From

The Institute of Medical Information Processing, Biometry, and Epidemiology

of the Ludwig-Maximilians-Universität, Munich, Germany

Chair of Epidemiology: Prof. Dr. Dr. H.-Erich Wichmann

and

The Institute of Epidemiology, Helmholtz Zentrum München, German Research

Center for Environmental Health (GmbH)

Director: Prof. Dr. Dr. H.-Erich Wichmann

Genetic association analysis with survival phenotypes

Thesis Submitted for a Doctoral degree in Human Biology
at the Faculty of Medicine Ludwig-Maxiimilians-University,
Munich, Germany

by

Andrea Martina Müller

From

Munich, Germany

2009

With approval of the Medical Faculty

of the University of Munich

First Reviewer:	Prof. Dr. Dr. H.-Erich Wichmann
Second Reviewer:	Prof. Dr. Ulrich Mansmann
	Prof. Dr. Elke Holinski-Feder
Co-supervision:	Dr. Iris M. Heid
Dean:	Prof. Dr. med. Dr. h.c. M. Reiser, FACR, FRCR
Date of the oral examination:	17.03.2009

Acknowledgements

In first instance I want to thank Prof. Dr. Dr. H.-Erich Wichmann, head of the Institute of Epidemiology at the Helmholtz Zentrum München and Chair of Epidemiology at the Institute of Medical Information Processing, Biometry, and Epidemiology of the University of Munich, who not only encouraged and enabled the work on this thesis at his institute, but also offered a variety of opportunities to work in the field of genetic epidemiology on exciting projects with experienced partners.

Furthermore, I thank my direct supervisor Dr. Iris Heid from the working group "Genetic Epidemiology" at the Institute of Epidemiology at the Helmholtz Zentrum München for initialising this methodological work, her continuous support, invaluable advice, fruitful discussions and project coordination within the GenStat group.

I am also grateful to PD Thomas Illig, head of the working group "Biological Samples in Genetic Epidemiology" and interim head of the working group "Genetic Epidemiology" at the Institute of Epidemiology at the Helmholtz Zentrum München, who organised availability of genetic data for this thesis and encouraged close collaboration with the laboratory and other working groups.

Through the multidisciplinarity of the work in the group "Genetic Epidemiology" I enjoyed the possibility to get involved into different projects and learn from different fields of epidemiology, medicine as well as genetics, which was only possible through the support of Prof. Dr. Dr. H.-Erich Wichmann, Dr. Iris Heid and PD Thomas Illig.

Important issues for the discussion part of this thesis were brought up by Prof. Dr. Helmut Küchenhoff from the Institut für Statistik at the Ludwig-Maximilians-Universität of Munich and Prof. Dr. Heike Bickeböller from the Department of Genetic Epidemiology at the University of Göttingen, whom I want to thank as well as all partners who contributed their data for evaluation within this thesis.

Special thanks go to all my current and former colleagues, from whom I want to especially emphasize Claudia Lamina, who contributed to this work through helpful discussions on statistical and programming issues.

Last but not least I thank my family and friends who were always at hand with help and unbelievable patience.

Table of Contents

1 Introduction

1.1 *General Introduction*

One of the major goals of epidemiologic research is too improve insight into risk factors associated with disease and disease development the recent advances in genetics offer a good possibility to analyse whether subgroups of the general population suffer from a genetically determined increased baseline risk or predisposition to develop disease. Therefore, identification of genetic variants that show association to health conditions is of growing interest and gave rise to the field of genetic epidemiology.

For several monogenic disorders, genetic variants have already been successfully identified and research is now focusing on complex polygenic disorders with high prevalence in the general population, e.g. type 2 diabetes, myocardial infarction, atherosclerosis and related parameters. For a better understanding of the disease causing mechanisms, it is important not only to measure whether genetic variants are of influence but also to quantify their impact on changes of health parameters.

More and more population-based studies provide long follow-up in combination with genetic data. Therefore, it is now possible to not only analyse the risk to develop disease through case-control studies but also the time of disease onset in the general population through application of methods from survival analysis. Especially for age-related complex diseases this is of increasing interest. Quantification of the impact of covariates on the outcome within this type of analysis, however, is still an unsolved general problem of epidemiology and statistics.

The aim of this dissertation was to identify the criterion which suits best for quantification of the impact of genetic variants within time-to-onset or survival

analysis, similar to a percentage of explained variation in linear regression. Eligible criteria were compared in their performance through simulation studies and application to mortality data from the KORA studies.

