(4), 227–229.

Klump, K. L., & Gobrogge, K. L. (2005). A review and primer of molecular genetic studies of anorexia nervosa. *International Journal of Eating Disorders*, *37*, S43–S48.

Kwon, J., & Goate, A. (2000). The candidate gene approach. *Alcohol Research and Health*, *24*(3), 164–168.

Laessle, R. G., Wittchen, H. U., Fichter, M. M., & Pirke, K. M. (1988). The significance of subgroups of bulimia and anorexia nervosa: Lifetime frequency of psychiatric disorders. *International Journal of Eating Disorders*, *8*(5), 569–574.

Masellis, M., Basile, V., Meltzer, H., Lieberman, J., Sevy, S., Macciardi, F., et al. (1998). Serotonin subtype 2 receptor genes and clinical response to clozapine in schizophrenia patients. *Neuropsychopharmacology*, *19*(2), 123–132.

Melke, J., Westberg, L., Nilsson, S., Landen, M., Soderstrom, H., Baghaei, F., et al. (2003). A polymorphism in the serotonin receptor 3a (htr3a) gene and its association with harm avoidance in women. *Archives Of General Psychiatry*, *60*(10), 1017–1023.

Mittlböck, M., & Schemper, M. (1996). Explained variation for logistic regression. *Statistics*

in Medicine, *15*, 1987–1997.

Moffitt, T., Caspi, A., & Rutter, M. (2005). Strategy for investigating interactions between measured genes and measured environments. *Archives of General Psychiatry*, *62*, 473–481.

Moore, J., Gilbert, J., Tsai, C., Chiang, F., Holden, T., Barney, N., et al. (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of theoretical biology*, *241*(2), 252–261.

Nacmias, B., Ricca, V., Tedde, A., Mezzani, B., Rotella, C., & Sorbi, S. (1999). 5-HT2A receptor gene polymorphisms in anorexia nervosa and bulimia nervosa. *Neuroscience Letters*, *277*(2), 134–136.

Nelson, M., Kardia, S., Ferrell, R., & Sing, C. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Research*, *11*(3), 458–470.

Parsons, M., D'Souza, U., Arranz, M., Kerwin, R., & Makoff, A. (2004). The –1438A/G polymorphism in the 5-hydroxytryptamine type 2a receptor gene affects promoter activity. *Biological Psychiatry*, *56*(6), 406–410.

Patrick, L. (2002). Eating disorders: A review of the literature with emphasis on medical complicatinos and clinical nutrition. *Alternative Medicine Reiview*, *7*(3), 184–202.

Prentice, R. L., & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, *66*(3), 403–411.

Pritzel, M., Brand, M., & Markowitsch, H. J. (Eds.). (2003). *Gehirn und Verhalten. Ein Grundkurs der physiologischen Psychologie*. Berlin: Spektrum Akademischer Verlag.

R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org` (ISBN 3-900051-07-0)

Ramoz, N., Versini, A., & Gorwood, P. (2007). Eating disorders: an overview of treatment responses and the potential impact of vulnerability genes and endophenotypes. *Expert Opinion on Pharmacotherapy*, *8*, 2029–2044.

Ricca, V., Nacmias, B., Boldrini, M., Cellini, E., Bernardo, M. di, Ravaldi, C., et al. (2004). Psychopathological traits and 5-HT2A receptor promoter polymorphism (–1438 G/A) in patients suffering from anorexia nervosa and bulimia nervosa. *Neuroscience Letters*, *365*(2), 92–96.

Ricca, V., Nacmias, B., Cellini, E., Di Bernardo, M., Rotella, C., & Sorbi, S. (2002). 5-HT2A receptor gene polymorphism and eating disorders. *Neuroscience Letters*, *323*(2), 105–108.

Ritchie, M., Hahn, L., Roodi, N., Bailey, L., Dupont, W., Parl, F., et al. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics*, *69*(1), 138–147.

Rutter, M. (2006). *Genes and behavior. nature–nurture interplay explained.* Oxford: Blackwell Publishing.

Rybakowski, F., Slopien, A., Dmitrzak-Weglarz, M., Czerski, P., Rajewski, A., & Hauser, J. (2006). The 5-HT2A –1438 A/G and 5-HTTLPR polymorphisms and personality dimensions in adolescent anorexia nervosa: Association study. *Neuropsychobiology*, *53*(1), 33–39.

Schönemann, P. (1997). On models and muddles of heritability. *Genetica*, *99*, 97–108.

Serretti, A., Calati, R., Mandelli, L., & De Ronchi, D. (2006). Serotonin transporter gene variants and behavior: A comprehensive review. *Current Drug Targets*, *7*(12), 1659–1669.

Sheppard, P. (1975). Natural selection and heredity. In (4th ed., chap. 5). London: Hutchinson.

Sorbi, S., Nacmias, B., Tedde, A., Ricca, V., Mezzani, B., & Rotella, C. (1998). 5-HT2A promoter polymorphism in anorexia nervosa. *Lancet*, *351*(9118), 1785.

