Assessing prediction error of genetic variants in Cox regression models

Dissertation

zur Erlangung des
mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

vorgelegt von
Yesilda L. Balavarca Villanueva
aus Lima, Perú

Göttingen, 2012

D 7
Referent: Prof. Dr. Martin Schlather
Koreferentin: Prof. Dr. Heike Bickeböller
Tag der mündlichen Prüfung: 20.04.2012

Acknowledgments

I would like to thank all people who supported me during my work on this thesis. I am very grateful to my supervisor Prof. Dr. Heike Bickeböller for her advices and critical evaluation of my work, as well as for allowing me to attend international conferences and meetings during my period of research training. I am also thankful to Prof. Dr. Martin Schlather for his helpful comments and suggestions towards my thesis.

Special thanks goes to Prof. Anne Dickinson, as leader of the Marie-Curie Research Training Network MCRTN-CT-2004-512253, to Dr. Kim Pearce, and Jean Norden for providing information, for the collaborative work, and fruitful discussion in the statistical analysis of the EUROBANK-HSCT data. Likewise, I am grateful to Prof. Dr. Gottfried Fisher and Dr. Katarina Ludajic for our joint work, that also contributed to improve my knowledge in the field of clinical HSCT research.

Many thanks to all my colleagues of the Department of Genetic Epidemiology for the collaborative work and nice working environment. Thanks to Andrew Entwistle for his helpful suggestions concerning the English writing.

I wish to express my love and gratitude to my parents and sister in Perú, for their love and constant encouragement. Finally, I am very glad to have the friendship and support of my friends in Göttingen, thanks for spending enjoyable times together.

Contents

List of Figures **ix**

List of Tables **xi**

1 Introduction **1**

2 Human genetics and association studies **4**
2.1 Overview of genetics 4
2.1.1 DNA, chromosome, and gene 4
2.1.2 Inheritance of chromosomes 5
2.1.3 Genotype and Phenotype 6
2.1.4 Hardy-Weinberg equilibrium 6
2.1.5 Single nucleotide polymorphism 7
2.2 Association studies 7
2.2.1 Cohort studies 8
2.2.2 Measures of disease occurrence 10
2.2.3 Implications of genetics in association studies 12
2.2.4 Test of independence - Pearson's chi-squared test 13
2.2.5 Logistic regression model with multiple risk factors 14

3 Survival Analysis **16**
3.1 Framework of survival data 16
3.2 Terminology and notations 18
3.3 The Kaplan-Meier estimator and the log-rank test 19
3.4 The Cox regression model 22
3.4.1 Maximum likelihood estimation of beta 24
3.4.2 Estimation of the hazard ratio and survival function 25
3.4.3 Hypothesis test for association 27
3.4.4 Extended Cox regression model 28
3.5 Goodness of fit of the Cox model through an R^2 measure 28
3.5.1 Deviance residuals 30
3.5.2 Deviation of survival 31
3.5.3 The Schoenfeld residuals 36

3.5.4 R^2 measures in Cox models with SNP factors . . . 38
3.6 Validation of Cox models . . . 39

4 Prediction error estimators 41
4.1 Prediction error . . . 42
4.2 General description of prediction error estimators . . . 44
4.2.1 Apparent error estimator . . . 44
4.2.2 Bootstrap cross-validation estimator . . . 45
4.2.3 The 0.632 estimator . . . 48
4.2.4 The 0.632^+ estimator . . . 49
4.3 The 0.632 estimator based on survival probabilities . . . 50
4.4 The 0.632 estimator based on Schoenfeld residuals . . . 54
4.4.1 Schoenfeld residuals as measure of prediction errors . . . 55
4.4.2 Estimators based on the Schoenfeld residuals . . . 56
4.5 Schoenfeld residuals and Brier score as criteria to estimate prediction errors . . . 59
4.6 Measure for the gain in prediction, R^2_{Pred} . . . 60
4.6.1 The R^2_{Pred} measure . . . 61
4.6.2 Estimated R^2_{Pred} based on apparent errors . . . 61
4.6.3 Estimated R^2_{Pred} based on 0.632 and 0.632^+ estimators . . . 62
4.6.4 Partial gain in prediction . . . 63

5 Simulation Study 65
5.1 Simulation settings . . . 66
5.1.1 The Data . . . 66
5.1.2 Validation set . . . 69
5.2 Specification of the methods . . . 69
5.3 Results . . . 71
5.3.1 Gain in prediction with the original estimator of R^2_{Pred} . . . 71
5.3.2 Gain in prediction with alternative estimators of R^2_{Pred} . . . 75
5.3.3 Capability of estimator of R^2_{Pred} to identify the correct genetic model . . . 85

6 Prediction in Haematopoietic Stem Cell Transplantation 92
6.1 Study on Haematopoietic Stem Cell Transplantation (HSCT) . . . 93
6.1.1 Background on HSCT . . . 93
6.1.2 The HSCT translational network project . . . 95
6.2 Data description . . . 97
6.2.1 Clinical data . . . 98
6.2.2 Genotype data . . . 100
6.2.3 Follow-up and survival times . . . 102
6.3 Statistical analysis . . . 105
6.4 Results . . . 108
6.4.1 Effect of clinical factors on overall survival . . . 108

6.4.2 Association of single SNPs with overall survival 108
6.4.3 Association of multiple SNPs with overall survival 114
6.4.4 Joint risk score for overall survival 116

7 Summary and discussion **121**

Bibliography **125**

List of Figures

5.1 Criterion of the Schoenfeld residuals. Mean gain in prediction using a validation set, for the additive, dominant and recessive genetic models 72
5.2 Criterion of the Brier score. Mean gain in prediction using a validation set, for the additive, dominant and recessive genetic models 73
5.3 Criterion of the Schoenfeld residuals - Additive model. Comparison of mean estimates of gain in prediction using a validation set and estimators of prediction errors 77
5.4 Criterion of the Schoenfeld residuals - Dominant model. Comparison of mean estimates of gain in prediction using a validation set and estimators of prediction errors 78
5.5 Criterion of the Schoenfeld residuals - Recessive model. Comparison of mean estimates of gain in prediction using a validation set and estimators of prediction errors 79
5.6 Criterion of the Brier score - Additive model. Comparison of mean estimates of gain in prediction using a validation set and estimators of prediction errors 81
5.7 Criterion of the Brier score - Dominant model. Comparison of mean estimates of gain in prediction using a validation set and estimators of prediction errors 82
5.8 Criterion of the Brier score - Recessive model. Comparison of mean estimates of gain in prediction using a validation set and estimators of prediction errors 83
5.9 Criterion of the Schoenfeld residuals. Frequency of fitted genetic models identified as the correct genetic model 87
5.10 Criterion of the Brier score. Frequency of fitted genetic models identified as the correct genetic model 88
5.11 Criterion of the Schoenfeld residuals. Comparison of mean estimate of gain in prediction between true and misspecified genetic models . . 90
5.12 Criterion of the Brier score. Comparison of mean estimate of gain in prediction between true and misspecified genetic models 91

6.1 Overall Kaplan-Meier survival curve after HSCT 104
6.2 Kaplan-Meier survival curves by EBMT-score of patients after HSCT 109
6.3 Estimate of the *partial* $\hat{R}^2_{Pred,.632}$ of single biallelic SNPs 112

6.4 Boxplots of distribution of estimated gain in prediction provided by three models for overall survival 115
6.5 Kaplan-Meier survival curves by joint clinical and genetic scores of patients undergoing HSCT 118
6.6 Kaplan-Meier survival curves combined into five groups of the joint clinical and genetic scores of patients undergoing HSCT 119
6.7 Boxplots of distribution of estimated gain in prediction with a new score model for overall survival 120

List of Tables

2.1 Distribution of disease in exposed and unexposed cohorts to a risk factor . . . 10

3.1 Distribution of events at time $t_{(i)}$ in exposed and unexposed groups to a risk factor . . . 21

4.1 Summary table of differences between criteria to estimate prediction errors of Cox regression models . . . 59
4.2 Estimators of prediction error rates and of the gain in prediction for survival outcomes . . . 64

5.1 Frequency, mean, and variance of the genotype predictor under different MAFs and genetic models . . . 74

6.1 Risk scores from clinical EBMT factors established for patients undergoing HSCT . . . 97
6.2 Post-transplant outcomes in HSCT patients . . . 98
6.3 Clinical characteristics of patients prior to HSCT . . . 99
6.4 Non-HLA genes typed in patients and donors in the HSCT study . . . 101
6.5 Descriptives of the non-HLA biallelic SNPs genotyped on the HSCT study . . . 103
6.6 Descriptives of the non-HLA multiallelic genes or haplotypes genotyped on the HSCT study . . . 104
6.7 Distribution of EBMT-score and death status of HSCT patients . . . 106
6.8 Cox regression model for overall survival with the EBMT-score . . . 109
6.9 The top ten SNPs/haplotypes associated to overall survival . . . 110
6.10 The top ten SNPs/haplotypes contributing to gain in prediction of overall survival . . . 111
6.11 Cox regression model for overall survival, with multiple SNPs in addition to the clinical EBMT-score . . . 114
6.12 Risk scores from the joint clinical and genetic model for overall survival 116
6.13 Frequency of death per risk score from the joint clinical and genetic model . . . 117
6.14 Frequency of death per group of risk scores from the joint clinical and genetic model . . . 117

6.15 Cox regression model for overall survival, with risk score groups from the joint clinical and genetic model 118

Chapter 1

Introduction

The identification of genes responsible for the development and course of diseases has been a major topic within the field of genetics in the last years. The goal can be to understand the biological function of the genes in relation to the development of disease; or also to identify the genes potentially associated with the disease that alone or together with other risk factors, e.g. environmental or clinical, can help in the construction of disease risk scores for individuals. The latter is of importance in clinical practice for early diagnosis and prediction of disease, which also give the chance to study new therapeutic plans or prevent new cases.

On this matter, statistical genetic models have been a formal way to evaluate association between genes and disease. In most cases, the statistical significance of the effect size of genes on disease is mainly of interest. However, statistical significance may not always reflect the biological or clinical impact of the gene on the disease, i.e. effect size may result in statistical significance while it may not be of high enough relevance from the biological point of view. Or inversely, the clinical importance of a variable may not always be detected by looking at the significance of a statistical result, since it could be affected by many components such as sample size, frequency of the risk groups, or power of the test used in the analysis.

In this sense, we are interested in measuring the contribution of genes for the development of disease, beyond the significance of statistical results. The level of contribution should reflect the potential of a genetic model to predict the disease. Our work specifically focuses on the contribution of single nucleotide polymorphisms (SNPs), which are basic units for the study of association between gene and disease.

A measure that serves in the evaluation of the contribution of variables to a model is the coefficient of determination (R^2). The R^2 is used in regression models to measure the percentage of variation in disease outcome explained by the model. The R^2 also serves to evaluate the capability of a model to predict a future outcome, the higher the R^2 the better the capability of prediction of a model. In medical investigations in the genetic field, it is of great interest to have available genetic models for prediction purposes, that is, to predict the risk of disease of an individual

based on his/her genetic profile. To accomplish that purpose the validation of a fitted model needs to be proved. The validation of a statistical model requires testing it on new independent data, different from the data used for model fitting. A common limitation, however, is the unavailability of new independent data. On this matter, there are some techniques dealing with this problem, say, cross-validation (Stone 1974, Geisse 1975), bootstrap (Efron and Tibshirani 1993), or the 0.632 estimator (Efron 1983, Efron and Tibshirani 1997).

The objective of our work is to provide and study procedures for the estimation of R^2 based on validation of Cox regression models for survival data, particularly of genetic models with SNPs. The application of our study will help us to evaluate the contribution of single or multiple candidate SNPs on the overall survival of patients from the TRANS-NET study of haematopoietic stem cell transplantation (TRANS-NET 2008).

Investigations on validation procedures and contributions of biallelic genes in Cox regression models are limited. Müller et al. (2008) discussed various criteria that derive close versions of R^2 to measure contribution from survival models. They recommended the criterion based on the Schoenfeld residuals as the most advantageous method for studying the contribution of a SNP in genetic association studies. However, that study was performed for evaluation of goodness of fit, that is, evaluating the Cox model on the same data used for fitting that model. This procedure is known to overestimate the true performance of a model, and it is not advisable to use in the evaluation of models for prediction purposes.

We think, however, that the Schoenfeld residuals can be adapted and be used in the context of model validation, so that we can obtain a new form of the Schoenfeld residuals by testing the model on new independent data and not on the same data used for model fitting. In that case, we can validate genetic Cox regression models, by evaluating the gain in prediction due to the genetic variants. We propose this procedure by combining methodologies for estimation of prediction errors, such as the 0.632 estimator (Efron 1983, Efron and Tibshirani 1997), with the original criterion of the Schoenfeld residuals. Then, a redefined R^2, say R^2_{Pred}, can be estimated to measure the gain in prediction contributed by the genetic variants, in comparison to a reference model.

In our study we also included the technique implemented by Gerds and Schumacher (2007), where they estimate prediction errors with the criterion of the Brier score. This technique generates prediction error curves that allow the study of the time course of prediction errors.

Another question that arises during genetic association studies is about the most appropriate modelling of the pattern for inheritance of the disease: the additive, the dominant or the recessive pattern. We have chosen these patterns as they are typically investigated. Currently, among these three patterns we select the one whose model produces the lowest significance level of the estimated effect size, or also, the pattern whose model produces the highest contribution to the outcome, i.e. the highest R^2. However, again, the models are evaluated on the same data used for

model fitting and not on independent data required for validating a model. Thus, the present work also evaluates, through simulation studies, the capability of the $\hat{R}^2_{Pred}$ estimator for the identification of the most likely pattern of inheritance of disease.

The main objective of the thesis is to develop a statistical procedure under the Cox regression model for judging the contribution of SNPs on disease onset. The estimator of this contribution should be valid for the population under study, and not only for a particular sample. It should be obtained, therefore, from procedures evaluating the validation of genetic models. In addition, we aim to evaluate whether the estimator for the gain of prediction ($\hat{R}^2_{Pred}$) is useful to identify the mode of inheritance of the disease.

The first part of the thesis introduces background on genetic concepts, association studies, and the main statistical procedures related to our work. Chapter 2 gives some definitions on genetics and on association studies, chapter 3 describes survival analysis, Cox regression models, and reviews some procedures for goodness of fit of Cox models through R^2. Chapter 4 focuses on estimators of prediction errors. Among other estimators, it describes the derivation of the 0.632 estimator of prediction errors and its adaptation on Cox regression model by Gerds and Schumacher (2007). It also describes our procedure for estimation of prediction errors in Cox regression model based on the Schoenfeld residuals, and remarks on some differences with the work of Gerds and Schumacher (2007). The estimators used include the bootstrap cross-validation and 0.632 estimator for the study of model validation. Finally, it formulates the estimator for the gain in prediction in survival models, the estimator $\hat{R}^2_{Pred}$.

In chapter 5 we investigate through simulation studies the performance of the different techniques to estimate $\hat{R}^2_{Pred}$. We evaluate our approach which is based on the Schoenfeld residuals, and in addition, the approach of Gerds and Schumacher (2007), which is based on the Brier score. Also, the utility of the estimator $\hat{R}^2_{Pred}$ for the identification of the most appropriate mode of inheritance of disease is investigated.

Chapter 6 shows the application of the procedures in a real data set. The data are from the EUROBANK database from the TRANS-NET project, which collects clinical and genetic variables as well as post-transplant outcomes from patients undergoing haematopoietic stem cell transplantation in different European centers. Finally, we present summary and discussion in chapter 7.

Chapter 2

Human genetics and association studies

The first part of this chapter gives an overview of some definitions and terminology on genetics based on Marieb (2004), Hartl and Clark (1997), and Sham (1998), that will be the basis for the understanding of genetic applications in our work. The second part provides the features of statistical association studies in genetics, and points out differences from a classical statistical association study.

2.1 Overview of genetics

2.1.1 DNA, chromosome, and gene

The cell is the basic structural and functional unit of life. One of the main parts of the cell is the nucleus, that controls cellular activities. In the cell nucleus are the *chromosomes*. The chromosomes are organized structures of *DNA* (deoxyribonucleic acid), which is the genetic material. A *gene* is a segment of DNA in a chromosome.

A chromosome consists of a long double strand of DNA, the strands are bound to each other and wound around each other in the form of a double helix. Each strand is formed by a sequence of subunits called *nucleotides*, which are symbolized as A, T, G, or C according to their nitrogen-rich base adenine, thymine, guanine, or cytosine, respectively. The pair of strands are bound as a complementary sequence of nucleotides, where A always binds to T and G to C, i.e. A and T are complementary bases, as are C and G. Hence, any double strand of DNA contains an equal number of nucleotides A and T, as well as an equal number of nucleotides G and C.

A gene as a segment of DNA carries information in the sequence of its bases. Each sequence of three bases, called a *triplet*, codes for a particular *amino acid*. Amino

acids serve as building blocks of *proteins*. In the list of genetic code we find $4^3 = 64$ distinct triples, but only 20 distinct amino acids, because some triplets code for the same amino acid and there are some stop codons. A long sequence of triplets in a gene forms a long chain of amino acids, which produces a protein. Given that each type of amino acid has different properties, the overall structure of a protein determines its biological function.

2.1.2 Inheritance of chromosomes

The whole structure of DNA is packed in 23 pairs of chromosomes in the cell nucleus, from which 22 pairs are autosomes and 1 pair of sex chromosomes. That constitutes the genetic makeup of an individual, which consists of two sets of instructions that are inherited from the union of two *gametes*, the ovum (from the mother) and the sperm (from the father), each of which contributes a set of 23 chromosomes. These chromosomes guide the expression of the genetic traits. The genetic trait of sex, besides some other genetic traits, is determined by the pair of sex chromosomes.

A pair of chromosomes is called *homologous* if they contain the same genes for the same biological features. Homologous chromosomes contain two copies of every gene at the same position in the DNA sequence, where each gene comes from each parent. Then, an individual has 23 homologous chromosomes, with the exception of a male individual whose pair of sex chromosomes are of different types. So, male individuals have 22 homologous chromosomes and one pair of sex chromosomes.

Cell division is important for the growth of the body and for development and repair of tissues. Cell division occurs every time during life through a process called *mitosis*. During mitosis the 23 pairs of chromosomes are duplicated in the nucleus of the cell (mother cell), and then the cell is divided into two identical cells (daughter cells), each with its own nucleus with a copy of the 23 pairs of chromosomes. Hence, every cell nucleus in the body carries the whole genetic information of an individual.

A different process of cell division occurs in the cells for inheritance, called germ line cells. Germ line cells produce gametes through a process called *meiosis*. During meiosis it takes place first an event called *crossover*, that allows the exchange of genetic material between maternal and paternal chromosomes, i.e. between homologous chromosomes. Then, the cell is divided into two daughter cells, where only one random member of each homologous chromosome is present in the nucleus of each daughter cell. Thus, daughter cells have a reduced number of 23 chromosomes. After that, a second division process similar to mitosis occurs in each daughter cell. At the end the complete process produces a total of 4 daughter cells (gametes) which contain only a single set of 23 chromosomes. Therefore, meiosis brings two new facts: it reduces the number of chromosomes and it introduces genetic variability.

The union of two random gametes, an ovum (from a female) and a sperm (from a male), at the moment of fertilization forms a *zygote*, which is the first cell of a new individual. The zygote contains the combined maternal and paternal chromosomes,

that makes a complete number of 23 pairs of chromosomes. The mitotic division of the zygote begins right after fertilization and a new individual is developed, *the offspring*.

2.1.3 Genotype and Phenotype

A chromosome may contain thousands of genes. The physical location of a gene or a single variant of a gene along the chromosome is called *locus*. Given that chromosomes are paired (homologous), the corresponding variants are also paired. Each of the two genetic variants, at the same locus, in homologous chromosomes are called *alleles*. For a specific variant, an individual receives one allele from the mother and one allele from the father. If the two alleles have the same nucleotide sequences, the individual is *homozygous* for that locus, or *heterozygous* otherwise.

The pair of alleles at a specific locus is called *genotype*. The physical expression of the genotype, with or without environmental influences, is called *phenotype*.

The term *homozygous genotype* refers to the genotype with two identical copies of an allele. Otherwise, the genotype is called *heterozygous genotype*.

When the alleles of a gene present only two possible different nucleotide sequences, it is said that the gene is *biallelic*. A biallelic gene provides three possible genotypes, two homozygous and one heterozygous genotype. For example, let A and a represent the two alleles of a biallelic gene, an individual may have a homozygous genotype AA or aa, or a heterozygous genotype Aa. If the genotype Aa expresses the same phenotype as does AA, it is said that A is a *dominant allele* because it masks the expression of its partner, while a is said to be a *recessive allele* because it needs the two copies (the genotype aa) to express its corresponding phenotype. When *co-dominance* occurs, the heterozygous genotype Aa expresses a phenotype were the expressions of both alleles are visible.

For quantitative traits, A is an *additive allele* if the phenotype measure of Aa is the average between the phenotype measures of AA and aa, given that the measure of AA is higher than aa.

2.1.4 Hardy-Weinberg equilibrium (HWE)

Among other assumptions, the Hardy-Weinberg law is supported by the assumption that individuals in a population mated with each other at random, i.e. the paternal and maternal gametes are combined at random. This assumption is known as *random mating*.

The Hardy-Weinberg law defines the mathematical relation between the allele frequencies and the genotype frequencies in the population. We illustrate this with an example. Let p and q be the relative frequencies of alleles A and a of a biallelic gene (i.e. $p+q=1$). Under random mating, the relative frequencies of the genotypes AA,

Aa, and aa are p^2, $2pq$, and q^2, respectively. In an infinitely large population, as long as random mating is accomplished the allele frequencies will remain unchanged after one generation, and the genotype frequencies in all subsequent generations will also remain unchanged, showing the existence of an equilibrium. Therefore, the genotype frequencies p^2, $2pq$ and q^2 represent the *Hardy-Weinberg equilibrium (HWE)*. It can also be generalized to genes with more than two alleles.

The implication of the Hardy-Weinberg law is that, under random mating, the population maintains constant allele frequencies, and therefore preserves genetic variation.

However, there are sources that could cause deviations from HWE, apart from non-random mating in the population. Random changes can occur in the allele frequencies in a population (*genetic drift*), that may cause alleles to disappear and reduce genetic variation. These effects are introduced due to finite population size. The effect of genetic drift is larger in small populations. It can also happen, that individuals with different genotypes have different rates of early death, causing a distortion in the genotype frequency of the current population and possibly altering the chances of random mating. Other sources of deviation from HWE are mutation and migration (Hartl and Clark 1997).

2.1.5 Single nucleotide polymorphism (SNP)

Different individuals in a population will present variations in their DNA sequences at a particular locus. These variations can be common or rare in the population. Variants occurring $> 1\%$ frequency in the population are known as *genetic polymorphism*. If that variation occurs because of differences in a single nucleotide (A, T, C, or G) in the DNA sequence, then it is called *single nucleotide polymorphism (SNP)*. SNPs are the most common genetic variation. The public SNP database of the National Center for Biotechnology Information (NCBI) has released to date more than 16 million of SNPs in human beings (Database of Single Nucleotide Polymorphisms (dbSNP) 2011, Sherry et al. 2001).

For a given SNP and its alleles, *minor allele* refers to the allele with the lower frequency in the population, and *minor allele frequency (MAF)* refers to such lower frequency, i.e. frequency of the minor allele. The most frequent allele in the population is called *wildtype allele*. A homozygous genotype containing two wildtype alleles is called *wildtype genotype*.

2.2 Association studies

In biomedical studies, researchers investigate the relationship between factors and disease outcome. *Factors* are characteristics that distinguish groups of individuals in the population of study. If the frequency of disease occurring on the individuals

differ statistically among the groups, then there is an *association* between the factor and the disease. In this case, the factor is called *risk factor*. According to the context of the study, the risk factors are also known as *predictors*, *covariates*, or *exposure factors*.

Various types of study designs are used according to particular objectives and other settings of the study. The present thesis is placed into the context of cohort study designs, therefore we only deal with this type of study in the next sections.

2.2.1 Cohort studies

A *cohort* can be defined as a group of individuals from the population who are followed for a period of time and whose outcome of interest, e.g. disease/no disease, is regularly evaluated, together with other characteristics.

Cohort studies are carried out to investigate associations of exposure factors with development of disease or with any other outcome of interest. The main characteristic of cohort studies is that the exposure of individuals is known from the beginning of the study, and the development of disease is investigated as a future outcome during the follow-up of the individuals. The counterpart of this design is the case-control study, where the disease status of the individuals is known at the beginning of the study, and their past exposure to risk factors is investigated.

Thus, in cohort studies, individuals under study differ in their measurement of exposure factor, e.g. for a factor with two categories: exposed/unexposed individuals or greater/lesser extent of exposure. These individuals are subsequently evaluated to record their disease status over a period of time. The objective is to measure and compare the frequency of disease between the exposed and unexposed individuals at the end of the study, and determine if the exposure factor plays a role in the development of disease. Eventually, not only the knowledge of development of disease is important to the study, but also the time at which it was first observed, or the course of disease.

The example above refers to a simple cohort study, where only two groups of individuals are recognized according to the exposure factor. However, more than two groups could also come up according to the exposure factor on the cohort. For example in genetic studies, it could be of interest to study the association of allele A -from a specific SNP- with the disease, where allele A is the exposure factor for the disease in the population. Here, individuals with genotypes AA and Aa form one group, and individuals with genotype aa form another group. However, we could also be interested in the association of the different genotypes with the disease, i.e. genotype as the exposure factor. Hence, we get three groups of individuals, each group with only genotypes AA, Aa, or aa.

There are two main types of cohort studies, that differ on the moment when we take knowledge of the exposure of the individual. In a *prospective cohort study*, the

information about the exposure factor is collected from the beginning of the study, and therefore we can identify the groups of exposure from the beginning of the study as well. Also, the time of follow-up for evaluation of disease development falls within the period of study. On the contrary, a *retrospective cohort study* uses existing data. The exposure status of the individuals as well as the identification of groups in the cohort are taken from recorded or historical data. These data might have been recorded to carry out a work not necessarily related to the current cohort study, but they might contain relevant data for the current study. Hence, the usual tasks for a prospective cohort study, e.g. identification of groups of exposure, follow-up of the individuals, and regularly evaluation of disease, were done before the beginning of the study.

An example of a prospective cohort study is the work presented in chapter 6. In short, a population of patients who underwent hematopoietic stem cell transplantation (HSCT) were recruited from 1983, the beginning of the study. The patients were followed and evaluated over years to register their survival status (alive/dead) until 2009, the end of the study. As the patients entered the study, blood samples were taken, from which the genotype for various genes were recorded. At the end of the study, these genes were evaluated as potential risk factors associated with survival of patients after transplantation.

Retrospective cohort studies have the advantage of being less costly, and of obtaining results faster. That is because the cohort data are already available at the beginning of the study, and no waiting time for observation of the outcome will be required. However, given that the data were collected for a different purpose, the disadvantage of retrospective cohort studies is that the data may not contain important information for the current cohort study. A retrospective cohort study depends on what is recorded in the data, the user will not have the chance to modify it, to access the individuals to collect additional data, or to verify the original data sources to improve the quality of the records.

We also mention two other special types of cohort studies, based on the way the groups of a cohort are obtained in relation to the exposure factor (Rothman 2002). The *special-exposure cohort study* focuses on individuals who share a particular exposure factor which is uncommon in the population. For example, an occupational group that is exposed to chemical substances. These are individuals specifically identified for the study because they are not spread over the population. The experience of disease of these exposed individuals could then be compared to individuals without the exposure. By taking a cohort from the general population it would be difficult to get enough individuals in the exposed group since the exposure is uncommon, unless we are performing a very large study.

On the contrary, for common exposure factors it is more appropriate to use a *general-population cohort study*. The general-population cohort is determined by the research question. A subset of individuals from that general-population are followed, regardless of the exposure factor, and the groups of individuals under the

Table 2.1: Distribution of disease in exposed and unexposed cohorts to a risk factor[a]

Risk Factor (X)	Disease (Y)		Total
	Yes (1)	No (0)	
exposed (1)	d_1	$n_1 - d_1$	n_1
unexposed (0)	d_0	$n_0 - d_0$	n_0
Total	d	$n - d$	n

[a] The letters in the cells represent absolute counts within the respective subgroups.

exposure factor are identified. The cohort study we presented in chapter 6 is a type of general-population cohort study, where the general-population is composed by individuals undergoing HSCT, and the exposure factors are a set of clinical and genetic factors.

2.2.2 Measures of disease occurrence

The basic measure of disease occurrence is the *risk rate*, which is also simply called as *risk* and it is the term we will use in this thesis. The risk can be generally defined as the probability that the disease occurs in a group of individuals. The risk measure is used to get other derivative measures quantifying the association between risk factor and disease. Here, we will present only two measures related with cohort studies, that are important for the understanding of concepts and procedures in the next chapters. For other measures and details we refer to Lachin (2000), Jewell (2004), and Rothman (2002).

The *relative risk or risk ratio (RR)* compares the risk of disease in an exposed group in relation to an unexposed group. The RR assesses how much the risk factor affects the disease risk in the exposed group with reference to the disease risk in the unexposed group.

Let Y be the disease under study, where $Y = 1$ if the disease is present, and $Y = 0$ otherwise; and let X be a binary risk factor, where $X = 1$ if the group is exposed, and $X = 0$ otherwise. Let Table 2.1 represent the distribution of disease at the interval time $[0, \tau]$, where τ is the end time of the study. Then,

$$RR = \frac{P(Y=1|X=1)}{P(Y=1|X=0)} = \frac{d_1/n_1}{d_0/n_0}\,, \tag{2.1}$$

where $P(Y=1|X=1)$ denotes the probability of disease in the exposed group, i.e. the disease risk in the exposed group; likewise, $P(Y=1|X=0)$ indicates the disease risk in the unexposed group.

Another important measure of association is the *hazard rate*, which is also simply called *hazard* and it is the term we will use in this thesis. The hazard accounts for the time feature of cohort studies. The hazard is the risk of getting the disease at a specific time point t within the time period of follow-up $[0, \tau]$ of the cohort. The hazard is also defined as the instantaneous rate of disease at time t. A mathematical definition and expressions for the hazard are given later in section 3.2.

A particular fact to consider is that the computation of the hazard is done with the population still at risk of disease, i.e. computation of a hazard at time t considers only non-diseased individuals by that time. Once an individual gets the disease at any time t, he/she should be removed before computing hazards at subsequent time points. We can also say that the hazard is computed considering only *individuals at risk* at the specific time t. Individuals at risk are individuals in the cohort who did not present the event by time t, this and other related concepts are described in section 3.2.

The hazard ratio measures the hazard of an exposed cohort in relation to the hazard of an unexposed cohort at a time t,

$$HR(t) = \frac{\lambda_1(t)}{\lambda_0(t)} = \frac{d_1(t)/n_1(t)}{d_0(t)/n_0(t)}, \tag{2.2}$$

where $\lambda_1(t)$ and $\lambda_0(t)$ are the hazards at time t among individuals at risk in the exposed and unexposed cohorts, respectively.

The hazard ratio can vary, increase or decrease, at different time points within the time interval $[0, \tau]$ of study. Under the assumption of no variability of the hazard ratios over time, it can be expressed as a constant measure. This is also known as *proportional hazards*.

$$HR(t) = \frac{\lambda_1(t)}{\lambda_0(t)} = HR, \quad \text{for} t \in [0, \tau]. \tag{2.3}$$

Both measures, RR and HR, take on values in the interval $(0, \infty)$. A measure value equal to 1 indicates no difference in disease risk between the exposed and unexposed group, which implies that the exposure factor plays no role in the development of disease. A measure value different from 1 indicates that the exposure factor has an influence on the development of disease. In the latter case, a value greater than 1 indicates that the exposed group is at higher risk than the unexposed group, whereas a value less than 1 indicates that the unexposed group in the population is at higher risk than the exposed group.

2.2.3 Implications of genetics in association studies

Some issues, that are not seen in standard association studies, characterize association studies with genetic variables. In the following we mention these issues by assuming a biallelic SNP as a genetic factor under study, having allele A as the predisposing allele for the disease, and allele a as the wildtype allele. The predisposing allele is also called the *risk allele*, and it is meant to have an influence on the development of the disease. In genetic association studies, the minor allele is usually assumed as the risk allele.

For a more extensive explanation on these and other related considerations we refer to Cordell and Clayton (2005) and Lunetta (2008).

Risk factor

A biallelic SNP could be studied as potential risk factor for association with disease. The SNP factor can be defined as a variable at the allelic or at the genotype level.

At the allelic level, the SNP factor can be defined as a binary variable standing for the presence or absence of the risk allele for the disease.

At the genotype level, the SNP factor is most commonly defined as follows:

i) as a categorical variable with three categories denoting the three genotypes (AA, Aa, and aa) of the biallelic SNP. It is necessary to create two dummy variables, each for a different genotype, except for the reference genotype. Each dummy variable (coded as 1/0) codes for the presence/absence of the respective genotype. That allows studying the influence of the genotypes as independent categories of exposure.

ii) as an ordinal variable (coded as 0, 1, 2) denoting the number of copies of the risk allele carried in the genotypes, so that an additive influence of the alleles on the disease is assumed.

iii) as a binary variable by assuming a dominant or recessive influence of the risk allele on the disease. In the latter case, a binary variable for a dominant allele is coded as 1 for (AA, Aa), and 0 for aa; a binary variable for a recessive allele is coded as 1 for AA, and 0 for (Aa, aa).

Hardy-Weinberg equilibrium - HWE

It is advised that before starting the association study, each SNP in the data set should be tested for HWE (see definition in section 2.1.4). Testing for HWE implies to evaluate whether the condition of random mating holds in the population. Under random mating, the frequencies of the three genotypes AA, Aa, and aa are expected to be np^2, $2npq$, and nq^2, respectively, where n is the number of individuals. Deviations from these frequencies can be an indication of non-random mating, but also

of population stratification, or non-random genotyping error, or missing genotypes, all of which can lead to spurious associations.

To test for HWE we use the goodness of fit chi-squared test χ^2. It compares the observed genotype frequencies in the data with the expected genotype frequencies under HWE. The expected frequencies are obtained by estimating the allele proportions (p and q, where $p+q=1$) from the data, and computing the frequencies under HWE, i.e. $np^2, 2npq$, and nq^2. Then, we test a SNP for HWE with the chi-square test with 1 degree of freedom, $\chi^2_{(1)}$.

Lunetta (2008) states that the HWE should be tested only in full samples that were not ascertained on any specific phenotype under study, i.e. not based on the outcome of interest. For instance, in case-control studies, HWE can be tested in the control group if the trait is rare, i.e. if the trait has very low frequency in the population. However, HWE can be tested neither in the case nor in the control group if the trait is common. The genotypes associated with the disease are expected to be present in the case group at higher frequencies than under HWE. The latter affects the frequencies of the genotypes in the control group only if the disease is common.

2.2.4 Test of independence - Pearson's chi-squared test

The Pearson's chi-squared test χ^2 is used to assess the association between a risk factor and disease. Given the 2x2 contingency table 2.1, the χ^2 test compares the observed frequencies, from the data, with the expected frequencies under the null hypothesis of no association between risk factor and disease, i.e. statistical independence between risk factor and disease.

The χ^2 statistic is

$$\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i}, \tag{2.4}$$

where O_i is the observed frequency in the ith cell of the table and E_i is the expected frequency under the null hypothesis. The statistic follows a χ^2 distribution with 1 degree of freedom.

From Table 2.1, the observed frequencies O_i are d_1, d_0, $n_1 - d_1$, and $n_0 - d_0$. The respective expected frequencies E_i are estimated as $n_1 d/n$, $n_0 d/n$, $n_1(n-d)/n$, and $n_0(n-d)/n$.

For risk factors with more than two groups, the test statistics is applied similarly as above, but now the χ^2 statistic follows a χ^2 distribution with $(R-1)(C-1)$ degrees of freedom. R and C are the number of rows and columns in the contingency table, respectively. In a genetic study of association of the three genotypes of a biallelic SNP with the disease, we have a 3x2 contingency table, and thus, a 2 degrees of freedom test of association.

2.2.5 Logistic Regression models to estimate risk ratio with multiple risk factors

The effect of multiple risk factors on disease can be estimated with a logistic regression model. The risk factors can be of categorical and quantitative type. The logistic regression model provides estimates of the odds ratio (OR), that is a measure used in case-control designs. However, the OR approximates the RR when the outcome is rare, this makes the logistic model applicable for cohort studies in those situations. For common outcomes, some alternatives exist such as stratified analysis and the log-binomial model. Studies about different model alternatives and their performance for estimating RR have been published (Skov et al. 1998, McNutt et al. 2003, Cummings 2004, Deddens and Petersen 2004).

The logistic regression model is a popular method of extensive use in medical studies. It is useful for analysis of cohort data with rare outcomes. To model the association between risk factors and disease, the logistic regression model uses the logit link function of the probability of disease π_i of an individual i as a linear function of his/her exposure factors $\boldsymbol{X_i} = (X_{i1}, \ldots, X_{iK})'$, for K exposure factors. Let Y_i=1/0 denote the presence/absence of disease for individual i, then $\pi_i = P(Y_i = 1 \mid \boldsymbol{X_i})$.

The logit link of π_i is

$$logit(\pi_i) = \log\big(\frac{\pi_i}{1-\pi_i}\big), \tag{2.5}$$

and the logistic regression model has the form

$$logit(\pi_i) = \alpha + \boldsymbol{\beta' X_i}. \tag{2.6}$$

In the logistic regression model in (2.6) π_i is a function of the vector of exposure factors X_i, the coefficient of the intercept α, and the vector of regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$. The coefficients α and β are the parameters to be estimated. The model provides an approximation to RR of disease of an individual i through the estimate of $\exp(\boldsymbol{\beta' X_i})$.

Under the logistic regression model the probability π_i can be obtained as

$$\pi_i = \frac{\exp(\alpha + \boldsymbol{\beta' X_i})}{1 + \exp(\alpha + \boldsymbol{\beta' X_i})} = \frac{1}{1 + \exp(-(\alpha + \boldsymbol{\beta' X_i}))}, \tag{2.7}$$

and

$$1 - \pi_i = 1 - \frac{\exp(\alpha + \boldsymbol{\beta' X_i})}{1 + \exp(\alpha + \boldsymbol{\beta' X_i})} = \frac{1}{1 + \exp(\alpha + \boldsymbol{\beta' X_i})}. \tag{2.8}$$

The estimate of the parameter $\boldsymbol{\beta}$ can be obtained from the likelihood as a function

of π_i. The likelihood $L(\pi)$ considers the joint probability of developing disease of all individuals in the data. Each individual outcome is a Bernoulli variable Y_i, where $Y_i = 1$ for disease and $Y_i = 0$ for no disease. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ be the vector of the probabilities of disease of all individuals $i = 1, \ldots, n$. The likelihood is then expressed as

$$L(\pi) = \prod_{i=1}^{n} \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}. \tag{2.9}$$

By taking the natural-log of the likelihood, $\ell = Log(L)$, we get

$$\ell(\pi) = \sum_{i=1}^{n} Y_i \log \pi_i + \sum_{i=1}^{n} (1 - Y_i) \log(1 - \pi_i),$$

and in terms of the model parameters $\theta = (\alpha, \beta)$ it is

$$\ell(\theta) = \sum_{i=1}^{n} Y_i(\alpha + \boldsymbol{\beta'} \boldsymbol{X_i}) - \sum_{i=1}^{n} \log(1 + \exp(\alpha + \boldsymbol{\beta'} \boldsymbol{X_i}).$$

The solution for θ is the maximum likelihood estimate which must be obtained by an iterative procedure such as the Newton-Raphson algorithm because a solution as for a close-form expression does not exist. The solution gives estimates for each parameter, i.e. α, β_k, for $k = 1, \ldots, K$. The Wald test and likelihood ratio test can be used to test the null hypothesis of no association between an exposure factor and disease, $H_0 : \beta_k = 0$, which in turn indicates that $RR_k = \exp(\beta_k) = 1$, and therefore the null hypothesis can also be formulated in terms of the RR, $H_0 : RR_k = 1$. For a global null hypothesis of no association we formulate $H_0 : \boldsymbol{\beta} = 0$, which implies that no factor in the model is associated with the disease.

Chapter 3

Survival Analysis

Survival analysis refers to statistical methods to study time to the occurrence of an event, e.g. disease, death, etc., from a cohort. The existing methodologies allow to study data as a whole or in subgroups, and furthermore to investigate associations between risk factors and time to the occurrence of an event.

In this chapter we give a general overview over the features characterizing survival data and introduce definitions and notations used in survival analysis. Next, two important methods of extensive use in medical studies are presented: the log-rank test and the Cox regression model. The latter is the main method involved in our work, and therefore it is important for the development of the subsequent sections and chapters of this thesis.

In addition, we provide descriptions of some procedures judging the fit of the data by the Cox regression model. Finally, we summarize the findings of the paper by Müller et al. (2008) about the study of these methods in the context of genetic association studies with survival traits.

3.1 Framework of survival data

Survival data result from cohort studies with a limited period of follow-up. A sample of individuals from the study population is followed for a period of time until the outcome of interest occurs. This outcome is called *event*, the event could be any specific experience of an individual in the study, for example, diagnosis of any disease, death, recovery from disease, a physical status (e.g. obesity), etc. The event is also known as *failure*, but failure is mostly related to negative events, so that positive experiences such as recovery from disease would not be well denoted with the term failure. For our writing we prefer to use the general term event.

One key feature of survival data is that it considers the *time to event*, and it is the main variable of study. Hence, survival data are also known as *time-to-event data*.

Time to event, also called *survival time*, is the time since the starting follow-up time of an individual after enrolment to the study until the time of the first observation of the event of interest. The individuals can have a common starting time of follow-up. However, in realistic clinical settings individuals enrol in the study at different times during the study period, and have different starting times of follow-up. For instance, in the cohort study of haematopoietic stem cell transplantation (HSCT) described in chapter 6, the day of transplantation is the zero time point, i.e. the starting time of a patient for the observation of different specific post-transplant outcomes. Therefore, patients had different starting follow-up times according to the day of transplantation.

The survival time can be recorded in a specific time scale, for example, in years, months, or days. It may also refer to the earlier age at which an individual presents an event.

A second key feature of survival data is *censoring*. An individual is censored when his/her survival time is unknown. Censoring occurs i) because of the end of study, ii) because individuals are lost to follow-up, or iii) due to withdrawals from the study before the occurrence of the event. Specifically, censoring occurs for individuals i) who did not yet present the event by the end of the study period, ii) who, after being followed for some time after enrolment, were not reached anymore from some time on to the end of the study, iii) who voluntarily withdraw from the study at some time during the follow-up period.