The introductory chapter of this thesis provides background information on general methodology in epidemiology with a focus on study types and survival data analysis. Furthermore, the basics of genetics and genetic association studies are described. In the main chapter of this thesis (chapter 2), possible criteria for judging the impact of genetic variants within survival data analysis are presented and subsequently investigated through simulation studies and application to mortality data from a large cohort study, the KORA study. The discussion of the results, conclusion and an outlook is given in chapter three.

1.2 *Epidemiologic studies*

1.2.1 Common study types in epidemiology

The aim of epidemiologic studies is to describe and investigate diseases and the factors influencing them. While clinical studies focus on investigation of treatment success, a broad variety of epidemiologic studies aims to identify and describe prognostic factors, i.e. factors that influence the probability of occurrence of disease or its development.

To investigate the question of disease development longitudinal or cohort studies are the appropriate study design. They start with a baseline investigation and collect follow-up information through regular re-examination or questionnaires. Therefore, cohort studies offer the possibility to investigate incidence or development of disease or health related factors.

1.2.2 Terminology

In the following the disease or health parameter investigated is generally called *phenotype*. Risk factors, i.e. factors that influence this phenotype, are either called *environmental* or *genetic* factors or covariates. It should be noted that all non-genetic factors, including e.g. environmental exposures like fine dust particles, but also life-style parameters like smoking and even age and sex, are generally termed environmental factors. *Association analysis* quantifies the relation between phenotype and environmental and/or genetic factors through statistical analysis. Estimated *effect sizes* describe the relative change in the phenotype due to different covariate values. In association analysis, it is common to define a subset of environmental or genetic covariates as *adjustment covariates* beside the covariate of primary interest. Adjustment covariates are supposed to influence the phenotype. If they also influence the covariate of primary interest, they are called *confounders* and need to be accounted for in analysis.

1.2.3 Statistical methods for analysis of association in epidemiologic studies

1.2.3.1 Methods for cross-sectional and case-control studies

The statistical model necessary for evaluation of the association between the phenotype Y and m environmental and/or genetic covariates $X_{1,\ldots,}X_m$, depends on the distribution of the phenotype. If the phenotype is normally distributed, which is often the case in cross-sectional studies, linear regression can be directly applied:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m$$

Here, $\beta_1,\ldots,\beta_m$ represent the true effect sizes for each covariate $X_{1,\ldots,}X_m$, while β_0 gives the baseline level of the phenotype given all covariates are 0. The estimation of $\beta_{1,\ldots,}\beta_m$ is the primary aim of the association analysis. In order to distinguish between true and estimated effect sizes, the latter are termed $\hat{\beta}_0,\hat{\beta}_1,...,\hat{\beta}_m$ and give the relative change in the level of the phenotype per unit increase of the covariate.

Sometimes, however, the phenotype is quantitative but not normally distributed. Often, a simple transformation $f(Y)$, then, yields a normalised phenotype and replaces Y in the upper regression model. An example would be CRP, a prominent inflammatory factor modelled as phenotype in association analysis investigating coronary artery diseases, which generally requires a log-transformation to log(CRP) and, therefore, yields a so-called loglinear model.

If the phenotype is a disease indicator, as e.g. in case-control studies, logistic regression analysis is performed. The disease indicator is then transformed into a probability to develop disease given the observed covariate values. Let the outcome Y be the indicator for the observed disease state (1=disease, 0=no disease), $X_{1,\ldots,}X_m$ be a set of covariates and $\beta_1,\ldots,\beta_m$ the vector of associated effect sizes. Then the logistic regression model with $p(Y=1|\,X_1,...,X_m)$ denoting the probability of disease given the covariates $X_{1,\ldots,}X_m$ is written:

$$p(Y=1|\,X_1,...,X_m)=\frac{1}{1+\exp(-\beta_0-\beta_1X_1-...-\beta_mX_m)}=\frac{\exp(\beta_0+\beta_1X_1+...+\beta_mX_m)}{1+\exp(\beta_0+\beta_1X_1+...+\beta_mX_m)}$$

Note that the transformation of probabilities $p(Y=1|\,X_1,...,X_m)$ into *logits:* $Logit(Y\mid X_1,...,X_m)=\ln\left(\frac{p(Y=1|\,X_1,...,X_m)}{1-p(Y=1|\,X_1,...,X_m)}\right)$ would yield a linear model. Therefore, the logistic model, as well as the loglinear model, falls into the class of generalised linear models, where linear regressions are evaluated on the transformed version of the phenotype. In logistic regression models, estimated effect sizes $\hat{\beta}$ are often

interpreted in their transformed version $\exp(\hat{\beta})$, which are known as odds ratios (OR) and act multiplicatively on the probability to get the disease of interest, whereas effects on the β scale are interpreted as additive effects on the logits. Odds ratios are interpreted as the relative change of odds to get the disease per unit change in the respective covariate.