Spielman, R., McGinnis, R., & Ewens, W. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *American Journal of Human Genetics*, *52*(3), 506–516.

Spigset, O., Andersen, T., Hägg, S., & Mjöndal, T. (1999). Enhanced platelet serotonin 5-HT2A receptor binding in anorexia nervosa and bulimia nervosa. *European Neuropsychopharmacology*, *9*(6), 469–473.

Spooner, D., Treuren, R. van, & Vicente, M. de. (2005). *Molecular markers for genebank management.* Rome, Italy: International Plant Genetic Resources Institute. IPGRI Technical Bulletin No. 10.

Spurlock, G., Heils, A., Holmans, P., Williams, J., D'Souza, U. M., Cardno, A., et al. (1998). A family based association study of T102C polymorphism in 5HT2A and schizophrenia plus identification of new polymorphisms in the promoter. *Molecular psychiatry*, *3*(1), 42–49.

Stoneking, M. (2001). Single nucleotide polymorphisms: From the evolutionary past... *Nature*, *409*, 821–822.

Striegel-Moore, R. H., & Frank, D. L. (2003). Epidemiology of binge eating disorder. *International Journal of Eating Disorders*, *34*(S1), S19–S29.

Thapar, A., Harold, G., Rice, F., Langley, K., & Michael, O. (2007). The contribution of

gene-environment interaction to psychopathology. *Development and Psychopathology*, *19*(4), 989–1004.

Venables, W. N., & Ripley, B. D. (2001). *Modern applied statistics with S-Plus* (Third Edition ed.). New York, NY: Springer.

Wade, T. D., Gillespie, N., & Martin, N. G. (2007). A comparison of early family life events amongst monozygotic twin women with lifetime anorexia nervosa, bulimia nervosa, or major depression. *International Journal of Eating Disorders*, *40*(8), 679–686.

Walitza, S., Wewetzer, C., Warnke, A., Gerlach, M., Geller, F., Gerber, G., et al. (2002). 5-HT2A promoter polymorphism –1438G/A in children and adolescents with obsessive-compulsive disorders. *Molecular Psychiatry*, *7*(10), 1054–1057.

Witte, J. S., Gauderman, W., James, & Thomas, D. C. (1999). Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: Basic family designs. *American Journal of Epidemiology*, *148*(9), 693–705.

Zaider, T. I., Johnson, J. G., & Cockell, S. J. (2000). Psychiatric comorbidity associated with eating disorder symptomatology among adolescents in the community. *International Journal of Eating Disorders*, *28*(1), 58–67.

Lebenslauf

Allgemeines

Name	DI (FH) **Ingo Nader**
Wohnsitz	**Leopold Scheidl-Gasse 13, 2000 Stockerau, Österreich**
Telefonnummer	**0676 / 94 54 087, 02266 / 68756**
email	**ingo.nader@univie.ac.at, neith@deimos-tec.org**
Geburtsdatum	**19. Oktober 1978**
Geburtsort	**Wien**
Staatsbürgerschaft	**Österreich**
Familienstand	**ledig**

Ausbildung

seit 01.10.2002	**Diplomstudium Psychologie** an der Universität Wien (erste Diplomprüfung am 15.10.2004)
01.10.1998 – 26.06.2002	**Fachhochschule für Produktions- und Automatisierungstechnik**, Wien
Juni 1997	**Matura am Realgymnasium mit Schwerpunkt Informatik** in Stockerau

Bisherige Tätigkeiten

seit 01.10.2003	**Studienassistent am Institut für psychologische Grundlagenforschung, Bereich Methodenlehre** der Universität Wien
01.02. 2005 – 1.03.2005	**Freier Projektmitarbeiter** für Uni Wien / Verein DISCimus Statistische Datenauswertung des Projektes *Computerkids* (Untersuchung von Ao. Univ.-Prof. Doz. Dr. phil. Georg Gittler)
01.10.2001 – 31.01.2002	**TGM VAEE** (Versuchsanstalt für Elektrotechnik und Elektronik; Praktikumssemester im Rahmen der FH-Ausbildung) Aufgaben: Entwicklung eines Gerätes zur Temperatur- und Luftfeuchtigkeitsmessung und Speicherung der Daten (Hardware und Mikroprozessorprogrammierung)
06.03.2000 – 30.06.2000	**Renault Automation Comau, Werk Evry, Frankreich** (Auslands-Praktikumssemester im Rahmen der FH-Ausbildung) Aufgaben: Transfer von Daten aus der Produktionsdatenmanagement-Anwendung *Matrix* in die Applikation des Kundendienstes
Sommer 1995 und 1996	**Software Manufaktur GmbH** (Ferialpraktikum)

Eidesstattliche Erklärung

Ich bestätige, dass ich die vorliegende Diplomarbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe, und dass die Arbeit in gleicher oder ähnlicher Form noch in keiner anderen Prüfungsbehörde vorgelegen hat. Alle Ausführungen der Arbeit, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Wien, im Oktober 2008

Ingo Nader