Even if the event was not observed during the follow-up period, the three cases of censoring above bring some information about the *actual survival time* of the censored individuals. We know that it is longer than the time to the *last contact* of the individual, i.e. it is longer than the time to the end of study, longer than the time to loss to follow-up, or longer than the time to withdrawal. In these situations, the time to the last contact is recorded instead of the actual survival time. We say that the data are *right-censored* because the information on survival time is incomplete on the right side of the follow-up period. Hence, the actual survival time is shortened. This is known as *censored survival time*. If the event was observed and the actual survival time is recorded, then it is considered an *uncensored survival time*. Our work considers survival data with right-censored survival times.

Survival data can also be *left-censored*. This occurs when the survival time of an individual is incomplete on the left side of the follow-up period. The time of first exposure is unknown, and the follow-up period starts at some time after the exposure. Kleinbaum (1996) gives an example of cohort of patients with HIV infection, whose follow-up time started at the time of the first positive test for the HIV virus. Here, the time of first exposure to the HIV virus was unknown. Thus, the survival times of these patients were left-censored.

3.2 Terminology and notations

In this section we introduce the mathematical terminology and notations used for the analysis of survival data in this thesis. $\boldsymbol{T}$ is presented as a random variable of the survival time of an individual, and it is always positive, $\boldsymbol{T} \geq 0$. A specific value of interest for $\boldsymbol{T}$ is denoted by t.

The *indicator of censoring*, $\boldsymbol{\delta}$, is a random variable with value 1 or 0, depending on whether the actual or the censored survival time was recorded. Another way to view this is to assume a random variable $\boldsymbol{C}$ that denotes the time to censoring, then

$$\boldsymbol{\delta} = I(\boldsymbol{T} \leq \boldsymbol{C}) = \begin{cases} 1 & \text{if } \boldsymbol{T} \leq \boldsymbol{C}, \text{ i.e. actual survival time was recorded} \\ 0 & \text{if } \boldsymbol{T} > \boldsymbol{C}, \text{ i.e. survival time was censored.} \end{cases} \tag{3.1}$$

The *survival function* $S(t)$ is defined as the probability of survival longer than time t,

$$S(t) = P(\boldsymbol{T} > t). \tag{3.2}$$

$S(t)$ is a decreasing function from 1 to 0, $S(0) = 1$, $S(\infty) = 0$. Survival probability of $S(t) = 1$ means nobody in the cohort has yet presented the event. This happens at the starting time of follow-up ($t = 0$). Survival probability of $S(t) = 0$ means everybody in the cohort presented the event, nobody survived. This happens in the model at $t = \infty$, which indicates that for very long periods of follow-up all the individuals will necessarily present the event at some time. In practice long periods of follow-up are not common. Usually, there is a limit date for the end of the study, and $S(t)$ will tend to 0 only if the event to be observed occurs for everyone and if it occurs within a time length considerably shorter than the duration of the study.

The *hazard function* $\lambda(t)$, is defined as the instantaneous rate per unit time for the occurrence of the event. The hazard function is computed by considering only *individuals at risk* at the specific time t, where time may be considered as a continuous random variable or it may be discretized . An individual is at risk at time t if his/her survival time is greater than t.

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq \boldsymbol{T} < t + \Delta t | \boldsymbol{T} \geq t) \qquad \text{if } \boldsymbol{T} \text{ is continuous.} \tag{3.3}$$

$$\lambda(t) = P(\boldsymbol{T} = t | \boldsymbol{T} \geq t) \qquad \text{if } \boldsymbol{T} \text{ is discrete.} \tag{3.4}$$

Then, the *cumulative hazard function* up to time t is,

$$\Lambda(t) = \int_0^t \lambda(u) \mathrm{d}u. \tag{3.5}$$

The cumulative hazard for discrete times is computed by replacing the integral in equation (3.5) by the sum over the discrete times $\boldsymbol{T}$.

There is a mathematical relationship between $S(t)$ and $\lambda(t)$, that we demonstrate next. Let $f(t)$ be the *probability density function* for the occurrence of the event,

$$f(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq \boldsymbol{T} < t + \Delta t) \qquad \text{if } \boldsymbol{T} \text{ is continuous,} \tag{3.6}$$

$$f(t) = P(\boldsymbol{T} = t) \qquad \text{if } \boldsymbol{T} \text{ is discrete,} \tag{3.7}$$

and $F(t)$ be the *cumulative distribution function*, $F(t) = P(\boldsymbol{T} \leq t)$. Then, from equation (3.2) we have that $S(t) = 1 - F(t)$. By taking the derivative of $S(t)$ we get

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -f(t). \tag{3.8}$$

Then, the hazard function in (3.3) can also be expressed as

$$\begin{aligned} \lambda(t) &= \frac{f(t)}{S(t)} \qquad \text{from equations (3.3) and (3.6)} \\ &= -\frac{\frac{\mathrm{d}S(t)}{\mathrm{d}t}}{S(t)} = -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t}, \end{aligned} \tag{3.9}$$

and therefore,

$$S(t) = \exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right) = \exp\left(-\Lambda(t)\right). \tag{3.10}$$

3.3 The Kaplan-Meier estimator of survival and the log-rank test

It is useful to compare the survival function $S(t)$ between two or more groups in the cohort. One practical way to do this is with the Kaplan-Meier estimator (Kaplan and Meier 1958), that allows estimating survival curves and displaying them graphically.

The Kaplan-Meier (KM) estimator is a non-parametric approach for estimating the survival distribution $\hat{S}(t)$.

Let $i = 1, \ldots, n$ be the index for individuals in the survival data, and $\mathcal{D}$ the subset of individuals with uncensored survival times. Let t_i be the survival time of the ith individual for $i \in \mathcal{D}$. The KM survival curve is constructed by ordering the survival times t_i to yield $t_{(i)}$, and subsequently estimating the survival probability at each of the times $t_{(i)}$.

The KM survival probability at $t = 0$ is always $\hat{S}(0) = 1$, and remains so until just before the first event is observed. The KM survival estimator at $t_{(i)}$ is given by

$$\begin{aligned}\hat{S}(t_{(i)}) &= \prod_{i \in \mathcal{D}} P(\boldsymbol{T} > t_{(i)} | \, \boldsymbol{T} \geq t_{(i)}) \\ &= \hat{S}(t_{(i-1)}) \times P(\boldsymbol{T} > t_{(i)} | \, \boldsymbol{T} \geq t_{(i)}), \end{aligned} \tag{3.11}$$

and the probability in equation 3.11 is obtained by

$$\begin{aligned} P(\boldsymbol{T} > t_{(i)} | \boldsymbol{T} \geq t_{(i)}) &= 1 - P(\boldsymbol{T} = t_{(i)} | \boldsymbol{T} \geq t_{(i)}) \\ &= 1 - \hat{\lambda}(t_{(i)}) \\ &= 1 - \frac{d_{t_{(i)}}}{n_{t_{(i)}}}, \end{aligned}$$

where $d_{t_{(i)}}$ and $n_{t_{(i)}}$ are the observed number of events and the number of individuals at risk, respectively, at time $t_{(i)}$. The individuals at risk at $t_{(i)}$ are individuals who have not yet presented the event and are still on follow-up by time $t_{(i)}$, i.e. individuals with $\boldsymbol{T} \geq t_{(i)}$. Hence, $n_{(i)}$ decreases as the time $t_{(i)}$ increases.

The *KM survival curve* is a step function with jumps at $t_{(i)}$. It represents the survival curve over time of the whole set of individuals. We could also compute the survival curves by groups of individuals. For example, we might estimate separate curves $\hat{S}(t|X)$ for groups of exposed and unexposed individuals for any given risk factor X.

We can also obtain the so called *reverse KM estimator*, by reversing the coding of the indicator of censoring $\boldsymbol{\delta}$, i.e. create a new complement indicator $\boldsymbol{\delta_c}$, such that $\boldsymbol{\delta_c} = 0$ if the event is observed and $\boldsymbol{\delta_c} = 1$ if censoring occurs; then, compute the KM survival function as described above with the indicator of censoring $\boldsymbol{\delta_c}$. This reverse estimator is used in applications that adjust estimated measures by the loss of individuals during the follow-up time (Schemper and Henderson 2000, Schumacher et al. 2007). It can be viewed as the survival probability for censoring. We will also use this estimator later in this thesis.

Two or more groups of individuals can be formally compared to evaluate whether there are statistical differences between their KM survival curves. The *log-rank test* serves this purpose. It tests the null hypothesis of no association between risk factor and survival. The null hypothesis also means that there is no difference between

Table 3.1: Distribution of events at time $t_{(i)}$ in exposed and unexposed groups to a risk factor[a]

Risk Factor (X)	Events at time $t_{(i)}$		Total
	Yes	No	
exposed (1)	$d_{1t_{(i)}}$	$n_{1t_{(i)}} - d_{1t_{(i)}}$	$n_{1t_{(i)}}$
unexposed (0)	$d_{0t_{(i)}}$	$n_{0t_{(i)}} - d_{0t_{(i)}}$	$n_{0t_{(i)}}$
Total	$d_{t_{(i)}}$	$n_{t_{(i)}} - d_{t_{(i)}}$	$n_{t_{(i)}}$

[a] The letters in the cells represent absolute counts within the respective groups.

survival curves of the groups. In terms of genetic association studies with SNPs, the log-rank test tests the null hypothesis of no difference between survival curves of the different genotype groups, i.e. $H_0 : S_{AA}(t) = S_{Aa}(t) = S_{aa}(t)$, where $S_g(t)$ represents the survival curve for the genotype group g.

The log-rank test is a large-sample chi-square test that uses the difference between observed and expected number of events at times $t \,|\, \boldsymbol{\delta} = 1$. It requires proportional hazards between the different comparison groups. Two groups have proportional hazards if the ratio of their hazards is constant over time. The latter implies that the survival curves of the groups do not cross each other over time.

For a comparison of two survival curves, a 2x2 Table similar to Table 2.1 is constructed with the distribution of the number of events by risk factor. For survival data, the distribution of the number of events is obtained at each uncensored survival time, i.e. $t_{(i)} \,|\, \delta_i = 1$ (see Table 3.1).

The *exact log-rank statistic* can be viewed as the Cochran-Maentel-Haenzel statistic (Agresti 2002) stratified by times $t_{(i)} \,|\, i \in \mathcal{D}$. Let $O_{1t_{(i)}}$ and $E_{1t_{(i)}}$ be the observed and expected number of events in the exposed group at time $t_{(i)}$. The statistic assumes a hypergeometric distribution for $O_{1t_{(i)}}$. Then, if $O_{1t_{(i)}} = d_{1t_{(i)}}$, from Table 3.1 we have:

$$E_{1t_{(i)}} = E(O_{1t_{(i)}}) = n_{1t_{(i)}} \, \frac{d_{t_{(i)}}}{n_{t_{(i)}}}, \tag{3.12}$$

and

$$Var(O_{1t_{(i)}}) = \frac{n_{1t_{(i)}} \; n_{0t_{(i)}} \; d_{t_{(i)}} \; (n_{t_{(i)}} - d_{t_{(i)}})}{n_{t_{(i)}}^2 \; (n_{t_{(i)}} - 1)}. \tag{3.13}$$

The exact log-rank statistic is

$$\text{Log-rank statistic}_{exact} = \frac{\Big(\sum_{i\in\mathcal{D}} \big(O_{1t_{(i)}} - E_{1t_{(i)}}\big)\Big)^2}{\sum_{i\in\mathcal{D}} Var(O_{1t_{(i)}})},$$

which can also be expressed with the overall sums over time,

$$\text{Log-rank statistic}_{exact} = \frac{\Big(\sum_{i\in\mathcal{D}} O_{1t_{(i)}} - \sum_{i\in\mathcal{D}} E_{1t_{(i)}}\Big)^2}{\sum_{i\in\mathcal{D}} Var(O_{1t_{(i)}})}. \tag{3.14}$$

Under the null hypothesis of no association between risk factor and survival, the statistic in (3.14) follows a chi-square distribution with one degree of freedom. The exact log-rank statistic can be obtained by using either the exposed or the unexposed groups, they will give the same result. In our description above we have used the exposed group.

We can also test the association of survival with a risk factor with $G > 2$ groups, although the procedure becomes complex because it involves the covariances of $O_{g\,t_{(i)}}$ for $g = 1, \ldots, (G-1)$ (Kleinbaum 1996).

An approximation to the exact log-rank statistic in (3.14) is the formula of the classic chi-square form that considers independence between the groups (Kleinbaum 1996),

$$\text{Log-rank statistic}_{\chi^2} = \sum_{g=1}^{G} \frac{\Big(\sum_{i\in\mathcal{D}} O_{g\,t_{(i)}} - \sum_{i\in\mathcal{D}} E_{g\,t_{(i)}}\Big)^2}{\sum_{i\in\mathcal{D}} E_{g\,t_{(i)}}}, \tag{3.15}$$

this statistic follows a chi-squared distribution with $(G-1)$ degrees of freedom.

3.4 The Cox regression model

The Cox regression model is a semi-parametric method that models the association between risk factors and survival times. The main assumption of the method is that constant hazard ratios hold for all the factors in the model, this is well known as *proportional hazard (PH) assumption*, hence the Cox model is also known as *Cox proportional-hazard model*.

Two main aspects of the Cox regression model are to be noted. First, it models in terms of the hazards; and second, it provides hazard ratios, which allow comparing event occurrence between groups of individuals, and furthermore, allow judging the impact of risk factors on occurrence of events.

The formula of the Cox regression model is,

$$\lambda(t, \boldsymbol{X}) = \lambda_0(t)\ \exp(\boldsymbol{\beta'X}), \tag{3.16}$$

where

t	is a specific survival time,
$\boldsymbol{X}=(\boldsymbol{X_1},\ldots,\boldsymbol{X_K})'$	is the vector of K risk factors, where $\boldsymbol{X_k}=(X_{1k},\ldots,X_{nk})$, for $k=1,\ldots,K$, and for n individuals,
$\lambda(t, \boldsymbol{X})$	is the hazard as a function of t and $\boldsymbol{X}$,
$\lambda_0(t)$	is the baseline hazard as a function of t,
$\boldsymbol{\beta}=(\beta_1,\ldots,\beta_K)'$	is the vector of regression coefficients.

The Cox model has two types of parameters, the baseline hazard $\lambda_0(t)$ and the vector of coefficients $\boldsymbol{\beta}$. The baseline hazard $\lambda_0(t)$ is the hazard not dependent on the risk factors $\boldsymbol{X}$, i.e. the hazard when $\boldsymbol{X_k}=0$, $\forall k=1,\ldots,K$. Given the assumption of proportional hazards, the baseline hazard function describes the shape of the hazard functions of all other groups of exposure.

A particular characteristic of the Cox model is that we neither have to assume any specific functional form nor have to estimate the parameter $\lambda_0(t)$. We only need to estimate the parameter $\boldsymbol{\beta}$. Because not all the parameters require estimation, the model is called semi-parametric.

From the Cox model (3.16) we get the expression for the hazard ratio (HR) only in terms of the factors $\boldsymbol{X}$ and the respective parameters $\boldsymbol{\beta}$. The HR is constant over time, this comes from the basic assumption of the Cox regression model, the proportionality of the hazards regardless of the survival time.

$$HR = \frac{\lambda(t, \boldsymbol{X})}{\lambda_0(t)} = \exp(\boldsymbol{\beta'X}). \tag{3.17}$$

The HR is interpreted as a measure of the effect of risk factors on the survival time of the individuals. Specifically, the HR gives the risk of event for individuals with exposure $\boldsymbol{X}$ with respect to the risk of event for individuals with no exposure.

According to expression (3.17), to estimate HR we require to estimate the parameter $\boldsymbol{\beta}$, that can be estimated through maximum likelihood estimation.

3.4.1 Maximum likelihood estimation of β

The estimation of the parameter $\boldsymbol{\beta}$ is mathematically derived by maximizing a *likelihood function*, denoted by L. The likelihood describes the joint probability of all observations in the data as a function of the unknown parameter $\boldsymbol{\beta}$.

The likelihood function for the Cox model is a *partial likelihood function (PL)* because the computation involves individual likelihoods L_i only of individuals with uncensored survival time, i.e. $L_i, \forall\, \delta_i = 1$, or also $\forall i \in \mathcal{D}$.

The computation of a classical likelihood function would involve the likelihood of all individuals in the data set. Given the special feature of censoring in survival data this is not possible. However, the information of individuals with censored survival times ($\delta_i = 0$) is not left out completely. The hazards of these individuals are involved in the computation of the likelihood L_i, for $\delta_i = 1$, as long as they are at risk at time t_i.

Given that an event occurred at time t_i, and given a set of individuals at risk $\mathcal{R}_{t_i}$, an individual likelihood L_i is the probability that the event was observed on that individual i,

$$\begin{aligned} L_i &= \frac{\lambda(t_i, \boldsymbol{X_i})}{\sum_{j \in \mathcal{R}_{t_i}} \lambda(t_i, \boldsymbol{X_j})} \\ &= \frac{\lambda_0(t_i)\ \exp(\boldsymbol{\beta' X_i})}{\sum_{j \in \mathcal{R}_{t_i}} \lambda_0(t_i)\ \exp(\boldsymbol{\beta' X_j})} \\ &= \frac{\exp(\boldsymbol{\beta' X_i})}{\sum_{j \in \mathcal{R}_{t_i}} \exp(\boldsymbol{\beta' X_j})}, \end{aligned} \tag{3.18}$$

The partial likelihood is

$$\begin{aligned} PL &= \prod_{i \in \mathcal{D}} L_i \\ &= \prod_{i \in \mathcal{D}} \frac{\exp(\boldsymbol{\beta' X_i})}{\sum_{j \in \mathcal{R}_{t_i}} \exp(\boldsymbol{\beta' X_j})}. \end{aligned} \tag{3.19}$$

In order to get the estimates of $\boldsymbol{\beta}$, the partial likelihood is maximized. For compu-

tational ease the natural-log is used, $\ell = log(PL)$. We then maximize

$$\ell = \sum_{i \in \mathcal{D}} \Big(\boldsymbol{\beta' X_i} - \log \big(\sum_{j \in \mathcal{R}_{t_i}} \exp(\boldsymbol{\beta' X_j}) \big) \Big) \tag{3.20}$$

by solving the score equations

$$U(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = 0. \tag{3.21}$$

The score in terms of the model is expressed as

$$\begin{aligned} U(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} &= \sum_{i \in \mathcal{D}} \Big(\boldsymbol{X_i} - \frac{\sum_{j \in \mathcal{R}_{t_i}} \boldsymbol{X_j} \exp(\boldsymbol{\beta' X_j})}{\sum_{j \in \mathcal{R}_{t_i}} \exp(\boldsymbol{\beta' X_j})} \Big) \\ &= \sum_{i \in \mathcal{D}} (\boldsymbol{X_i} - \bar{\boldsymbol{X}}_{\boldsymbol{i}}), \end{aligned} \tag{3.22}$$

where $\bar{\boldsymbol{X}}_{\boldsymbol{i}}$ is a weighted average of $\boldsymbol{X_i}$, with weights $\exp(\boldsymbol{\beta' X_j})$, over the set of individuals $j \in \mathcal{R}_{t_i}$. The solution for $U(\boldsymbol{\beta}) = 0$ is the ML estimator $\hat{\boldsymbol{\beta}}$.

The variance of $\hat{\boldsymbol{\beta}}$ can be obtained by using the second derivative of l (Therneau and Grambsch 2000),

$$var(\hat{\boldsymbol{\beta}}) = \Big[- \frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^2} \Big]^{-1}, \tag{3.23}$$

where

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^2} = \sum_{i \in \mathcal{D}} \Big(- \frac{\sum_{j \in \mathcal{R}_{t_i}} (\boldsymbol{X_i} - \bar{\boldsymbol{X}}_{\boldsymbol{i}})^2 \exp(\boldsymbol{\beta' X_j})}{\sum_{j \in \mathcal{R}_{t_i}} \exp(\boldsymbol{\beta' X_j})} \Big). \tag{3.24}$$

3.4.2 Estimation of the hazard ratio and survival function

The ML estimator $\hat{\boldsymbol{\beta}}$ allows to estimate HRs directly. It also allows the estimation of the survival function, but still requires knowing the baseline survival, i.e. the survival function for a model not dependent on covariates, as we show next.

With the ML estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_K)$, the estimator $\widehat{\boldsymbol{HR}}$ is deduced from equation (3.17),

$$\widehat{\boldsymbol{HR}} = \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{X}). \tag{3.25}$$

Specifically, the risk of event of the ith individual can be estimated as

$$\widehat{HR_i} = \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{X_i}), \qquad \text{for } i = 1, \ldots, n,$$

and the risk of event due to the kth risk factor can be estimated as

$$\widehat{HR_k} = \exp(\hat{\beta}'_k), \qquad \text{for } k = 1, \ldots, K.$$

From the Cox model in (3.16) we formulate the estimator of the hazard,

$$\hat{\lambda}(t, \boldsymbol{X}) = \hat{\lambda}_0(t) \ \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{X}). \tag{3.26}$$

Hence, the estimator of the cumulative hazard is

$$\hat{\Lambda}(t, \boldsymbol{X}) = \hat{\Lambda}_0(t) \ \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{X}), \tag{3.27}$$

and the estimator of the general survival function in equation (3.10) is

$$\hat{S}(t) = \exp(-\hat{\Lambda}(t)). \tag{3.28}$$

Then, the estimator of a survival function given covariates is

$$\begin{aligned} \hat{S}(t|X) &= \exp(-\hat{\Lambda}(t, \boldsymbol{X})) \\ &= \exp(-\hat{\Lambda}_0(t) \ \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{X})) && \text{from equation (3.27)} \qquad (3.29) \\ &= \hat{S}_0(t)^{\exp(\hat{\boldsymbol{\beta}}' \boldsymbol{X})} && \text{from equation (3.28).} \qquad (3.30) \end{aligned}$$

There are two approaches to estimate the baseline survival function $\hat{S}_0(t)$: the Breslow estimator and the Kalbfleisch-Prentice estimator. Here we provide only the respective expressions. For details we refer to Therneau and Grambsch (2000).

The Breslow estimator of $S_0(t)$ is:

$$\hat{S}_0(t) = \exp(-\hat{\Lambda}_0(t)), \tag{3.31}$$

where the baseline cumulative hazard $\hat{\Lambda}_0(t)$ is

$$\hat{\Lambda}_0(t) = \sum_{t_i<t} \Big(\frac{d_{t_i}}{\sum_{j \in \mathcal{R}_{t_i}} \exp(\hat{\boldsymbol{\beta}} \boldsymbol{X_j})} \Big).$$

The Kalbfleisch-Prentice estimator of $S_0(t)$ is

$$\hat{S}_0(t) = \prod_{t_i<t} \hat{\alpha}_i \,, \tag{3.32}$$

where $\hat{\alpha}_i$, for $i \in \mathcal{D}$, are the ML estimators that satisfies

$$S(t|X) = \prod_{t_i<t} \alpha_i^{\exp(\boldsymbol{\beta' X})}.$$

3.4.3 Hypothesis test for association

To determine the statistical significance of $\hat{\boldsymbol{\beta}}$ we can use formal tests such as the *Wald test* or the *likelihood ratio (LR) test*. Both procedures test the null hypothesis of no association between risk factors and survival, i.e. $H_0 : \boldsymbol{\beta} = 0$ or $H_0 : HR = 1$, or for a specific risk factor, $H_0 : \beta_k = 0$.

Under this null hypothesis, the Wald statistic is computed by dividing the estimate $\hat{\beta}_k$ by its standard error $se(\hat{\beta}_k)$. It follows approximately a standard normal distribution, that corresponds to the Z distribution ($N(0,1)$),

$$\text{Wald statistic} = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)} \sim Z \,.$$

The LR statistic uses the log-likelihood of the restricted and unrestricted models. Given the Cox model $\lambda(t, \boldsymbol{X}) = \lambda_0(t) \exp(\boldsymbol{\beta' X})$, and given the null hypothesis $H_0 : \boldsymbol{\beta} = 0$, the restricted model $M0$ is reduced to the baseline hazard function $M0 : \lambda(t, \boldsymbol{X}) = \lambda_0(t)$. The unrestricted model $M1$ is a Cox model such that $\beta_k \neq 0$ for any $k = 1, \ldots, K$.

Given a null hypothesis for a specific factor k, $H_0 : \beta_k = 0$, the restricted model is $M0 : \lambda(t, X) = \lambda_0(t) \exp(\boldsymbol{\beta' X})$, with the kth component of $\boldsymbol{\beta}$ equal to zero, $\beta_k = 0$. The unrestricted model $M1$ is a Cox model such that $\beta_k \neq 0$.

The LR statistic is computed by $-2\,(\log L_{M0} - \log L_{M1})$, it follows a chi-square distribution with degrees of freedom (df) equal to the difference of df of the two models.

3.4.4 Extended Cox regression model

The proportional-hazard (PH) assumption of the Cox model implies that estimated survival curves of different risk groups do not cross each other over time.

If the PH assumption is not met, alternative procedures such as stratified analysis or an extended Cox model can be used. Stratified analysis allows fitting the Cox model with factors satisfying the PH assumption, and stratifying the model by the factor(s) not meeting the PH assumption. The stratified Cox model allows for different baseline hazards for each stratum, and produces the same constant hazard ratios of factors in the model for all the strata.

An extended Cox model allows including time dependent factors. These are factors which are not fixed over time. For example, let us consider an additional therapy to follow by some patients who underwent hematopoietic stem cell transplantation. We may want to evaluate whether the therapy has an influence on the death of these patients. Then, a factor $\boldsymbol{X_k}$ that stands for receiving therapy enters as a time dependent risk factor in the Cox model. By assuming that $\boldsymbol{X_k}$ is the only factor in the Cox model we have $\lambda(t, \boldsymbol{X_k}) = \lambda_0(t)\ \exp(\beta_k' X_k(t))$, where $X_k(t)$ equals 1 or 0 if the patient received or did not receive therapy by time t, respectively.

The extended Cox model allows to include both fixed and time dependent factors in the model. Then, the PH assumption is not valid anymore, and the hazard ratios are formulated as time dependent parameters. For a discussion on these alternatives refer to Kleinbaum (1996) and Therneau and Grambsch (2000).

3.5 Goodness of fit of the Cox model through an R^2 measure

An important step after fitting a model is to verify whether that model fits the data well. In ordinary linear regression models, the coefficient of determination, R^2, is the most popular measure for judging the good fit of a model. R^2 is interpreted as the fraction of explained variation by the model. In its general form, it is computed by comparing the sum of squared residuals generated from a model with the factors under consideration (covariate model) with that from a model without those factors (null model). The residuals of the data are obtained for each individual by the difference between the observed and expected value of the outcome, where the latter value is provided by the model. Thus, the residuals measure the accuracy of the model to estimate the outcome.

The corresponding residual of an ith outcome is $(Y_i - \hat{Y}_i)$. Then, the sum of squared residuals (SSR) of the model is

$$SSR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2,$$

where Y_i and $\hat{Y}_i$ are the observed outcome and the expected outcome given by the model, respectively.

The R^2 measure is defined as

$$R^2 = 1 - \frac{SSR_X}{SSR_0},$$

where the SSR_X and SSR_0 are the SSR from the covariate and null model, respectively.

The evaluation of the goodness of fit of a Cox model to survival data will indicate how well the factors in the model, and its corresponding estimated effects through the HR, represent the information from the data. Furthermore, it will also indicate how well these factors help explaining the variation of developing events at distinct times from individuals in the data.

For Cox regression models, however, the calculation of residuals, and hence of R^2, is not straightforward because of the censoring. As we have described, survival data include incomplete data that are accounted by an indicator of censoring δ. On this matter, some suggestions have been proposed to formulate complementary versions of R^2 for Cox regression models.

We are particularly interested in studying R^2 versions for application on genetic association studies, that is, to evaluate the goodness of fit of Cox regression models with genetic factors, especially SNPs. An interesting study on this particular topic was carried out by Müller et al. (2008).

Müller et al. (2008) aimed at identifying the most appropriate criterion for evaluating the goodness of fit of genetic covariates on survival outcome. They particularly focused on binary or trichotomous covariates, typical variable types of a genetic variant such as a SNP. The criteria should allow for interpretation of R^2 as explained variation and approximate the range [0,1] as for classical R^2 for linear models. Values of $R^2 = 0$ indicate no explained variation by the model, i.e. complete absence of association between the SNP factor in the model and survival. Values of $R^2 = 1$ indicate fully explained variation, i.e. the SNPs in the model are the responsible factors for occurrence of events, perfect association. Moreover the criteria should take into account the effect sizes (HR) of the covariates.

Three criteria were chosen for the investigation of the R^2 versions: the deviance residuals (Therneau et al. 1990), the deviation of survival (Schemper and Henderson 2000), and the Schoenfeld residuals (Schoenfeld 1982).

In the next sections of this chapter, we briefly describe the criteria used and the findings in the study by Müller (2004) and Müller et al. (2008). Although they considered two different versions for the criterion of deviation of survival, we describe only the one that performed better and which is relevant to our work.

3.5.1 Deviance residuals

Before defining the deviance residuals we first introduce the martingale residuals (Therneau et al. 1990). The martingale residuals are obtained for each individual i by

$$\hat{M}_i = \delta_i - \hat{\Lambda}(t_i)\,, \tag{3.33}$$

it can be viewed as the difference between the observed and the expected number of events of individual i at time t_i.

The sum of squared residuals is,

$$SSR_{\hat{M}} = \sum_{i=1}^{n} \hat{M}_i{}^2.$$

The above applies to a null model. Similarly, by using $\hat{\Lambda}(t_i, X)$ in equation (3.33) we can obtain $\hat{M}_{i|X}$ and $SSR_{M|X}$ for a covariate model. Then, the R^2 is formulated as

$$R^2_{\hat{M}} = 1 - \frac{SSR_{\hat{M}|X}}{SSR_{\hat{M}}}.$$

The martingale residuals tend to have a highly skewed distribution over the range $[-\infty, 1]$. This results from the fact that $\hat{\Lambda}(t_i)$ increases with longer survival times, and therefore $\hat{M}_i$ increases negatively with longer times t_i. Also, $SSR_{M|X}$ will tend to generate higher extreme values than SSR_M. Then, the R^2 version from this criterion will tend to generate negative values as well.

The deviance residual is a normalized transformation of the martingale residual (Therneau et al. 1990). It is similar to the deviance residual for a Poisson model (Agresti 2002). Considering the observed (δ_i) and the estimated ($\hat{\Lambda}(t_i)$) number of events, the deviance residual is

$$\widehat{dev.res}_i = sign\left(\delta_i - \hat{\Lambda}(t_i)\right)\sqrt{d_i}\,,$$

where

$$\begin{aligned} d_i &= -2\left(\delta_i \log\left(\frac{\hat{\Lambda}(t_i)}{\delta_i}\right) + \left(\delta_i - \hat{\Lambda}(t_i)\right)\right) \\ &= -2\left(\delta_i \log\left(\frac{\delta_i - \hat{M}_i}{\delta_i}\right) + \hat{M}_i\right). \end{aligned}$$

Thus, the deviance residuals is expressed by

$$\widehat{dev.res}_i = sign(\hat{M}_i)\sqrt{-2\left(\hat{M}_i + \delta_i \log\left(\frac{\delta_i - \hat{M}_i}{\delta_i}\right)\right)},$$

and the sum of squared residuals is

$$SSR_{\widehat{dev.res}} = \sum_{i=1}^{n}(\widehat{dev.res}_i)^2.$$

The R^2 version from this criterion has not been very satisfactory since it still generates negative values although less than the martingale residuals. An alternative proposed by Stark (1997) is to compute the $K_{d.norm}$, which is an R^2 version that compares the absolute difference between the deviance residuals from the covariate and null model with respect to the null model,

$$K_{d.norm} = \frac{\sum_{i=1}^{n} |\widehat{dev.res}_i - \widehat{dev.res}_{i|X}|}{\sum_{i=1}^{n} |\widehat{dev.res}_i|},$$

where $dev.res_{i|X}$ and $dev.res_i$ are the deviance residuals for the covariate and null model, respectively.

This version generates no negative R^2 values. However, the difference between the residuals of the two models could still be very high, so that the R^2 exceeds the limit of 1. That occurs especially for low censoring and high coefficients with high variance of the covariates.

3.5.2 Deviation of survival

Schemper and Henderson (2000) proposed a criterion based on the mean absolute deviation of survival at specific times t. Let $S_i(t)$ be the true survival status of individual i at time t such that $S_i(t) = 1$, if $t_i > t$; or $S_i(t) = 0$, if $t_i \leq t$. Let $S(t)$ be the survival probability at time t. The mean absolute deviation is defined as

$$M(t) = \frac{1}{n}\sum_{i=1}^{n} |\, S_i(t) - S(t)\,|.$$

By replacing the survival status $S_i(t)$ for all the n individuals we have:

$$\begin{aligned} M(t) &= S(t)\,|\,1 - S(t)\,| + (1 - S(t))\,|\,0 - S(t)\,| \\ &= S(t)\,(\,1 - S(t)) + (1 - S(t))\,S(t) \\ &= 2\,S(t)\,(\,1 - S(t)). \end{aligned}$$

This criterion is a time dependent measure. Since it is more practical to use an overall measure for the full follow-up time, a weighted average over t was suggested,

$$\begin{aligned} D(\tau) &= \frac{\displaystyle\int_0^\tau M(t)\,f(t)\,\mathrm{d}t}{\displaystyle\int_0^\tau f(t)\,\mathrm{d}t} \\ &= \frac{2\displaystyle\int_0^\tau \Big(S(t)\big(1 - S(t)\big)\Big)\,f(t)\,\mathrm{d}t}{\displaystyle\int_0^\tau f(t)\,\mathrm{d}t}. \end{aligned}$$

Let $D_X(\tau)$ be the mean absolute deviation for a covariate model, i.e. a model with survival function given by $S(t|X)$ instead of $S(t)$. Then, the R^2 version of the mean deviation of survival, denoted by $V(\tau)$ as a reference to the variance in survival for the null model (i.e. $S(t)\big(1 - S(t)\big)$), is

$$V(\tau) = 1 - \frac{D_X(\tau)}{D(\tau)}.$$

Given the presence of censoring in the data, the estimation of $D(\tau)$ is not simple. Schemper and Henderson (2000) formulated it by considering three possible situations at each observed and ordered survival time $t_{(j)}$, for $j \in \mathcal{D}$.

Let $S_i(t_{(j)})$ be the survival status at time $t_{(j)}$ for the ith individual, then

$$S_i(t_{(j)}) = \begin{cases} 1 & \text{if } i \text{ got no event up to time } t_{(j)} \;\; (t_i > t_{(j)}) \\ 0 & \text{if } i \text{ got the event at or before time } t_{(j)} \;\; (t_i \le t_{(j)}\,|\,\delta_i = 1) \\ \text{NA} & \text{no assigned, if } i \text{ is censored at or before time } t_{(j)} \;\; (t_i \le t_{(j)}\,|\,\delta_i = 0). \end{cases}$$

Then, a different term applies for each of the above possibilities $S_i(t_{(j)})$, and the

mean absolute deviation is given by the sum over the three possibilities,

$$\begin{aligned}\hat{M}(t_{(j)}) &= \frac{1}{n}\sum_{i=1}^{n} |\, S_i(t_{(j)}) - \hat{S}(t_{(j)}) \,| \\ &= \frac{1}{n}\sum_{i=1}^{n} \Big[\, I(t_i > t_{(j)})(1-\hat{S}(t_{(j)})) \\ &\quad + \delta_i\, I(t_i \le t_{(j)})\hat{S}(t_{(j)}) \\ &\quad + (1-\delta_i)\, I(t_i \le t_{(j)})\Big\{(1-\hat{S}(t_{(j)}))\frac{\hat{S}(t_{(j)})}{\hat{S}(t_i)} + \hat{S}(t_{(j)})(1-\frac{\hat{S}(t_{(j)})}{\hat{S}(t_i)})\Big\}\Big].\end{aligned}$$

For individuals with no assigned survival status $S_i(t_{(j)})$, an extrapolation was applied, assuming that they have identical risks to those from the uncensored individuals at $t_{(j)}$.

The overall measure is formed as a weighted average of $\hat{M}(t_{(i)})$, where the weights W_j are estimated as the number of deaths $d_{t_{(j)}}$ corrected by censoring,

$$\hat{D}(\tau) = \frac{\sum_{j\in\mathcal{D}} W_j\, \hat{M}(t_{(j)})}{\sum_{j\in\mathcal{D}} W_j}, \qquad W_j = \frac{d_{t_{(j)}}}{\hat{G}(t_{(j)})}.$$

The correction of censoring is done through the reverse Kaplan-Meier estimator (see page 20) denoted here as $\hat{G}(t_{(j)})$. The estimate of $\hat{G}(t_{(j)})$ gives the probability of not being censored up to time $t_{(j)}$. Then, W_j can be interpreted as the number of deaths if no censoring had occurred up to time $t_{(j)}$.

The estimate of $\hat{D}_X(\tau)$ for the covariate model is obtained in a similar way. By using $\hat{S}(t_{(j)}|X_i)$ instead of $\hat{S}(t_{(j)})$ we obtain $\hat{M}(t_{(j)}|X)$, and then $\hat{D}_X(\tau)$. Then,

$$\hat{V}(\tau) = 1 - \frac{\hat{D}_X(\tau)}{\hat{D}(\tau)}.$$

A similar criterion: the Brier score

Here we introduce a criterion that was not used in the study of Müller et al. (2008), but it is a similar criterion to the deviation of survival. It is the criterion of the *Brier score* (Brier 1950, Graf et al. 1999), and it is one of the main approaches to estimate prediction errors in survival models, which we discuss in the following chapters of this thesis.

The Brier score is based on the concept of the so-called verification score first proposed by Brier (1950), a measure applied for the study of misclassification error of models with binary or categorical outcomes. This measure was later adopted for the study of estimation errors from survival models (Graf et al. 1999; Gerds and Schumacher 2006, 2007).

The measure is formulated as the usual concept of residuals, that is, the squared deviation between observed and estimated outcomes under the model. In this case the outcome is survival, $S(t)$. To account for the loss of information due to censoring, Graf et al. (1999) formulated this measure as a weighted average of squared deviations, where the weights $W_C(\,t, \hat{G}\,)$ are summing to the total sample size n. Latter, Gerds and Schumacher (2006) extended these weights to allow for non-random censoring, $W_C(\,t, \hat{G}, \boldsymbol{X_i}\,)$. To use a similar notation as above, we denote this measure as SSR_{br}, where the subscript br denotes the Brier score. We present this measure as formulated by Gerds and Schumacher (2006),

$$SSR_{br}(t) = \frac{1}{n} \sum_{i=1}^{n} \left(S_i(t) - \hat{S}(t)\right)^2 \; W_C(\,t, \hat{G}, \boldsymbol{X_i}\,), \tag{3.34}$$

where

$$S_i(t) = \begin{cases} 1 & \text{if } i \text{ got no event up to time } t \; (t_i > t) \\ 0 & \text{otherwise,} \end{cases}$$

and

$$W_C(\,t, \hat{G}, \boldsymbol{X_i}\,) = \frac{I(t_i \leq t)\,\delta_i}{\hat{G}(t_i - \,|\boldsymbol{X_i})} + \frac{I(t_i > t)}{\hat{G}(t\,|\boldsymbol{X_i})}. \tag{3.35}$$

$W_C(\,t, \hat{G}, \boldsymbol{X_i}\,)$ is a weighting scheme determined only by the empirical estimate of the survival function for censoring ($\hat{G}$), the reverse KM estimator (see page 20). $W_C(\,t, \hat{G}, \boldsymbol{X_i}\,)$ does not involve the use of $\hat{\beta}$ estimates, or any other estimate based on the Cox model. Thus, the weights are independent of the fitted Cox model. Also, the weights allow adjustments for censoring conditional on the covariates $\boldsymbol{X_i}$ by using the conditional reverse KM estimator $\hat{G}(t\,|\boldsymbol{X_i})$, i.e. adjustment for non-random censoring.

The first term of $W_C(\,t, \hat{G}, \boldsymbol{X_i}\,)$ determines the weights for individuals with observed events at their respective survival times t_i–, where t_i– $\leq t$. The weights for these individuals are the inverse of the survival probability to censoring at the time when their events were observed, and given their covariates $\boldsymbol{X_i}$. Hence, the weights assigned to these individuals remain constant after their respective survival times t_i–. The second term of $W_C(\,t, \hat{G}, \boldsymbol{X_i}\,)$ determines the weights for individuals at risk at

time t. The weights for these individuals are the inverse of the survival probability to censoring at time t, given their covariates.

It can be noted that the weights apply for each individual i separately. Also, individuals censored before the specified time t are not considered, they get weights $W_C(\, t, \hat{G}, \boldsymbol{X_i}\,) = 0$. It means that an individual contributes to the estimate of survival deviations only until his/her time of censoring, if it occurs.

For a covariate model, $SSR_{br}(t|\boldsymbol{X})$ can be computed by replacing $\hat{S}(t)$ by $\hat{S}(t|\boldsymbol{X})$,

$$SSR_{br}(t|\boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^{n} \left(S_i(t) - \hat{S}(t|\boldsymbol{X_i})\right)^2 \; W_C(\, t, \hat{G}, \boldsymbol{X_i}\,). \tag{3.36}$$

The respective R^2 version can be defined as a time dependent measure,

$$\hat{R}^2_{br}(t) = 1 - \frac{SSR_{br}(t|\boldsymbol{X})}{SSR_{br}(t)}. \tag{3.37}$$

An overall measure of SSR can also be defined over the full follow-up time τ, which can be obtained by a cumulative $SSR_{br}(t|\boldsymbol{X})$ (Gerds and Schumacher 2007, Binder and Schumacher 2008). For the purpose of our work we will compute residuals at times of observed events, then

$$SSR_{br|X}(\tau) = \sum_{t \in t_{\mathcal{D}}} SSR_{br}(t|\boldsymbol{X}). \tag{3.38}$$

$SSR_{br}(\tau)$ can be computed similarly. Then, the R^2 version is

$$\hat{R}^2_{br}(\tau) = 1 - \frac{SSR_{br|X}(\tau)}{SSR_{br}(\tau)}. \tag{3.39}$$

Two main differences between the criterion based on the Brier score and the deviation of survival may be observed. First, the criterion of the Brier score assigns weights $W_C(\, t, \hat{G}, \boldsymbol{X_i}\,) = 0$ to individuals censored before time t, and therefore information of these censored individuals are left out from the computation of $SSR_{br}(t|\boldsymbol{X})$. In the criterion of deviation of survival, the weights are independent of censoring, therefore all individuals are considered; and moreover the deviation of survival at some time t after censoring of an individual is extrapolated based on the information of uncensored individuals. Second, the Brier score allows for adjustment of non-random censoring, while the criterion of deviation of survival does not.