1.2.3.2 Methods for cohort studies

If the phenotype is measured in a cohort with several repeated measurements at each follow-up time point, longitudinal analysis tools including mixed models with fixed effects or random effects are applied. The exact model depends on the investigator's focus and possible and necessary assumptions. It is for example possible to analyse inter- as well as intra-individual variability or simple changes in overall mean values of the phenotype. As longitudinal analysis with repeated measurements is a very complex field, which is not the focus of this thesis, no further description is given here.

The second special phenotype available in cohorts with follow-up information is called *survival phenotype* and measures the time until a certain event occurs. Note that the name "survival phenotype" originates from mortality data analysis but may refer to any time-to-event phenotype, e.g. time to onset of disease (i.e. disease-free survival), relapse or surgery. Due to the study design, the event of interest has usually not yet occurred for all subjects at follow-up. These subjects with incomplete survival times are called censored observations with survival times censored at follow-up. An example close to a real study situation is given in Figure 1. Here, recruiting of study participants is realised over a period of 3 years. After 4 years, follow-up starts. Gathering of follow-up information is here presented to take one

year. Nine subjects are still under observation at follow-up. The event of interest has not yet occurred. Therefore, they have censored survival times at follow-up while all other subjects have complete survival times, where the exact time of the event is known. The right panel shows, how individual observation lengths are distributed. These are the times modelled in survival analysis.

Analysis of the time until occurrence of an event, while accounting for this uncertainty in the data, requires specialised methodology for analysis. In this case, the outcome is not a single variable but a composition between the observed failure time and the indicator whether the event of interest has occurred or not (*status indicator*). The methods from *survival analysis* allow for censoring in the data through definition of time dependent risk sets. Therefore, censored individuals are taken into account at least as long as they are under observation.

Survival analysis is characterised by three major functions of interest:

- *failure function F*(*t*): cumulated probability to have an event until time *t*
- *survival function S*(*t*): the probability to not have an event until time *t*
- *hazard function λ*(*t*): instantaneous probability to have an event at time *t*+*δt*, or the *cumulated hazard* as its integral *Λ*(*t*)

These three functions are related as follows:

$$1 - F(t) = S(t) = \exp(-\Lambda(t))$$

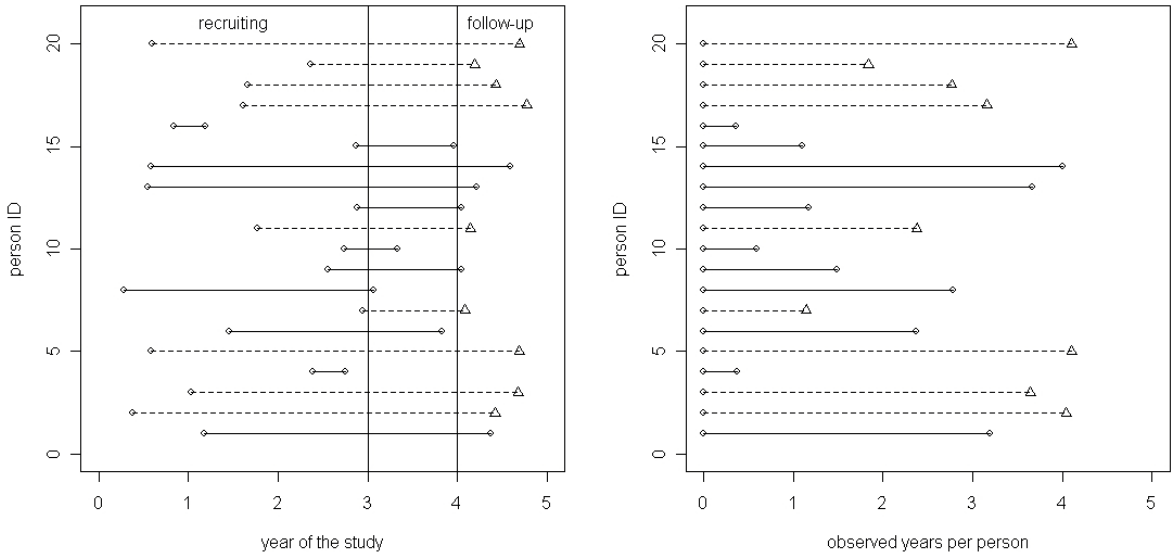

Figure 1: A longitudinal study including twenty persons during a recruiting time of three years and a follow-up after 4-5 years. Dashed lines mark censored subjects that are still under observation at follow-up. Solid lines mark subjects with complete survival times, where the precise time of occurrence of the event of interest is known. The left panel shows individual observation times during the study period. The right panel shows length of individual observation times.

In the following *censoring times* denote failure times from censored individuals and *event times* are failure times from individuals with an event. In survival analysis, estimation is usually performed through non-parametric methods or the semi-parametric Cox proportional hazards model. As estimation is generally based on risk sets and events at event times, individuals with censoring before the first event time do not occur in any risk set and can generally be excluded.

Non-parametric survival

The survival function can be estimated nonparametrically by the Kaplan-Meier estimator [Kaplan and Meier, 1958]. For each failure time t_i on the time axis the

probability of an event is calculated based on the number of events $d(t_i)$ relative to the number of individuals at risk $R(t_i)$:

$$S_{KM}(t) = \prod_{i=o}^{t} \left(1 - \frac{d(t_i)}{R(t_i)}\right)$$

Note that it is usually assumed that the number of distinct time points equals the number of individuals. In this case, $d(t_i)$ may also be replaced by the individual status indicator δ_i which is 0 for censored observations and 1 for observations with an event.

$S_{KM}(t)$ is often visualised as Kaplan-Meier step function plotted over time with steps at each observed event time (e.g. Figure 2 for the example data given in Figure 1). Kaplan-Meier curves do not need to end at $S_{KM}(t)=0$. In Figure 2, for example, 37% of the initial population still remain at risk after the last observed event occurred and are displayed as censored (cross at the right end of the step function). Kaplan-Meier curves can also be calculated for subgroups which can then be tested for significant discrepancies through logrank tests. The nonparametric estimation, however, does not allow for adjustment for covariates or analysis of continuous covariates.

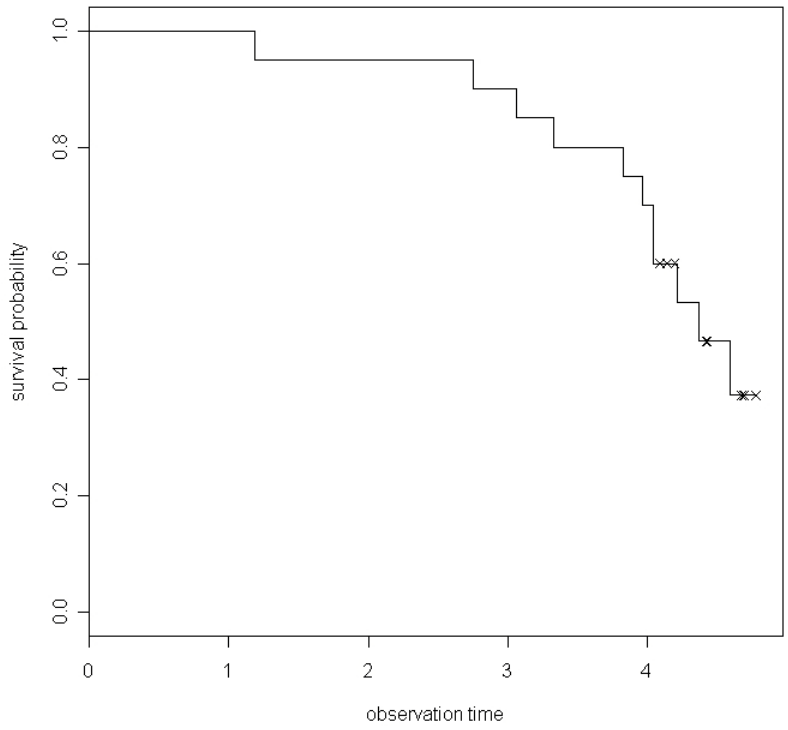

Figure 2: Kaplan-Meier plot for example data shown in Figure 1. The solid line shows the survival step function. Steps only occur at event times. Censored observations are displayed as crosses at censoring time points along the step function and contribute to the height of the function as long as they are under observation.