3.5.3 The Schoenfeld residuals

This measure was proposed by Schoenfeld (1982). It estimates residuals by comparing the observed value of a covariate with its expectation under the Cox model. Given an observed event at time t, the Schoenfeld residuals compare the covariate value of the failing individual with the expected covariate value of a failing individual at time t according to the model.

The Schoenfeld residuals can be estimated only for individuals with uncensored survival times, i.e. only for $i \in \mathcal{D}$.

The Schoenfeld residuals are estimated separately for each covariate in the Cox model. Let X_{ik} be the value of the kth covariate for an ith individual such that $i \in \mathcal{D}$, and let $\mathcal{R}_{t_i}$ be the set of individuals at risk at time t_i. Then, the Schoenfeld residuals corresponding to the kth covariate are formulated as

$$sch_{ik}(\hat{\boldsymbol{\beta}}) = X_{ik} - E(X_{ik}|\mathcal{R}_{t_i}, \hat{\boldsymbol{\beta}}) \qquad \text{for } i \in \mathcal{D}, \quad \text{and } k = 1, \ldots, K,$$

where $E(X_{ik}|\mathcal{R}_{t_i}, \hat{\boldsymbol{\beta}})$ is the expected value of the covariate according to the Cox model, which is a function of the $\hat{\boldsymbol{\beta}}$ estimate from the Cox model.

At this specific step, we can view X_{ik} as a random variable with probability

$$\pi_j(\hat{\boldsymbol{\beta}}) \;=\; \frac{\exp(\hat{\boldsymbol{\beta}}' \boldsymbol{X_j})}{\sum_{j \in \mathcal{R}_{t_i}} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{X_j})} \qquad \text{for } j \in \mathcal{R}_{t_i} \text{ and } i \in \mathcal{D}.$$

Note that this probability is also the individual likelihood in equation 3.18.

Then, the expected value of the covariate is

$$\begin{aligned} E(X_{ik}|\mathcal{R}_{t_i}, \hat{\boldsymbol{\beta}}) \;&=\; \sum_{j \in \mathcal{R}_{t_i}} X_{jk} \;\; \pi_j(\hat{\boldsymbol{\beta}}) \\ &= \sum_{j \in \mathcal{R}_{t_i}} X_{jk} \; \frac{\exp(\hat{\boldsymbol{\beta}}' \boldsymbol{X_j})}{\sum_{j \in \mathcal{R}_{t_i}} \exp(\hat{\boldsymbol{\beta}}' \boldsymbol{X_j})}. \end{aligned} \tag{3.40}$$

The estimation of $E(X_{ik}|\mathcal{R}_{t_i}, \hat{\boldsymbol{\beta}})$ assumes that each individual $j \in \mathcal{R}_{t_i}$ will present the event at time t_i with probability $\pi_j(\hat{\boldsymbol{\beta}})$, which is a function of the estimate $\hat{\boldsymbol{\beta}}$ from the Cox model.

For a null model, the covariate is assumed to play no role on survival, that means

$\hat{\boldsymbol{\beta}} = 0$. Then, the residuals $sch_{ik}(0)$ should be computed with expected value

$$\begin{aligned} E(X_{ik}|\mathcal{R}_{t_i}, 0) &= \sum_{j \in \mathcal{R}_{t_i}} X_{jk} \ \pi_j(0) \\ &= \sum_{j \in \mathcal{R}_{t_i}} X_{jk} \ \frac{1}{\sum_{j \in \mathcal{R}_{t_i}} 1} \\ &= \sum_{j \in \mathcal{R}_{t_i}} X_{jk} \ \frac{1}{n_{t_i}}, \end{aligned} \tag{3.41}$$

where n_{t_i} is the number of individuals at risk at time t_i. In this case, all the individuals $j \in \mathcal{R}_{t_i}$ will have the same probability $\pi_j(0)$ to present the event at time t_i, and the expected value will simply be the average of covariate values over all individuals at risk. The weighted average of squared residuals $SSR_{sch|\hat{\beta}}$ is

$$SSR_{sch|\hat{\beta}} = \frac{1}{n_{\mathcal{D}}} \sum_{i \in \mathcal{D}} sch_{ik}^2(\hat{\boldsymbol{\beta}})$$

where $n_{\mathcal{D}}$ is the number of individuals in the subset $\mathcal{D}$. In a similar manner we compute $SSR_{sch|0}$. Then, the R^2 version of the Schoenfeld residuals (O'Quigley and Flandre 1994) for a model with only one covariate (K=1) is

$$\begin{aligned} \hat{R}^2_{sch} &= 1 - \frac{SSR_{sch|\hat{\beta}}}{SSR_{sch|0}} \\ &= 1 - \frac{\sum_{i \in \mathcal{D}} sch_{ik}^2(\hat{\boldsymbol{\beta}})}{\sum_{i \in \mathcal{D}} sch_{ik}^2(0)}. \end{aligned} \tag{3.42}$$

Schoenfeld residuals for multiple factors

For a model with more than one covariate, the concept of *prognostic index* is used instead of the value of the covariate. The prognostic index is the linear combination between the value of the covariates and their respective coefficients $(\hat{\boldsymbol{\beta}}'\boldsymbol{X})$. Therefore, for each individual $i \in \mathcal{D}$ there is a vector of residuals $sch_i(\hat{\boldsymbol{\beta}}) = (sch_{i1}(\hat{\boldsymbol{\beta}}), \ldots, sch_{iK}(\hat{\boldsymbol{\beta}}))$, and the residuals based on the prognostic index is $\hat{\boldsymbol{\beta}}' sch_i(\hat{\boldsymbol{\beta}})$, and the squared residuals is $\left(\hat{\boldsymbol{\beta}}' sch_i(\hat{\boldsymbol{\beta}})\right)^2$.

Moreover, the sum of squared residuals is formulated as a weighted sum, where the

weights $W_{KM}(t_i)$ are the length of the step of the marginal Kaplan-Meier survival curve at times t_i (O'Quigley and Xu 2001).

The weights $W_{KM}(t_i)$ represent the event distribution adjusted by the presence of censoring in the data. Hence, the larger the censoring at time t the larger the step height of the Kaplan-Meier curve, and thus, the larger the weights assigned to the residuals at this time. In the absence of censoring $W_{KM}(t_i)$ is constant at all times t_i, for $i \in \mathcal{D}$. $\sum_{i\in\mathcal{D}} W_{KM}(t_i) = 1$ if no individuals remain by the time of the last observed event.

Then, the general adjusted $SSR_{sch|\hat{\beta}}$ for multiple factors is

$$SSR_{sch|\hat{\beta}} = \frac{\sum_{i\in\mathcal{D}} W_{KM}(t_i) \left\{\hat{\boldsymbol{\beta}}' sch_i(\hat{\boldsymbol{\beta}})\right\}^2}{\sum_{i\in\mathcal{D}} W_{KM}(t_i)}.$$

and

$$\begin{aligned}\hat{R}^2_{sch} &= 1 - \frac{SSR_{sch|\hat{\beta}}}{SSR_{sch|0}} \\ &= 1 - \frac{\sum_{i\in\mathcal{D}} W_{KM}(t_i) \left\{\hat{\boldsymbol{\beta}}' sch_i(\hat{\boldsymbol{\beta}})\right\}^2}{\sum_{i\in\mathcal{D}} W_{KM}(t_i) \left\{\hat{\boldsymbol{\beta}}' sch_i(0)\right\}^2}. \end{aligned} \tag{3.43}$$

3.5.4 R^2 measures in Cox models with SNP factors

Müller et al. (2008) evaluated through simulation studies the performance of the R^2 versions from the three criteria: the deviance residuals ($K_{d.norm}$), the deviation of survival ($\hat{V}(\tau)$), and the Schoenfeld residuals ($\hat{R}^2_{sch}$). The application was done on genetic association studies, specifically for Cox regression models with SNP factors. These factors could enter the model either as binary (dominant or recessive effect of the allele) or trichotomous covariate (additive effect of the allele). The criteria were evaluated on the performance of the R^2 for varying percentage of censoring, for strength of the association with effect size, and for the limit in the range [0,1].

In general, they found that by the percentage of censoring in the data the $K_{d.norm}$ was the most affected and the $\hat{R}^2_{sch}$ the least affected. The former can be explained because the $K_{d.norm}$ does not correct for censoring while the others do. On the other hand, the R^2 versions from all the three criteria increased with higher effect

sizes (HR) of the SNPs. R^2 from additive effects of alleles were higher than the respective dominant or recessive effects. However, $\hat{R}^2_{sch}$ showed stronger dependence with effect sizes than the others, it increased faster with higher effect sizes, a fact that also makes $\hat{R}^2_{sch}$ cover best the upper range of 1.

Therefore, Müller et al. (2008) recommended the criterion of the Schoenfeld residuals as the most appropriate measure for evaluation of goodness of fit in association studies with genetic models. However, the identification of variables contributing well to the occurrence of an outcome based on a data set is not a guarantee that these variables will predict well the outcome. It may occur that the variables fit very well the particular data used for modelling the outcome, but it may not be that good if a different set of data, i.e. with a different set of individuals, is fitted. Then, the model would not be appropriate for prediction.

For prediction purposes, we require a model that fits good the data at hand as well as any subset of data from the same population of study. Then, further evaluation is required to validate a prediction model for the outcome.

3.6 Validation of Cox models

One important goal in medical studies is *prognostic modelling*. Prognostic models serve to make predictions about a future outcome of an individual based on *predictors* in the model. Predictors are variables that should have been identified during the fitting model procedure to play a role on the outcome of individuals. They should have been identified by analysing a representative sample from the population, whose individual's outcomes are to be predicted. During the fitting of a model these variables are called "covariates" or "risk factors". In the context of prognostic models, they are called "predictors".

A fitted model may prove to be a good fit for the data sample, this is determined in the evaluation of goodness of fit of the model. However, to prove if a model is a good prognostic model, it should be *validated*. Validation of a model is done through the estimation of *prediction errors*. By estimating prediction errors we test the fitted model to new data, independent from the data sample used in the fitting procedure.

By validating a model we prove that the risk factors predict well not only the outcomes for the particular individuals in a cohort, but also the outcome for any other individual from the study population. In genetic association studies, validated genetic models can be used to predict the risk of disease of an individual patient based on the genotypes carried at the specific loci in the genetic model. It can also be useful to predict the risk of a patient undergoing a specific therapy based on a joint influence of the genotypes with clinical or environmental predictors.

Validation procedures in Cox regression models has not been widely investigated yet. One approach was suggested by Gerds and Schumacher (2007), who used the criterion of the Brier score to estimate prediction errors. In the context of genetic association studies, we want to focus on the investigation of prognostic Cox models with SNP factors. On this matter, the results of the study by Müller et al. (2008) suggest that the criterion of the Schoenfeld residuals could be an acceptable approach. Hence, we developed an approach to estimate prediction errors with the criterion of the Schoenfeld residuals.

The methodologies for estimators of prediction errors, as well as our approach for application on Cox regression models using the criterion of the Schoenfeld residuals, are described in the next chapter.

Chapter 4

Prediction error estimators

Model fitting is useful for understanding the information from a population based on some available representative data of that population. By fitting a model we wish to identify variables associated with an outcome based on the information provided by individuals in the data. A fitted model can be a good representation of that information in the data, but that is not a guarantee that the model will represent well the information of individuals who are not in the data.

Model validation is the step to judge if a model will fit well the data of individuals in the population. Hence, we evaluate the validity of the model for a population and not only for the data at hand. A model is valid if it can closely predict the outcome of the individuals based on independent variables in the model, these variables are called *predictors*.

Through prediction errors we can evaluate the performance of a model with respect to predictions. A good model for prediction is expected to produce low prediction errors. The prediction errors are computed as the squared difference between a future response and its prediction from the fitted model. In practice, the future response can be taken from new data independent from the data used for fitting the model. However, there are not always new data available. Then, some techniques have been proposed to remedy the lack of new data.

Among some proposed techniques to replace the unavailability of data to estimate prediction errors are cross-validation (Stone 1974), bootstrap (Efron and Tibshirani 1993), the 0.632 estimator (Efron 1983), and the 0.632+ estimator (Efron and Tibshirani 1997). Both the cross-validation and the bootstrap estimator estimate prediction errors by using, in different ways, the available data to repeatedly simulate two subsets: the training sample to fit the model and the validation set to validate the model. The 0.632 estimator and the 0.632+ estimator were introduced as improvements to the two first aforementioned estimators.

Gerds and Schumacher (2007) extended the applicability of these estimators of prediction errors to survival data. Based on the definition of the Brier score (Brier 1950),

they computed prediction errors as the squared difference between the true survival status (taken from the validation set) and the estimated survival probabilities under the model (fitted with the training sample).

We propose the applicability of these estimators of prediction errors to survival data by using the definition of the Schoenfeld residuals (Schoenfeld 1982). By adapting the Schoenfeld residuals to prediction errors, we use the squared difference between the true value of a covariate (taken from the validation set) and the expected value of the covariate under the model (fitted with the training sample). We propose this approach to be applied on genetic association studies, because according to the investigation on genetic association studies by Müller et al. (2008), the criterion of the Schoenfeld residuals was the best to evaluate the goodness of fit of a genetic survival model (see section 3.5.4). Hence, we wish to provide the approach of estimation of prediction errors with the criterion of the Schoenfeld residuals.

Our main interest is to evaluate the improvement in prediction of a genetic model with respect to a reference model, e.g. model without the genetic factors. We can use an R^2 estimator as the goodness of fit but based on estimators of prediction errors. It will provide the fraction of contribution of the genetic factors to the prediction of the survival outcome, i.e. the gain in prediction by considering the genetic variables as risk factors for the survival outcome, in comparison to a reference model.

In the first section, this chapter introduces the concept of prediction errors, describes the existing estimators, and presents the adaptation of these estimators to survival data based on the Brier score as developed by Gerds and Schumacher (2007). The second section presents our development, the adaptation of the estimators of prediction errors to survival data based on the Schoenfeld residuals. Finally, we formulate the R^2 estimator based on estimators of prediction errors.

4.1 Prediction error

The *prediction error* is a measure of the capability of a model to predict the future response of new observations. In linear regression models, prediction error is defined as the squared difference between the value of a future response and its prediction under the model.

Brier (1950) introduced the concept of verification score to estimate the error of misclassification of the model to a set of categories. For a particular set of two categories, the verification score was defined as the mean squared difference between the occurrence of an event and the estimated probability of occurrence of that event. They used as an example the dichotomous outcome for the event of rain, the occurrence of rain was coded as rain=1 or no-rain=0, and the probability of raining was estimated from previous climatological observations. In that context, the verification score is equivalent to the above definition of prediction error, here the occurrence of rain (1/0) was the future response and the estimated probability of raining was

the predicted value. The verification score from Brier (1950) is known as *the Brier score*.

In any case, estimating prediction errors requires a *prediction rule* to predict the future outcome, and a set of independent new observations, which is called *validation set*, to test the performance of the prediction rule. The prediction rule can be either a model, a function or any other algorithm constructed with a training sample.

Estimating prediction errors is straightforward for models such as the linear regression models since the predicted outcome is obtained directly from the model, that is a linear function of the predictors and the estimated parameters of the model, i.e. the prediction rule is the model itself.

For models such as the logistic regression model, the predicted binary outcome is estimated by a probability π that is a non-linear function of the predictors and the parameters $\theta = (\alpha, \beta)$ of the model (see section 2.2.5), i.e. the prediction rule is not the model itself but a function of its parameters.

The expected prediction error of a prediction rule is formulated as

$$Err = E\left\{ Y_0 - r\left(\boldsymbol{X_0}\right)\right\}^2, \tag{4.1}$$

where Y_0 and $\boldsymbol{X_0} = (X_{01}, \ldots, X_{0K})$, are the values of the outcome and K predictors, respectively, of a random individual 0 drawn from the study population. r is the prediction rule for the outcome Y_0. $r(\boldsymbol{X_0})$ is the predicted outcome given the predictor $\boldsymbol{X_0}$.

A value of $Err = 0$ indicates the full validity of the prediction rule, that occurs if $Y_0 = r(\boldsymbol{X_0})$. The larger the distance between Y_0 and $r(\boldsymbol{X_0})$ the larger the prediction errors, and therefore the less valid the prediction rule.

In practice we can estimate Err in terms of a prediction error rate by averaging the prediction errors for all individuals in a validation set. Let V_0 be a validation set, and n_{V_0} be the size of V_0, then

$$\widehat{Err} = \frac{1}{n_{V_0}} \sum_{i=1}^{n_{V_0}} \left\{ Y_i - r\left(\boldsymbol{X_i}\right)\right\}^2. \tag{4.2}$$

Then, estimating Err requires two independent sets of data: the training sample to estimate the prediction rule r, and the validation set providing the outcomes Y_i and predictors $\boldsymbol{X_i} = (X_{i1}, \ldots, X_{iK})$, for $i = 1, \ldots, n_{V_0}$, to test the performance of the prediction rule. The estimate $\widehat{Err}$ is the average squared difference between the observed outcome Y_i and the predicted outcome $r(\boldsymbol{X_i})$ over all individuals in the validation set. The predicted outcome is obtained from the prediction rule given the predictor set $\boldsymbol{X_i}$.

However, a new data set as validation set is not always available, or to collect new data will be costly or it will require great investment of time. Thus, some tech-

niques have been suggested to estimate Err without demanding new data. These techniques estimate Err using the available data in different ways, by considering the availability of a training sample to estimate the prediction rule, and of a validation set to evaluate the prediction rule. The techniques that we will consider here are: the apparent error ($\overline{err}$), the bootstrap error ($\widehat{Err}_B$), the cross-validation error ($\widehat{Err}_{CV}$), the bootstrap cross-validation ($\widehat{Err}_{B0}$), the 0.632 estimator ($\widehat{Err}_{.632}$), and the 0.632+ estimator ($\widehat{Err}_{.632+}$), which are described in the following sections.

4.2 General description of prediction error estimators

Before describing the estimators of prediction errors, we introduce some useful notations used in this and subsequent sections:

Q	available data of size n
n	sample size of Q
V_0	validation set independent from Q
n_{V_0}	sample size of V_0
Q^*	a bootstrap sample of size n drawn at random with replacement from Q
Q^0	a validation set. It is obtained as the subset of Q that is not included in the bootstrap sample Q^*
Q_b^*	a bth bootstrap sample Q^*, for $b = 1, \ldots, B$
Q_b^0	a bth validation set Q^0, for $b = 1, \ldots, B$
n_{b0}	sample size of Q_b^0
r	the prediction rule estimated with Q
r_b^*	the prediction rule estimated with Q_b^*.

4.2.1 Apparent error estimator

The apparent error estimates Err by using the available data for both estimating the prediction rule and testing its performance. That means, the available data Q are used as both training sample and validation set.

$$\overline{err} = \frac{1}{n} \sum_{i=1}^{n} \{ Y_i - r(\boldsymbol{X_i}) \}^2,$$

where $(Y_i, \boldsymbol{X_i})$ are the outcome and predictor values, respectively, of the ith individual in Q. r is the estimated prediction rule with the training sample Q. $r(\boldsymbol{X_i})$ is the predicted outcome given the predictor $\boldsymbol{X_i}$.

The drawback of this estimator of prediction errors is that it uses the same data as both training sample and validation set. The $\overline{err}$ corresponds to the general

definition for evaluation of goodness of fit, which is known to underestimate Err, i.e. to produce negative bias (Efron 1983, Gerds and Schumacher 2007). The prediction rule will predict the outcomes from the training sample more accurately than the outcomes from a validation set independent from the training sample. Therefore, $\overline{err}$ will produce lower prediction errors than expected under independence of the training sample and validation set.

4.2.2 Bootstrap cross-validation estimator

In this part we first describe the techniques of bootstrap and cross-validation because they have the basic ideas of the bootstrap cross-validation. These techniques use the available data to repeatedly simulate training samples and validation sets, so that we can estimate prediction errors with the availability of two different data sets.

Bootstrap

Given the available data Q, the simplest bootstrap approach to estimate prediction errors consists of the following (Efron and Tibshirani 1993):

1. Draw at random and with replacement a sample of size n from Q. We obtain a bootstrap sample Q^*.
2. Estimate the prediction rule using the bootstrap sample Q^*. Then, test the rule on the original data Q and estimate the prediction error. Hence, Q^* is the training sample and Q is the validation set.
3. Repeat steps 1. and 2. B times to obtain a set of B estimates of prediction errors.
4. Compute the average over the B estimates to obtain the bootstrap estimate of prediction errors.

For an extended theory on bootstrap methods we refer to Efron and Tibshirani (1993).

The bootstrap estimator of Err is

$$\widehat{Err}_B = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n}\sum_{i=1}^{n}\{\, Y_i - r_b^*\,(\boldsymbol{X_i})\,\}^2,$$

where $(Y_i, \boldsymbol{X_i})$ are the outcome and predictor values, respectively, of the ith individual in Q, and $r_b^*(\boldsymbol{X_i})$ is the predicted outcome under the estimated prediction rule r_b^*.

Even if $\widehat{Err}_B$ is an improved estimator to the $\overline{err}$, Efron (1983) found that it still underestimates Err, although it has acceptable variance. The underestimated error

can be explained by the fact that part of the individuals in the validation set are also in the training sample. Hence, the prediction rule is partially tested on the same individuals used to estimate that prediction rule.

Cross-validation

Cross-validation (Stone 1974) implies to split the available data Q into a training sample and a validation set, so that we have two independent sets of data: one to estimate and one to evaluate the prediction rule.

The cross-validation consists of the following:

1. Split Q into H equal-sized parts. Get Q_h, for $h = 1, \ldots, H$.
2. Take a hth subset apart. Estimate the prediction rule with the set Q_{-h}. Then, test the prediction rule using the Q_h subset and estimate the prediction error. Hence, Q_{-h} is the training sample and Q_h is the validation set.
3. Repeat step 2. for each subset Q_h to obtain a set of H estimates of prediction errors.
4. Compute the average over the H estimates to obtain the cross-validation estimate of prediction errors.

This is also called H-fold cross-validation. For an extended theory on cross-validation estimators we refer to Stone (1974) and Efron and Tibshirani (1993).

The cross-validation estimator of Err is

$$\widehat{Err}_{CV} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h}\sum_{i\in h}\{\, Y_i - r_{-h}\,(\boldsymbol{X_i})\,\}^2,$$

where $(Y_i, \boldsymbol{X_i})$ are the outcome and predictor values, respectively, of the individual i in the validation set Q_h, and $r_{-h}(\boldsymbol{X_i})$ is the predicted outcome under the estimated prediction rule r_{-h}. n_h is the sample size of Q_h.

In the studies of Efron (1983) and Efron and Tibshirani (1997), the cross-validation was applied with different H-fold partitions. They found that this approach tends to produce higher variability when H increases although it approximates Err very well. For instance, with the leave-one-out cross-validation ($H = n$), it estimates quite unbiased Err but with very high variability, whereas for a 5-fold ($H = 5$) or 10-fold cross-validation ($H = 10$) it gives lower variability but overestimates Err, i.e. estimates with positive bias.

Bootstrap cross-validation

The bootstrap cross-validation is a combination of the bootstrap and the cross-validation described above. It consists of the following:

1. Draw at random and with replacement a sample of size n from Q. We obtain a bootstrap sample Q^*.
2. Estimate the prediction rule using the bootstrap sample Q^*. Then, test the prediction rule on the set Q^0, which is the subset of Q not included in Q^*, and estimate the prediction error. Hence, Q^* is the training sample and Q^0 is the validation set.
3. Repeat steps 1. and 2. B times. We obtain a set of B estimates of prediction errors.
4. Compute the average over the B estimates to obtain the bootstrap cross-validation estimate of prediction errors.

The bootstrap cross-validation estimator of Err is

$$\widehat{Err}_{B0} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n_{b0}}\sum_{i\in Q_b^0}\{\, Y_i - r_b^*\,(\boldsymbol{X_i})\,\}^2, \tag{4.3}$$

where $(Y_i, \boldsymbol{X_i})$ are the outcome and predictor values, respectively, of the individual i in the bth validation set Q_b^0, $r_b^*(\boldsymbol{X_i})$ is the predicted outcome under the estimated prediction rule r_b^*, and n_{b0} is the sample size of Q_b^0, for $b = 1, \ldots, B$.

Notice that the sets Q_b^0, for $b = 1, \ldots, B$, may differ in size from one bootstrap sample to the other.

In the original papers $\widehat{Err}_{B0}$ was defined slightly different than in equation (4.3). Efron (1983) defined $\widehat{Err}_{B0}$ as a general average over all the individual errors obtained from all the sets Q_b^0, for $b = 1, \ldots, B$. Efron and Tibshirani (1997) defined $\widehat{Err}_{B0}$ by first averaging the prediction errors for each individual i over the sets $\{Q_b^0 \,|i \in Q_b^0\}$, and then by averaging over all average individual errors i. Efron and Tibshirani (1997) stated that these two definitions agree as $B \to \infty$, and they produced nearly the same results in their simulations.

Equation (4.3) follows the definition of $\widehat{Err}_{B0}$ as presented by Gerds and Schumacher (2007) in the estimation of prediction errors for survival data. It first takes the average prediction error over all individuals $i \in Q_b^0$, for each $b = 1, \ldots, B$. Then, it takes the average over the B average prediction errors. As stated above, these slight differences in the calculation should lead to similar results. We find the latter definition is a more natural calculation of the mean errors and it is the definition we use in the next sections.

As a type of cross-validation approach the bootstrap cross-validation also tends to estimate Err with positive bias (Efron and Tibshirani 1997, Gerds and Schumacher 2007). If we consider only the original elements of Q in the bootstrap sample Q_b^*, the sample size of Q_b^* is reduced in comparison to the original sample size n. Considering the experience with the cross-validation approach, the positive bias is produced when part of the data was excluded from the training sample. Since Q_b^* in this approach also excludes part of the data, this approach tends to estimate Err with positive bias.

4.2.3 The 0.632 estimator

The 0.632 estimator ($\widehat{Err}_{.632}$) was proposed by Efron (1983) as a correction to the negative bias of the apparent error estimator ($\overline{err}$) and the positive bias of the bootstrap cross-validation estimator ($\widehat{Err}_{B0}$). The former occurs because the prediction rule is estimated and evaluated on the same data, the latter occurs because for each bootstrap sample the prediction rule is estimated using less information than contained in the original data (Efron 1983, Efron and Tibshirani 1993, Schumacher et al. 2007).

According to the study of Efron (1983), $\widehat{Err}_{B0}$ is computed from bootstrap samples containing about 0.632 times the total information of the original data. This was the argument to formulate $\widehat{Err}_{.632}$ as a linear combination of $\overline{err}$ and $\widehat{Err}_{B0}$, with weight 0.632 for $\widehat{Err}_{B0}$.

$$\widehat{Err}_{.632} = (1 - 0.632)\,\overline{err} \;+\; 0.632\,\widehat{Err}_{B0}. \tag{4.4}$$

Efron (1983) compared the performance of the bootstrap, the cross-validation, the 0.632, and other related alternatives for estimation of Err. The 0.632 estimator appeared to have the best performance at approximating Err with less bias and variance than other alternatives.

Efron and Tibshirani (1997) argued that highly overfitting rules do not benefit from the 0.632 estimator. Highly overfitting rules tend to exaggeratedly predict individual outcomes in the data rather than generalize the relation between outcome and predictors. These rules produce severe negative bias of the apparent error estimate ($\overline{err} \to 0$). In that case, the prediction error should be estimated only from the bootstrap cross-validation estimator, i.e. $\widehat{Err} = \widehat{Err}_{B0}$. However, the 0.632 estimator yields to $\widehat{Err}_{.632} = 0.632\,\widehat{Err}_{B0}$, which underestimates the prediction error. Thus, an alternative estimator that corrects for overfitting of the prediction rule was proposed, this was the 0.632+ estimator ($\widehat{Err}_{.632+}$).

4.2.4 The 0.632+ estimator

Efron and Tibshirani (1997) proposed the 0.632+ estimator as a correction to the 0.632 estimator. The purpose was to make the estimator valid for highly overfitting rules, for which the $\widehat{Err}_{.632}$ estimate results in an underestimated Err.

The correction affects the constant 0.632 used as a weight in the 0.632 estimator (equation (4.4)). The 0.632+ estimator ($\widehat{Err}_{.632+}$) makes the weight vary and depend on the amount of overfitting of the prediction rule, such that $\widehat{Err}_{.632+}$ assigns larger weights to $\widehat{Err}_{B0}$ when the amount of overfitting is large. The amount of overfitting is measured by ($\widehat{Err}_{B0} - \overline{err}$). However, the 0.632+ estimator uses in the calculation the *relative overfitting rate* (R), that is a scaled amount of overfitting.

The latter requires first the definition of the *no-information error rate* (γ). The no-information error rate is the expected prediction error when the rule is tested on data where the outcomes are independent of the predictors.

$$\gamma = E_{ind}\,\{\,Y_i - r\,(\boldsymbol{X_i})\,\}^2, \qquad \text{for } i = 1, \ldots, n,$$

where E_{ind} indicates expectation under independence of outcomes Y and predictors $\boldsymbol{X}$.

Since the rule is estimated with a training sample, γ will tend to be larger as the rule tends to overfit the training sample.

To estimate γ, data with independent outcomes and predictors can be obtained by permuting the respective variables Y and $\boldsymbol{X}$ of the available data, so that each outcome Y_i is combined with every $\boldsymbol{X_j}$ for $i, j = 1, \ldots, n$. Then, the estimator $\hat{\gamma}$ is

$$\hat{\gamma} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \{\,Y_i - r\,(\boldsymbol{X_j})\,\}^2. \tag{4.5}$$

Since $\hat{\gamma}$ gives an estimated error of the prediction rule under no relation between outcomes and predictors, the *no-information value* ($\hat{\gamma} - \overline{err}$) gives an estimate of the maximum possible amount of overfitting of the prediction rule.

The relative overfitting rate $\hat{\mathsf{R}}$ is then defined as the amount of overfitting with respect to the no-information value.

$$\hat{\mathsf{R}} = \frac{\widehat{Err}_{B0} - \overline{err}}{\hat{\gamma} - \overline{err}}. \tag{4.6}$$

The $\hat{\mathsf{R}}$ values are in the range [0,1]. $\hat{\mathsf{R}} = 0$ if $\widehat{Err}_{B0} = \overline{err}$, that means no overfitting rule. $\hat{\mathsf{R}} = 1$ if the amount of overfitting ($\widehat{Err}_{B0} - \overline{err}$) equals the no-information value ($\hat{\gamma} - \overline{err}$), that indicates highly overfitting rule.

The weight $\hat{w}$ is then formulated as

$$\hat{w} = \frac{0.632}{1 - 0.368\,\hat{\mathsf{R}}} \quad , \qquad \hat{w} \in [0.632, 1], \tag{4.7}$$

and the 0.632+ estimator is

$$\widehat{Err}_{.632+} = (1 - \hat{w})\,\overline{err} + \hat{w}\,\widehat{Err}_{B0}. \tag{4.8}$$

For larger amount of overfitting (i.e. when $\hat{\mathsf{R}}$ approximates 1), $\hat{w}$ tends to 1. That means, under highly overfitting rules the bootstrap cross-validation estimator $\widehat{Err}_{B0}$ has weights larger than 0.632.

For reduced amount of overfitting (i.e. when $\hat{\mathsf{R}}$ approximates 0), $\hat{w}$ tends to 0.632. That means, under no overfitting rules the bootstrap cross-validation estimator has weights 0.632. The latter means that under no overfitting rules $\widehat{Err}_{.632+} = \widehat{Err}_{.632}$.

In some instances $\hat{\mathsf{R}}$ may fall out of the range [0,1]. That happens when $\hat{\gamma} \leq \overline{err}$ or $\overline{err} < \hat{\gamma} \leq \widehat{Err}_{B0}$. To avoid these situations $\widehat{Err}_{B0}$ and $\hat{\mathsf{R}}$ are redefined:

$$\widehat{Err}_{B0}^{\,\prime} = min\,(\widehat{Err}_{B0}, \hat{\gamma}\,), \tag{4.9}$$

$$\hat{\mathsf{R}}^{\prime} = \begin{cases} \dfrac{\widehat{Err}_{B0}^{\,\prime} - \overline{err}}{\hat{\gamma} - \overline{err}} & \text{if } \hat{\gamma},\ \widehat{Err}_{B0}^{\,\prime} > \overline{err} \\ \\ 0 & \text{otherwise.} \end{cases} \tag{4.10}$$

For the computation of $\widehat{Err}_{.632+}$, $\hat{\mathsf{R}}^{\prime}$ and $\widehat{Err}_{B0}^{\,\prime}$ are then plugged into equations (4.7) and (4.8), respectively. Note that if $\overline{err} < \hat{\gamma} \leq \widehat{Err}_{B0}$, then $\hat{\mathsf{R}}^{\prime} = 1$ and $\widehat{Err}_{.632+} = \hat{\gamma}$.

4.3 The 0.632 and 0.632+ estimators based on survival probabilities

Gerds and Schumacher (2007) extended the methodology of the 0.632+ estimator to the application on right-censored survival data. They used the criterion of the Brier score (Brier 1950) to estimate the prediction errors. Hence, the prediction errors were computed as the squared difference between the true survival status of an individual and the predicted survival probability based on the Cox regression model.

The basis of the Brier score as applied on survival data was described in section 3.5.2, page 33. In this section we describe and formulate the use of the Brier score to estimate prediction errors, and consider the Cox regression model as the prediction rule.

The expected prediction error of a prediction rule in survival data is formulated as a time dependent prediction error

$$Err(t) = E\ \{\ Y_0(t) - r\ (t|\boldsymbol{X_0})\ \}^2, \tag{4.11}$$

where $Y_0(t)$ is the survival status at time t, and $\boldsymbol{X_0}$ is the vector of predictors of a random individual drawn from the study population. Hence,

$$Y_0(t) = S(t) = \begin{cases} 1 & \text{if individual 0 got no event until time } t\ (t_0 > t) \\ 0 & \text{otherwise.} \end{cases} \tag{4.12}$$

The prediction rule r is the survival probability as a function of the parameters of the Cox regression model (see equation (3.29)). $r(t|\boldsymbol{X_0})$ is the predicted survival probability at time t given the predictor $\boldsymbol{X_0}$. Hence, $r(t|\boldsymbol{X_0}) = \hat{S}(t|\boldsymbol{X_0})$, which is estimated from the survival probability in equation (3.29) evaluated on $\boldsymbol{X_0}$, i.e. given $\boldsymbol{X_0}$.

It can be realized that the prediction error in (4.11) considers the survival status $S(t)$ as a binary variable with values 1/0, whereas the predicted survival $\hat{S}(t)$ is a continuous variable in the range [1,0].

By following the expression of $\widehat{Err}$ in equation (4.2), the estimate $\widehat{Err}$(t) can be obtained as the average of the prediction errors over all individuals in a validation set. Let n_{V_0} the size of a validation set, then

$$\widehat{Err}(t) = \frac{1}{n_{V_0}} \sum_{i=1}^{n_{V_0}} \{\ Y_i(t) - r\ (t|\boldsymbol{X_i})\ \}^2\ W_C(\ t, \hat{G}, \boldsymbol{X_i}\ ). \tag{4.13}$$

However, because of the special characteristics of survival data, two new features can be observed in this estimator of prediction error (equation (4.13)), these are: i) the time dependent estimate of the prediction errors, that accounts for the time frame t to survival, and ii) the adjustment of the prediction errors by a weight function $W_C(\ t, \hat{G}, \boldsymbol{X_i}\ )$ that is the reverse Kaplan-Meier function (see page 34).

Some reminds first, the weight function $W_C(\ t, \hat{G}, \boldsymbol{X_i}\ )$ is independent of the Cox model that in this case is the prediction rule r. $\hat{G}$ is a survival function of the censoring variable conditional on the covariates $\boldsymbol{X_i}$, so that $\hat{G} = \hat{G}(t\,|\boldsymbol{X_i})$. These weights were introduced to reduce the bias due to censoring from the estimates of prediction errors.

Since the $\widehat{Err}$(t) is a time dependent estimator, it allows generating prediction error curves against time (Gerds and Schumacher 2007, Schumacher et al. 2007).

In the absence of a validation set the different estimators described in sections 4.2.1 to 4.2.4 can be applied to estimate $Err(t)$. In the following we present these estimators as adapted by Gerds and Schumacher (2007).

The apparent error estimator

Given the available data set Q of size n, the apparent error is computed by considering Q as both the training sample and validation set:

$$\overline{err}\,(t) = \frac{1}{n}\sum_{i=1}^{n}\{\, Y_i(t) - r\,(t|\boldsymbol{X_i})\,\}^2 \;\; W_C(\,t, \hat{G}, \boldsymbol{X_i}\,), \tag{4.14}$$

where $(Y_i(t), \boldsymbol{X_i})$ are the survival status at time t ($S_i(t)$) and the predictor values, respectively, of the ith individual in Q. The $Y_i(t)$ take values 1 or 0 as defined in equation (4.12).

r is the prediction rule, the Cox regression model fitted with the training sample Q. $r(t|\boldsymbol{X_i})$ is the predicted survival probability $\hat{S}(t|\boldsymbol{X_i})$ according to the fitted Cox regression model and tested at a predictor value $\boldsymbol{X_i}$. The weights $W_C(\,t, \hat{G}, \boldsymbol{X_i}\,)$ are based on the reverse KM survival function from the data set Q.

It can be noticed that the apparent error estimator is the same as the estimator of the sum of squared deviation $SSR_{br}(t|\boldsymbol{X})$ in equation (3.36). As it has been indicated, residuals are obtained to evaluate the goodness of fit of a model to the data at hand. That means, to fit and evaluate the Cox model, which here is the prediction rule, using the same data. The estimator $SSR_{br}(t|\boldsymbol{X})$ estimate residuals using the survival probabilities of the individuals. Hence, the apparent error estimator is equivalent to the estimator of goodness of fit $SSR_{br}(t|\boldsymbol{X})$.

The bootstrap cross-validation estimator

Given the available data set Q of size n, let Q_b^* and Q_b^0 be the training sample and the validation set, respectively, drawn from Q (as described on page 47) for $b = 1, \ldots, B$. Then, the bootstrap cross-validation estimator is

$$\widehat{Err}_{B0}\,(t) = \frac{1}{B}\sum_{b=1}^{B}\;\frac{1}{n_{b0}}\sum_{i\in Q_b^0}\{\, Y_i(t) - r_b^*\,(t|\boldsymbol{X_i})\,\}^2 \;\; W_C(\,t, \hat{G}, \boldsymbol{X_i}\,), \tag{4.15}$$

where $(Y_i(t), \boldsymbol{X_i})$ are the survival status at time t and the predictor values, respectively, of the individual i in the validation set Q_b^0. $r_b^*(t|\boldsymbol{X_i})$ is the predicted survival probability given the predictor $\boldsymbol{X_i}$ based on the Cox regression model fitted

with the bootstrap sample Q_b^*. The weights $W_C(\, t, \hat{G}, \boldsymbol{X_i}\,)$ are based on the reverse Kaplan-Meier function from the data set Q.

The 0.632 estimator

By following the definition in section 4.2.3, the 0.632 estimator is the linear combination of the apparent error and bootstrap cross-validation with weight 0.632. Here, the 0.632 estimator keeps the time dependent feature of the two component estimators (equations (4.14) and (4.15)),

$$\widehat{Err}_{.632}\,(t) = (1 - 0.632)\;\overline{err}\,(t) + 0.632\;\widehat{Err}_{B0}\,(t). \tag{4.16}$$

The 0.632+ estimator

By following the definition in section 4.2.4, the 0.632+ estimator is the linear combination of the apparent error and bootstrap cross-validation with weight $\hat{w}$ (see equation (4.7)). Here, given the time dependent feature of the bootstrap cross-validation (equation (4.15)), the combination involves the use of a time dependent weight $\hat{w}(t)$. The weight $\hat{w}(t)$ is an estimate that depends on the relative overfitting of the prediction rule at a specific time t.

$\hat{w}(t)$ requires the adaptation of the no-information error rate $\hat{\gamma}$ (see equation (4.5)) to a time dependent estimator $\hat{\gamma}(t)$. The no-information error rate tests the prediction rule on data where the survival status Y_i is independent from the predictors $\boldsymbol{X_i}$. Hence,

$$\hat{\gamma}(t) = \frac{1}{n^2}\;\sum_{i=1}^{n}\sum_{j=1}^{n}\{\, Y_i(t) - r\,(t|\boldsymbol{X_j})\,\}^2\; W_C(\, t, \hat{G}, \boldsymbol{X_i}\,).$$

Then, the relative overfitting rate as defined in equation (4.6) is

$$\hat{\mathsf{R}}(t) = \frac{\widehat{Err}_{B0}\,(t) - \overline{err}\,(t)}{\hat{\gamma}(t) - \overline{err}\,(t)}.$$

The estimate of the time dependent weight $\hat{w}(t)$ is

$$\hat{w}(t) = \frac{0.632}{1 - 0.368\,\hat{\mathsf{R}}(t)}\quad , \qquad \hat{w}(t) \in [0.632, 1], \tag{4.17}$$

and the 0.632+ estimator is

$$\widehat{Err}_{.632+}\,(t) = (1 - \hat{w}(t))\;\overline{err}\,(t) + \hat{w}(t)\;\widehat{Err}_{B0}\,(t). \tag{4.18}$$

If $\hat{\gamma}(t) \leq \overline{err}\,(t)$ or $\overline{err}\,(t) < \hat{\gamma}(t) \leq \widehat{Err}_{B0}\,(t)$, the relative overfit $\hat{\mathsf{R}}(t)$ falls out of the range [0,1]. To avoid these particular situations the estimators $\widehat{Err}_{B0}\,(t)$ and $\hat{\mathsf{R}}(t)$ are redefined,

$$\widehat{Err}'_{B0}(t) = min\,(\widehat{Err}_{B0}\,(t), \hat{\gamma}(t)\,),$$

$$\hat{\mathsf{R}}'(t) = \begin{cases} \dfrac{\widehat{Err}'_{B0}(t) - \overline{err}\,(t)}{\hat{\gamma}(t) - \overline{err}\,(t)} & \text{if } \hat{\gamma}(t),\ \widehat{Err}'_{B0}(t) > \overline{err}\,(t) \\ 0 & \text{otherwise.} \end{cases}$$

Then, $\hat{\mathsf{R}}'(t)$ and $\widehat{Err}'_{B0}(t)$ are plugged in equations (4.17) and (4.18), respectively, to estimate $\widehat{Err}_{.632+}\,(t)$.

Further details on this estimator are in section 4.2.4.

4.4 The 0.632 and 0.632+ estimators based on Schoenfeld residuals

In this section we propose and describe our approach of estimating prediction errors of prediction rules from survival data. We propose the use of the criterion of the Schoenfeld residuals to estimate the prediction errors. This approach was motivated by the study of Müller et al. (2008) whereby they showed the criterion of the Schoenfeld residuals was the most appropriate for evaluation of goodness of fit of genetic survival Cox models (see section 3.5.4). Here, we show the use of the Schoenfeld residuals as measures of prediction errors to estimate the prediction performance of a Cox regression model.

Some specific notations used to describe prediction errors in this section are:

$\mathcal{D}$	set of uncensored individuals in the available data Q
$\mathcal{D}_{V_0}$	set of uncensored individuals in a validation set
$\mathcal{D}_b^0$	set of uncensored individuals in the subset Q_b^0
$\mathcal{R}_{t_i}$	set of individuals at risk at time t_i in the available data Q
$\mathcal{R}^*_{b,t_i}$	set of individuals at risk at time t_i in the bootstrap sample Q_b^*
$\hat{\beta}$	vector of estimated parameters from the Cox model fitted with the available data Q
$\hat{\beta}_b^*$	vector of estimated parameters from the Cox model fitted with the bootstrap sample Q_b^*.

4.4.1 Schoenfeld residuals as measure of prediction errors

The basis of the computation of the Schoenfeld residuals (section 3.5.3) is the difference between the true value of a covariate and its expectation under the fitted Cox regression model. Hence, the Schoenfeld residuals accomplishes the computational form of prediction errors as defined in equation (4.2).

For this adaptation, we formulate the prediction errors by taking the definition of sum of squared Schoenfeld residuals with multiple factors ($SSR_{sch|\hat{\beta}}$ in equation (3.43)). It is based on the prognostic index, has the adjustment for censoring data, and it is also applicable for a single factor.

In this context, the error rate measure to be formulated with the Schoenfeld residuals (Err) is not the same measure previously formulated with the Brier score ($Err(t)$, see section 4.3). Both approaches evaluate the prediction capability of the model through definitions of prediction errors. However, the former evaluates the model in terms of error at identifying the true category of the covariate(s), given the time to event for an individual; and the latter evaluates the model in terms of error at identifying the true survival status, given the covariates and some specific time during the follow up period. In both cases, this identification by the model is on the continuous scale.

Given the characteristics of the Schoenfeld residuals, the estimates of prediction errors are obtained only from uncensored individuals, i.e. individuals whose events were observed during the follow up period (see equation (3.43)).

The error rate Err is expressed as

$$Err \;=\; E\left\{\,\hat{\boldsymbol{\beta}}\,'\left(\,\boldsymbol{X_0} - E(\boldsymbol{X_0}|\mathcal{R}_{t_0},\hat{\boldsymbol{\beta}})\,\right)\right\}^2,$$

where $\boldsymbol{X_0} = (X_{01},\ldots,X_{0K})$ is the vector of K covariates of a random individual drawn from the study population. t_0 is the observed time to event of that individual. $E(\boldsymbol{X_0}|\mathcal{R}_{t_0},\hat{\boldsymbol{\beta}}) = (E(X_{01}|\mathcal{R}_{t_0},\hat{\boldsymbol{\beta}}),\ldots,E(X_{0K}|\mathcal{R}_{t_0},\hat{\boldsymbol{\beta}}))$ is the vector of expected values of the covariates as a function of the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ of the Cox model, given the set of individuals at risk $\mathcal{R}_{t_0}$ at time t_0.

As seen in equation (4.2), an estimator $\widehat{Err}$ can be computed by averaging the prediction errors over individuals i in a validation set. In this case, given that the Schoenfeld residuals are obtained only for uncensored individuals, Err is estimated only with individuals $i \in \mathcal{D}_{V_0}$, the set of uncensored individuals in the validation set.

By following the formula of squared Schoenfeld residuals in equation (3.43), the average can be rather computed as a weighted average of prediction errors over all individuals $i \in \mathcal{D}_{V_0}$. The weights $W_{KM}(t_i)$ were introduced to correct for censoring, and we keep it here. The weights are the length of the steps of the Kaplan-Meier survival curve from the training sample (see page 37).

The estimator $\widehat{Err}$ is then

$$\widehat{Err} = \frac{\sum_{i \in \mathcal{D}_{V_0}} W_{KM}(t_i) \left\{ \hat{\boldsymbol{\beta}}' \left(\boldsymbol{X_i} - E(\boldsymbol{X_i} | \mathcal{R}_{t_i}, \hat{\boldsymbol{\beta}}) \right) \right\}^2}{\sum_{i \in \mathcal{D}_{V_0}} W_{KM}(t_i)}. \tag{4.19}$$

Each element of the vector $E(\boldsymbol{X_i}|\mathcal{R}_{t_i}, \hat{\boldsymbol{\beta}}) = (E(X_{i1}|\mathcal{R}_{t_i}, \hat{\boldsymbol{\beta}}), \ldots, E(X_{iK}|\mathcal{R}_{t_i}, \hat{\boldsymbol{\beta}}))$ is obtained as in equation (3.40).

For the case of having biallelic gene predictors, and depending on the modelled effect assumed for the gene covariate (dominant, recessive or additive), the computation of Err uses either the values of $\{0, 1\}$ or $\{0, 1, 2\}$ for $\boldsymbol{X_i}$. Naturally, the expected values of the covariate are continuous values.

4.4.2 Estimators of prediction errors based on the Schoenfeld residuals

In the absence of a validation set, the techniques described in sections 4.2.1 to 4.2.4 can be used to get $\widehat{Err}$. We now give the corresponding formulae.

The apparent error estimator

Let $\mathcal{D}$ be the subset of uncensored individuals from the available data Q. The apparent error is computed by considering Q as both the training sample and validation set, then

$$\overline{err} = \frac{\sum_{i \in \mathcal{D}} W_{KM}(t_i) \left\{ \hat{\boldsymbol{\beta}}' \left(\boldsymbol{X_i} - E(\boldsymbol{X_i} | \mathcal{R}_{t_i}, \hat{\boldsymbol{\beta}}) \right) \right\}^2}{\sum_{i \in \mathcal{D}} W_{KM}(t_i)}, \tag{4.20}$$

where $\boldsymbol{X_i}$ is the vector of covariates for the ith individual in $\mathcal{D}$. t_i is the observed time to event for that ith individual, i.e. event time for $i \in Q \mid \delta_i = 1$. $\hat{\boldsymbol{\beta}}$ is the vector of parameter estimates for the K covariates in the Cox regression model fitted with the training sample Q.

$E(\boldsymbol{X_i}|\mathcal{R}_{t_i}, \hat{\boldsymbol{\beta}})$ gives a vector of expected values of the covariates, given the fitted Cox model, at an event time t_i. $W_{KM}(t_i)$ is the length of the step at time t_i of the Kaplan-Meier survival curve derived from the training sample Q.

It can be noticed that this formulation of apparent error resembles the formulation of sum of squares $SSR_{sch|\hat{\beta}}$ in equation (3.43), although in that case it was expressed

as a weighted sum of the residuals, and here it is expressed as a weighted mean. As it has been indicated, residuals are obtained to evaluate the goodness of fit of a model to the data at hand. Since the formulation of apparent error uses the same data at hand to fit and evaluate the model, the apparent error of the Schoenfeld residuals is also a measure for goodness of fit of a model.

The Bootstrap cross-validation estimator

Let Q_b^* be a bootstrap sample of size n from the available data Q, and Q_b^0 the respective validation set as defined on page 47, for $b = 1, \ldots, B$. Let $\mathcal{D}_b^0$ be the subset of uncensored individuals of Q_b^0, i.e. $i \in Q_b^0 \mid \delta_i = 1$.

The bootstrap cross-validation is computed by considering Q_b^* and Q_b^0 as the training sample and validation set, respectively. Then,

$$\widehat{Err}_{B0} = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{i\in\mathcal{D}_b^0} W_{KM}(t_i)\left\{\hat{\boldsymbol{\beta}}_{\boldsymbol{b}}^{*\prime}\left(\boldsymbol{X_i} - E(\boldsymbol{X_i}|\mathcal{R}_{b,t_i}^*, \hat{\boldsymbol{\beta}}_{\boldsymbol{b}}^*)\right)\right\}^2}{\sum_{i\in\mathcal{D}_b^0} W_{KM}(t_i)}, \tag{4.21}$$

where $\boldsymbol{X_i}$ is the vector of covariate for the ith individual in $\mathcal{D}_b^0$. t_i is the time to event for that ith individual. $\hat{\boldsymbol{\beta}}_{\boldsymbol{b}}^*$ is the vector of K parameter estimates for the covariates in the fitted Cox model with the training sample Q_b^*. $\mathcal{R}_{b,t_i}^*$ is the set of individual at risk in the training sample at time to event t_i.

$E(\boldsymbol{X_i}|\mathcal{R}_{b,t_i}^*, \hat{\boldsymbol{\beta}}_{\boldsymbol{b}}^*)$ gives a vector of expected values of the covariates, given the fitted Cox model from the training sample Q_b^*, and given the individuals at risk $\mathcal{R}_{b,t_i}^*$.

For each bootstrap sample b the weighted average of the Schoenfeld residuals is computed with weights $W_{KM}(t_i)$. These weights are obtained from the Kaplan-Meier survival curve derived from the available data Q.

Since the validation set Q_b^0 is obtained by collecting the remaining individuals from Q not included in the bootstrap sample Q_b^*, and considering that bootstrap samples are drawn independently from each other, different validation sets will be obtained for new bootstrap samples b. Therefore, different sets of times t_i are obtained for newly drawn bootstrap samples b. Thus, the weights $W_{KM}(t_i)$ are also quantities that vary with each new bootstrap sample.

Also, the Kaplan-Meier survival curve, used here to derive $W_{KM}(t_i)$, is the same as in the apparent error estimator (equation (4.20)). However, as said above, the observed event times t_i vary with different samples, and therefore, the weights are not the same as in the apparent error estimator.

The 0.632 estimator

The 0.632 estimator is the linear combination of the apparent error and bootstrap cross-validation estimators from equations (4.20) and (4.21)

$$\widehat{Err}_{.632} = (1 - 0.632)\,\overline{err} + 0.632\,\widehat{Err}_{B0}. \tag{4.22}$$

The details on the background of this estimator are described in section 4.2.3.

The 0.632+ estimator

The 0.632+ estimator is the linear combination of the apparent error and bootstrap cross-validation (equations (4.20) and (4.21)) with weight $\hat{w}$ that depend on the rate of overfitting of the Cox regression model (see section 4.2.4).

$\hat{w}$ also requires the computation of the no-information error rate $\hat{\gamma}$ that tests the prediction rule on data where the covariates $\boldsymbol{X_i}$ are independent from the event times t_i. The estimator $\hat{\gamma}$ is obtained as

$$\hat{\gamma} = \frac{1}{n_{\mathcal{D}} \sum_{i \in \mathcal{D}} W_{KM}(t_j)} \sum_{j \in \mathcal{D}} \sum_{i \in \mathcal{D}} W_{KM}(t_j) \left\{ \hat{\boldsymbol{\beta}}' \left(\boldsymbol{X_i} - E(\boldsymbol{X_i} | \mathcal{R}_{t_j}, \hat{\boldsymbol{\beta}}) \right) \right\}^2,$$

where $n_{\mathcal{D}}$ is the number of individuals in the subset $\mathcal{D}$, i.e. the number of individuals with observed event time in the available data Q ($i \in Q \mid \delta_i = 1$).

Then, the rate of overfitting is

$$\hat{\mathsf{R}} = \frac{\widehat{Err}_{B0} - \overline{err}}{\hat{\gamma} - \overline{err}}.$$

The rate $\hat{\mathsf{R}}$ is then used in the computation of weights $\hat{w}$,

$$\hat{w} = \frac{0.632}{1 - 0.368\,\hat{\mathsf{R}}} \quad , \qquad \hat{w} \in [0.632, 1],$$

and the 0.632+ estimator is,

$$\widehat{Err}_{.632+} = (1 - \hat{w})\,\overline{err} + \hat{w}\,\widehat{Err}_{B0}. \tag{4.23}$$

Sometimes $\hat{\mathsf{R}}$ may fall out of the range [0,1], then some corrections as shown in equations (4.9) and (4.10) should be applied to avoid these situations.

Further details and description of the 0.632+ estimator are in section 4.2.4.

Table 4.1: Summary table of differences between criteria to estimate prediction errors of Cox regression models

Characteristics	Criteria	
	Schoenfeld residuals	Brier score[a]
Main variable	covariate	survival status
Given variable	time to event	covariate
Individuals	with observed events	all available
Time dependent estimates	no (see equation 4.19)	yes (see equation 4.13)

[a] Approach adapted by Gerds and Schumacher (2007).

4.5 Schoenfeld residuals and Brier score as criteria to estimate prediction errors

In section 4.3 we have presented the approach to estimate prediction errors via the criterion of the Brier score adapted by Gerds and Schumacher (2007). In the present section we want to highlight some differences with the approach that we propose above in section 4.4, the estimate of prediction errors via the criterion of the Schoenfeld residuals. Table 4.1 summarizes these differences.

First, in the approach with the Schoenfeld residuals, the main variable to measure the prediction error rate of the model is the covariate, whereas in the approach with the Brier score it is the survival status.

Second, in the approach with the Schoenfeld residuals, the predicted value is a function of the time to event, whereas in the approach with the Brier score the predicted value is a function of the covariate(s).

Third, the approach with the Schoenfeld residuals uses only uncensored individuals of the validation set, i.e. individuals with observed time to event, while the approach with the Brier score uses all individuals regardless of their censoring status. Hence, the Schoenfeld residuals approach uses only part of the data used with the Brier score. The former approach makes the calculations less heavy, especially when small percentages of events are observed, since less data will be involved in the calculations. Moreover, even if this approach uses less data, it uses all data involved with the main variable of study, the time to event. We have mentioned in the paragraph above that the approach with the Schoenfeld residuals evaluate prediction as a function of the time to events.

Last, the approach with the Schoenfeld residuals summarizes the individual prediction errors into a single estimator ($\widehat{Err}$), while the approach with the Brier score gives estimators per time ($\widehat{Err}(t)$), which can also be later summarized into a single estimator (see page 62). Computing estimators per time makes the Brier score

approach heavy, since it has to be run many times, although the advantage is that it allows constructing the course of estimated prediction errors over time.

Thus, using the criterion of the Brier score to estimate prediction errors seems to be more informative because it allows examining the rate of errors over time, but it can be costly in terms of time and complexity of the approach. Using the criterion of the Schoenfeld residuals can be simpler because less data are used and the computation is performed only once, but it does not provide a time course of errors. However, this approach does use and compute errors over time, these are the times at observed events, which are the most interesting variable in survival analysis.

4.6 Measure for the gain in prediction in survival models: R^2_{Pred}

In clinical studies it is of interest to find appropriate models for the development of an outcome, e.g. disease or death, which could be affected by either a single or multiple factors. When the model is intended to be used for predictions it should be validated. A model for prediction is used for example to prevent the occurrence of disease or to decide on the most appropriate treatment for an individual.

In the previous sections of this chapter, we have defined and formulated estimators of prediction errors to evaluate models for prediction. In this section, we formulate the R^2 measure for the gain in prediction due to consideration of predictors in the Cox regression model. We denote this measure as R^2_{Pred}, where the subscript *Pred* denotes prediction.

An estimate of gain in prediction due to predictors in the model can be measured by comparing the relative difference of prediction error rate ($\widehat{Err}$) of the Cox model, with respect to the *null model* (model without predictors). An estimate of gain in prediction due to a particular subset of predictors in the model can be measured by using the rate not predicted by the model (i.e. the complement gain in prediction) with respect to that of a *reference model* (model without the particular subset of predictors). In that case we compute a partial gain in prediction, *partial* R^2_{Pred}.

The R^2_{Pred} can also be used in practice to evaluate and compare candidate models for prediction of survival outcomes.

We can obtain estimators of the R^2_{Pred} measure by using the different estimators of prediction errors described in the previous sections 4.3 and 4.4, for estimates using the criterion of the Brier score and the Schoenfeld residuals, respectively. The R^2_{Pred} measure is expected to correct overestimation that results from R^2 measures of goodness of fit.

Regarding the predicted values of the outcomes or covariates under null models, we remark that if we use the criterion of the Brier score, the Kaplan-Meier estimator (equation (3.11)) can be used as a prediction rule to predict the survival probabilities.

If we use the criterion of the Schoenfeld residuals, the equation (3.41) can be used as prediction rule to predict the covariate values.

4.6.1 The R^2_{Pred} measure

The R^2_{Pred} measure is defined as the relative difference of the prediction error rate of a Cox model with predictors, with respect to the null model (model without predictors). This measure gives the fraction of contribution of the predictors to the improvement in prediction of the survival outcome. The R^2_{Pred} measure is

$$R^2_{Pred} = 1 - \frac{Err}{Err^{\,\text{null}}},$$

where Err is the expected prediction error of a Cox model with the predictors, and Err^{null} is the expected prediction error of the null model.

An estimate of R^2_{Pred} can be obtained by using the estimates of prediction error, $\widehat{Err}$,

$$\hat{R}^2_{Pred} = 1 - \frac{\widehat{Err}}{\widehat{Err}^{\,\text{null}}}. \tag{4.24}$$

As seen in the previous sections, $\widehat{Err}$ uses an independent validation set to test the performance of the Cox model to the prediction of the survival outcome (see equations (4.13) and (4.19)). When there is no available data set to estimate $\widehat{Err}$, they can alternatively be estimated by the techniques and estimators described in sections 4.2.1 to 4.2.4. For Cox regression models the estimators are specifically formulated in sections 4.3 and 4.4, for estimates using the criteria of the Brier score and the Schoenfeld residuals, respectively. In the following sections we remark some important issues on these estimators.

4.6.2 $\hat{R}^2_{Pred}$ based on apparent errors

The estimator of prediction based on estimates of the apparent error is,

$$\hat{R}^2_{Pred,\,app} = 1 - \frac{\overline{err}}{\overline{err}^{\,\text{null}}},$$

where $\overline{err}$ and $\overline{err}^{\,\text{null}}$ are the apparent error estimates of prediction of the predictor model and the null model, respectively.

As described in section 4.2.1, the apparent error estimator $\overline{err}$ uses the same data to estimate a prediction rule and to evaluate its performance for prediction, i.e. it

uses the same data as both training sample and validation set. In the context of Cox models, $\hat{R}^2_{Pred,app}$ can be obtained either with the criterion of the Schoenfeld residuals or with the Brier score, by using the respective apparent errors (equations (4.20) or (4.14), respectively). The $\overline{err}$ also corresponds to the sum of squared residuals used to estimate the R^2 for the goodness of fit of a model. Hence, the $\hat{R}^2_{Pred,app}$ corresponds to the R^2 estimate of the goodness of fit of a Cox regression model, either evaluated with the Schoenfeld residuals ($\hat{R}^2_{sch}$, equation (3.43)) or with the Brier score ($\hat{R}^2_{br}$, equation (3.37)).

4.6.3 $\hat{R}^2_{Pred}$ based on 0.632 and 0.632+ estimators

It is known that measures based on the apparent error estimators underestimate the expected error, measures based on the bootstrap cross-validationtechniques such as the 0.632 and the 0.632+ estimators can be more appropriate.

The estimator of prediction based on estimates of the 0.632 estimator is,

$$\hat{R}^2_{Pred,.632} = 1 - \frac{\widehat{Err}_{.632}}{\widehat{Err}^{null}_{.632}},$$

where $\widehat{Err}_{.632}$ and $\widehat{Err}^{null}_{.632}$ are the 0.632 estimators of prediction errors for a predictor model and a reference model, respectively. The estimator of prediction based on estimates of the 0.632+ estimator ($\hat{R}^2_{Pred,.632+}$) can be computed similarly.

The R^2_{Pred} estimators, $\hat{R}^2_{Pred,.632}$ and $\hat{R}^2_{Pred,.632+}$, should be the appropriate estimators of the gain in prediction due to the predictors of interest in the Cox model. These estimators are obtained using techniques that account for evaluation of prediction rules on independent data, which should be more appropriate than techniques evaluating prediction on the same data used for fitting the model. In the next chapter, we evaluate through simulation studies the performance of these estimators.

Specifications on $\hat{R}^2_{Pred}$ with the criterion of the Brier score

As indicated in section 4.3, the prediction error estimator with the criterion of the Brier score is a time dependent estimator. The respective $\hat{R}^2_{Pred}$ can also be formulated as a time dependent estimator, for example, with the 0.632 estimator we have

$$\hat{R}^2_{Pred,.632}(t) = 1 - \frac{\widehat{Err}_{.632}(t)}{\widehat{Err}^{null}_{.632}(t)}. \tag{4.25}$$

Otherwise, it can be computed as an overall $\hat{R}^2_{Pred,.632}$ estimator, by using the cumulative prediction errors to the end time τ. For the interest of our study we will

estimate prediction errors at times of observed events. It is of our own interest to compare and treat both criteria, the Schoenfeld residuals and the Brier score, as similar as possible, where the only difference between them is the main variable taken to measure the prediction errors. Then, by following the expression given in equation (3.38),

$$\widehat{Err}_{.632}(\tau) = \sum_{t \in t_{\mathcal{D}}} \widehat{Err}_{.632}\,(t), \tag{4.26}$$

where the weights $W_{KM}(t_i)$ are the length of the step of the marginal Kaplan-Meier survival curve at times t_i.

Similarly, we can compute $cum\,\widehat{Err}^{\,\text{null}}_{.632}(\tau)$. Then, the overall estimator $\hat{R}^2_{Pred}$ with the Brier score is

$$\hat{R}^2_{Pred,.632}(\tau) = 1 - \frac{\widehat{Err}_{.632}\,(\tau)}{\widehat{Err}^{\,\text{null}}_{.632}\,(\tau)}\,.$$

4.6.4 Partial gain in prediction, the *partial* $\hat{R}^2_{Pred}$

The *partial gain in prediction* is the rate of prediction attributed to a particular subset of predictors in the model, with respect to a *reference model* (model excluding the particular subset of predictors). It can be measured by the *partial* $\hat{R}^2_{Pred}$. It uses the rate unpredicted by the model with the whole set of predictors ($1-\hat{R}^2_{Pred}$), with respect to that of the reference model ($1-\hat{R}^2_{Pred}{}^{\text{ref}}$).

The partial gain in prediction can be estimated as

$$partial\ \hat{R}^2_{Pred} = 1 - \frac{1-\hat{R}^2_{Pred}}{1-\hat{R}^2_{Pred}{}^{\text{ref}}}\,,$$

where $\hat{R}^2_{Pred}$ and $\hat{R}^2_{Pred}{}^{\text{ref}}$ are the estimates of gain in prediction with the predictor model (containing the whole set of covariates) and with the reference model, respectively. The *partial* $\hat{R}^2_{Pred}$ can be estimated using $\hat{R}^2_{Pred}$ with techniques previously described. Hence we can estimate *partial* $\hat{R}^2_{Pred,app}$, *partial* $\hat{R}^2_{Pred,.632}$ or *partial* $\hat{R}^2_{Pred,.632+}$.

Table 4.2 gives an overview of the estimators of prediction errors as well as the different approaches and notations presented in this chapter.

Table 4.2: Estimators of prediction error rates (Err) and of the gain in prediction (R^2_{Pred}) for survival outcomes

Estimator of Err	Criteria to define Err		R^2_{Pred} [b]
	Schoenfeld residuals	Brier score[a]	
with a validation set[c]	$\widehat{Err}$	$\widehat{Err}$(t)	$\hat{R}^2_{Pred}$
without a validation set[d]			
apparent error	$\overline{err}$	$\overline{err}(t)$	$\hat{R}^2_{Pred,app}$
bootstrap cross-validation	$\widehat{Err}_{B0}$	$\widehat{Err}_{B0}(t)$	$\hat{R}^2_{Pred,B0}$
0.632 estimator	$\widehat{Err}_{.632}$	$\widehat{Err}_{.632}(t)$	$\hat{R}^2_{Pred,.632}$
0.632+ estimator	$\widehat{Err}_{.632+}$	$\widehat{Err}_{.632+}(t)$	$\hat{R}^2_{Pred,.632+}$

[a] Approach adapted by Gerds and Schumacher (2007).
[b] Notation for time independent R^2_{Pred}. Note that, although we are using a unique notation for R^2_{Pred}, different estimates will result depending on the criterion used to define Err. Time dependent R^2_{Pred} can be obtained with the criterion of the Brier score (equation (4.25)). We can also estimate a *partial* R^2_{Pred} (section 4.6.4), which is not included in this Table.
[c] This provides original estimates of Err and R^2_{Pred}, see equations (4.19) and (4.13).
[d] Techniques when no validation set is available, see sections 4.4 and 4.3.

Chapter 5

Simulation Study

In this chapter we want to evaluate the appropriateness of the estimators of R^2_{Pred} to measure the gain in prediction in survival models.

We have seen in section 4.6 that the estimator $\hat{R}^2_{Pred}$ requires the estimates of prediction error rates ($\widehat{Err}$) of two models. Under the unavailability of a true validation set, i.e. a data set independent from the data set used for model fitting, the estimates of $\widehat{Err}$ cannot be directly computed. They can be obtained by alternative techniques that solve the lack of a true validation set (see section 4.2 for a general view of the techniques, and sections 4.4 and 4.3 for specific formulations on survival data). Hence, we evaluate how well the alternative estimates of $\hat{R}^2_{Pred}$ that result from the techniques to estimate prediction error rates, compare to the original estimate of $\hat{R}^2_{Pred}$ that uses a true validation set to estimate the prediction error rates of the models (see section 4.6).

We specially evaluate the techniques of the apparent error, the 0.632, and the 0.632+ estimators. We did not include the bootstrap cross-validation estimator because the 0.632 estimator is the technique solving the positive bias of that estimator, as well as the negative bias of the apparent error estimator. However, we included the apparent error estimator because it is the estimator commonly used as a first alternative for evaluation of model validation, and it is our goal to show the benefit of using a more appropriate estimator to validate models for prediction purposes.

Our simulation study evaluates the gain in prediction when we consider biallelic SNPs in Cox regression models for survival outcomes. We simulated survival data with single SNP factors under various frequencies of the minor allele and various effect sizes of the SNP factor on the occurrence of event. Also, we evaluated the $\hat{R}^2_{Pred}$ under the assumption that the risk allele follows a true additive, dominant or recessive genetic model.

In addition, since in real situations the true mode of inheritance of a disease is unknown, i.e. the disease can be inherited as an additive, dominant, or recessive mode, we evaluated whether an estimator $\hat{R}^2_{Pred}$ would be useful to identify the true mode of inheritance.

In all the cases, the estimators $\hat{R}^2_{Pred}$ have been computed by estimating prediction errors under both criteria: the Schoenfeld residuals and the Brier score (see sections (4.4) and (4.3)). The former method is the new approach we introduced in this thesis in the context of evaluating prediction for Cox models. The latter method has recently been used in applications on survival data (Schumacher et al. 2007, Binder and Schumacher 2008, Porzelius et al. 2010). Here we test both criteria particularly for applications on survival data with SNP factors. Some similarities and differences in the simulation results from both criteria are also highlighted.

5.1 Simulation settings

5.1.1 The Data

We simulated data sets in order to approach realistic scenarios of cohort studies for association of a biallelic SNP, the predictor, with a survival outcome. We considered biallelic SNPs with additive, dominant or recessive genetic model of the risk allele.

Data sets of $n = 1,000$ unrelated individuals were simulated as for a medium size cohort study. We generated the SNPs with different minor allele frequencies (MAF): 10%, 25%, 35%, or 50%. We did not try MAF smaller than 10% to avoid convergence problems during the fitting of the Cox regression model. Even that, we experienced convergence problems very often when the Cox model included a SNP with recessive genetic model, then we simulated the recessive models with a minimum MAF of 15%.

We also simulated different effect sizes of the risk allele. In genetic association studies, the effect sizes of the risk alleles do not tend to be high. Then, we simulated risk alleles with moderate hazard ratios (HR) of 1.25, 1.5 or 2.0 with respect to the wildtype allele.

The data sets were simulated with a total of 60% of censoring. This percentage was meant to include the three forms of censoring occurring in a cohort study: the losses to follow-up, the withdrawals from the study, and discontinued follow-up due to the end of the study. It was assumed that censoring occurred completely at random, i.e. censoring was independent of the predictor. The study of Müller et al. (2008) showed that the percentage of censoring did not influence much on changes in the estimates of R^2, unless they are very high, e.g. >80%. Thus, here we did not investigate effects of variation in censoring for small to moderate censoring.

To simulate each data set, we considered the minor allele as the risk allele. Then, assuming the set of parameters listed above for MAF, HR, and censoring, we proceeded as follows:

We generated a genotype vector $\boldsymbol{X}$ of size $n=1,000$, where $\boldsymbol{X}$ take on values {0, 1, 2} for the $\{wildtype, heterozygous, homozygous\}$ genotype, respectively. The values were assigned randomly to n unrelated individuals with genotype probabilities p_j, for $j = 0, 1, 2$. The probabilities p_j were computed as expected under Hardy-Weinberg equilibrium (see section 2.1.4):

$$
\begin{aligned}
p_0 &= (1-MAF)^2 \\
p_1 &= 2\times(1-MAF)\times MAF \\
p_2 &= MAF^{\,2}\,.
\end{aligned}
\tag{5.1}
$$

Next, the vector of time to event $\boldsymbol{T}$ was generated using the derivation of Bender et al. (2005), which is given in equation (5.2). In the next lines we show the reasoning for this derivation. Let $F(T)$ be the cumulative distribution function such that $F(T) = 1 - S(T)$. Let U_1 be a random variable which is uniformly distributed over the interval [0,1], $U_1 \sim$ Uniform[0,1]. By statistical theory (Mood et al. 1974) it holds that $F^{-1}(U_1) = T$, and $U_1 = F(T)$. In addition, it holds that if $U_1 \sim$ Uniform[0,1], then $U = 1 - U_1$ has the same distribution. Then, $U = 1 - F(T) = S(T) \sim$ Uniform[0,1].

Thus, by taking the conditional survival function under the Cox regression model (equation (3.10)) it follows that,

$$
U = S(T|X) = \exp\,\big(-\,\Lambda_0\,(T)\;\exp\,(\hat{\beta}^{\,\prime}\boldsymbol{X})\,\big) \sim \text{Uniform}[0,1]\,,
$$

Deriving the cumulative hazard we have,

$$
\Lambda_0\,(T) = -\,\frac{\log(U)}{\exp\,(\hat{\beta}^{\,\prime}\boldsymbol{X})}\,.
$$

if $\lambda_0(t) > 0,\ \forall\, t$, then

$$
T = \Lambda_0^{-1}\Big(-\,\frac{\log(U)}{\exp\,(\hat{\beta}^{\,\prime}\boldsymbol{X})}\Big)\,. \tag{5.2}
$$

According to this equation, the simulation of survival time T requires the knowledge of the cumulative baseline hazard function Λ_0. For simplicity we assumed $\boldsymbol{T}$ was exponentially distributed because it provides a constant baseline hazard function λ_0 (Bender et al. 2005), and the cumulative baseline hazard function is $\Lambda_0(T) = \lambda_0\, T$. Hence, $T = \Lambda_0^{-1}\,(\lambda_0\, T)$, and from equation (5.2) we have

$$
\lambda_0\, T = -\,\frac{\log(U)}{\exp\,(\hat{\beta}^{\,\prime}\boldsymbol{X})}\,,
$$

and

$$
T = -\,\frac{\log(U)}{\lambda_0\;\exp\,(\hat{\beta}^{\,\prime}\boldsymbol{X})}\,, \tag{5.3}
$$

where the denominator $\lambda_0\;\exp\,(\hat{\beta}^{\,\prime}\boldsymbol{X})$ is also a constant hazard that depends only

on the genotypes, i.e. $\lambda_0 \ \exp(\hat{\beta}' \boldsymbol{X}) = \lambda(\boldsymbol{X}), \forall t$.

Given that in our simulations we considered only a single gene predictor, i.e. $K = 1$, we generated T such that

$$T = -\frac{\log(U)}{\lambda_0 \ HR^{X_m}}, \tag{5.4}$$

where U was a random draw from the uniform distribution in the interval [0,1], HR was the specific parameter of the effect size assumed for the association of the risk allele with the event, and X_m was a recoded value of X according to the genetic model we assumed for the risk allele, i.e. the mode of inheritance of the event. We also made λ_0 vary according to the assumed genetic model.

Hence, for an additive genetic model: $X_m = X$, and $\lambda_0 = 0.12$;

for a dominant genetic model: $X_m = \begin{cases} 0 & \text{if } X{=}0 \\ 1 & \text{if } X{=}1{,}2 \end{cases}$, and $\lambda_0 = 0.30$;

for a recessive genetic model: $X_m = \begin{cases} 0 & \text{if } X{=}0{,}1 \\ 1 & \text{if } X{=}2 \end{cases}$, and $\lambda_0 = 0.30$.

The baseline hazards of $\lambda_0 = 0.12$ and $\lambda_0 = 0.30$ were chosen from the interval [0,1], with the only condition of not exceeding the limit of 1 for any of the hazards $\lambda(t, \boldsymbol{X})$, which should fall in the interval [0,1] too. For instance, considering the set of simulated parameters $HR = \{1.25, 1.5, 2.0\}$ for the dominant and recessive models, the maximum simulated hazard was of $\lambda(t, \boldsymbol{X}) = 0.30 \times 2.0 = 0.60$, whereas for the additive model it was of $\lambda(t, \boldsymbol{X}) = 0.12 \times 2.0^2 = 0.48$. However, the values of this parameter seem not to be influential in the results (data not shown).

Drawing t repeatedly for each X_m of the n individuals provided the time to event vector $\boldsymbol{T}$.

Next, the time to censoring vector $\boldsymbol{C}$ was generated as a totally random variable from a uniform distribution on the interval $[0, \text{t.censor}]$. The upper limit t.censor was assigned as to produce 60% of censoring. This is roughly the amount of censoring we observed in our study data. The upper limit t.censor can be viewed as the end time of the cohort study. The time t.censor was chosen through some pilot simulations, in which we assigned to t.censor various values equal to various quantiles of the vector $\boldsymbol{T}$, and then generated the respective vector $\boldsymbol{C}$. Then, we derived the indicator of censoring $\boldsymbol{\delta} = I(\boldsymbol{T} \leq \boldsymbol{C})$. We selected the value of t.censor that produced roughly 60% of censoring, i.e. 60% of 0's in the vector $\boldsymbol{\delta}$. The quantile for the 66.5th percentile of vector $\boldsymbol{T}$ ($q_{0.665}$ of $\boldsymbol{T}$) approximated the required 60% of censoring.

Then, we generated the time to censoring vector $\boldsymbol{C}$ from Uniform $[0, \text{t.censor}]$, with t.censor = ($q_{0.665}$ of $\boldsymbol{T}$). Finally, the observed survival time vector was obtained by $\boldsymbol{T}^* = min(\boldsymbol{T}, \boldsymbol{C})$, and the censoring status vector by $\boldsymbol{\delta} = I(\boldsymbol{T} \leq \boldsymbol{C})$.

The data sets were generated with 100 replications.

5.1.2 Validation set

For each simulated data set of size $n=1,000$, a second data set of $n_v=10,000$ individuals was also generated to serve as a validation set. Hence, the validation set was independent from the first simulated data set, and it was large enough to approximate the true underlying model for the whole population. The validation set was used to reproduce the estimated prediction errors under the availability of an independent data set to test the performance of the Cox regression models for prediction of the survival outcome. Subsequently, the gain in prediction ($\hat{R}^2_{Pred}$) derived from these prediction errors was estimated and served as the original estimate to compare and evaluate the performance of the alternative estimators with techniques under evaluation in this simulation study. In this case, given the large size of the validation set, the original estimator $\hat{R}^2_{Pred}$ can also be viewed as an approximation to the true estimate R^2_{Pred}.

5.2 Specification of the methods

prediction error estimators

The $\hat{R}^2_{Pred}$ estimates (equation (4.24)) were obtained from the simulated data sets based on three estimators of prediction error rates: the apparent error, the 0.632 and the 0.632+. These estimators have been described in section 4.3 for estimates with the criterion of the Brier score (Gerds and Schumacher 2007), and in section 4.4 for estimates with the Schoenfeld residuals, which is our proposal for application in genetic association studies.

In the case of the criterion of the Brier score, the overall Kaplan-Meier estimator of survival (equation (3.11)) was used to predict survival probabilities under the null model. In the case of the criterion of the Schoenfeld residuals, the expected value of a covariate under a null model was estimated as in equation (3.41).

The time dependent prediction errors produced with the criterion of the Brier score (equation (4.13)) were computed at times t when events were observed, i.e. $t = t_i \,|\, i \in \mathcal{D}$. These times t were chosen such that we evaluated prediction errors at the same time points as done with the Schoenfeld residuals.

On the other hand, since we simulated the censoring vector $\boldsymbol{C}$ as a totally random variable, the reverse Kaplan-Meier estimator $\hat{G}$ of survival to censoring was an estimator independent of the predictors $\boldsymbol{X_i}$. Therefore, the weighting scheme $W_C(\, t, \hat{G}, \boldsymbol{X_i}\,)$ that adjusts for the presence of censoring in the estimates of prediction errors (equation (4.13)), was also independent of the predictors. Then, the weights were computed with an overall reverse Kaplan-Meier estimator $\hat{G}(t)$,

$$W_C(\, t, \hat{G}\,) = \frac{I(t_i \leq t)\, \delta_i}{\hat{G}(t_i-)} + \frac{I(t_i > t)}{\hat{G}(t)}. \tag{5.5}$$

Bootstrap samples

The estimates of prediction error rates with the bootstrap cross-validation estimator $\widehat{Err}_{B0}$ (equatios (4.15) and (4.21)) were computed with B =50 and 200 bootstrap samples for the criteria of the Brier score and Schoenfeld residuals, respectively. We also obtained some previous estimates with the Brier score on B=100 bootstrap samples (data not shown), and no noticeable differences were observed in the results. In addition, it has been shown that there is not much difference in estimating the errors with small or large number of bootstrap samples (Efron and Tibshirani 1997). Although this has not been specifically shown for survival data, B=100 is being used (Schumacher et al. 2007, Binder and Schumacher 2008).

Hence, given the intensive computation with the Brier score, we decided to run the complete set of simulations of this criterion on $B = 50$ bootstrap samples. On the other hand, we decided to use larger number of bootstrap samples with the criterion of the Schoenfeld residuals ($B = 200$) because it is less computationally intensive and it is the first time we are testing this techniques for evaluation of prediction errors in Cox models.

Mean and standard deviation of $\hat{R}^2_{\boldsymbol{Pred}}$

To report the results of $\hat{R}^2_{Pred}$, we took the mean $\hat{R}^2_{Pred}$ over the 100 replications of each simulated scenario. Likewise, the respective standard deviations were also estimated based on the 100 replications.

Identification of the correct mode of inheritance

To judge how good the different estimators can identify the correct genetic model for the risk allele, we recoded the generated genotype vector $\boldsymbol{X}$ three times such that each mode of inheritance (additive, dominant, and recessive) was assumed for the predictor. That means, in this evaluation, we ignored the true simulated mode of inheritance and we assumed it was unknown. Then, for each data set we recoded the genotype vector $\boldsymbol{X}$ as:

i) a variable for an additive model of the risk allele: $X_a = X$,

ii) a variable for a dominant model of the risk allele: $X_d = \begin{cases} 0 & \text{if } X{=}0 \\ 1 & \text{if } X{=}1{,}2 \end{cases}$,

iii) a variable for a recessive model of the risk allele: $X_r = \begin{cases} 0 & \text{if } X{=}0{,}1 \\ 1 & \text{if } X{=}2 \end{cases}$.

Hence, three separated Cox regression models were fitted under the assumption of an additive, a dominant, or a recessive mode, i.e. we fitted Cox models with

the genetic variable X_a, X_d, or X_r, respectively. Then, we selected the mode of inheritance that yield the highest $\hat{R}^2_{Pred}$ from the three Cox models. Further, we compared and obtained the percentage of selected modes over the 100 replications that were identical to the true simulated mode of inheritance. We used this frequency to judge the capability of an estimator to identify the true mode of inheritance.

All the procedures were implemented using R software v.2.12.0 (R Development Core Team 2008), package survival (Therneau and Lumley 2008) for fitting Cox regression models, and package pec (Gerds 2009) to estimate prediction errors based on the Brier score.

5.3 Results

5.3.1 Gain in prediction with the original estimator $\hat{R}^2_{Pred}$

Figures 5.1 and 5.2 show the mean $\hat{R}^2_{Pred}$ values obtained by using the validation sets to estimate prediction errors with the criteria of the Schoenfeld residuals and Brier score, respectively. The pattern of the $\hat{R}^2_{Pred}$ values tended to increase with higher MAF and higher HR in the additive and recessive genetic models. However, in the dominant genetic model, the $\hat{R}^2_{Pred}$ increased with higher HR, and with MAF up to 35% with the criterion of the Schoenfeld residuals (Figure 5.1, dominant model), and with MAF up to 25% with the Brier score (Figure 5.2, dominant model).

The latter pattern was also observed in the study of Müller et al. (2008), the authors explained that there is a relation between the $\hat{R}^2$ values and the variances of the genetic covariates. In our results we also observed that relation now for prediction, the higher the variance of the predictor the higher the gain in prediction $\hat{R}^2_{Pred}$. We show this relation in the next lines.

Let $\boldsymbol{p} = (p_0, p_1, p_2)'$ be the vector of genotype frequencies as a function of the MAF given by HWE (see expression (5.1)), and let $\boldsymbol{c}$ be the vector of genotype codes according to the specific genetic model, i.e. $\boldsymbol{c} = (0, 1, 2)'$, $\boldsymbol{c} = (0, 1, 1)'$, or $\boldsymbol{c} = (0, 0, 1)'$, for the additive, dominant, or recessive genetic model, respectively.

The mean of the predictor can be expressed as

$$\begin{aligned} E(\boldsymbol{X}) &= \sum_{j=0}^{2} c_j \, p_j \\ &= \boldsymbol{c}' \, \boldsymbol{p} \, . \end{aligned} \tag{5.6}$$

Hence, the mean of the predictor is:
$p_1 + 2\,p_2 = 2MAF$, $p_1 + p_2 = 2MAF - MAF^2$, and $p_2 = MAF^2$, for the additive, dominant, and recessive genetic model, respectively.