Semi-parametric survival (Cox proportional hazards regression)

In case of continuous covariates or if adjustment for covariates is required, the semiparametric Cox proportional hazards model [Cox, 1972] has become standard. This model assumes a general *baseline hazard* $\lambda_0(t)$ for all subjects given all covariates *X* are zero, and may be interpreted as a time-dependent baseline risk which is shifted for each subject corresponding to its observed covariate values. The Cox model is written:

$$\lambda(t|X) = \lambda_0(t)\exp(\beta' X)$$

Constraints on the shape of the baseline hazard through specification of an underlying general survival distribution would allow for application of fully parametric models. But a priori knowledge is rare and often some distribution, e.g. Weibull or exponential, has to be assumed in case of parametric survival analysis. The semiparametric Cox model incorporates the baseline hazard $\lambda_0(t)$ as a nonparametric term without any constraints on its shape except that it accounts for all subjects, while all covariates enter the model as parametric terms.

Estimation in Cox proportional hazards regression is based on the partial likelihood function which is the part of the full likelihood that is independent of the underlying baseline hazard [Cox, 1975]. It is assumed that censoring is uninformative in the sense that censored observations do not contribute additional information to the estimation. The logarithm of the partial likelihood is described as:

$$\ln PL(\beta, X) = \sum_{Y_i uncensored} \left(X_i \beta - \ln \sum_{t_j \geq t_i} \exp(X_i \beta) \right)$$

Here, it is assumed that the number of distinct failure times *k* equals the number of subjects *n* and the index *i* is defined for *i*=1,…,*n*. In case of tied failure times (*k*<*n*) Breslow's approximation is applied with $S_i = \sum_{j \in D_i} X_j$ and D_i the set of indexes with equal failure time:

$$\ln PL(\beta, X) = \sum_{i01}^{k} \left(S_i \beta - d_i \ln \sum_{t_j \geq t_i} \exp(X_i \beta) \right)$$

The score function $U(\beta, X)$ resulting from the first derivation of the logarithm of the partial likelihood is set to 0 for estimation of effect sizes $\hat{\beta}$. It is written:

$$U(\beta, X) = \sum_{Y_i uncensored} \left(X_i - \frac{\sum_{t_j \geq t_i} X_j * \exp(X_i \beta)}{\sum_{t_j \geq t_i} \exp(X_i \beta)} \right)$$

The fraction in this definition may be interpreted as the expectation of the covariate calculated over all subjects that are still at risk.

Effect estimates are interpreted as $\hat{\beta}$ or as *hazard ratios*, $HR = \exp(\hat{\beta})$, which are comparable to odds ratios from logistic regression but have to be assumed to be constant over time. This general assumption of constant hazard ratios over time is the reason for the full name of the model: *Cox proportional hazards regression*. Proportionality of hazards can be tested based on scaled Schoenfeld residuals [Grambsch and Therneau, 1994]. The hazard ratio describes the factor for the hazard corresponding to each unit increase in the associated covariate. Like odds ratios in logistic regression, hazard ratios act multiplicatively.

The survival function from Cox regression is defined:

$$S_{Cox}(t|X) = \exp(-\Lambda_0(t))^{\exp(\beta' X)}$$

It is also possible to visualise the survival function resulting from Cox regression by Kaplan-Meier plots. The average survival function is then displayed for all covariates taking their mean values.

1.3 *Background in genetics*

1.3.1 The human genome

The genetic information of humans is coded in the form of DNA and stored in the cell nuclei. Each nucleus contains 22 pairs of homologous chromosomes, the autosomes, as well as two sex chromosomes X and Y, which generally combine as XY in men and as XX in women. These chromosomes carry the major part of the genetic information. The DNA is a macromolecule composed of four nucleotide bases, which are either classified as purines (thymine (T) or adenine (A)) or pyrimidines (guanine

(G) or cytosine (C)). Each DNA chain consists of two strands of nucleotide bases that are arranged in reverse and complementary fashion. The two strands are connected through hydrogen bonds between the complementary base pairs (two bonds between purines and three between pyrimidines). DNA is organised into more highly coiled structures by folding around histone and other proteins, which play a role in gene regulation. (Figure 3).