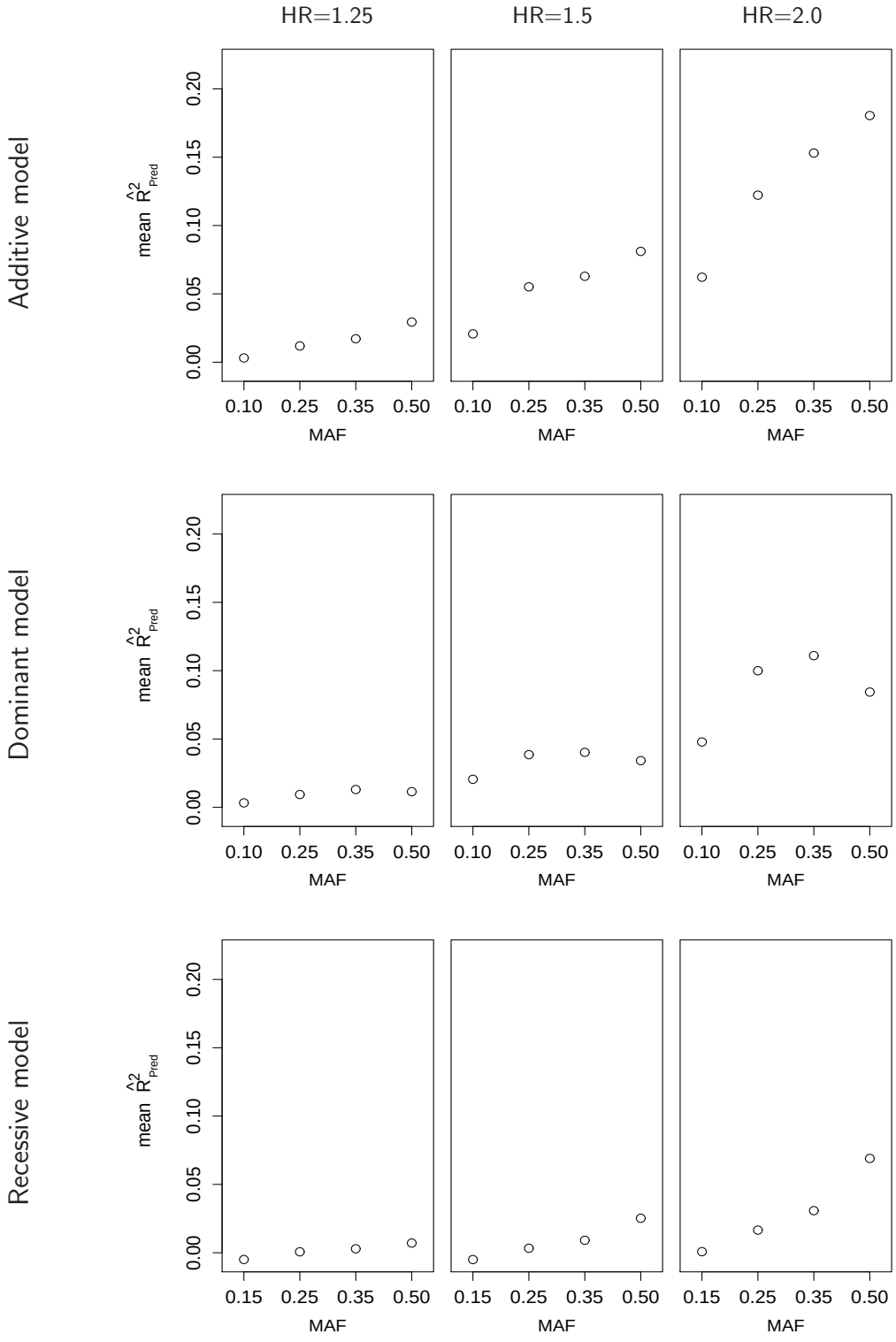

Figure 5.1: Criterion of the **Schoenfeld residuals**. Mean gain in prediction ($\hat{R}^2_{Pred}$) using a validation set, for the additive, dominant and recessive genetic models; and for different minor allele frequencies (MAF) and hazard ratios (HR)

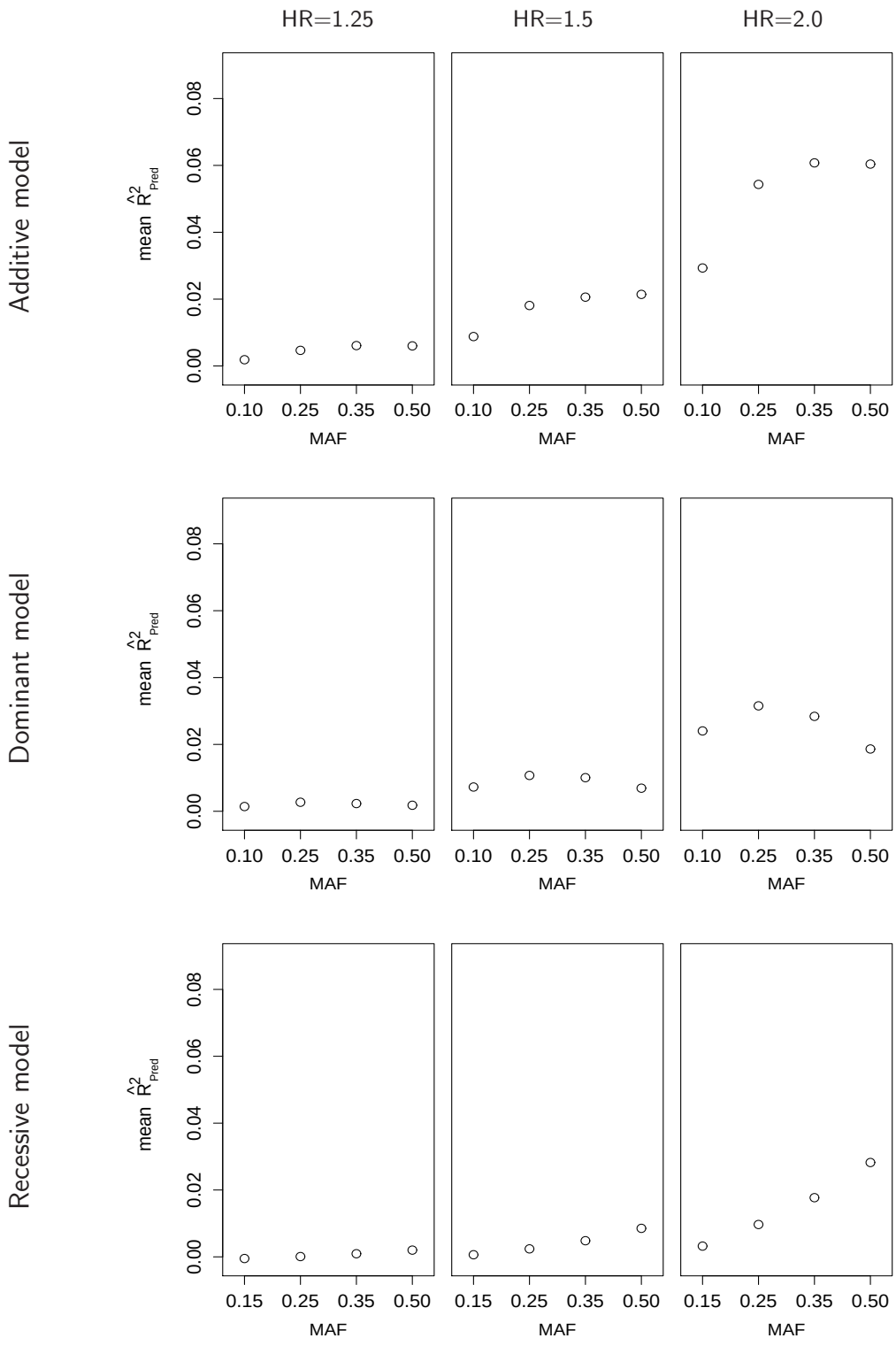

Figure 5.2: Criterion of the **Brier score**. Mean gain in prediction ($\hat{R}^2_{Pred}$) using a validation set, for the additive, dominant and recessive genetic models; and for different minor allele frequencies (MAF) and hazard ratios (HR)

Table 5.1: Frequency, mean, and variance (%) of the genotype predictor under different minor allele frequencies (MAFs), and for three genetic models

MAF	genotype frequency[a]			mean, variance of the genotype[b]					
	p_0	p_1	p_2	additive		dominant		recessive	
10[c]	81.00	18.00	1.00	20,	18.00	19.00,	15.39		-
15[c]	72.25	25.50	2.25		-		-	2.25,	2.20
25	56.25	37.50	6.25	50,	37.50	43.75,	24.61	6.25,	5.86
35	42.25	45.50	12.25	70,	45.50	57.75,	24.40	12.25,	10.75
50	25.00	50.00	25.00	100,	50.00	75.00,	18.75	25.00,	18.75

[a] genotype frequencies under Hardy-Weinburg equilibrium, see expression (5.1).
[b] mean and variance of the genotype, according to the assumed genetic model for the risk allele: additive, dominant, and recessive; see expressions (5.6), and (5.7).
[c] a minimum MAF of 10% was simulated for additive and dominant models, whereas a minimum MAF of 15% was simulated for recessive models.

The variance of the predictor can be expressed as

$$\begin{aligned} Var(\boldsymbol{X}) &= \sum_{j=0}^{2} \big(c_j - E(\boldsymbol{X})\big)^2 p_j \\ &= \big(\boldsymbol{c} - E(\boldsymbol{X})\big)^{2\prime} \boldsymbol{p}\,. \end{aligned} \tag{5.7}$$

Hence, the variance of the predictor is:
p_1, $(1-p_0)\,p_0$, and $p_2\,(1-p_2)$, for the additive, dominant, and recessive genetic model, respectively, where p_0, p_1, and p_2 are the genotype frequencies depending on MAF (see expression (5.1)).

Table 5.1 shows the theoretical frequencies and variances of the predictor under the simulated $MAFs$ and under the assumed genetic models in this study. The variance of the predictor increased with higher MAF and the highest variance was obtained at MAF=50% in the additive and recessive models. In the dominant model the variance of the predictor increased until MAF=25%, this variance was only slightly higher than the variance at 35%. A more detailed overview showed the highest variance at $MAF \approx 29\%$ (data not shown). The pattern of the $\hat{R}^2_{Pred}$ values were in agreement with these patterns of the variance of the predictor. In the case of the dominant model, the highest $\hat{R}^2_{Pred}$ value was obtained at MAF=35% and 25% with the approach of the Schoenfeld residuals and Brier score, respectively. The latter were in agreement with the theoretical highest variance at MAF=29%.

On the other hand, the $\hat{R}^2_{Pred}$ values decreased depending on the genetic model. The $\hat{R}^2_{Pred}$ values from the recessive model were lower than from the dominant model, which in turn were lower than from the additive model. These results are also related to the variance of the SNP predictor. As seen in Table 5.1, under the same MAF,

the predictor had higher variance if it is of an additive genetic model, followed by a dominant, and then by a recessive genetic model. At MAF=50%, the dominant and recessive models showed the same variance, which was also in agreement with the pattern observed in our results for the $\hat{R}^2_{Pred}$ values (Figures 5.1 and 5.2). The latter is natural since the genotype distribution in both genetic models is 75% for one genotype group and 25% for the second group. Additionally, the gain in prediction $\hat{R}^2_{Pred}$ also increased with higher effect sizes HR of the predictor.

The main difference between the $\hat{R}^2_{Pred}$ estimates obtained with the two approaches, the Schoenfeld residuals and Brier score, was that, with higher MAF and higher HR, the $\hat{R}^2_{Pred}$ estimates from the Schoenfeld residuals increased faster than from the Brier score. That makes the $\hat{R}^2_{Pred}$ estimates from the Schoenfeld residuals to appear higher than from the Brier score and be more noticeable at higher HR and MAF.

The maximum gains of prediction (expressed in percentages, $\hat{R}^2_{Pred} \times 100$) with the criterion of the Schoenfeld residuals were 18.4%, 11.3% and 6.3%, respectively, from an additive, dominant and recessive genetic model; while the maximum gain with the criterion of the Brier score were 6.5%, 3.5%, and 3.0%, respectively.

Thus, the gain in prediction due to the genetic predictor was more noticeable when they were estimated with the Schoenfeld residuals. We have also seen that both criteria produced the same information pattern of $\hat{R}^2_{Pred}$ for varying HR and MAF of the SNPs. Then, by considering that in genetic association studies the effect sizes of genes are usually small or moderate, the use of the Schoenfeld residuals can be more advantageous to evaluate prediction of the Cox model. The gain in prediction with the approach of the Schoenfeld residuals will be differentiated better than with the approach of the Brier score. However, the latter does not disqualify the Brier score as a useful alternative for evaluation of prediction.

The criterion of the Schoenfeld residuals evaluates error in prediction of the covariate values of a failing individual. This can be an advantage if we wish to build a predictor to classify patients according to their risk of failure. The criterion of the Schoenfeld residuals focuses the study on this predictor and determines how well the data support it as a predictor of failures in the data.

5.3.2 Gain in prediction with alternative $\hat{R}^2_{Pred}$ estimators

We evaluated how well the gain in prediction could be estimated by the estimators obtained with techniques when no validation set is available. These techniques are the apparent error ($\hat{R}^2_{Pred,app}$), the 0.632 ($\hat{R}^2_{Pred,.632}$), and the 0.632$^+$ ($\hat{R}^2_{Pred,.632^+}$) estimators.

We compared the mean gain in prediction from each estimator with the mean gain in prediction from the original estimator ($\hat{R}^2_{Pred}$), whose results were discussed in the

previous section 5.3.1. The results of the comparisons are presented for the three simulated genetic models: additive, dominant, and recessive; and for both criteria: the Schoenfeld residuals and the Brier score.

Estimators with the criterion of the Schoenfeld residuals

With the criterion of the Schoenfeld residuals, all the estimators ($\hat{R}^2_{Pred,app}$, $\hat{R}^2_{Pred,.632}$, and $\hat{R}^2_{Pred,.632^+}$) approximated well the original estimator $\hat{R}^2_{Pred}$. However, the $\hat{R}^2_{Pred,.632}$ seemed to be the best, this holds for the additive, dominant, and recessive genetic models (Figures 5.3, 5.4, and 5.5).

The $\hat{R}^2_{Pred,app}$ estimator tended to overestimate $\hat{R}^2_{Pred}$ for all effect sizes. This is consistent with the theory that the capability of a model for prediction is overestimated when it is evaluated on the same data used for model fitting (Schemper and Stare 1996).

The $\hat{R}^2_{Pred,.632^+}$ estimator also overestimated the $\hat{R}^2_{Pred}$estimator. In addition, this estimator behaved similarly to the $\hat{R}^2_{Pred,app}$ estimator. This can be explained because the no-information error rate ($\hat{\gamma}$) from the predictor model was mostly smaller than the bootstrap cross-validation error and approximated the apparent error estimates, i.e. $\overline{err} < \hat{\gamma} \leq \widehat{Err}_{B0}$ and $\hat{\gamma} \approx \overline{err}$. Hence, we had to apply the corrections as shown in equations (4.9) and (4.10). That led to estimates of $\widehat{Err}_{.632^+} = \hat{\gamma} \approx \overline{err}$, which in turn led to estimates of $\hat{R}^2_{Pred,.632^+} \approx \hat{R}^2_{Pred,app}$. As deduced from our result of the no-information error rate, the $\hat{\gamma}$ estimator used data that still kept the existing relation between predictors and time to events in the data. Even if these are results with a single predictor, we assume that it will not be better in the case of multiple predictors. Thus, a different concept of estimator of no-information error rate might be needed for this criterion. However, an alternative estimator may not be better than the 0.632 estimator, which performed quite well in the context of our simulations.

The original purpose of the 0.632+ estimator was to account for overfitting of the model and, therefore, to correct the possible overoptimism of the 0.632 estimator (see section 4.2.4). In that sense, under our simulated scenarios with only one predictor, we expected only small overfitting, and the 0.632+ estimates should approximate the 0.632 estimates, i.e. $\hat{R}^2_{Pred,.632^+} \approx \hat{R}^2_{Pred,.632}$. However, our results did not support this approximation. Thus, even if the 0.632+ estimator gave an acceptable mean estimate $\hat{R}^2_{Pred,.632^+}$, it behave different as expected, and was not better than the 0.632 estimator.

Therefore, the estimators of $\hat{R}^2_{Pred}$ obtained from the criterion of the Schoenfeld residuals worked generally well in the context and scenarios considered in our simulation settings. However, the 0.632 estimator showed to be the best because it gave better approximations to the original estimates $\hat{R}^2_{Pred}$ than the 0.632+ and the apparent error estimators.

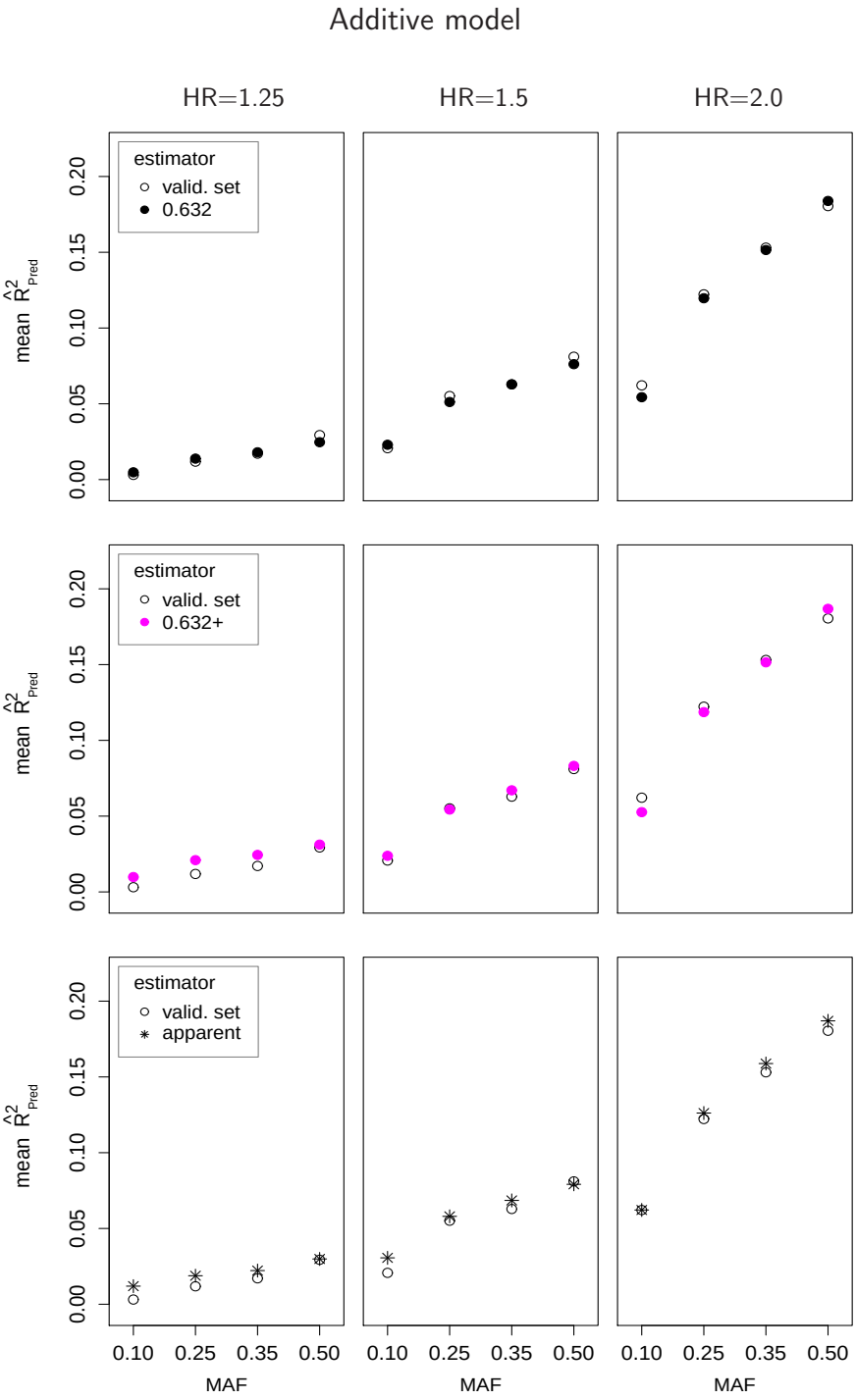

Figure 5.3: Criterion of the **Schoenfeld residuals**. Comparison of mean estimates of gain in prediction ($\hat{R}^2_{Pred}$) using a validation set and estimators of prediction errors. The 0.632 ($\hat{R}^2_{Pred,.632}$), the 0.632+ ($\hat{R}^2_{Pred,.632+}$), and the apparent error ($\hat{R}^2_{Pred,app}$) estimators. Estimates of gain in prediction are plotted for different minor allele frequencies (MAF) and hazard ratios (HR)

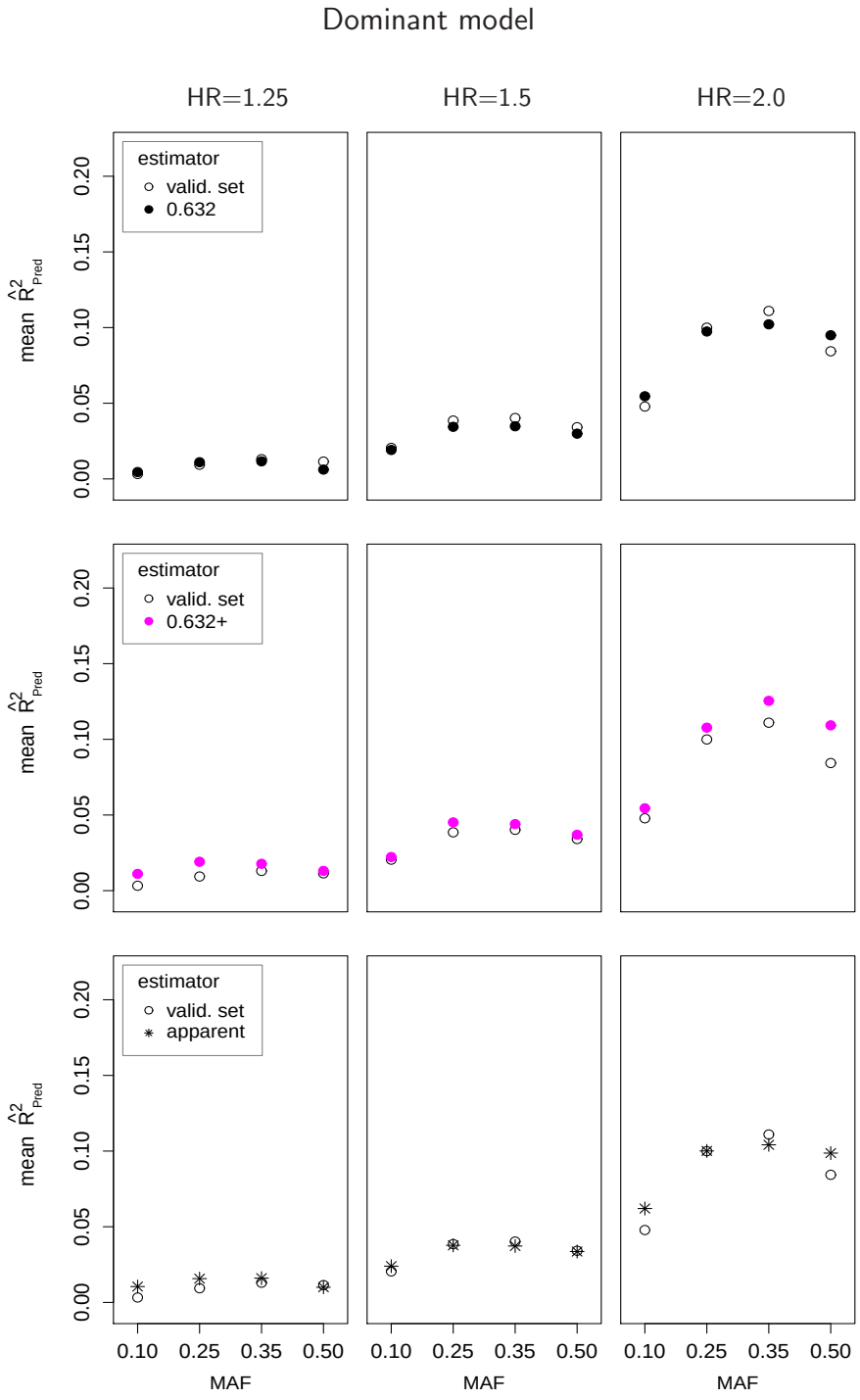

Figure 5.4: Criterion of the **Schoenfeld residuals**. Comparison of mean estimates of gain in prediction ($\hat{R}^2_{Pred}$) using a validation set and estimators of prediction errors. The 0.632 ($\hat{R}^2_{Pred,.632}$), the 0.632+ ($\hat{R}^2_{Pred,.632+}$), and the apparent error ($\hat{R}^2_{Pred,app}$) estimators. Estimates of gain in prediction are plotted for different minor allele frequencies (MAF) and hazard ratios (HR)

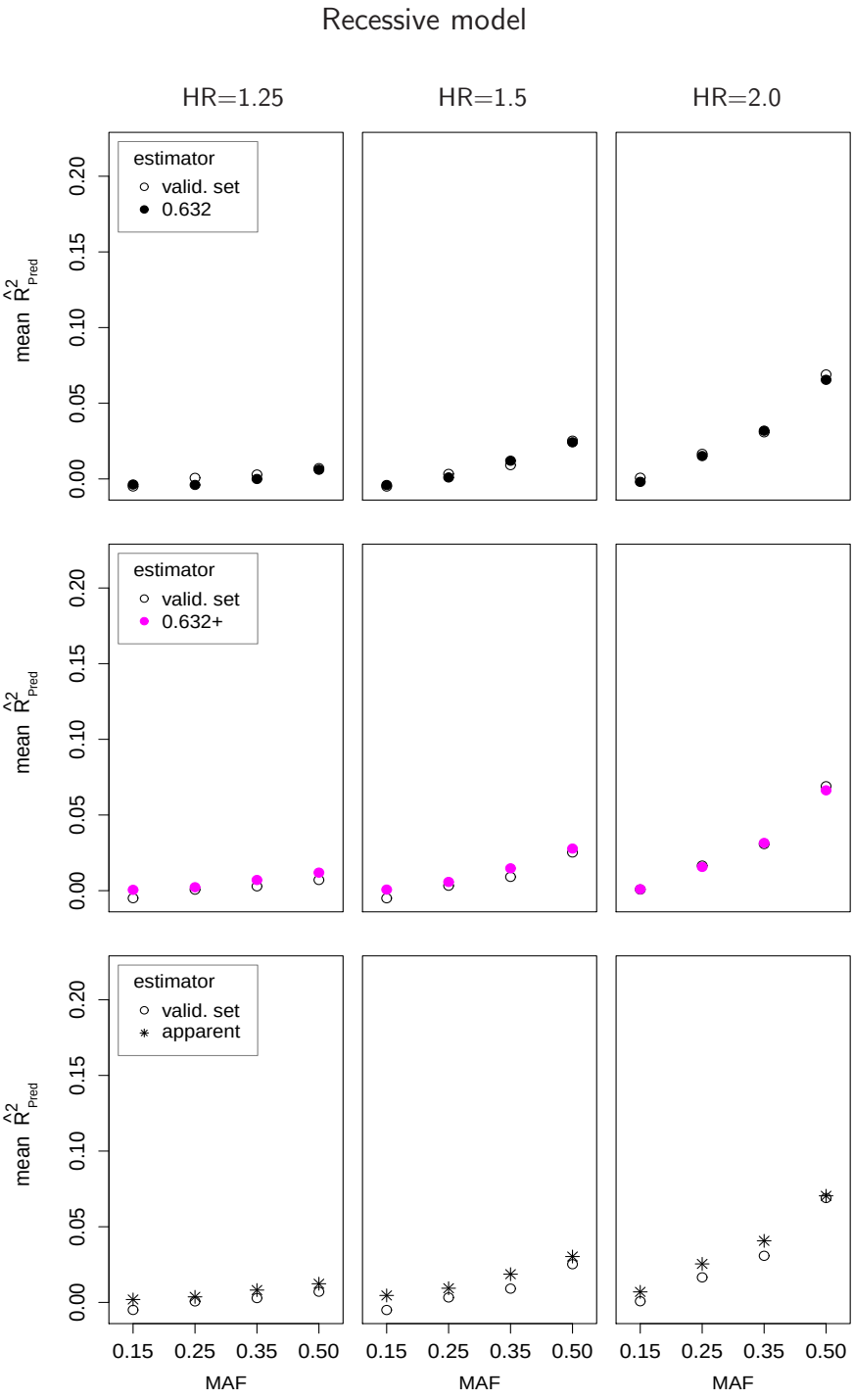

Figure 5.5: Criterion of the **Schoenfeld residuals**. Comparison of mean estimates of gain in prediction ($\hat{R}^2_{Pred}$) using a validation set and estimators of prediction errors. The 0.632 ($\hat{R}^2_{Pred,.632}$), the 0.632+ ($\hat{R}^2_{Pred,.632+}$), and the apparent error ($\hat{R}^2_{Pred,app}$) estimators. Estimates of gain in prediction are plotted for different minor allele frequencies (MAF) and hazard ratios (HR)

Estimators with the criterion of the Brier score

With the criterion of the Brier score, all the estimators ($\hat{R}^2_{Pred,app}$, $\hat{R}^2_{Pred,.632}$, and $\hat{R}^2_{Pred,.632^+}$) approximated well the original estimator $\hat{R}^2_{Pred}$. The differences between them were minor. This holds for the additive, dominant, and recessive genetic models (Figures 5.6, 5.7, and 5.8).

Our findings are in line with the findings of Gerds and Schumacher (2007) from an empirical study to estimate prediction errors with the Brier score. They found that the apparent error estimates were almost identical to bootstrap-based estimates (bootstrap cross-validation and 0.632^+ estimates), and they concluded that the apparent error could reproduce well the true prediction errors, in that case they assumed the bootstrap-based estimates were a reference of true prediction errors. Although they focused their study on the estimators of prediction errors, and not on the gain in prediction, their conclusion agrees with ours, since the estimators for the gain in prediction, that we studied, are derived from the estimators of prediction errors. Indeed, we found that the estimators $\hat{R}^2_{Pred,app}$, $\hat{R}^2_{Pred,.632}$, and $\hat{R}^2_{Pred,.632^+}$ are all numerically similar.

However, by comparing these estimators to the original $\hat{R}^2_{Pred}$ estimator, a very close inspection led us to realize that the $\hat{R}^2_{Pred,.632}$ estimator was more exact than the others. The $\hat{R}^2_{Pred,app}$ estimator tended to slightly overestimate $\hat{R}^2_{Pred}$, while the $\hat{R}^2_{Pred,.632^+}$ estimator tended to slightly underestimate $\hat{R}^2_{Pred}$. Hence, since $\hat{R}^2_{Pred,.632^+} < \hat{R}^2_{Pred,.632}$, the 0.632^+ estimator accomplished the purpose of correcting the possible overoptimism of the 0.632 estimator. But, in our study we considered only one predictor, then no obvious overoptimism was expected, and indeed the $\hat{R}^2_{Pred,.632^+}$ estimates were only slightly smaller than the $\hat{R}^2_{Pred,.632}$ estimates.

Therefore, the estimators of $\hat{R}^2_{Pred}$ obtained from the criterion of the Brier score worked well under our simulated scenarios. Even if only minor differences were observed between the estimators, the 0.632 estimator was particularly the best approximation to the original estimator $\hat{R}^2_{Pred}$.

Test for the difference of the mean $\hat{R}^2_{Pred}$ between estimators

After visual inspection of boxplots across the 100 repetitions, we found that the $\hat{R}^2_{Pred}$ from the different estimators generally followed a normal distribution. This held for estimates from the original and the alternative estimators, and for both criteria of estimation.

We performed Z-tests with a significance level of 0.05 for the difference in the mean gain in prediction among the different estimators within each criterion. The results confirmed our findings described above. We found no evidence of significant differences between the mean $\hat{R}^2_{Pred}$ and $\hat{R}^2_{Pred,.632}$ in most of the scenarios. Moreover, we

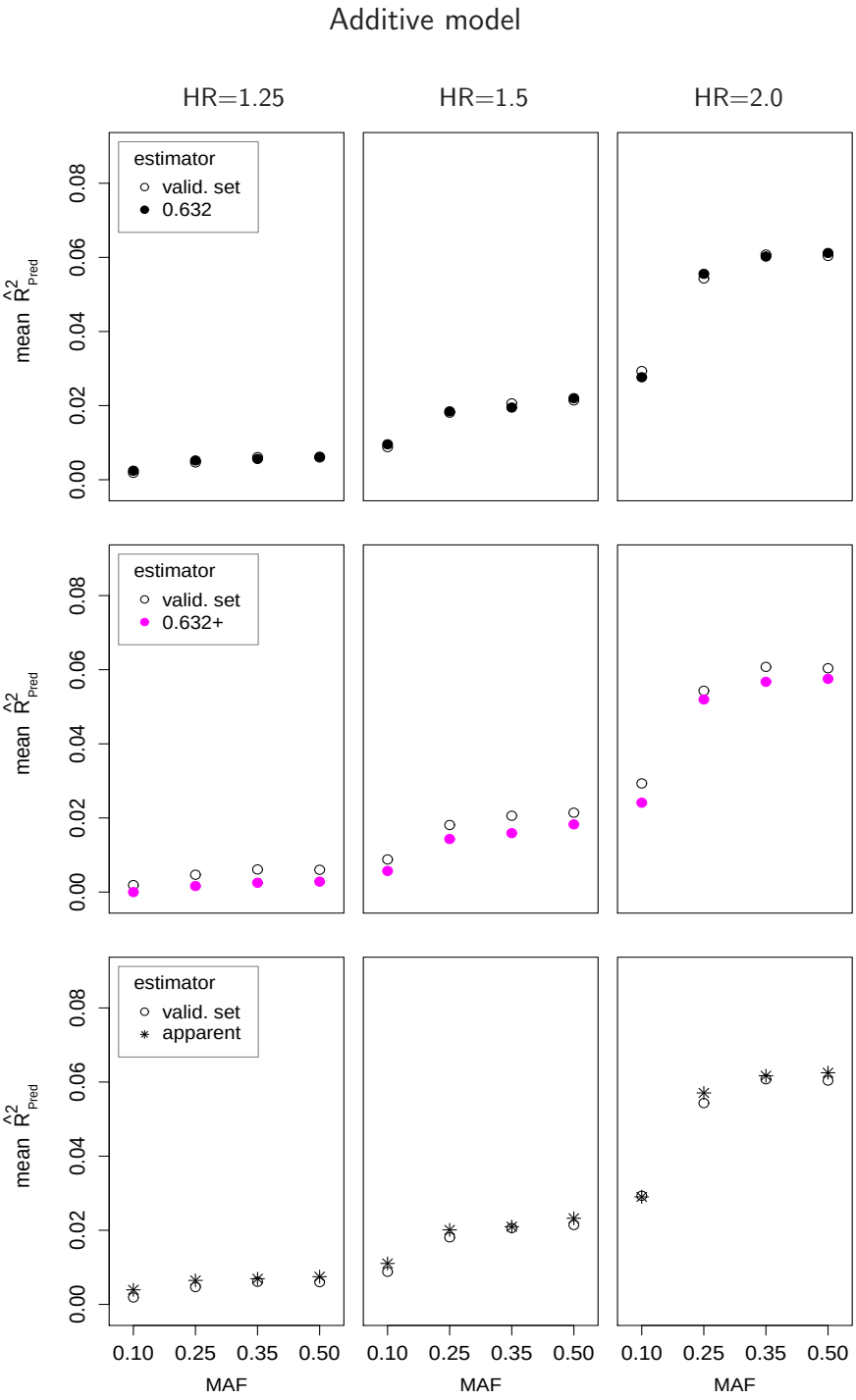

Figure 5.6: Criterion of the **Brier score**. Comparison of mean estimates of gain in prediction ($\hat{R}^2_{Pred}$) using a validation set and estimators of prediction errors. The 0.632 ($\hat{R}^2_{Pred,.632}$), the 0.632+ ($\hat{R}^2_{Pred,.632+}$), and the apparent error ($\hat{R}^2_{Pred,app}$) estimators. Estimates of gain in prediction are plotted for different minor allele frequencies (MAF) and hazard ratios (HR)

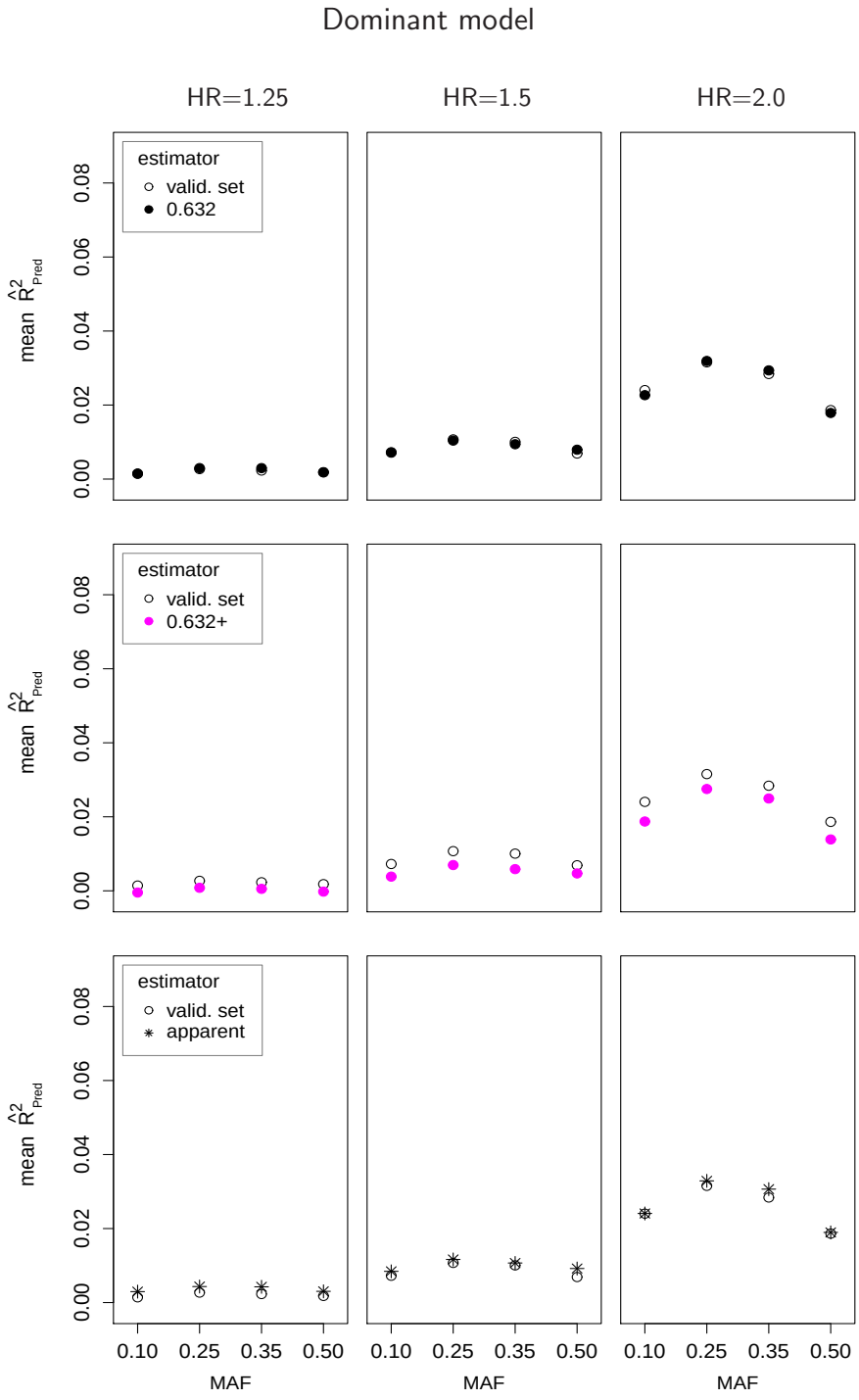

Figure 5.7: Criterion of the **Brier score**. Comparison of mean estimates of gain in prediction ($\hat{R}^2_{Pred}$) using a validation set and estimators of prediction errors. The 0.632 ($\hat{R}^2_{Pred,.632}$), the 0.632+ ($\hat{R}^2_{Pred,.632+}$), and the apparent error ($\hat{R}^2_{Pred,app}$) estimators. Estimates of gain in prediction are plotted for different minor allele frequencies (MAF) and hazard ratios (HR)

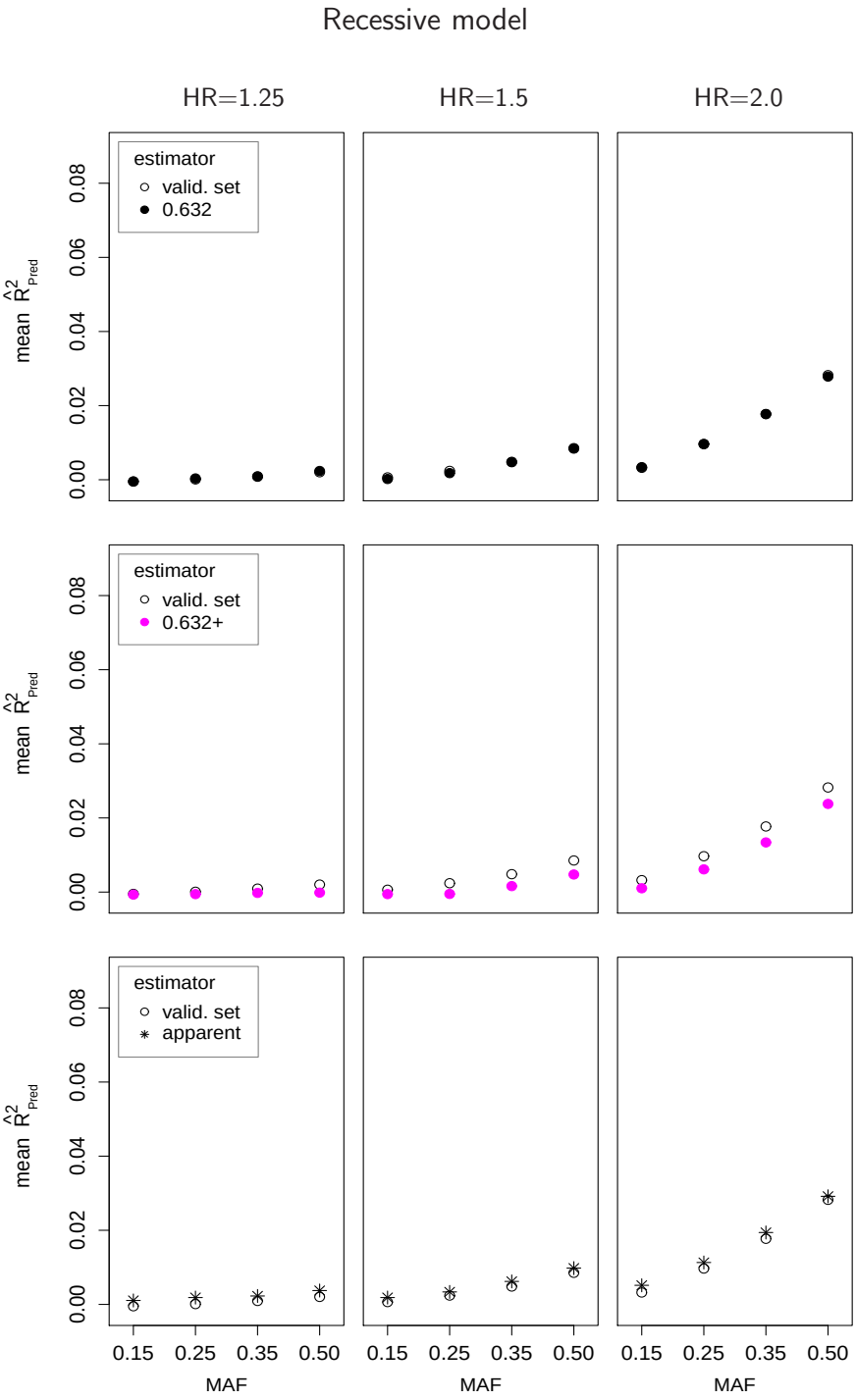

Figure 5.8: Criterion of the **Brier score**. Comparison of mean estimates of gain in prediction ($\hat{R}^2_{Pred}$) using a validation set and estimators of prediction errors. The 0.632 ($\hat{R}^2_{Pred,.632}$), the 0.632+ ($\hat{R}^2_{Pred,.632+}$), and the apparent error ($\hat{R}^2_{Pred,app}$) estimators. Estimates of gain in prediction are plotted for different minor allele frequencies (MAF) and hazard ratios (HR)

found evidence of significant differences between the mean $\hat{R}^2_{Pred}$ estimator and the $\hat{R}^2_{Pred,app}$ and the $\hat{R}^2_{Pred,.632+}$ estimators in various of the scenarios.

Furthermore, there was also evidence of significant differences in the mean gain of prediction between pairs of estimators: $\hat{R}^2_{Pred,.632}$, $\hat{R}^2_{Pred,.632+}$, and $\hat{R}^2_{Pred,app}$, in various of the scenarios. An exception to the latter was, no evidence of significance in the means between the $\hat{R}^2_{Pred,app}$ and the $\hat{R}^2_{Pred,.632+}$ estimators with the criterion of the Schoenfeld residuals. This also confirms our description of findings with estimators with Schoenfeld residuals described above.

Standard errors of the estimators

Regarding the standard errors, they generally increased slightly with higher *HR* and higher *MAF*.

The standard error of estimators from the Schoenfeld residuals were higher than from the Brier score. The difference can be explained by the fewer observations used in the estimates with the criterion of the Schoenfeld residuals. Usually, computations of an estimator using less data make it more variable compared to computations with more data. But also, higher estimate values produce higher standard errors. Hence, since the estimators with the Schoenfeld residuals produced higher estimates it tended to produce higher standard errors as well.

Among the estimators with the Schoenfeld residuals, the standard errors for the $\hat{R}^2_{Pred,.632}$ and for the $\hat{R}^2_{Pred,.632+}$ estimators were similar, while for $\hat{R}^2_{Pred,app}$ it was always slightly higher. The latter can be explained since the $\hat{R}^2_{Pred,app}$ is estimated with a specific sample at each replication, while the $\hat{R}^2_{Pred,.632}$ and $\hat{R}^2_{Pred,.632+}$ are estimated with bootstrap techniques whose mean estimates make these estimations be more representative for any data set. The highest standard error (expressed in percentages, for example, using $\hat{R}^2_{Pred,.632}\times 100$) was $\approx 5\%$ for the additive and dominant models, and $\approx 3\%$ for the recessive model.

Among the estimators with the Brier score, the standard errors for the three estimators ($\hat{R}^2_{Pred,.632}$, $\hat{R}^2_{Pred,.632+}$, and $\hat{R}^2_{Pred,app}$) were similar. Also, there were not much differences in standard errors between the three genetic models. The highest standard errors were $\approx 2\%$ for all the genetic models.

For small MAFs and very small HRs, some of the individual $\hat{R}^2_{Pred}$ estimates among the 100 replications were negative. For example, for $MAF = 0.10$ and $HR = 1.25$, 27%, 66%, and 8% gave negative $\hat{R}^2_{Pred}$ values with the Brier score, while they were 36%, 8%, and 16% for the Schoenfeld residuals. Hence, the Schoenfeld residuals performed slightly better than the Brier score. The negative values occurred when the predictor model did not improve the prediction of the null model, but it rather performed worse than the null model, i.e. the prediction error of the predictor model was higher than the prediction error of the null model. This occasional behaviour was also observed in the work by Müller et al. (2008). In theory the $\hat{R}^2_{Pred}$ should fall in the range [0,1], where $\hat{R}^2_{Pred}$=0 means the predictor and null models performed

equally in terms of prediction, and $\hat{R}^2_{Pred}$=1 means the predictor model improved the prediction of the null model. If we assume the covariate model cannot perform worse than the null model, then a negative value can be considered as 0.

In our study, we did not evaluate the individual values but the mean values. Then, we kept and considered all the estimates as they were, both negative and positive, to estimate the mean $\hat{R}^2_{Pred}$. We observed that some of those mean estimates in the recessive model were slightly negative. That indicates that models that include SNP predictors with recessive genetic models, small effect sizes, and small frequencies of the risk allele will hardly show an improvement on the prediction of the survival outcome.

The facts discussed in the two previous paragraphs held not only for the $\hat{R}^2_{Pred}$ estimator but also for the $\hat{R}^2_{Pred,app}$, $\hat{R}^2_{Pred,.632}$, and $\hat{R}^2_{Pred,.632+}$ estimators.

If we modify the individual negative values to 0, it would produce increased mean estimates, and we did not apply it here. However, in practice when we work with only one data set, a negative estimate of the gain in prediction can be assumed to be 0, i.e. the prediction model performs as good as the null/reference model.

In conclusion, the 0.632 estimator showed overall to be more appropriate than the 0.632+ and the apparent error estimators to approximate the original $\hat{R}^2_{Pred}$ estimate for the gain in prediction with a SNP predictor in Cox regression models. Moreover, from the two approaches to estimate the gain in prediction $\hat{R}^2_{Pred}$, the Schoenfeld residuals is preferred since it yields higher estimates and allows us to differentiate better the improvement in prediction of the models in comparison to the Brier score. The higher estimates with the criterion of the Schoenfeld residuals come from the fact that it accomplishes better the limits on the range [0,1] of the R^2 as a measure of explained variation (Müller et al. 2008).

Hence, for clinical applications when the interest is to predict outcomes on individual genetic basis the evaluation of prediction capability of a model is important. We have found that the Schoenfeld residuals with the 0.632 estimator can be used as a tool for evaluation of prediction in that context. Moreover, this estimator aids to distinguish better the effect of a predictor on the outcome. This is an advantage in genetic studies where the effects of SNPs are usually small.

5.3.3 Capability of $\hat{R}^2_{Pred}$ to identify the correct genetic model

In genetic studies it is important to identify the genetic variants contributing the most to the development of a trait. However, for the case of biallelic SNPs it is also important to identify the most appropriate genetic model of the risk allele for the development of the trait. In this section we illustrated how good selection of a particular genetic model was, based on $\hat{R}^2_{Pred}$ estimators.

From each data set we computed three estimators of $\hat{R}^2_{Pred}$, each from a Cox model including the SNP predictor in the form of an additive, a dominant or a recessive

genetic model. Since we simulated data based on a specific genetic model, we had knowledge of the true genetic model, and a Cox model including the true one was fitted there. Then, we fitted one correct and two misspecified models, in terms of the genetic model assumed for the predictor.

For instance, if a data set was simulated with a SNP with an additive effect on the survival trait, we fitted the correct Cox regression model with such genetic model of the SNP (i.e. $X_a = X$=0, 1, 2, according to the number of risk alleles of the genotype), and in addition we fitted the two misspecified Cox regression models: one with the dominant model of the SNP (i.e. X_d=1 if X=1 or 2, and X_d=0, otherwise), and one with the recessive model of the SNP (i.e. X_r=1 if X=2, and X_r=0, otherwise).

As often done in practice, the highest $\hat{R}^2_{Pred}$ produced from the three fitted models was selected, and the corresponding genetic model was assumed to be the correct one. The capability of the estimator $\hat{R}^2_{Pred}$ to identify the correct genetic model was computed as the frequency of the correct genetic model selected among the 100 replications.

Since the 0.632 estimator was shown to be the most appropriate estimator to approximate $\hat{R}^2_{Pred}$, we analysed the capability of the $\hat{R}^2_{Pred,.632}$ to identify the true genetic model.

Identification of the correct genetic model with the $\hat{R}^2_{Pred,.632}$ estimator

We found that the $\hat{R}^2_{Pred,.632}$ estimator identified very often the correct genetic model. The $\hat{R}^2_{Pred,.632}$ from the correct genetic model was frequently higher than $\hat{R}^2_{Pred,.632}$ from the misspecified models, and the frequency of correct identification increased with higher hazard ratios. This holds for $\hat{R}^2_{Pred,.632}$ estimations using either the criterion of the Schoenfeld residuals (Figure 5.9) or the criterion of the Brier score (Figure 5.10).

The number of times to identify the correct additive genetic model with the criterion of the Brier score were higher than with the Schoenfeld residuals, this could be because of the higher variability seen with the Schoenfeld residuals than with the Brier score. There were no obvious differences between the two criteria in the identification of a correct dominant or recessive model.

In additional evaluations, we found that when a misspecified genetic model was selected, the value of the correct estimate was almost kept, i.e. the excess in the estimated value produced by a misspecified model was minor. According to Figure 5.9, from the criterion of the Schoenfeld residuals, a dominant model would be selected many times instead of the correct additive model of the risk allele. That means that in those cases, the $\hat{R}^2_{Pred,.632}$ of the dominant model (the wrong model here) was higher than the additive model (the correct model here). However, the increased gain in prediction given by the wrong dominant model was on average no larger than

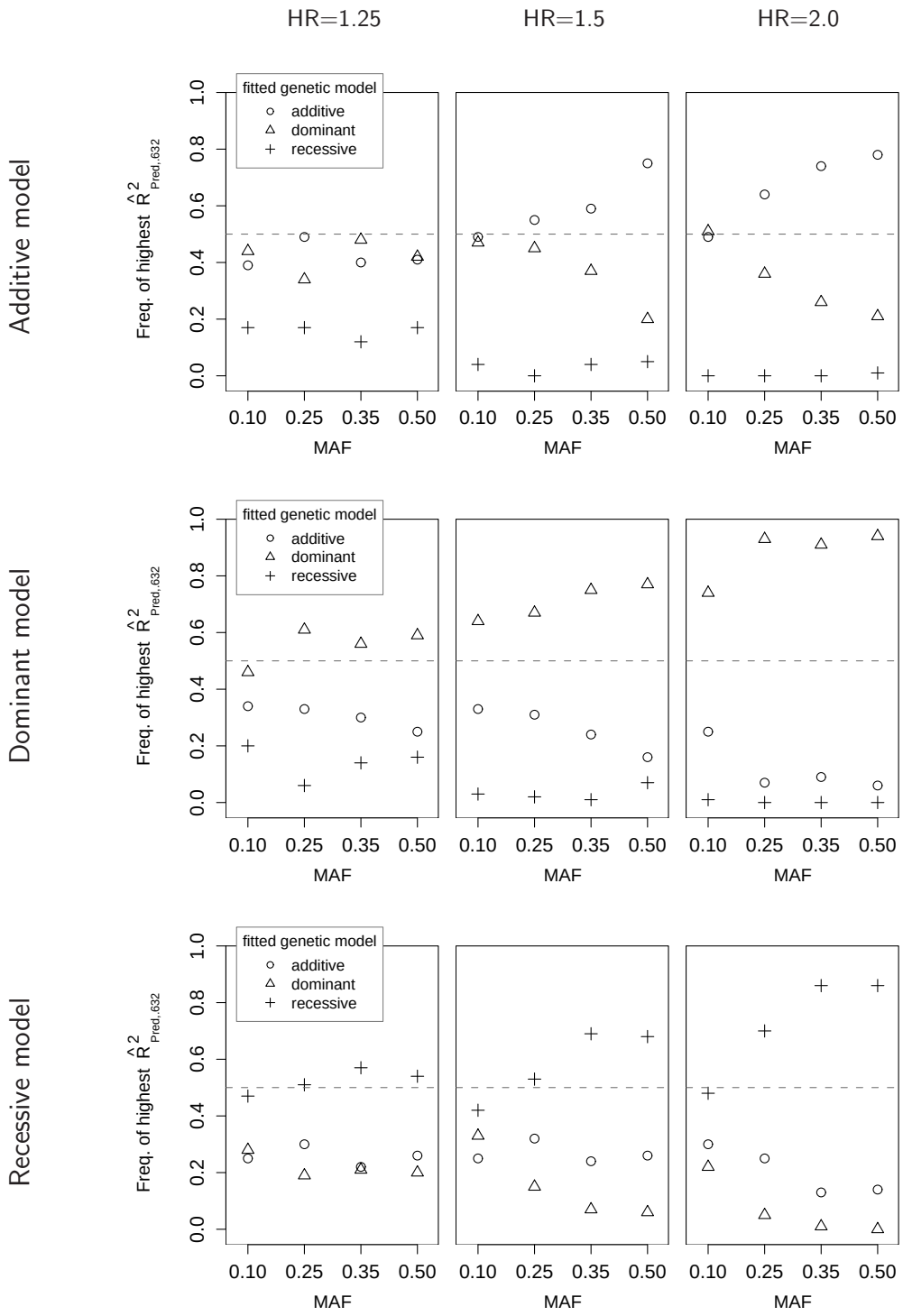

Figure 5.9: Criterion of the **Schoenfeld residuals**. Frequency of fitted genetic models identified as correct[a] genetic models associated with the outcome. Frequencies are plotted for different minor allele frequencies (MAF) and hazard ratios (HR)

[a] A genetic model was taken as correct if it provided the highest mean estimate of gain in prediction ($\hat{R}^2_{Pred,.632}$) among the three fitted genetic models: additive, dominant and recessive. The correct simulated genetic model is indicated on the left side of the plots.

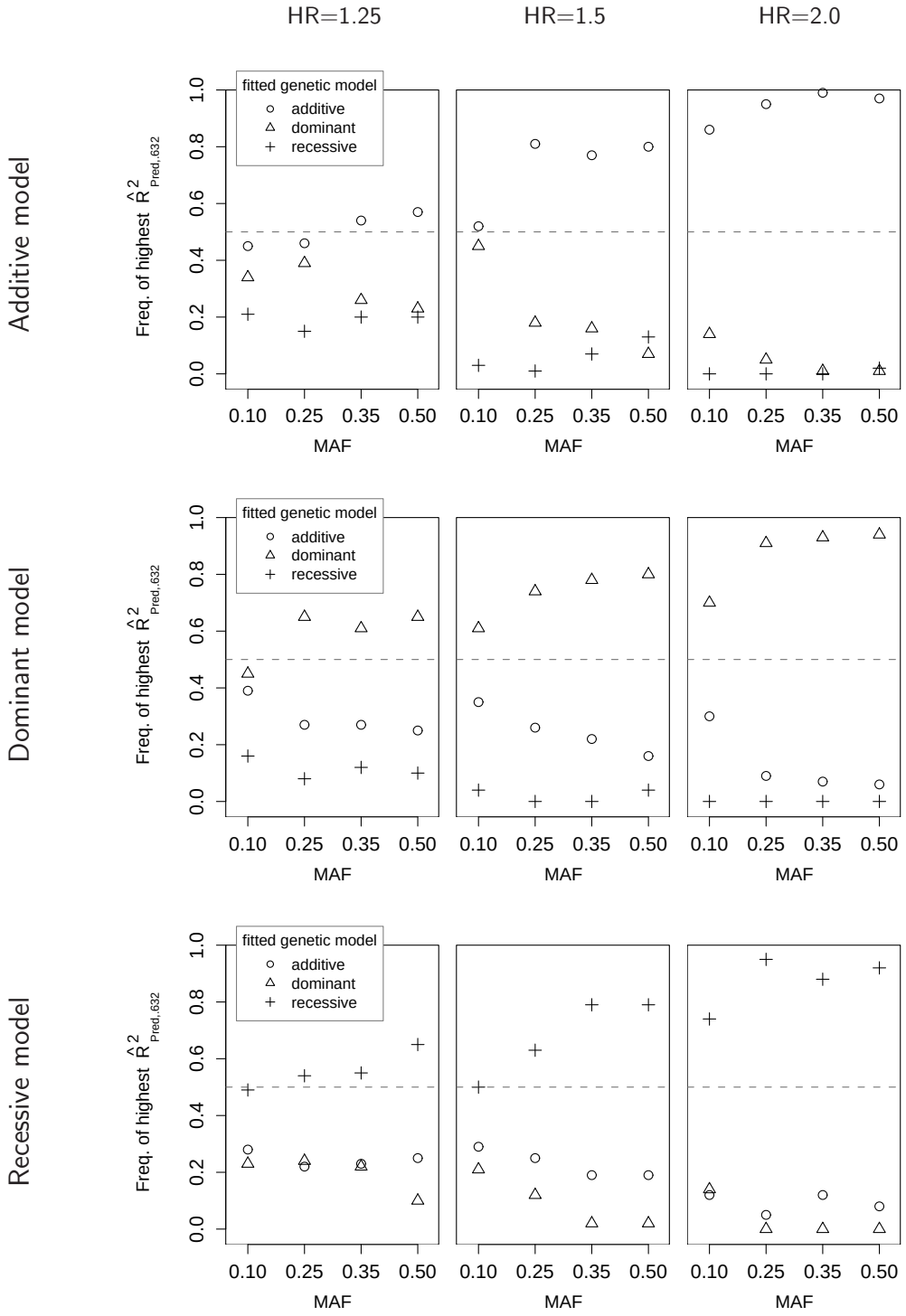

Figure 5.10: Criterion of the **Brier score**. Frequency of fitted genetic models identified as correct[a] genetic models associated with the outcome. Frequencies are plotted for different minor allele frequencies (MAF) and hazard ratios (HR)

[a] A genetic model was taken as correct if it provided the highest mean estimate of gain in prediction ($\hat{R}^2_{Pred,.632}$) among the three fitted genetic models: additive, dominant and recessive. The correct simulated genetic model is indicated on the left side of the plots.

2%. In general, selecting a misspecified genetic model involved to estimate $\hat{R}^2_{Pred,.632}$ with a positive bias of 2% when we used the criterion of the Schoenfeld residuals, and of 0.5% when we used the criterion of the Brier score (data not shown).

The slightly larger bias produced with the criterion of the Schoenfeld residuals in comparison with the Brier score should not be interpreted as a disadvantage, it is rather understandable since it also produces larger $\hat{R}^2_{Pred,.632}$ values.

Therefore, the wrong selection of the correct genetic model did not necessarily mean big deviation from the correct contribution of the genetic variant, which is an important conclusion for purposes of application in real genetic data.

The latter does not mean that we can pick up either of the genetic models arbitrarily. If we do so, we might be underestimating the prediction of the gene unnecessarily, as it can be seen from Figures 5.11 and 5.12. By assuming a misspecified recessive model when the correct genetic model is additive or dominant would cause large losses in prediction capability of the gene, especially when, in both cases, the recessive model has almost no chance to produce the highest $\hat{R}^2_{Pred,.632}$ (see Figures 5.9 and 5.10, additive and dominant models). By assuming a misspecified dominant model when the correct genetic model is recessive would also cause large losses than necessary, especially when we see that the dominant model is the one with the least chance to produce the highest $\hat{R}^2_{Pred,.632}$ (see Figures 5.9 and 5.10, recessive models).

If we wish to assume a particular genetic model without previous selection of the highest $\hat{R}^2_{Pred,.632}$, it would be safer to assume an additive genetic model, i.e. to keep the original distribution by the number of alleles of the predictor. That would mean to have the least losses in prediction capability of the gene, compared to any other alternative. However, the drawback of using an additive model, arbitrarily, in Cox regression models, is that convergence problems in the estimation of the parameters ($\boldsymbol{\beta}$) of the predictors might appear, especially for small samples or for samples with sparse events.

In conclusion, using the $\hat{R}^2_{Pred,.632}$ estimator to identify the most appropriate genetic model of a SNP is advantageous. The $\hat{R}^2_{Pred}$ has big chances to collect the correct genetic model by selecting the highest $\hat{R}^2_{Pred,.632}$ from the three possible genetic models considered in this work.

In cases when a misspecified genetic model is collected, the deviation to the increment of the true prediction measure might only be minor. On the contrary, choosing a genetic model arbitrarily might cause meaningful losses on the true prediction measure of a SNP. Moreover, both criteria, the Schoenfeld residuals and the Brier score, proved to work well as tools to identify the correct genetic model, with a slight advantage of the Brier score in the identification of an additive model.

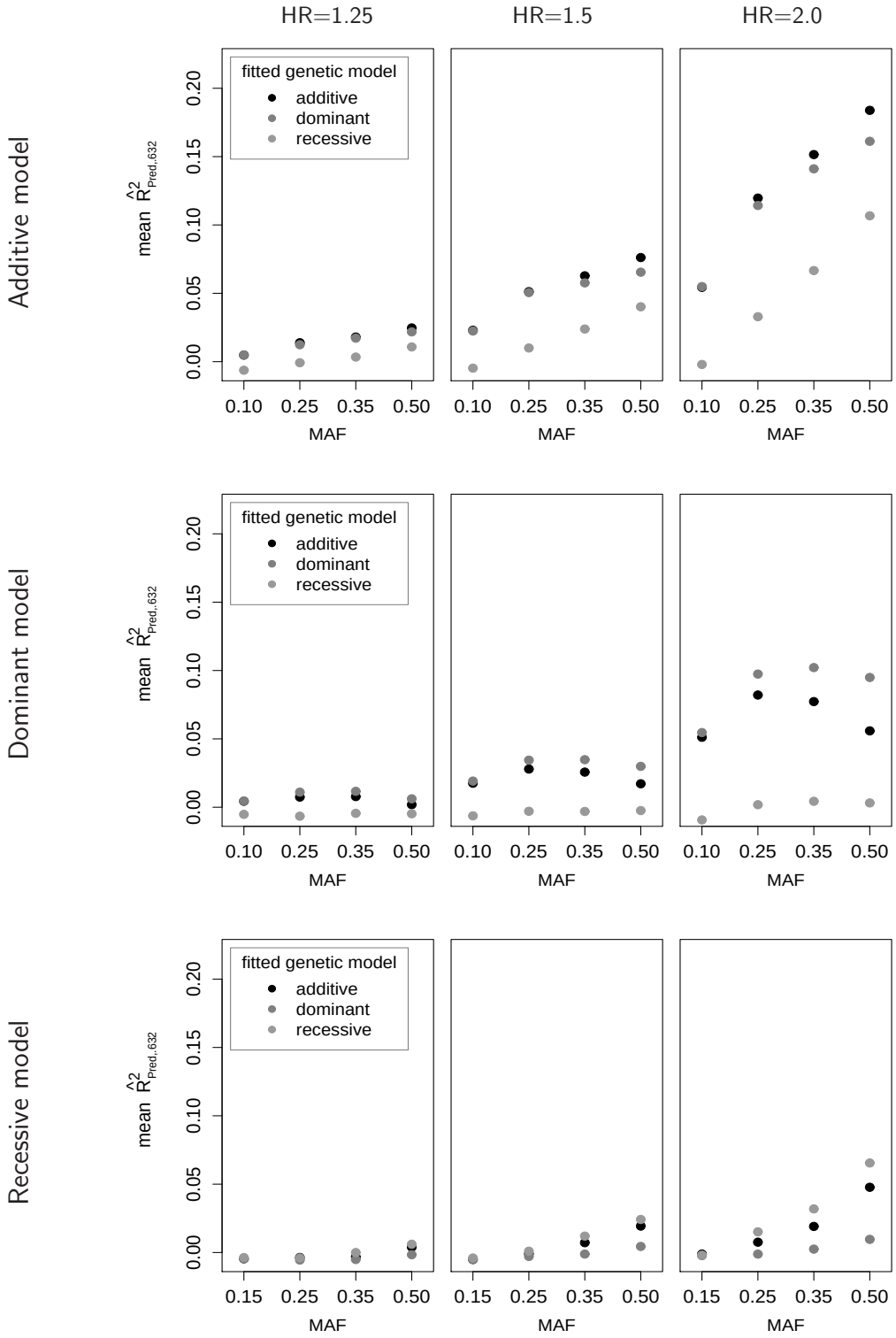

Figure 5.11: Criterion of the **Schoenfeld residuals**. Comparison of mean estimate of gain in prediction (by using $\hat{R}^2_{Pred,.632}$) between true (indicated at the left side of each Figure) and misspecified genetic models in the fit of the genetic Cox model. Estimates of gain in prediction are plotted for different minor allele frequencies (MAF) and hazard ratios (HR)

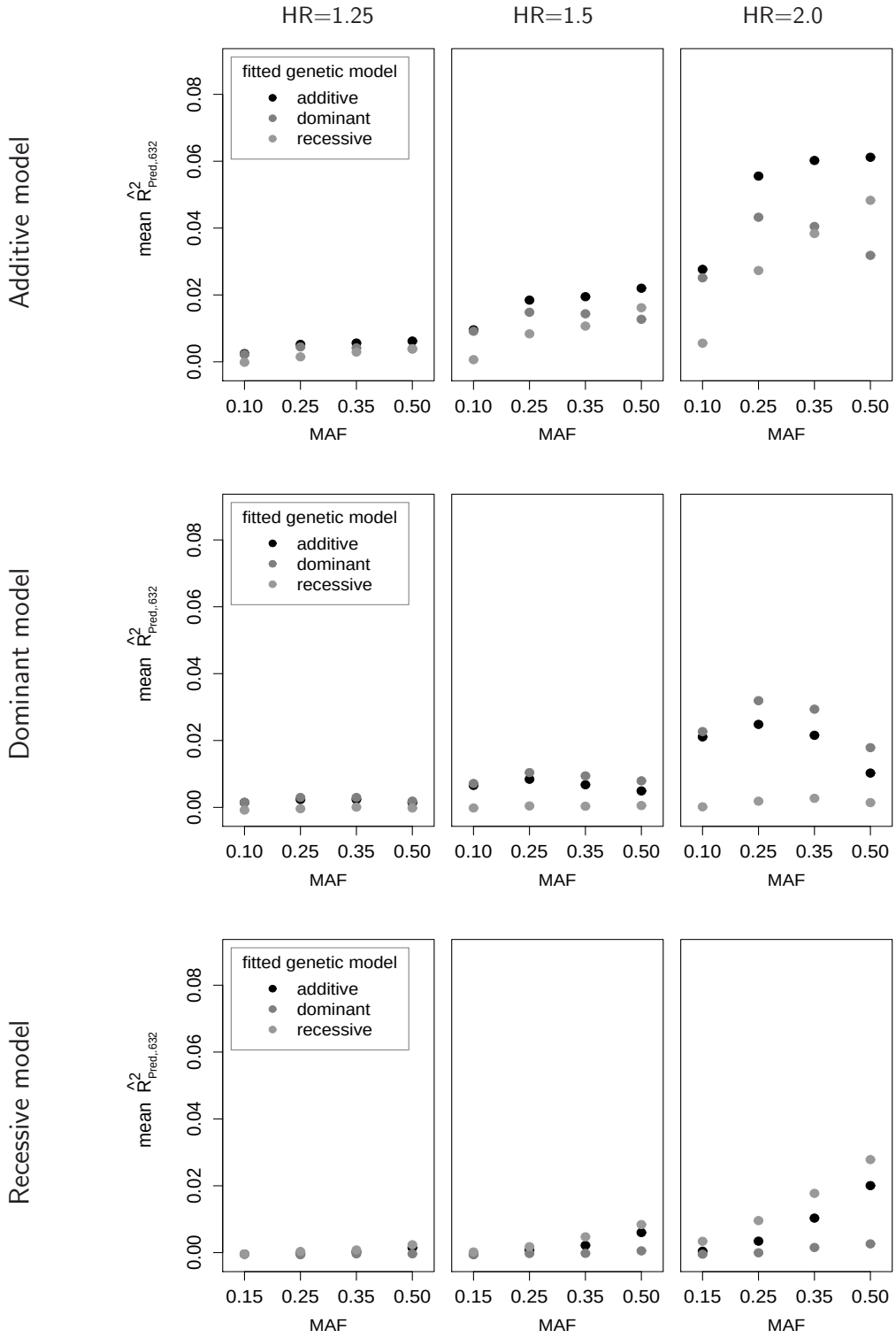

Figure 5.12: Criterion of the **Brier score**. Comparison of mean estimate of gain in prediction (by using $\hat{R}^2_{Pred,.632}$) between true (indicated at the left side of each Figure) and misspecified genetic models in the fit of the genetic Cox model. Estimates of gain in prediction are plotted for different minor allele frequencies (MAF) and hazard ratios (HR)

Chapter 6

Prediction in Haematopoietic Stem Cell Transplantation

In this chapter we presented a cohort study for the overall survival of patients undergoing haematopoietic stem cell transplantation (HSCT). We investigated the association between some candidate SNPs and the overall survival of the patients. The associations were investigated by fitting Cox regression models to measure the effect of single SNPs as well as the effect of multiple SNPs on overall survival. In all cases the models were adjusted by the effect of clinical factors summarized into clinical risk scores.

The goal of this study was to identify relevant SNPs that may contribute, together with established clinical factors, to the prediction of overall survival of HSCT patients.

We also evaluated the gain in prediction of overall survival after incorporating genetic factors in the Cox models, compared to using only the clinical risk scores. The gain in prediction was estimated by using the $\hat{R}^2_{Pred,.632}$ estimator based on the criterion of Schoenfeld residuals. We decided to apply this specific estimator because of the results shown in our simulation study (chapter 5): first, the $\hat{R}^2_{Pred,.632}$ estimator approximates the correct measure of gain in prediction ($\hat{R}^2_{Pred}$) better than the other competing estimators (the $\hat{R}^2_{Pred,app}$ and the $\hat{R}^2_{Pred,.632^+}$); and second, the estimators based on the criterion of the Schoenfeld residuals tend to distinguish the gain of prediction from a predictor model better than the estimators based on the competing criterion (the Brier score).

In the first section of this chapter we present the background on the topic of HSCT and on a translational project for study of HSCT. Then, we describe the data in the second section, and describe the statistical methods used for analysis of the data in the third section. Finally, we present the results and the respective discussion in the last section.

6.1 Study on Haematopoietic Stem Cell Transplantation (HSCT)

Medical treatment to cure cancer of the blood and immune system have been the concern of medical scientists in this field. HSCT is an alternative that has gained in importance as a treatment in recent years. However, it still needs more development to be successful. Main problems derived from this treatment are post-transplant complications, outcomes such as infection, or the so called graft-versus-host disease, and subsequently mortality. The research community on HSCT has determined and validated five clinical factors influencing the success of the transplants. Among others, these clinical factors are the age of the patient, the relation between patient and donor of the stem cells, and the stage of the disease. Clinical risk scores, derived from the clinical factors, have been constructed to aid the prediction and prevention of post-transplant complications in patients undergoing HSCT.

Based on previous studies, it is also believed that genetic factors play an important role on the success of the transplants. So far, these studies have been performed with single SNPs or for small groups of patients, and there is not yet a consensus on the SNPs involved with the success of the transplants. Moreover, multiple effect of SNPs on post-transplant complications has not been extensively studied either.

In the following, we give some definitions and background on haematopoietic stem cell transplantation. Then, we introduce a translational HSCT project that intends to deal, among other things, with the study of effect of multiple SNPs on HSCT.

6.1.1 Background on HSCT

Stem cells are precursor cells that can divide and differentiate into specialized cells types in the body such as a muscle cell, a red blood cell, or a brain cell. They can self-renew to produce more stem cells and to replenish others (Marieb 2004, Stem Cell Basics: Introduction 2009). There are three types of human stem cells: embryonic, embryonic germ, and adult. The embryonic and embryonic germ stem cells develop at early stage of development, approximately five days and five to nine weeks, respectively. Adult stem cells are found in developed tissue, for example the haematopoietic stem cells (Marieb 2004).

Haematopoietic stem cells (HSCs) are a type of stem cells that form blood and immune cells. They have the properties to replenish all blood cell types and to self-renew (Müller-Sieburg et al. 2002, Hematopoietic Stem Cells 2001). HSCs are found mainly in the bone marrow of adults, that is the spongy tissue in the interior of bones, and also in peripheral, circulating blood. HSCs are nowadays used as a therapy. HSC transplants are used as a therapy mainly in patients with haematological malignant diseases, i.e. cancer of the blood and immune systems, such as leukaemia and lymphoma, which result from the uncontrolled proliferation of white blood cells (Hematopoietic Stem Cells 2001).

Haematopoietic stem cell transplantation (HSCT) is a medical procedure to replace the cells destroyed by radiotherapy or chemotherapy in patients with haematological malignant diseases. Sources to collect HSCs are the bone marrow or peripheral blood. The HSCs are collected from the same patient (autologous transplant) or from a donor (allogeneic transplant). In the case of allogeneic transplant, a donor can be related (usually a sibling) or unrelated to the patient (Hematopoietic Stem Cells 2001, Copelan 2006). In this context, the transplanted HSCs are also called *allogeneic graft*, and the patient undergoing the HSCs transplant is called *host*.

The donor tissue type has to be compatible with that of the patient, since severe immune reactions may occur, the severity of which is dependent on the extent of incompatibility. On this matter, matching of the human leukocyte antigen (HLA) between patients and donors is crucial.

Antigen is a substance that is recognized as foreign and activates the immune system (Marieb 2004). The HLA is the major histocompatibility complex (MHC) in humans. The MHC helps the immune system protecting the body by recognizing proteins from the own individual, and proteins from foreigners such as viruses and bacteria. MHC consists of three classes genes located on chromosome 6, Class I (the main genes are HLA-A, -B, and -C), Class II (the main genes are of the types HLA-DP, -DQ, and -DR), and Class III (these genes encode components of the so called complement system) (HLA gene family 2009).

Class I genes produce proteins that are found on almost all cell's surface, these proteins bind to peptides (fragments of proteins) from inside the cells and display them to T cells. T cells are white blood cells known as lymphocytes. Cytotoxic T cells (T_C cells, a type of T cells) destroy cells whose peptides are recognized as foreigners. Hence, under HLA mismatch between donor and patient, T_C cells from the patient recognize peptides of the donor cells as foreigners and cause *graft rejection*. Likewise, T_C cells from the donor recognize peptides of the patient cells as foreigners and cause *Graft-versus-Host Disease (GvHD)* (Graft-versus-host disease 2011, HLA gene family 2009).

Class II genes produce proteins that are found on the cell's surface of certain cells of the immune system. These proteins bind to peptides from outside the cells and display them to the T cells. Recognition of foreigner peptides stimulate production of helper T cells (T_H cells, a type of T cells), which in turn stimulate proliferation of B cells. B cells produce antibodies to the foreigner peptides, the antigens. Antibodies cannot destroy antigens, but they can inactivate and tag them for destruction (Marieb 2004, HLA gene family 2009).

Two types of GvHD can appear: acute and chronic GvHD (aGvHD and cGvHD). aGvHD is normally observed within few weeks (the first 100 days) after transplant. It affects the liver, skin, and gastrointestinal tract (Graft-versus-host disease 2011). cGvHD is normally observed after 100 days. Organs commonly affected include the skin, mouth, liver, eyes, gastrointestinal tract, lung, and oesophagus (Lee and Flowers 2008).

Typing of HLA locus to identify full HLA-matched donors has become a standard

of care. However, it is usually difficult to find full HLA-matched unrelated donors. HSC transplants are still a problem for the association with GvHD on that group of patients, more than with sibling donors (Riddell and Appelbaum 2007).

Types of treatments are used to prepare a patient for stem cell transplantation, these treatments are called *conditioning regimens* (conditioning regimen 2011), which influence the prevention of GvHD. Myeloablative regimens such as total-body irradiation are designed to kill all residual cancer cells in the body of the patient. This causes immunosuppression (reduction of the activity of the immune system), that reduces chances of graft rejection, and favours engraftment in allogeneic transplantation. The drawback of myeloablative regimens is that since it lowers activity of the immune system, there is lower reaction against foreigners, such as viruses, which increases the chance of fatal post-transplant infections, i.e. it increases *transplant related mortality (TRM).*

Non-myeloablative regimens use doses of chemotherapy and radiation much lower than those of myeloablative regimens. These regimens rely on a graft-versus-tumour (GvT) effect to kill tumour cells with donor T cells. Reduced-intensity regimens vary between myeloablative to non-myeloablative. The advantage of both reduced-intensity and non-myeloablative regimens is the decreased toxicity as well as a lower chance of infections and TRM. However, *relapse* (returning of the disease) increases since remaining tumoural cells proliferate again, unless tumoural cells were eliminated by the GvT effect (Conditioning Regimens 2011).

Another way of preventing GvHD has been focused on T cell removal from donor stem cells. This is called *T cell depletion.* On one side, it helps reducing the occurrence of GvHD, since there is no reaction of the graft to the host cells, but it increases the possibility of graft rejection. On the other side, since there is no activity of the immune system from donor stem cells, no GvT effect takes place, which increases the chances of relapse (Riddell and Appelbaum 2007).

Even if conditioning regimens and T cell depletion can improve post-transplant problems with GvHD, it has not been demonstrated that these ways of preventions improve survival (Riddell and Appelbaum 2007). The overall survival of patients after HSCT is not yet encouraging. The survival rate is about 40-60% at 5 years after transplant. As described above, adverse clinical outcomes such as GvHD, infections are further TRM are problems that diminish the success of the transplants.

6.1.2 The HSCT translational network project

The European community and the University of Newcastle-upon-Tyne (UK), agreed in December, 2004, on the implementation of the project: *Identification of genomic and biological markers as predictive/diagnostic/therapeutic tools for use in allogeneic stem cell transplantation: Translational research towards individualized patient medicine.* From here onwards we referred this project as the *HSCT TRANSNET project.* The project was implemented for a period of 48 months.

As a mean of overcoming the situation of low survival of the patients after HSCT, as well as occurrences of GvHD, the HSCT TRANS-NET project aimed to identify and verify over cohorts from different transplant centers, novel prediction indicators and diagnostic markers for the assessment of patients and donors. The novel aspects should lead to considerations of new concepts and clinical practices for individual assessment of patients previous to the transplant. The latter should also implicate improvement of therapies and clinical protocols (TRANS-NET 2008).

Participant institutions of the project were the University of Newcastle-upon-Tyne (UK) as the leading institution, the Geneva University Hospital (Switzerland), the Universitätsmedizin of the Georg-August-University Göttingen (Germany), University of Oslo (Norway), University of Glasgow (UK), Klinikum der Universität Regensburg (Germany), Charles University, Prague (Czech Republic), Helmholtz Zentrum München (GmbH) (Germany), St James's Hospital and Trinity College Dublin (Ireland), INSERM U396, Paris (France), Medical University of Vienna (Austria), and Universitätsklinikum Düsseldorf (Germany).

The HSCT TRANS-NET project was also in collaboration with the European Group for Blood and Marrow Transplantation (EBMT). The EBMT Group is a non-profit organization that aims to promote all aspects associated with the transplantation of haematopoietic stem cells (EBMT 2011). In previous studies, the Chronic Leukaemia Working Party of the EBMT Group determined five clinical risk factors that influence, among other outcomes, the overall survival of HSCT patients with chronic myeloid leukaemia (CML) (Gratwohl et al. 1998, 2006). These clinical factors are: age of patient by the time of transplantation, patient/donor sex combination, donor type, stage of disease, and time from diagnosis to transplantation. From here on we shall refer to these clinical factors as the *EBMT factors*. The EBMT factors have been further validated as clinical risk factors not only on patients with CML disease but also on patients with other haematological diseases such as acute leukaemia, myelodysplastic syndrome, or non-Hodgkin lymphoma, (Gratwohl et al. 2009), and these clinical factors are considered well established in the community.

Table 6.1 contains details on the categories and respective risk scores of the established clinical EBMT factors. The risk of death of a patient undergoing HSCT increases with the age, it is highest for the group of patients older than 40 years. Male patients receiving stem cells from female donors have increased risk compared to other patients, also patients with unrelated donors have higher risk of death than patients with sibling donors. Moreover, the risk of death increases for patients with worse clinical disease stage at the date of transplantation, and for those patients with larger waiting time from the date of diagnosis to HSC transplant.

On the side of genetic studies, it is argued that SNPs in non-HLA genes can influence immune responses and can have an impact on individual patient outcomes during HSCT (Mullally and Ritz 2007). Various studies for association of non-HLA genes with HSCT outcomes have been conducted. Some of the studies were, however, performed on small and specific group of patients, e.g. only on patients with

Table 6.1: Risk scores per category from clinical EBMT factors[a] established for patients undergoing HSCT

EBMT factors	Categories	Risk score
Age of patient at transplantation (years)	< 20	0
	20 - 40	1
	> 40	2
Patient/donor sex combination	Male/Female	1
	other	0
Donor type	HLA-identical sibling	0
	HLA-matched unrelated donor	1
Stage of disease at transplantation	Early	0
	Intermediate	1
	Late	2
Time from diagnosis to transplant	≤ 12 months	0
	> 12 months	1

[a] Clinical factors and risk scores established by the EBMT Group (Gratwohl et al. 1998, 2006, and 2009).

HLA-matched sibling donors, some reviews on these studies have been published (Dickinson et al. 2004, 2008; Mullally and Ritz 2007). However, there is not yet a confirmed conclusion about the influence of non-HLA genes on outcomes based on larger studies and/or on a more extended group of HSCT patients.

As part of the HSCT TRANS-NET research, clinical data (including EBMT factors) and genetic data, specifically non-HLA genes, from patients and donors were collected from the transplant centers participating in this project. The data were compiled in the HSCT-EUROBANK database (EUROBANK 2008).

In our study we aim to evaluate genetic data from the HSCT TRANS-NET project to identify non-HLA genes of potential relevance to prediction of survival after HSCT, in addition to the established clinical EBMT factors. Hence, our study also intends to provide risk groups, derived from the joint clinical and genetic factors, for the identification of patients with high and low risk of death after haematopoietic stem cell transplantation.

6.2 Data description

We accessed the EUROBANK database through our partners in Newcastle, who prepared a database as for our study. We received this database on January 2010, with data on 993 patients. Data on a patient also included data from his/her respective donor. Since our study focused on the overall survival following HSCT of patients with malignant haematological diseases, we only selected patients with malignant haematological disease (n=930), and moreover patients with available survival status and available survival times. Hence, data on 888 patients were available.

Table 6.2: Post-transplant outcomes in HSCT patients

Outcomes	Categories	Frequency	(%)
Acute Graft versus	No aGvHD	252	(33)
Host Disease (aGvHD)	Grade I	131	(17)
	Grade II	213	(28)
	Grade III	106	(14)
	Grade IV	56	(7)
	NA[a]	4	(1)
Chronic Graft versus	No cGvHD	249	(33)
Host Disease (cGvHD)	Limited	84	(11)
	Extensive	118	(16)
	Not determined[b]	147	(19)
	NA[a]	164	(21)
Relapse	No relapse	545	(72)
	Relapse	216	(28)
	NA[a]	1	(0)
Transplant Related	Alive	363	(48)
Mortality (TRM)	No TRM	165	(22)
	TRM	228	(30)
	TRM not determined	2	(0)
Overall death	Alive	363	(48)
	Dead	399	(52)
Total		762	(100)

[a] Not available. [b] Extension of disease was not determined.

6.2.1 Clinical data

For the analysis of clinical data there were 762 patients whose EBMT factor data were deemed to be complete. Table 6.2 summarizes the clinical post-transplant outcomes of these patients. The observed outcomes included acute and chronic GvHD, relapse, transplant related mortality and death. By the end of the study, a total of 66% of patients had developed acute GvHD, 46% chronic GvHD, 28% had suffered relapse, 30% with transplant related mortality, and 52% had died overall.

Table 6.3 contains the clinical characteristics of the patients. Even though the study was focused on adult patients, we considered twelve patients younger than 18 years old contained in the data since they followed the same protocol and treatment as an adult patient. The youngest and oldest patient were 16 and 68 years of age, respectively. The mean and median age of the patients was 40 years.

Twenty-two percent of the transplants were from a female donor to a male patient. The type of donors was almost balanced between HLA-identical siblings (52%) and HLA-matched unrelated donors (48%). Most of the patients had the diagnosis acute

Table 6.3: Clinical characteristics of patients prior to HSCT

Factors	Categories	Frequency	(%)
Age of patient at transplantation (years)[a]	< 20	32	(4)
	20 - 40	346	(46)
	> 40	384	(50)
Age of donor (years)	< 20	34	(4)
	20 - 40	424	(56)
	> 40	286	(38)
	NA[b]	18	(2)
Patient/donor sex combination[a]	Other sex combinations	598	(78)
	Male patient, female donor	164	(22)
Donor type[a]	HLA-matched sibling	395	(52)
	HLA-matched unrelated donor	367	(48)
Diagnosis of haematological disease	Acute leukaemia (AL)	391	(51)
	Chronic myeloid leukaemia (CML)	172	(23)
	Lymphoma	112	(15)
	other diagnoses[c]	87	(11)
Source of stem cells	Bone Marrow	356	(47)
	Peripheral blood	388	(51)
	Bone Marrow/Peripheral blood	3	(0)
	NA[b]	15	(2)
Cytomegalovirus status	Positive, either patient or donor	511	(67)
	Both negative	227	(30)
	NA[b]	24	(3)
Stage of disease at transplantation[a]	Early	328	(43)
	Intermediate	196	(26)
	Late	238	(31)
Time from diagnosis to transplant[a]	≤ 12 months	439	(58)
	> 12 months	323	(42)
T-cell depletion	T-cell depletion	263	(35)
	No T-cell depletion	499	(65)
Conditioning regimen	Myeloablative	503	(66)
	Reduced-intensity	259	(34)
Center of transplant	Vienna/Prague[d]	225	(30)
	Regensburg/Münich[e]	217	(28)
	Newcastle	204	(27)
	Rostock	73	(9)
	Paris	43	(6)
Total		762	(100)

[a] Clinical EBMT factor (Gratwohl et al. 2009). [b] Not available.
[c] Other diagnoses included: plasma cell neoplasia, myelodysplasia syndrome, and chronic myelomonocytic leukaemia. [d] 29% patients were treated in Vienna and 1% in Prague.
[e] 27.5% patients were treated in Regensburg and 0.5% in Münich.

leukaemia (AL, 51%). Forty-three percent of the patients presented early stages of the haematological disease, followed by late stage (31%), and intermediate stage of disease (26%).

The transplant centers included: Vienna, (28%), Regensburg (28%), Newcastle (27%), Rostock (10%), Paris (6%), and Prague (1%). The centers of Vienna and Prague worked in close collaboration by exchanging the medical staff assessing GvHD in order to ensure consistent evaluation of the disease (Ludajic et al. 2008). Likewise, the centers of Regensburg and Münich worked in close collaboration to each other.

6.2.2 Genotype data

Genotype data were typed by the respective transplant centers. A total of 743 patients had available genotype data (including that of their respective donors). The genotype data included genotypes on 20 non-HLA genes from patients and donors. The genes were typed on 1, 2, or 3 different candidate SNPs (Table 6.4).

We had a total of 31 biallelic SNPs, 1 multiallelic SNP, 2 haplotypes, and 1 microsatellite. *Haplotype* is obtained from the combination of alleles of a sequence of adjacent SNPs; for example, the GCR haplotype (GCR_HT) in the data is a combination of alleles of 3 single SNPs of the GCR gene. *Microsatellite* is a repeated sequence of nucleotides, for example, a microsatellite repeat sequence of n times CA, $(CA)_n$. In our data, the IFNG gene locus is a microsatellite of $(CA)_{10}$ (allele 1) to $(CA)_{14}$ (allele 5) (Calvo et al. 2002).

Some of our candidate SNPs are involved in the alteration of cytokine production. *Cytokines* are chemical mediators enhancing the immune response. SNPs from the interleukin genes (IL1, IL6, and IL10), tumour necrosis factor (TNF), and interferon γ (IFNG) belong to this group. SNPs from the estrogen receptor gene (ESR1), vitamin D receptor (VDR), and nucleotide binding oligomerization domain containing 2 (NOD2) are involved in host defence and the immune system (Dickinson et al. 2004, 2008; Mullally and Ritz 2007). It may also happen that the heat shock protein gene (HSP70) has a direct role in GvH reactions (Novota et al. 2008), and it has also been stated that HSP70 can produce powerful cytokines affecting the functional capability of the immune cells (Jarvis et al. 2003).

Previous studies have shown that these SNPs were associated with at least one of the HSCT outcomes, such as aGvHD, cGvHD, TRM, or overall survival. However, the studies have been performed for specific subgroup of patients, or in small studies, e.g. only on patients with HLA-matched sibling donors (Dickinson et al. 2004, 2008; Mullally and Ritz 2007; Jarvis et al. 2003).

Thus, in our study, we intended to analyse the SNPs in association with overall survival for the full cohort of patients. The analysis was performed to evaluate the effect of the SNPs on overall survival, in addition to the effect of the predetermined clinical EBMT factors.

Table 6.4: Non-HLA genes typed in patients and donors in the HSCT study

Full gene name	Gene name	Chr.[a]	SNP rs number	SNP label[b]
Cluster of differentiation 14	CD14	5	rs2569190	CD14
Cluster of differentiation 91	CD91	12	rs1799986	CD91
Complement component 3	C3	19	rs2230199	C3
Estrogen receptor	ESR1	6	rs2234693	ESR1.1
			rs9340799	ESR1.2
Glucocorticoid receptor	GCR	5	rs33389	GCR.1
			rs33388	GCR.2
			rs6198	GCR.3
			haplotype[c]	GCR_HT
Heat shock protein	HSP70-hom	6	rs2075800	HSP70-hom.1
			rs2227956	HSP70-hom.2
Interferon gamma	IFNG	12	microsat.[d,e]	IFNG
Interleukin 1 receptor antagonist	IL1RN	2	rs419598	IL1RN
Interleukin 4	IL4	5	rs2243250	IL4
Interleukin 6	IL6	7	rs1800797	IL6.1
			rs1800796	IL6.2
			rs1800795	IL6.3
Interleukin 10	IL10	1	rs1800896	IL10.1
			rs1800872	IL10.2
			haplotype[f]	IL10_HT
Interleukin 12	IL12B	5	rs3212227	IL12B
Interleukin 13	IL13	5	rs1800925	IL13.1
			rs20541	IL13.2
			rs1881457	IL13.3
Low density lipoprotein receptor	LOX1	12	rs11053646	LOX1
Myelin and lymphocyte protein	MAL	11	rs8177374	MAL
Multi drug resistance receptor	MDR1	7	rs1045642	MDR1.1
			rs2032582[d]	MDR1.2
Nucleotide binding oligomerization domain containing 2	NOD2	16	rs2066844	NOD2.1
			rs2066845	NOD2.2
			rs2066847	NOD2.3
Tumour necrosis factor	TNF	6	rs1800629	TNF
Tumour necrosis factor receptor	TNFRSF1B	1	rs1061622	TNFRSF1B
Vitamin D receptor	VDR	12	rs731236	VDR.1
			rs7975232	VDR.2

[a] Chromosome number. [b] SNP/haplotype labels used in Tables and Figures in this chapter.
[c] Haplotype GCR from SNPs: rs33389, rs33388, and rs6198. [d] Multiallelic.
[e] Microsatellite, where the alleles are defined by repeated CA sequences.
[f] Haplotype IL10 from SNPs: rs1800896, rs1800871 (not available in our data), and rs1800872.

The main drawback for the analysis of these genetic data were many missing genotypes. Not all the patients were typed for each of the SNPs. Percentage of missing genotypes for each SNP varied from 9% to 44%, which reduced the data from 743 to between 678 and 417 patients (see Tables 6.5 and 6.6, column: Sample size).

SNPs with the highest percentages of missing data corresponded to genes HSP70-hom, IL12B, and MDR1. Also, even if the single GCR SNPs did not have very high percentage of missing data, the GCR haplotype did have because it accumulated missing data from the three typed GCR SNPs. High percentage of patients and donors from the center Regensburg had missing genotypes on these SNPs. The patients and donors from the center Rostock were not typed on these SNPs, except on the GCR haplotype. High percentages of missing genotype data causes loss of a large part of the data if multiple SNPs are jointly evaluated during the survival analysis with Cox regression models. That is because the Cox regression analysis requires complete data in all the variables included in the model, otherwise the data are deleted. In view of that, we excluded from the analysis the SNPs with very high percentages of missing genotypes (these were all typed SNPs from genes: HSP70-hom, IL12B, and MDR1; and the haplotype from gene GCR).

Moreover, chi-square tests performed to test Hardy-Weinberg equilibrium (HWE) (section 2.2.3, page 12) indicated that, the genotype frequencies of SNPs labelled as IL1RN and IL6.2 in patients and donors, VDR.2 in patients, and MAL in donors, do not follow genotype distribution according to the HWE (p-value$<$0.05). These SNPs should rather be excluded since their results might lead to spurious association. Even so, we kept and evaluated them during the analysis. However, results from these SNPs should be taken with care.

6.2.3 Follow-up and survival times

The 762 patients with complete data on the clinical EBMT factors, received stem cell transplants in the period from November 1983 to December 2005, with follow-up until November 2009. The follow-up time was between 7 months to 20 years, with a median of 5-6 years. A total of 399 (52%) deaths were observed during the period of study.

Figure 6.1 shows the overall survival of the patients over months after transplant. The median survival time of the patients was 3 years (i.e. 50% survival probability at 3 years after transplant), and moreover, they had 48.5% survival probability after 4 years of transplant. The events of death are rather infrequent after 4 years of transplants, and the survival probability remains above 40% at later years.

We considered the survival times as number of days from date of transplant to death. To estimate prediction errors with the Schoenfeld residuals we need continuous and untied survival times. There were a few cases of tied survival times, which we untied manually. To untie survival times, we added subsequently a fraction 0.01 to tied

Table 6.5: Descriptives of the non-HLA biallelic SNPs genotyped on the HSCT study (overall size of the genotype data, n=743)

SNP label[a]	Allele pairs	MA[b]	MAF %[c]		Missing %[d]		Sample size[e]	
			patient	donor	patient	donor	patient	donor
CD14	G/A	A	49	47	28	27	536	542
CD91	C/T	T	15	16	31	25	514	560
C3	C/G	G	22	22	30	26	523	550
ESR1.1	C/T	C	46	44	13	15	646	635
ESR1.2	G/A	G	40	37	13	13	646	644
GCR.1	C/T	T	14	15	30	24	521	561
GCR.2	T/A	T	44	46	29	24	527	562
GCR.3	G/A	G	17	17	28	24	533	561
HSP70-hom.1[f]	G/A	A	35	32	42	44	433	417
HSP70-hom.2[f]	T/C	C	18	19	33	34	495	488
IL1RN	T/C	C	24	24	11	10	663	671
IL4	T/C	T	15	16	17	17	617	618
IL6.1	G/A	A	39	39	24	20	561	593
IL6.2	C/G	C	7	6	22	17	582	618
IL6.3	G/C	C	42	39	11	9	664	678
IL10.1	G/A	G	47	46	18	17	609	619
IL10.2	A/C	A	24	26	17	16	620	626
IL12B[f]	A/C	C	21	20	41	37	440	471
IL13.1	C/T	T	19	18	27	20	543	592
IL13.2	A/G	A	21	22	22	20	577	596
IL13.3	C/A	C	19	19	21	17	584	613
LOX1	G/C	G	6	8	30	22	517	579
MAL	T/C	T	15	17	32	26	507	550
MDR1.1[f]	C/T	C	47	46	38	35	460	485
NOD2.1	C/T[g]	T[g]	5	5	20	20	592	595
NOD2.2	C/G[g]	G[g]	2	1	20	20	592	595
NOD2.3	- /C[g]	C[h]	3	3	20	20	592	595
TNF	G/A	A	15	15	20	21	592	587
TNFRSF1B	T/G	G	24	23	17	14	618	637
VDR.1	T/C	C	36	40	15	11	630	659
VDR.2	C/A	C[i]	49	47	13	9	649	676

[a] For genetic details regarding the label, see Table 6.4. [b] Minor allele.
[c] Minor allele frequency, based on the available sample size (two last columns).
[d] Percentages based on n=743. [e] Available sample size, it excludes missing genotypes.
[f] Excluded from the statistical analysis because of high missing%.
[g] Alleles of NOD2 SNPs were not specified in the data. The alleles were identified from SNPedia (Cariaso and Lennon 2011). [h] Minor allele found by Heliö et al. (2003).
[i] Minor allele differs between patients (allele A) and donors (allele C).

Table 6.6: Descriptives of the non-HLA multiallelic genes or haplotypes genotyped on the HSCT study (overall size of the genotype data, n=743)

Label[a]	Alleles or haplotypes	MA[b]	MAF %[c]		Missing %[d]		Sample size[e]	
			patient	donor	patient	donor	patient	donor
IFNG	1-5[g]	1,5,4[j]	0.3	0.2	12	11	655	662
MDR1.2[f]	A,G,T	A	1	1	39	35	456	483
GCR_HT[f]	.[h]	GCT[k]	9	8	36	29	474	529
IL10_HT	.[i]	ATA	24	25	20	20	591	598

[a] For genetic details regarding the label, see Table 6.4. [b] Minor or rare allele/haplotype.
[c] Minor allele/haplotype frequency, based on the available sample size (two last columns).
[d] Percentages based on n=743. [e] Available sample size, it excludes missing genotypes.
[f] Excluded from the statistical analysis because of high missing%.
[g] Alleles 1 to 5. Allele 1 corresponds to the microsatellite repeat sequence of 10 times CA, and alleles 2-5 to 11-14 times CA, respectively, (Calvo et al. 2002).
[h] haplotypes: ACA, ATA, ACT, GCA, GCT. [i] haplotypes: ACC, ATA, GCC.
[j] Rare alleles in patients were allele 1 (0.3%), 5 (2.1%), and 4 (4%); the frequent alleles were allele 2 (47.5%) and 3 (46%). The frequencies were similar for donors.
[k] Minor haplotypes in patients were GCT and GCA, both with frequencies 9%.

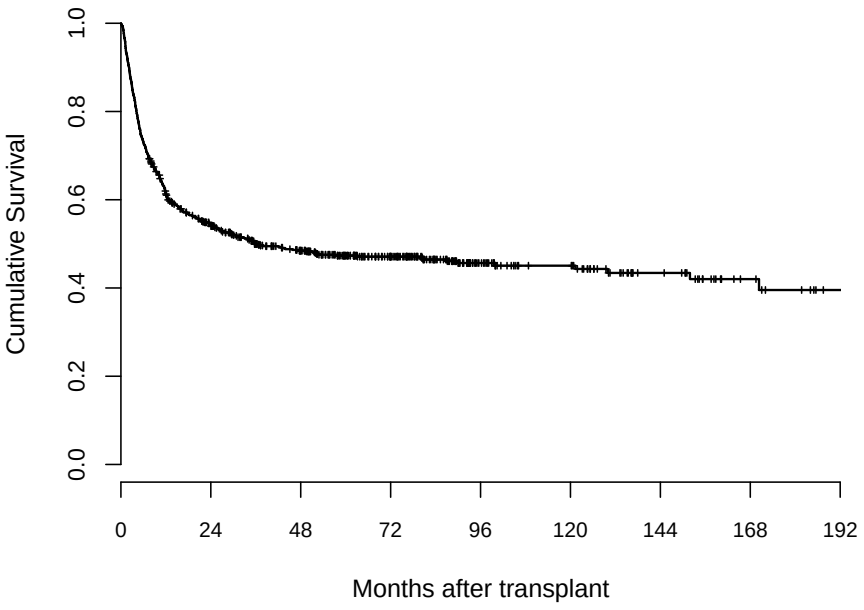

Figure 6.1: Overall Kaplan-Meier survival curve after HSCT

survival times. There were 76 cases of two patients with tied survival times, 19 cases of three patients, 5 cases of four patients, and 1 case of six patients with tied survival times (this latter was at 36 days after transplant).

6.3 Statistical analysis: Cox models with clinical and genetic factors for overall survival

We investigated the association of SNPs with overall survival by simple Cox regression models (with single SNPs), and by multiple Cox regression models (with joint effects of multiple SNPs). In both cases we estimated the effects of the SNPs in addition to the effect of the clinical EBMT-score.

Association of EBMT factors with overall survival

The clinical EBMT factors were included in the Cox models as a discrete variable of the clinical risk score of death for a patient undergoing HSCT. A clinical risk score was obtained as a sum of the single risk scores of the clinical EBMT factors of a patient, hence we call the clinical risk score *the clinical EBMT-score*. The single scores per category of factors are shown in Table 6.1. These scores have been established after studies by the EBMT Group (Gratwohl et al. 1998, 2006, 2009).

The distribution of the clinical EBMT-scores of patients in our data are shown in Table 6.7. The scores varied from 0 to 7 risk points. Most of the patients (25%) had a score 4 for the risk to death, followed by patients with scores 3 (19%), 2 (18%), and 5 (16%). Overall, the frequency of deaths increased with higher EBMT-scores of the patients.

We tested the effect of the clinical EBMT-scores on overall survival of the patients by using the Cox regression model with this single variable. The score entered the model either as a categorical or as a discrete variable. In case of a categorical variable, we considered 6 categories, where the two first scores 0-1 were combined to use it as a reference group with acceptable sample size (91 patients). With this categorical variable we could evaluate the agreement of higher risk scores with higher hazard ratios to death. In case of a discrete variable of the EBMT-score, this resulted in a single estimate that was the hazard ratio by a unit increment on the risk score of a patient. The discrete EBMT-score was later used in the analysis of a joint model with genetic factors.

Further, we evaluated the gain in prediction of overall survival with the clinical EBMT-score with respect to no predictors. The gain in prediction was measured with the $\hat{R}^2_{Pred,.632}$ estimator (see section 4.6.3).

Table 6.7: Distribution of EBMT-score and death status of HSCT patients

EBMT-score	alive (%[a])	dead (%[a])	Total (%[a])	(%[b])
0	4 (57)	3 (43)	7 (100)	(1)
1	48 (57)	36 (43)	84 (100)	(11)
2	73 (54)	63 (46)	136 (100)	(18)
3	79 (54)	68 (46)	147 (100)	(19)
4	85 (46)	101 (54)	186 (100)	(25)
5	45 (37)	77 (63)	122 (100)	(16)
6	28 (40)	43 (61)	71 (100)	(9)
7	1 (11)	8 (89)	9 (100)	(1)
Total	363 (48)	399 (52)	762 (100)	(100)

[a] Percentage of alive/dead status within score.
[b] Overall percentage of EBMT-scores.

Association of single SNPs with overall survival

The SNPs considered in this analysis are listed in Table 6.5. We investigated the association of the SNPs with the overall survival of the patients by modelling the effects of the single SNPs, in addition to the effects of the clinical EBMT-score on overall survival.

We fitted Cox regression models with the two factors, where the clinical EBMT-score was included as a discrete variable with values [0-7], and the SNP was included as a factor assuming a specific genetic model: additive, dominant, or recessive. The effects of the SNPs were evaluated by estimating hazard ratios from the Cox regression models. The significance of the effects of the SNPs on overall survival was judged with the log-likelihood ratio test (LRT) at a 0.05 level. We used the LRT of comparing the covariate model, that contained the EBMT-score and the SNP, with respect to a reference model with only the EBMT-score factor.

We also evaluated the gain in prediction on overall survival provided by the SNPs. We used the *partial* $\hat{R}^2_{Pred,.632}$ estimator (see section 4.6.4) to estimate the partial gain in prediction with a model including a single SNP, with respect to a model with only the clinical EBMT-score (the reference model).

We also investigated which of the three assumed genetic models for each SNP contributed the most to the gain in prediction of overall survival. We compared and selected the genetic model producing the highest *partial* $\hat{R}^2_{Pred,.632}$ estimates among the three estimates.

The associations of microsatellites or haplotypes with the overall survival of the patients were studied in a similar fashion as with the SNPs. The IL10 haplotype was recoded into binary variables coding for the presence/absence of specific haplotypes, e.g. to define patients with IL10 haplotype ACC, we created a binary variable coding for the absence/presence of haplotype ACC in patients (i.e. the presence of at least one haplotype ACC was coded as 1, and 0 otherwise), and also, another binary

variable coding for the absence/presence of haplotypes ACC/ACC in patients (i.e. the presence of the haplotype pair ACC/ACC was coded as 1, and 0 otherwise).

The analyses were performed by taking the respective non-missing genotype data available in each case. The respective sample sizes are shown in the column sample size of Table 6.5.

Association of multiple SNPs with overall survival

In addition to finding relevant single SNPs for overall survival, we aim to identify a set of SNPs that might jointly affect the overall survival of the patients following HSCT, in addition to the established clinical EBMT scores. For this analysis we considered the SNPs listed in Tables 6.5 and 6.6.

We used a stepwise procedure to build a multiple Cox regression model for the association of SNPs on overall survival. We used the Wald statistic test for variable removal (p-value=0.1), and the score statistic for variable entry (p-value=0.05). The chosen tests and p-values are in line with the statistical procedures for model building recommended by the EBMT organization as a mean to unify statistical analyses in the context of the EBMT studies (Labopin and Iacobelli 2003).

The stepwise procedure was set with an initial model containing the clinical EBMT-score as a discrete variable with values [0-7]. The SNPs listed in Tables 6.5 and 6.6 (SNPs, microsatellites, and haplotypes) were treated as candidate variables to enter the model, except the SNPs with the highest missing genotypes. Each biallelic SNP was recoded into three different variables, each variable represented a particular genetic model: additive, dominant, or recessive. The three genetic models of a SNP were candidate variables to enter into the model, but only one of them entered the model. Hence, a SNP entered the model with its most significant genetic model.

Candidate SNPs with less than 5 death events in any of the genotype groups were excluded to avoid estimation problems in the Cox model.

The fitting and estimation procedure of the Cox regression method implemented in statistical software packages, such as SAS and R, do consider only complete data on all the candidate variables considered to enter the Cox model. Given that in our data there are missing genotypes for different SNPs and in different set of patients, running the automatic stepwise procedure would drastically reduce the data from the beginning of the procedure. In an attempt to do this, we ended up with only 10% of the data to perform the analysis. That would be useless to get any reliable result.

To avoid the loss of data due to missing genotypes, we implemented a strategy for the selection procedure. At each step of the procedure, we used complete data (i.e. non-missing data) from the set of variables in the model at that step. That implicated to perform stepwise Cox model by using non-missing data from a set of fewer variables, and therefore larger data were available than from considering the whole set of candidate variables. The data were still reduced, but it occurred gradually at

subsequent steps as new variables entered the model. Hence, we profited of building the model with larger data from the beginning of the procedure in comparison to the automatic stepwise procedure.

It is important to remark that the strategy above assumes random missing data. If missingness occurred at random, the information contained in the reduced data will be similar to the information from the original data. By running our implemented strategy we assumed that the information obtained at one step of the procedure was preserved at the subsequent steps.

To get a final model we stopped the stepwise procedure when either no more SNPs could accomplish the entering and removal criteria, or when the inclusion of an additional SNP reduced the size of the data to less than 50% of the initial size.

Finally, the gain in prediction of overall survival with the final model was evaluated by the $\hat{R}^2_{Pred,.632}$ and *partial* $\hat{R}^2_{Pred,.632}$ estimators.

6.4 Results

6.4.1 Effect of clinical EBMT factors on overall survival

Table 6.8 shows the Cox regression model with the effect of the categorical clinical EBMT-score on overall survival. The survival probabilities of groups of patients per EBMT-score are shown with the Kaplan-Meier curves in Figure 6.2. In general, patients with increasing EBMT risk scores showed an increase of the hazards of overall survival. Patients with scores 2 and 3 performed pretty similar with a hazard ratio of approximately 1.2 with respect to the reference group 0-1. Patients with score 5 performed slightly worse than patients with score 6, the Kaplan-Meier curves showed that this occurred mainly during the first two years after transplantation, at later times they have similar performance. The estimated gain in prediction of overall survival with respect to a null model was $\hat{R}^2_{Pred,.632}$=3.1%.

On the other hand, the hazard ratio of the effect of a discrete EBMT-score on overall survival indicated that the hazard increased on average a fraction 0.16 for each unit increment on the EBMT risk score of a patient. This effect was significant according to the Wald-test (HR=1.16, 95% CI=1.09-1.24, p-value<0.001). Moreover, the estimated gain in prediction of overall survival with this EBMT-score, with respect to a null model, was $\hat{R}^2_{Pred,.632}$=5.1%.

6.4.2 Association of single SNPs with overall survival

We measured the association of single SNPs and haplotypes with overall survival by estimating the hazard ratio and significance from the respective Cox regression models. We found that the presence of IL10 haplotype ACC/ACC in donors had

Table 6.8: Cox regression model for overall survival with the EBMT-score[a] (n=762)

EBMT-score	Coefficient	SE (coef)[b]	p-value[c]	Hazard ratio	Confidence interval (95%)
As categorical variable[d]					
2	0.17	0.20	0.407	1.18	0.79 - 1.77
3	0.18	0.20	0.375	1.20	0.81 - 1.77
4	0.38	0.19	0.044	1.46	1.01 - 2.12
5	0.68	0.20	<0.001	1.97	1.34 - 2.90
6	0.62	0.22	0.005	1.85	1.20 - 2.87
7	1.19	0.39	0.002	3.27	1.53 - 7.01
As discrete variable					
EBMT-score	0.15	0.03	<0.001	1.16	1.09 - 1.24

[a] Gain in prediction of overall survival: $\hat{R}^2_{Pred,.632}$= 3.1% and 5.1% for the categorical and discrete variable, respectively. [b] Standard error of coefficient. [c] P-value of the Wald-test, significance of the hazard ratio of a score group of patients with respect to the reference group. [d] Reference group: 0-1.

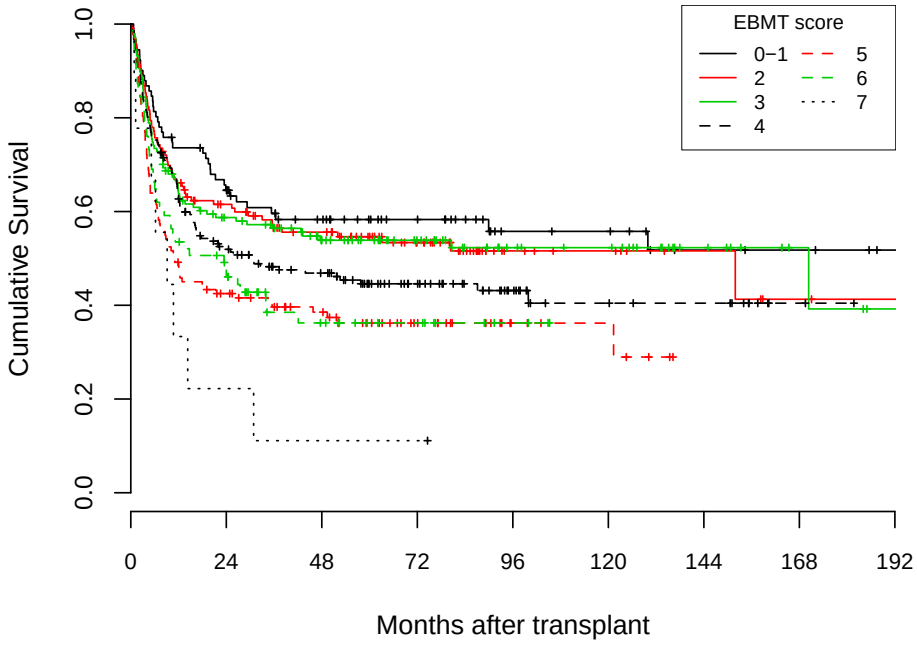

Figure 6.2: Kaplan-Meier survival curves by EBMT-score of patients after HSCT

Table 6.9: The top ten SNPs/haplotypes associated to overall survival (*ordered from the lowest to the highest p-value*)

Gene names[a]	SNP rs number[b]	p-value[c]	Hazard ratio	Confidence interval (95%)	$\hat{R}^2_{Pred,.632}$[d] *partial*, full
D-IL10	haplotype(ACC/ACC)	0.002	0.48	0.29 - 0.80	2.5, 7.6
P-MAL	rs8177374(T) a	0.021	1.33	1.05 - 1.68	1.0, 9.8
D-IL6	rs1800795(C) a	0.021	1.19	1.03 - 1.38	0.7, 4.7
P-ESR1	rs2234693(C) d	0.027	1.31	1.03 - 1.67	1.4, 6.3
D-IL13	rs1800925(T) r	0.032	0.46	0.20 - 1.03	3.3, 8.6
P-GCR	rs33388(T) a	0.035	0.83	0.70 - 0.99	0.3, 7.0
D-IL4	rs2243250(T) r	0.052	0.49	0.22 - 1.10	0.0, 4.4
D-IL6	rs1800797(A) a	0.056	1.18	1.00 - 1.39	0.1, 5.8
P-IL10	rs1800872(A) r	0.057	0.63	0.38 - 1.05	1.0, 6.2
D-IL10	rs1800896(G) a	0.063	1.16	0.99 - 1.35	0.0, 4.8

[a] The prefixes P-/D- before the gene names denote the genes in Patients/Donors.
[b] In parentheses is the minor allele of the respective SNP. The extensions a, d, or r denote the additive, dominant or recessive genetic model, respectively.
[c] p-value for the log-likelihood ratio test that compares a model with the additional gene factor with respect to a model with only the EBMT-score factor.
[d] Partial gain in prediction with a SNP/haplotype, with respect to the EBMT-score, and full gain in prediction due to the joint model of both predictors (SNP/haplotype and EBMT-score).

the highest association to overall survival of the patients after HSCT. The presence of IL10 haplotype ACC/ACC in donors tended to be protective for the patients (HR=0.48, 95% CI=0.29-0.80, p-value=0.002). The consideration of this haplotype improved the prediction of overall survival, *partial* $\hat{R}^2_{Pred,.632}$= 2.5%. Haplotype for IL10 in donors has also been previously identified as risk factor for overall survival in patients with CML, in addition to the EBMT-score (Dickinson et al. 2010), although that study reported the ATA/ACC as the protective haplotype. In other studies, the presence of IL10 haplotype ACC in patients was associated with severe acute GvHD III-IV (Dickinson et al. 2004).

Table 6.9 contains the top ten SNPs or haplotypes with the highest association according to the p-values of the LRT, with reference to the EBMT-score. These SNPs had p-values smaller than 0.063. The Table also contains the respective partial gain in prediction (*partial* $\hat{R}^2_{Pred,.632}$) of a SNP, with respect to the EBMT-score, as well as the full gain due to both types of predictors (EBMT-scoreand the SNPs). Our interpretation will be focused on the *partial* $\hat{R}^2_{Pred,.632}$ since it reflects the additional gain in prediction due to the new added predictors for overall survival.

We can observe that lower p-values of association did not necessarily yield higher prediction, as indicated by the *partial* $\hat{R}^2_{Pred,.632}$ values. The positive values of the *partial* $\hat{R}^2_{Pred,.632}$ indicate that these SNPs/haplotypes added to the prediction of overall survival in addition to the EBMT-score. The zero and negative values indicate that the SNP did not add to the gain in prediction of overall survival. The

Table 6.10: The top ten SNPs/haplotypes contributing to gain in prediction of overall survival (*ordered from highest to lowest partial* $\hat{R}^2_{Pred,.632}$)

Gene names[a]	SNP rs number[b]	$\hat{R}^2_{Pred,.632}$ [d] *partial*, full		p-value[c]	Hazard ratio	Confidence interval (95%)
D-IL13	rs1800925(T) r	3.3,	8.6	0.032	0.46	0.20 - 1.03
D-IL10	haplotype(ACC/ACC)	2.5,	7.6	0.002	0.48	0.29 - 0.80
D-IL13	rs1881457(C) r	1.4,	6.0	0.147	0.62	0.31 - 1.25
P-ESR1	rs2234693(C) d	1.4,	6.3	0.027	1.31	1.03 - 1.67
P-MAL	rs8177374(T) a	1.0,	9.8	0.021	1.33	1.05 - 1.68
P-IL10	rs1800872(A) r	1.0,	6.2	0.057	0.63	0.38 - 1.05
D-IL6	rs1800795(C) a	0.7,	4.7	0.021	1.19	1.03 - 1.38
P-IFNG	genotype (3/3)	0.6,	7.4	0.260	0.86	0.67 - 1.12
P-IL10	haplotype(ATA/ACC)	0.5,	5.9	0.277	1.19	0.88 - 1.61
D-GCR	rs33388(T) a	0.5,	6.7	0.415	1.07	0.91 - 1.26

[a] The prefixes P-/D- before the gene names denote the genes in Patients/Donors.
[b] In parentheses is the minor allele of the respective SNP. The extensions a, d, or r denote the additive, dominant or recessive genetic model, respectively.
[c] p-value for the log-likelihood ratio test that compares a model with the additional gene factor with respect to a model with only the EBMT-score factor.
[d] Partial gain in prediction with a SNP/haplotype, with respect to the EBMT-score, and full gain in prediction due to the joint model of both predictors (SNP/haplotype and EBMT-score).

SNPs listed in Table 6.9 added to the prediction of overall survival, except the IL4 rs2243250(T) in donors.

As a different viewpoint we also listed the top ten SNPs/haplotypes with the highest gain of prediction according to the *partial* $\hat{R}^2_{Pred,.632}$ (Table 6.10). The SNP with the highest gain in prediction of overall survival was the rs1800925(T) from the IL13 gene in donors (*partial* $\hat{R}^2_{Pred,.632}$=3.3%). This SNP also appeared in Table 6.9 as one of the top ten SNPs with the highest associations with overall survival (HR=0.46, 95% CI=0.20-1.03, p-value=0.032) .

Other SNPs/haplotypes with both high association (p-value) and high prediction (*partial* $\hat{R}^2_{Pred,.632}$) to overall survival were: IL10 haplotype(ACC/ACC) in donors, ESR1 rs2234693(C) in patients, MAL rs8177374(T) in patients, IL10 rs1800872(A) in patients, and IL6 rs1800795(C) in donors.

We should take note that, for studies where pre-selection of SNPs are performed, the decision concerning which measure to use for evaluation should be based on the aim of the study. For prediction purposes, it can also be helpful to consider a measure of prediction such as the $\hat{R}^2_{Pred,.632}$ and *partial* $\hat{R}^2_{Pred,.632}$ estimators in the selection of the SNPs.

Figure 6.3 compares the partial gain in prediction (*partial* $\hat{R}^2_{Pred,.632}$) attributed to each biallelic SNP under three possible genetic models: additive, dominant, and recessive, with respect to the EBMT-score. We can see that only a few SNPs added to the gain in prediction of overall survival in addition to the EBMT-score.

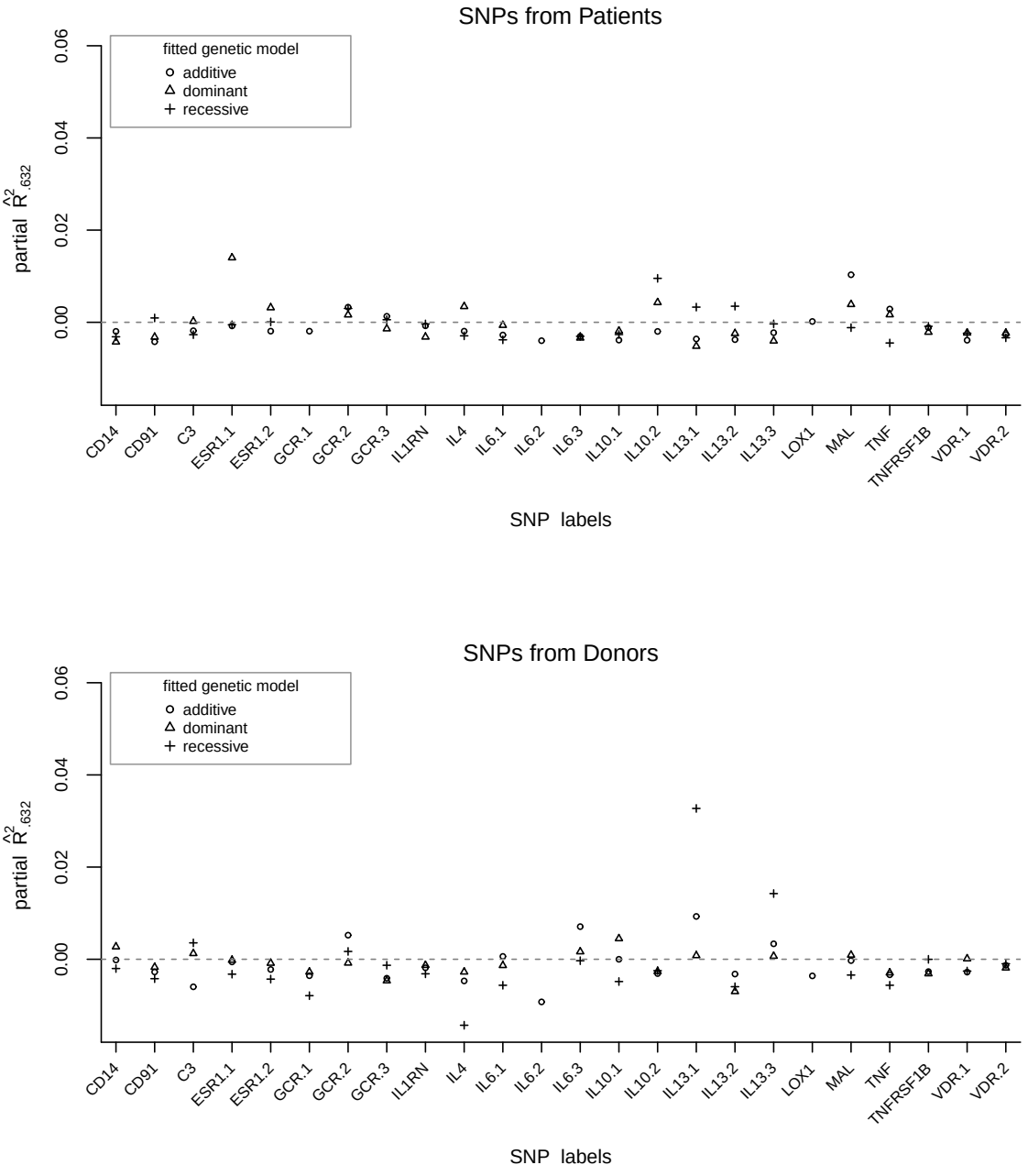

Figure 6.3: Estimate of the *partial* $\hat{R}^2_{Pred,.632}$ of single biallelic SNPs[a] under an additive, dominant or recessive genetic model[b]

[a] For details of the respective SNPs, see Table 6.4.

[b] Genetic model(s) for some SNPs could not be fitted because of small number of patients in at least one of the genotype groups.

We also observe that in most cases, the *partial* $\hat{R}^2_{Pred,.632}$ under the three genetic models of a SNP were similar, and in addition, most of these SNPs showed little or did not show improvement to the prediction, i.e. *partial* $\hat{R}^2_{Pred,.632} \approx 0$, e.g. the SNP rs33388(T) in patients (label GCR.2) showed an improvement of *partial* $\hat{R}^2_{Pred,.632} \approx 0.3\%$. However, even if the SNP had improved the prediction of overall survival, the particular selection of a genetic model is irrelevant in terms of prediction, since the three models performed similarly. There can be differences in terms of the effect on the outcome. For example, the SNP rs33388(T) in patients showed a significant evidence effect only under the additive genetic model (additive: HR=0.83, 95% CI=0.70-0.98, p-value=0.035; dominant: HR=0.80, 95% CI=0.62-1.02, p-value=0.073; recessive: HR=0.77, 95% CI=0.56-1.06, p-value=0.77).

By considering SNPs whose minimum gain in prediction was 0.5%, we identified 7 SNPs, which are also listed in Table 6.10. Three of the SNPs showed the highest gain in prediction under the additive model (SNP rs8177374(T) in MAL gene in patients, SNP rs33388(T) in GCR gene in donors, and SNP rs1800795(C) in IL6 gene in donors), one under the dominant model (SNP rs2234693(C) in ESR1 gene in patients), and three under the recessive model (SNPs rs1800925(T) and rs1881457(C) in IL13 gene in donors, and SNP rs1800872(A) in IL10 gene in patients.

From those contributing SNPs, the most relevant ones (*partial* $\hat{R}^2_{Pred,.632}>1\%$) are the SNP rs8177374(T) in MAL gene, and SNP rs2234693(C) in ESR1 gene in patients, as well as the SNPs rs1800925(T) and rs1881457(C) in IL13 gene in donors.

The single IL10 SNPs in donors, rs1800896(G) and rs1800872(A), added very little or did not add to the prediction of overall survival (*partial* $\hat{R}^2_{Pred,.632}$ was 0.4% and <0%, respectively). However, the IL10 haplotype (ACC/ACC) in the same donors did improve the prediction to overall survival (*partial* $\hat{R}^2_{Pred,.632}=2.5\%$).

In some cases, a SNP from patients provided higher gain in prediction in comparison to the same SNP from donors. Within this list are the SNPs from ESR1, IL4, and MAL gene. In other cases, a SNP from donors provided higher gain in prediction in comparison to the same SNP from patients, these are the SNP rs1800795(C) in IL6 gene, SNPs rs1800925(T) and rs1881457(C) in IL13 gene.

From all these views, we conclude that specification of an appropriate genetic model is important to evaluate the gain in prediction with some SNPs, while it may not be relevant with some others. The consideration of SNPs from both patients and donors seems to be important -besides the clinical EBMT factors- since they contributed differently to the prediction of overall survival of the patients.

From the point of view of both the impact of the effect of the SNP and the gain in prediction of overall survival, the SNP rs1800925 in IL13 gene, and the haplotype (ACC/ACC) in IL10 from donors, as well as the SNP rs2234693(C) in ESR1 gene, and the SNP rs8177374 in MAL gene from patients, showed to be relevant single SNPs, in addition to the EBMT-score.

Table 6.11: Cox regression model for overall survival, with multiple SNPs in addition to the clinical EBMT-score (n=419)

Factors[a]		Coefficient	SE (coef)[b]	p-value	Hazard ratio	Confidence interval (95%)
EBMT-score		0.20	0.04	<0.001	1.22	1.12 - 1.33
D-IL10	haplotype(ACC/ACC)	-0.72	0.31	0.020	0.49	0.26 - 0.89
P-MAL	rs8177374(T) a	0.29	0.13	0.026	1.34	1.04 - 1.74
P-ESR1	rs9340799(G) d	0.42	0.14	0.003	1.52	1.15 - 2.01
D-IL6	rs1800795(C) a	0.25	0.09	0.007	1.29	1.07 - 1.55

[a] The prefixes P-/D- before the gene names denote the genes in Patients/Donors.
[b] Standard error of coefficient.

6.4.3 Association of multiple SNPs with overall survival

We fitted a Cox regression model with multiple SNPs associated to overall survival, in addition to the EBMT-score. The model was built by performing a stepwise procedure with a set of candidate SNPs (see Tables 6.5 and 6.6). The model was obtained based on 56% of the data (n=419). Table 6.11 shows the SNPs that entered the Cox model, and their respective estimates of the hazard ratio.

In addition to the EBMT-score, the model included SNPs from four genes: IL10 and IL6 from donors, MAL and ESR1 from patients.

The presence of haplotype ACC/ACC of gene IL10 in donors was protective for death following HSCT (HR=0.49, 95% CI=0.26-0.89, p-value=0.020). A similar finding was reported in a multiple SNP study in CML patients (Dickinson et al. 2010), although they reported the ATA/ACC haplotype pair as protective for death. In other studies, the simple presence of haplotype ACC in gene IL10 in patients was associated with severe acute GvHD III-IV (Dickinson et al. 2004).

Also, the risk of death increased by the presence of each additional allele T carried by the patient in SNP rs8177374 of gene MAL (HR=1.34, 95% CI=1.04-1.74, p-value=0.026). The presence of allele G in SNP rs9340799 of gene ESR1 in patients increased the risk of death of the patients (HR=1.52, 95% CI=1.15-2.01, p-value=0.003). This SNP has previously been reported in association with survival and aGvHD in patients with HLA-matched siblings (Middleton et al. 2003). In our analysis of single SNPs, this SNP was ranked at the 11 and 17 position according to the ordering of p-values (0.077) and prediction with the *partial* $\hat{R}^2_{Pred,.632}$ (0.3%), respectively.

Moreover, the risk of death increased by the presence of each additional allele C carried by the donor in the SNP rs1800795 of gene IL6 (HR=1.29, 95% CI=1.07-1.55, p-value=0.007). The complement allele G of this SNP has been associated with an increment in the risk of acute and chronic GvHD in patients with HLA-matched siblings (Cavet et al. 2001), which means that the presence of allele C would be protective for GvHD. In our results we found that it can also represent an increased risk for post-transplant mortality in overall HSCT patients.

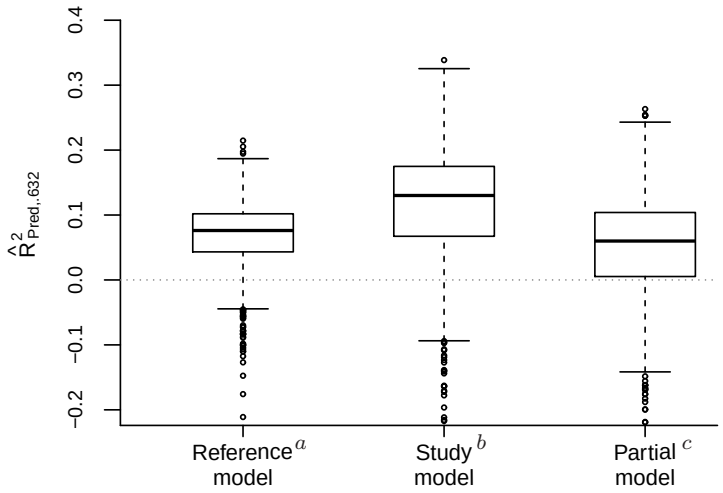

Figure 6.4: Boxplots of distribution of estimated gain in prediction[d] provided by three models for overall survival

[a] The *reference model* is the clinical model containing only the EBMT-score.

[b] The *study model* is the joint multiple model of SNPs/haplotype and EBMT-score (model in Table 6.11).

[c] The *partial model* refers to the Study model with respect to the Reference model (i.e. partial model of SNPs/haplotype with respect to the EBMT-score).

[d] Mean gain in prediction was $\hat{R}^2_{Pred,.632}$= 6.8% (Reference model), 11.6% (Study model), and *partial* $\hat{R}^2_{Pred,.632}$= 5.1% (gain due to SNPs/haplotype, Partial model). All the estimates were obtained from B=1000 repeated bootstrap samples drawn from available data of n=419 patients.

The $\hat{R}^2_{Pred,.632}$ estimator indicated that the gain in prediction of overall survival with this multiple model (EBMT-score and SNPs), with respect to a null model, was on average 11.6%. This estimate points out an improvement in prediction due to the SNPs in the model, in comparison to a model with only clinical EBMT-score, which in turn provided on average 6.8% of gain in prediction. The *partial* $\hat{R}^2_{Pred,.632}$ of the SNPs with respect to the EBMT-scorewas 5.1%, which reveals the additional gain in prediction due to the SNPs. Figure 6.4 shows the distribution of the $\hat{R}^2_{Pred,.632}$ estimates. All the estimates were computed based on n=419 available patients used in the identification of the multiple model.

Thus, the results showed that, single or multiple SNPs contributed to the improvement of prediction of overall survival, in addition to the EBMT-score alone, although the improvement became more important when multiple SNPs were modelled in association with overall survival.

Therefore, besides the clinical EBMT-score, the consideration of genetic factors previous to the transplant seems to be worth to evaluate and prevent the risk of death of a patient after haematopoietic stem cell transplantation.

Table 6.12: Risk scores per category from the joint clinical and genetic model for overall survival

Factors	Categories	Risk score
EBMT-score		$[0-7]^a$
D-IL10 haplotype	ACC/ACC	0
	other than ACC/ACC	4
D-IL6 rs1800795	GG	0
	GC	1
	CC	2
P-ESR1 rs9340799	AA	0
	GA or GG	2
P-MAL rs8177374	CC	0
	TC	1
	TT	2

[a] It is the clinical EBMT risk score, which is a discrete variable obtained from clinical scores, see Table 6.1.

6.4.4 Joint risk score based on clinical and genetic variables for overall survival

We constructed risk scores with the joint influence of clinical and genetic variables on overall survival (Table 6.12). The score of each variable was obtained from the coefficient of the Cox regression model, multiplied by an integer 5, and then rounding them to the closest integer. The integer 5 was chosen to keep the original scoring [0-7] for the clinical EBMT-score, as originally established by its authors (Table 6.1).

The scores were always assigned as positive values. The reference score value was always assigned as 0. Table 6.12 contains the risk score per category of genotype of SNPs/haplotype identified for a joint model.

The IL10 haplotype had the highest score among all the factors. The presence of a haplotype different from ACC/ACC in donors contributed with score 4 to the joint risk score of each patient.

A joint risk score for overall survival of each patient was obtained by summing up each individual risk score for the categories to which a patient pertained. This joint risk score ranged from 0 to 15 (Table 6.13). These scores were grouped into five distinguished categories (Table 6.14) according to the natural grouping by the Kaplan-Meier survival curves (Figure 6.5). These score categories can be viewed as a classification of patients into different risk levels of death, from low risk (scores 1-6) to high risk (scores 13-15).

We fitted a Cox model with the joint score factor with five categories (Table 6.15). We found that the score categories were significantly different, with increasing hazard ratios for higher risk score levels.

Table 6.13: Frequency of death per risk score from the joint clinical and genetic model

Joint risk score[a]	alive	(%[b])	dead	(%[b])	Total	(%[b])	(%[c])
1	1	(100)	0	(0)	1	(100)	(0)
3	2	(100)	0	(0)	2	(100)	(0)
4	6	(67)	3	(33)	9	(100)	(2)
5	5	(71)	2	(29)	7	(100)	(2)
6	17	(81)	4	(19)	21	(100)	(5)
7	27	(59)	19	(41)	46	(100)	(11)
8	24	(52)	22	(48)	46	(100)	(11)
9	32	(55)	26	(45)	58	(100)	(14)
10	33	(42)	45	(58)	78	(100)	(19)
11	23	(32)	48	(68)	71	(100)	(17)
12	15	(35)	28	(65)	43	(100)	(10)
13	3	(13)	20	(87)	23	(100)	(6)
14	4	(31)	9	(69)	13	(100)	(3)
15	0	(0)	1	(100)	1	(100)	(0)
Total	192	(46)	227	(54)	419	(100)	(100)

[a] The joint clinical and genetic risk score according to Table 6.12.
[b] Percentage of alive/dead status within score.
[c] Overall percentage of joint risk scores.

Table 6.14: Frequency of death per group of risk scores from the joint clinical and genetic model

Joint risk score[a]	alive	(%[b])	dead	(%[b])	Total	(%[b])	(%[c])
1-6	31	(78)	9	(22)	40	(100)	(9)
7-9	83	(55)	67	(45)	150	(100)	(36)
10	33	(42)	45	(58)	78	(100)	(19)
11-12	38	(33)	76	(67)	114	(100)	(27)
13-15	7	(19)	30	(81)	37	(100)	(9)
Total	192	(46)	227	(54)	419	(100)	(100)

[a] The joint clinical and genetic risk score according to Table 6.12.
[b] Percentage of alive/dead status within score.
[c] Overall percentage of joint risk scores.

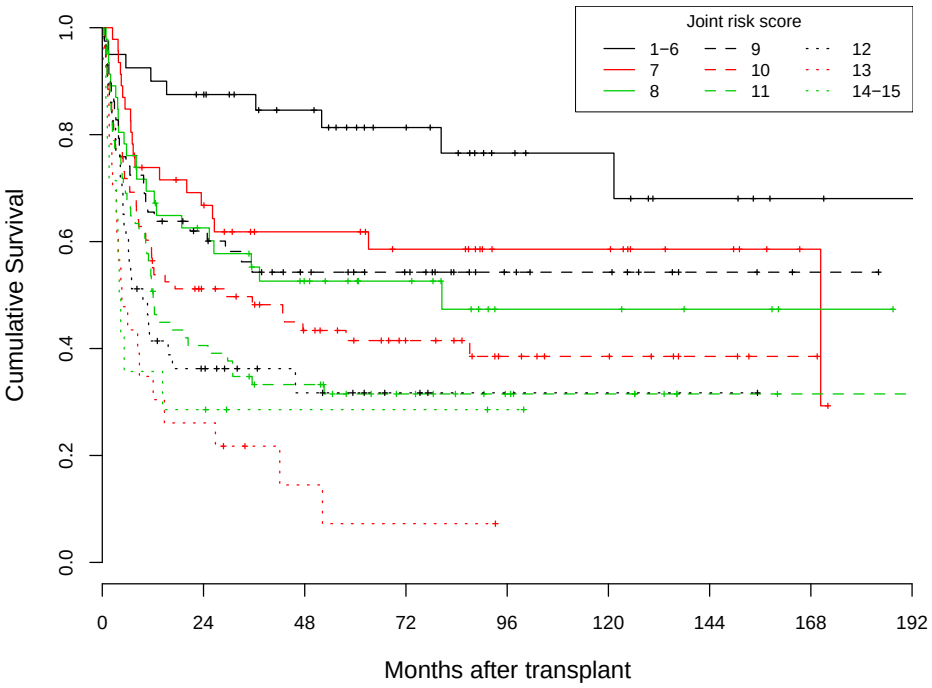

Figure 6.5: Kaplan-Meier survival curves by joint clinical and genetic scores of patients undergoing HSCT

The Kaplan-Meier survival curves (Figure 6.6) show well separated curves of these categories. The median survival time was 4 months for patients with joint risk score 13-15 (the highest risk group), 11 months for patients with joint risk score 11-12, and 2.5 years for patients with joint risk score 10. Patients with joint risk score 7-9 survived for a long period, the median survival time of this group was 14 years. More than 50% of patients with joint risk score 1-6 (the lowest risk group) were alive by the end of the study (Figure 6.6).

The estimated gain in prediction with the new scoring with five categories was 16.9%, which is higher than with the original factors (11.6%). This higher prediction can be explained by the fewer number of parameters needed to predict survival. Moreover,

Table 6.15: Cox regression model for overall survival, with risk score groups from the joint clinical and genetic model[a] (n=419)

Joint risk score[b]	Coefficient	SE (coef)[c]	p-value	Hazard ratio	Confidence interval (95%)
7-9	0.90	0.36	0.011	2.46	1.23 - 4.93
10	1.29	0.37	<0.001	3.62	1.77 - 7.42
11-12	1.54	0.35	<0.001	4.67	2.33 - 9.33
13-15	2.02	0.38	<0.001	7.57	3.58 - 16.00

[a] Estimated gain in prediction of overall survival: $\hat{R}^2_{Pred,.632}$=16.9%.

[b] Reference group: risk score 1-6. [c] Standard error of coefficient.

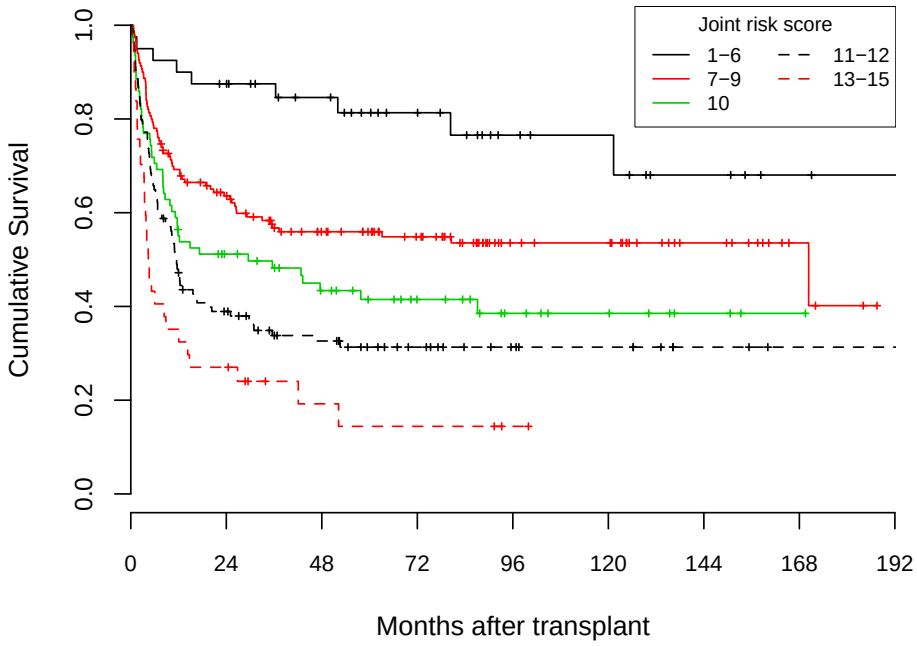

Figure 6.6: Kaplan-Meier survival curves combined into five groups of the joint clinical and genetic scores of patients undergoing HSCT

by assuming the new score as a discrete variable, only one parameter would be involved in the prediction of overall survival. The gain in prediction with the latter variable was estimated in 15.3% (Figure 6.7).

Moreover, we evaluated the new score on subgroup of patients in our data. Subgroups of patients were taken according to disease diagnosis, donor type, T-cell depletion, and conditioning regimen (details of the specific groups are in Table 6.3). The new score improved the prediction of overall survival in all the subgroups (data not shown), except on acute leukaemia patients, whose prediction was slightly smaller in comparison to the clinical EBMT-score.

We believe that the consideration of a risk score scheme including both clinical and genetic variables can contribute to the identification of risk group of patients undergoing HSCT. Our findings indicate that genes IL10, MAL, ESR1 and IL6 are involved in the reduced survival of HSCT patients, in addition to the clinical EBMT-score.

Classifying patients into few risk levels, here five, according to their clinical and genetic profiles might aid the prevention of failures in patients with high risk of death. Clinicians can consider these risk scores to identify and evaluate the chances of an individual patient to respond successfully to the transplantation. Hence, patients with high risk of failures (risk scores > 11) might have the possibility to receive medical treatment, other than HSCT, to improve their chances of survival.

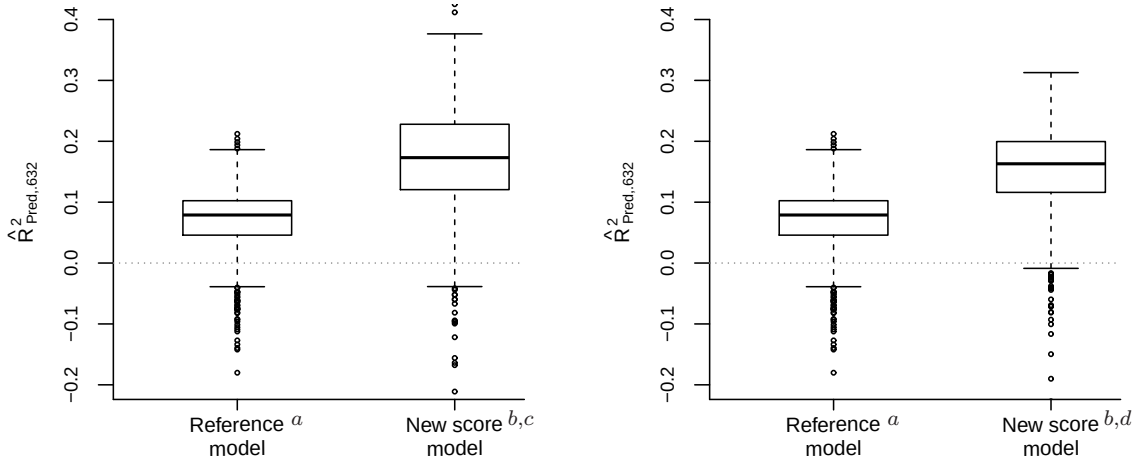

Figure 6.7: Boxplots of distribution of estimated gain in prediction[e] with a new score model for overall survival

[a] The *reference model* is the clinical model containing the EBMT-score.

[b] The *new score model* is the clinical-genetic model summarized with a joint risk score with five groups.

[c] The risk score modelled as a categorical variable (model in Table 6.15).

[d] The risk score modelled as a discrete variable taking on values [1-5] (1=lowest, 5=highest risk score group).

[e] Mean gain in prediction was $\hat{R}^2_{Pred,.632}$= 6.8% (Reference model), 16.9% (New score model, categorical), and 15.3% (New score model, discrete). All the estimates were obtained from B=1000 repeated bootstrap samples drawn from available data of n=419 patients.

Chapter 7

Summary and discussion

In the last years, genetic studies have been focused on the identification of genetic variants, specially SNPs, that can be associated with the development of disease. Fitting statistical models based on a set of data aid answering these research questions. However, recently the objective of some studies is not only to find models for the association of SNPs with the outcome, but also to make prediction of the outcome based on the fitted model. In clinical studies, attempts are done to build genetic risk scores to predict occurrence of disease. One way to measure the prediction capability of the model is through the coefficient of determination, R^2, which is obtained with the mean squared residuals of the model with respect to that of a null or reference model.

When we evaluate the goodness of fit of the model, we measure the residuals, i.e. the differences between the observed outcome in the data and its estimate given the model. When we evaluate the capability of prediction of the model, we measure the prediction errors, which are based on the differences between the observed outcome in a validation set and its estimate given the model. A common problem in genetic studies is the unavailability of a validation set, either because the trait is not common and collecting new data will take long time or because of the expenses to collect new data. Hence, there are techniques such as the 0.632 estimator that facilitate the evaluation of a model for prediction without requiring new data.

In this thesis we made use of the 0.632 and the 0.632+ estimator in combination with the criterion of the Schoenfeld residuals to provide a new approach for the evaluation of prediction capability of survival models. The Schoenfeld residuals measure the difference between the observed value of a covariate in the data, which is a predictor, and its expectation under the model. In our approach we adapted the Schoenfeld residuals to the concept of prediction errors. Hence, we measured the difference between the observed value of a covariate in a validation set and its expectation under the model.

The 0.632 estimator is a linear combination of the apparent error and the bootstrap cross-validation, using weight 0.632 for the bootstrap cross-validation. The apparent

error is the measure of prediction error to evaluate the goodness of fit of the model, while the bootstrap cross-validation estimator measures the prediction error to evaluate prediction of the model, which is done by repeatedly splitting the available data into training sample and validation set. A particular feature of our approach is that the prediction errors can be estimated only at times of observed events, and therefore, they are estimated only for individuals with observed events. The latter is the main difference of our approach with the approach based on the criterion of the Brier score (Gerds and Schumacher 2007), which estimates prediction errors based on the difference of survival status and estimated survival probability, under the model, for all available individuals.

We formulated an estimator of R^2 for prediction based on estimates of prediction errors. We denoted it by $R^2_{Pred},$. The R^2_{Pred} gives the gain in prediction due to predictors considered in the model. Hence, we compared through simulation studies the performance of the R^2_{Pred} estimators with two criteria, one using the Schoenfeld residuals (our approach) and the second using the Brier score. We carried out the comparisons by estimating the prediction capability of single SNPs. We simulated SNPs with different frequencies and effect sizes for the risk allele, that should be in accordance with real scenarios of genetic studies.

The results revealed that the R^2_{Pred} estimates with the criterion of the Schoenfeld residuals is higher than estimates with the criterion of the Brier score. This is in agreement with the findings of Müller et al. (2008), who found that the R^2 estimator with the Schoenfeld residuals for goodness of fit gives higher estimates than the Brier score. If the effect size of the predictor is large, the ratio of prediction error given by the model in comparison to the null model is smaller when the errors are measured with the criterion of the Schoenfeld residuals than with the Brier score. This advantage of the Schoenfeld residuals can be explained by the fact that, if the predictor is correctly identified, failing individuals in a validation set will be more prone to belong to the risk category(ies) of the predictor. In addition, if the model contains the correct predictor, the estimate of the covariate value will tend to the risk category value of the predictor. Hence, the prediction error of the covariate value will tend to decrease as the effect size of the predictor increases, and consequently, R^2_{Pred} increases as effect size increases.

On the other hand, since the survival status of every individual in the validation set is known, the estimator of prediction error with the Brier score averages the errors of failing and not failing individuals. Hence, if we consider only the risk group, the prediction error from the failing individuals might be leveled out by the errors from the not failing individuals, and therefore the overall mean prediction error does not get much differentiated from the null model, unless the effect of the predictor is very high.

Even if there are differences in the R^2_{Pred} estimates between both criteria, they describe the same desired behaviour pattern. That is, the R^2_{Pred} estimates increase with higher effect size (higher hazard ratio) and higher MAF. Hence, we can say that both criteria provide similar information about the gain in prediction but the

Schoenfeld residuals allow us to differentiate more clearly the contribution of the predictors to the outcome.

Another observation is that the $\hat{R}^2_{Pred,app}$, which is used for evaluating the goodness of fit of the model, underestimates the original estimate of prediction, $\hat{R}^2_{Pred}$. This was expected since it is known that evaluation of a model on the same data used for model fitting gives an overoptimistic view of the real predictive performance of a model. The $\hat{R}^2_{Pred,.632}$ estimator shows the best performance as estimator of the original $\hat{R}^2_{Pred}$ estimate. The $\hat{R}^2_{Pred,.632}$ closely estimates the $\hat{R}^2_{Pred}$ values in all the simulated scenarios. The $\hat{R}^2_{Pred,.632^+}$ estimates well the $\hat{R}^2_{Pred}$ values in the scenarios with moderate effect sizes but not for small effects in the case of the criterion of the Schoenfeld residuals, and it underestimates the $\hat{R}^2_{Pred}$ values in all the scenarios in the case of the criterion of the Brier score. Hence, we conclude that $\hat{R}^2_{Pred,.632}$ *is the best estimator to use for evaluation of gain in prediction, and moreover, estimating it with the criterion of the Schoenfeld residuals will give us a better view of the effect of the predictor on the outcome.*

We should remark, however, that our conclusions hold under the conditions of our simulations. We have simulated times to event with an exponential distribution, that assumes a constant rate of occurrence of events over time. Thus, our results might change if we vary the condition to a distribution with changing rate of events. In addition, motivated by our application study on HSCT, we carried out the simulation study for 60% of censoring. Only minor changes were observed when we performed the study for 40% of censoring (data not shown). Our final results and conclusions about the patterns by effect sizes and MAFs, as well as the comparison between criteria and estimators of $\hat{R}^2_{Pred}$ remained similar as with a 60% of censoring.

We did not evaluate the approaches under varying sample sizes. We used a sample size of n=1000 that can be accessible for a moderate clinical study. However, this issue may need further investigation.

As mentioned above, one important issue we deduce from our study is that the 0.632 estimator performs very well in estimating the original $\hat{R}^2_{Pred}$. Thus, it can be used for the study of prediction capability of a Cox model without going for a more complex estimator as the 0.632^+ estimator. This fact is important especially for researchers from the medical community or other users of statistical methods, who require straight forward methodologies for application in their practical research.

The Schoenfeld residuals use the relation between time to event and covariate values of failing individuals to estimate the prediction capability of a model. This feature of the Schoenfeld residuals is an important advantage for studies where the main interest is the construction of a new categorical variable explaining the outcome. The new variable will be directly evaluated based on observed time to event. This is the case, for example, when we construct genetic risk scores classifying patients according to different risk levels for the outcome. By evaluating the genetic risk score, the criterion of the Schoenfeld residuals determines how well the proposed risk score represents the occurrence of events over time. This is done by taking

into account the size effect of the risk groups, so that the more differentiated the groups the most relevant the risk score. The latter goes in favour of the estimator with the Schoenfeld residuals against the C-index (Harrell et al. 1996), that is a measure commonly used in clinical practice to evaluate risk scores, among other similar measures (Ripatti et al. 2010, Paynter et al. 2010). The C-index evaluates the concordance between times of events and the rank of the covariate values, it does not consider the distance between the covariate values, i.e. it does not consider the effect size of the predictor. Hence, the introduction of the 0.632 estimator with the Schoenfeld residuals should be a good alternative to measure more appropriately the gain in prediction with a predictor variable in the Cox model.

Our proposed estimator should be further evaluated on other conditions and scenarios not covered in our simulation study. One further aspect is, for example, to evaluate prediction under competing risk outcomes, when two or more outcomes can occur during the follow-up period of a patient, where the occurrence of one outcome can mask the occurrence of its competitors.

We used our proposed estimator of prediction in a practical clinical application to judge the use of a clinical-genetic risk score to predict survival of patients undergoing haematopoietic stem cell transplantation (HSCT). The results revealed that the EBMT-score, i.e. the clinical risk score regarded as established in the HSCT community, only poorly predict survival of HSCT patients. Consideration of a joint clinical-genetic risk score clearly improved the prediction, although it is still so low that it is very far from practical use. This study demonstrated that investigations for HSCT and also in other contexts should not be limited to model fitting but it should be evaluated in terms of prediction. Clearly, further investigations are needed to find a more satisfactory model to predict survival after HSCT. Having a good model for prediction based on a few risk categories can aid clinicians to decide on the best clinical-genetic profiles a patient should accomplish to be successfully treated with HSCT.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Bender, R., Augustin, T. and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models, *Statistics in medicine* **24**: 1713–1723.

Binder, H. and Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models, *BMC Bioinformatics* **9**: 14.

Brier, G. (1950). Verification of forecasts expressed in terms of probability, *Monthly Weather Review* **78**(1): 1–3.

Calvo, J., Martínez, N., Etxagibel, A., Calleja, S., Sáez-Torres, C., Sedeño, M., Julá, R., Muncunill, J., Matamoros, N. and Gayá, A. (2002). Allelic frequencies of polymorphic variants of cytokine genes (IL1A, IL1B, IL1RN, IL6, IL10, IL12p40, and IFNG) in a Spanish population, *Inmunología* **21**(2): 76–86.

Cariaso, M. and Lennon, G. (2011). SNPedia [online].
URL: `http://www.SNPedia.com` [Accessed 28 August 2011].

Cavet, J., Dickinson, A., Norden, J., Taylor, P., Jackson, G. and Middleton, P. (2001). Interferon-γ and interleukin-6 gene polymorphisms associate with graft-versus-host disease in HLA-matched sibling bone marrow transplantation, *Blood* **98**(5): 1594–1600.

conditioning regimen (2011). In *National Cancer Institute, Dictionary of Cancer Terms* [online].
URL: `http://www.cancer.gov/dictionary?CdrID=549700` [Accessed 12 November 2011].

Conditioning Regimens (2011). In *Medscape reference* [online].
URL: `http://emedicine.medscape.com/article/991032-overview#aw2aab6b7` [Accessed 12 November 2011].

Copelan, E. A. (2006). Hematopoietic Stem-Cell Transplantation, *The new England Journal of Medicine* **354**: 1813–1826.

Cordell, H. and Clayton, D. (2005). Genetic association studies, *Lancet* **366**: 1121–1131.

Cummings, P. (2004). Re:"Estimating the relative risk in cohort studies and clinical trials of common outcomes". Letter, *American Journal of Epidemiology* **159**(2): 213–215.

Database of Single Nucleotide Polymorphisms (dbSNP) (2011). Bethesda, MD: National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID:134).
URL: `http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi` [Accessed 22 August 2011].

Deddens, J. A. and Petersen, M. R. (2004). Re:"Estimating the relative risk in cohort studies and clinical trials of common outcomes". Letter, *American Journal of Epidemiology* **159**(2): 213–215.

Dickinson, A. (2008). Non-HLA genetics and predicting outcome in HSCT, *International Journal of Immunogenetics* **35**: 375–380.

Dickinson, A., Middleton, P., Rocha, V., Gluckman, E. and Holler, E. (2004). Genetic polymorphisms predicting the outcome of bone marrow transplants, *British Journal of Haematology* **127**: 479–490.

Dickinson, A., Pearce, K., Norden, J., O'Brien, S., Holler, E., Bickeböller, H., Balavarca, Y., Rocha, V., Kolb, Hans-Jochem, H., Ilona, Hromadnikova, P., Niederwieser, D., Brand, R., Ruutu, T., Apperley, J., Szydlo, R., Goulmy, E., Siegert, W., de Witte, T. and Gratwohl, A. (2010). Impact of genomic risk factors on outcome after hematopoietic stem cell transplantation for patients with chronic myeloid leukemia, *Haematologica* **95**(6): 922–927.

EBMT (2011). European Group for Blood And Marrow Transplantation [online].
URL: `http://www.ebmt.org` [Accessed 12 August 2011].

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation, *Journal of the American Statistical Association* **78**(382): 316–331.

Efron, B. and Tibshirani, R. (1993). *An introduction to the Bootstrap*, Chapman & Hall, New York London.

Efron, B. and Tibshirani, R. (1997). Improvement on cross-validation: The .632+ bootstrap method, *Journal of the American Statistical Association* **92**(438): 548–560.

EUROBANK (2008). Marie-Curie Research Training Network. TRANSNET. Identification of genomic and biological markers as predictive/diagnostic/therapeutic tools for use in allogeneic stem cell transplantation: Translational research towards individualised patient medicine [online]. URL: `http://www.eurotransplantbank.org/DesktopDefault.aspx` [Accessed 05 September 2008].

Geisse, S. (1975). The predictive sample reuse method with applications, *Journal of the American Statistical Association* **70**(350): 320–328.

Gerds, T. (2009). *pec: Prediction Error Curves for Survival Models*. R package version 1.1.0.

Gerds, T. and Schumacher, M. (2006). Consistent estimation of the expected Brier score in general survival models with right-censored event times, *Biometrical Journal* **48**: 1029–1040.

Gerds, T. and Schumacher, M. (2007). Efron-type measures of prediction error for survival analysis, *Biometrics* **63**: 1283–1287.

Graf, E., Schmoor, C., Sauerbrei, W. and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data, *Statistics in Medicine* **18**: 2529–2545.

Graft-versus-host disease (2011). In *Wikipedia* [online]. Wikimedia Foundation, Inc. URL: `http://en.wikipedia.org/wiki/Graft_versus_host_disease` [Accessed 12 November 2011].

Gratwohl, A., Brand, R., Apperley, J., Crawley, C., Ruutu, T., Corradini, P., Carreras, E., Devergie, A., Guglielmi, C., Kolb, H.-J. and Niederwieser, D. (2006). Allogeneic hematopoietic stem cell transplantation for chronic myeloid leukemia in Europe 2006: transplant activity, long-term data and current results. An analysis by the Chronic Leukemia Working Party of the European Group for Blood And Marrow Transplantation (EBMT), *Haematologica* **91**(4): 513–521.

Gratwohl, A., Hermans, J., Goldman, J., Arcese, W., Carreras, E., Devergie, A., Frassoni, F., Gahrton, G., Kolb, H., Niederwieser, D., Ruutu, T., Vernant, J., de Witte, T. and Apperley, J. (1998). Risk assessment for patients with chronic myeloid leukaemia before allogeneic blood or marrow transplantation, *Lancet* **352**: 1087–1092.

Gratwohl, A., Stern, M., Brand, R., Apperley, J., Baldomero, H., de Witte, T., Rocha, V., Passweg, J., Sureda, A., Ticheli, A. and Niederwieser, D. (2009). Risk Score for Outcome After Allogeneic Hematopoietic Stem Cell Transplantation, *Cancer* **115**: 4715–4726.

Harrell, F., Lee, K. and Mark, D. (1996). Multivariate prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Statistics in Medicine* **15**: 361–387.

Hartl, D. and Clark, A. (1997). *Principles of Population Genetics*, Sinauer Associates Sunderland, Massachusetts.

Heliö, T., Halme, L., Lappalainen, M., Fodstad, H., Paavola-Sakki, P., Turunen, U., Färkkilä, M., Krusius, T. and Kontula, K. (2003). CARD15/NOD2 gene variants are associated with familially occurring and complicated forms of Crohns disease, *Inflammatory Bowel Disease* **52**: 558–562.

Hematopoietic Stem Cells (2001). In *Stem Cell Information* [online]. Bethesda, MD: National Institutes of Health, U.S. Department of Health and Human services. URL: `http://stemcells.nih.gov/info/2001report/chapter5` [Accessed 08 August 2011].

HLA gene family (2009). In *Genetics Home Reference* [online]. Bethesda, MD: U.S. National Library of Medicine.
URL: `http://ghr.nlm.nih.gov/geneFamily/hla` [Accessed 12 November 2011].

Jarvis, M., Marzolini, M., Wang, X. N., Jackson, G., Sviland, L. and Dickinson, A. (2003). Heat shock protein 70: correlation of expression with degree of graft-versus-host response and clinical graft-versus-host disease, *Transplantation* **76**(5): 849–853.

Jewell, N. (2004). *Statistics for Epidemiology*, Chapman & Hall/CRC, Boca Raton.

Kaplan, E. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations, *Journal of the American Statistical Association* **53**(282): 457–481.

Kleinbaum, D. (1996). *Survival Analysis*, Springer, New York.

Labopin, M. and Iacobelli, S. (2003). Statistical guidelines for EBMT [online], *Guidelines*, The European Group for Blood And Marrow Transplantation.
URL: `http://www.ebmt.org/1WhatisEBMT/Op_Manual/OPMAN_StatGuidelines_oct2003.pdf` [Accessed 12 August 2011].

Lachin, J. (2000). *Biostatistical Methods: The Assessment of Relative Risks*, Wiley, New York.

Lee, S. J. and Flowers, M. E. D. (2008). Recognizing and Managing Chronic Graft-Versus-Host Disease, *American Society of Hematology* **2008**(1): 134–141.

Ludajic, K., Balavarca, Y., Bickeböller, H., Pohlreich, D., Kouba, M., Dobrovolna, M., Vrana, M., Rosenmayr, A., Fischer, G. F., Fae, I., Kalhs, P. and Greinix, H. T. (2008). Impact of HLA-DPB1 allelic and single amino acid mismatches on HSCT, *British Journal of Haematology* **142**: 436–443.

Lunetta, K. (2008). Genetic association studies, *Circulation* **118**: 96–101.

Marieb, E. (2004). *Human Anatomy and Physiology*, Pearson Benjamin Cummings, San Francisco.

McNutt, L.-A., Wu, C., Xue, X. and Hafner, J. P. (2003). Estimating the relative risk in cohort studies and clinical trials of common outcomes, *American Journal of Epidemiology* **157**(10): 940–943.

Middleton, P., Norden, J., Cullup, H., Cavet, J., Jackson, G., Taylor, P. and Dickinson, A. (2003). Oestrogen receptor alpha gene polymorphism associates with occurrence of graft-versus-host disease and reduced survival in HLA-matched sib-allo BMT, *Bone Marrow Transplantation* **32**(1): 41–47.

Mood, A., Graybill, F. and Boes, D. (1974). *Introduction to the Theory of Statistics*, McGraw-Hill:New York.

Mullally, A. and Ritz, J. (2007). Beyond HLA: the significance of genomic variation for allogeneic hematopoietic stem cell transplantation, *Blood* **109**(4): 1355–1362.

Müller, M. (2004). Goodness-of-fit criteria for survival data [online], *Discussion paper*, Sonderforschungsbereich 386: Statistische Analyse diskreter Strukturen. Ludwig-Maximilians-Universität München.
URL: `http://hdl.handle.net/10419/31016` [Accessed 27 June 2011].

Müller, M., Döring, A., Küchenhoff, H., Lamina, C., Malzahn, D., Bickeböller, H., Vollmert, C., Klopp, N., Meisinger, C., Heinrich, J., Kronenberg, F., Wichmann, H. and Heid, I. (2008). Quantifying the contribution of genetic variants for survival phenotypes, *Genetic Epidemiology* **32**(6): 574–585.

Müller-Sieburg, C., Cho, R., Thoman, M., Adkins, B. and Sieburg, H. (2002). Deterministic regulation of hematopoietic stem cell self-renewal and differentiation, *Blood* **100**(4): 1302–1309.

Novota, P., Sviland, L., Zinöcker, S., Stocki, P., Balavarca, Y., Bickeböller, H., Rolstad, B., Wang, X., Dickinson, A. and Dressel, R. (2008). Correlation of Hsp70-1 and Hsp70-2 gene expression with the degree of graft-versus-host reaction in a rat skin explant model, *Transplantation* **85**(12): 1809–1816.

O'Quigley, J. and Flandre, P. (1994). Predictive capability of proportional hazards regression, *Proceedings of the National Academy of Sciences* **91**: 2310–2314.

O'Quigley, J. and Xu, R. (2001). *Explained variation in Cox regression. In: Crowley J, editor*, Handbook of Statistics in Clinical Oncology. New York: Marcel Dekker, Inc.: p 397-410.

Paynter, N., Chasman, D., Paré, G., Buring, J., Cook, N., Miletich, J. and Ridker, P. (2010). Association Between a Literarute-Based Genetic Risk Score and Cardiovascular Events in Women, *The Journal of the American Medical Association* **303**(7): 631–637.

Porzelius, C., Schumacher, M. and Binder, H. (2010). Sparse regression techniques in low-dimensional survival data settings, *Statistics and Computing* **20**: 151–163.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0
URL: `http://www.R-project.org` [Accessed 30 May 2011].

Riddell, S. and Appelbaum, F. (2007). Graft-Versus-Host Disease: A Surge of Developments, *PLoS Medicine* **4**(7): 1174–1177.

Ripatti, S., Tikkanen, E., Orho-Melander, M., Havulinna, A., Silander, K., Sharma, A., Guiducci, C., Perola, M., Jula, A., Sinisalo, J., Lokki, M.-L., Nieminen, M., Melander, O., Salomaa, V., Peltonen, L. and Kathiresan, S. (2010). A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses, *The Lancet* **376**: 1393–1400.

Rothman, K. (2002). *Epidemiology: An Introduccion*, Oxford University Press, Oxford.

Schemper, M. and Henderson, R. (2000). Predictive accuracy and explained variation in Cox regression, *Biometrics* **56**: 249–255.

Schemper, M. and Stare, J. (1996). Explained variation in survival analysis, *Statistics in Medicine* **15**: 1999–2012.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model, *Biometrika* **69**(1): 239–241.

Schumacher, M., Binder, H. and Gerds, T. (2007). Assessment of survival prediction models based on microarray data, *Bioinformatics* **23**(14): 1768–1774.

Sham, P. (1998). *Statistics in Human Genetics*, Arnold, London.

Sherry, S., Ward, M., Kholodov, M., Baker, J., Phan, L., Smigielski, E. and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation, *Nucleic Acids Research* **29**(1): 308–311.

Skov, T., Deddens, J., Petersen, M. and Endahl, L. (1998). Prevalence proportion ratios: estimation and hypothesis testing, *International Journal of Epidemiology* **27**(2): 91–95.

Stark, M. (1997). *Beurteilungskriterien für die Güte von Modellen zur Analyse von Überlebenszeiten*, PhD thesis, Berlin: Logos Verlag.

Stem Cell Basics: Introduction (2009). In *Stem Cell Information* [online]. Bethesda, MD: National Institutes of Health, U.S. Department of Health and Human services.
URL: `http://stemcells.nih.gov/info/basics/basics1` [Accessed 08 August 2011].

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *Journal of the Royal Statistical Society. Series B (Methodological)* **36**(2): 111–147.

Therneau, T. and Grambsch, P. (2000). *Statistics for Biology and Health. Modeling Survival Data: Extending the Cox Model*, Springer-Verlag, New York.

Therneau, T., Grambsch, P. and Fleming, T. (1990). Martingale based residuals for survival models, *Biometrika* **77**: 147–160.

Therneau, T. and Lumley, T. (2008). *survival: Survival analysis, including penalised likelihood. (S original by Terry Therneau and ported by Thomas Lumley).* R package version 2.34-1.

TRANS-NET (2008). Marie-Curie Research Training Network. Identification of genomic and biological markers as predictive/diagnostic/therapeutic tools for use in allogeneic stem cell transplantation: Translational research towards individualised patient medicine [online].
URL: `http://www.trans-net.org.uk/index.htm` [Accessed 05 September 2008].