The genetic code is defined by the sequence of nucleotide bases, genes can be encoded on either of the two strands. Special strand-specific nucleotide sequences encode predefined functions, e.g. start and stop sequences surround DNA sequences that encode for transcription areas needed for protein biosynthesis and therefore also give the reading direction of a gene (i.e. a functional sequence in the DNA). Protein coding sequences within a gene are called exons, non-coding regions within genes and between exons are named introns. Another important region within a gene is the promoter, a control point for transcription and therefore substantial for regulation of gene expression.

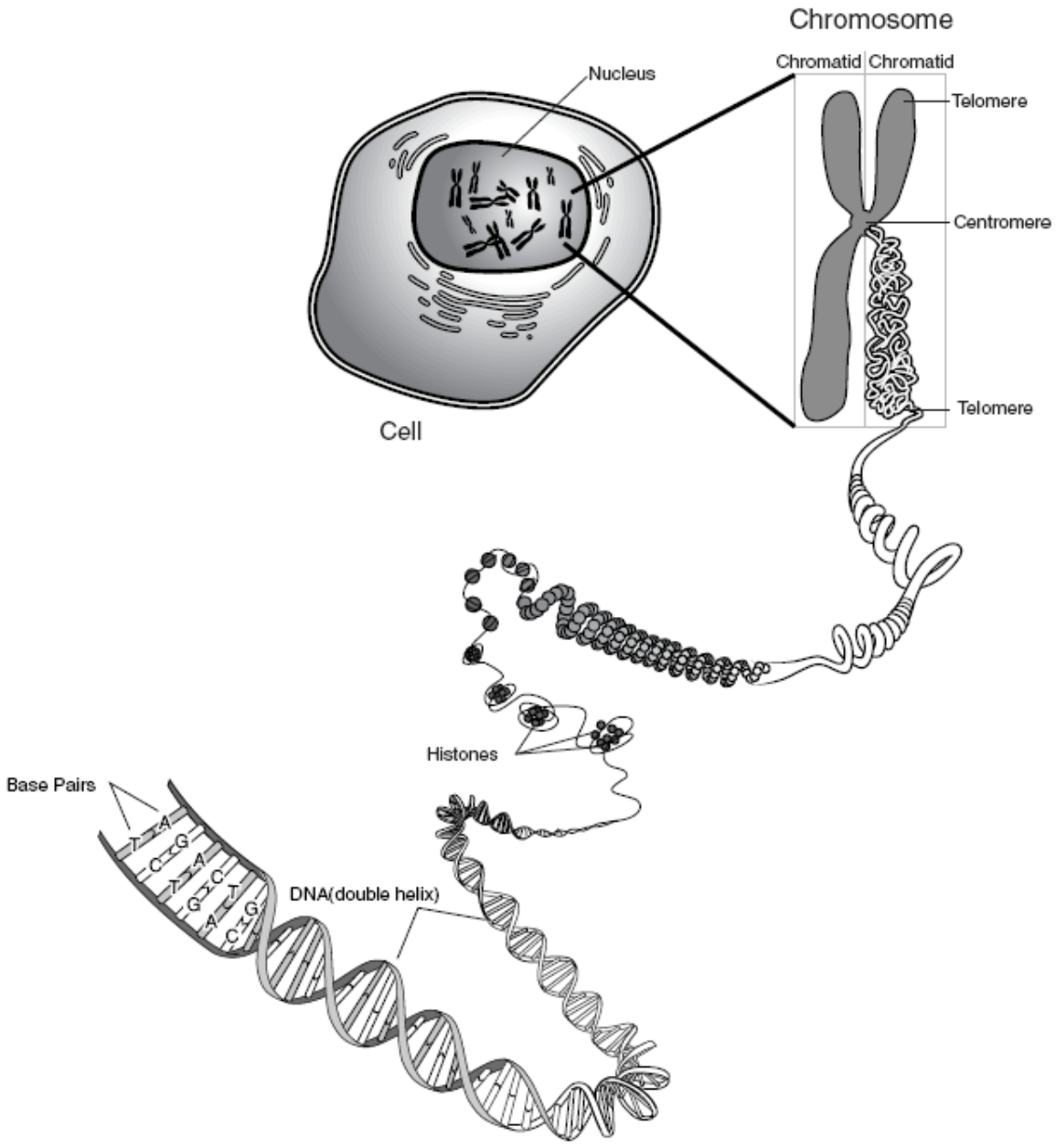

Figure 3: Chromosomes are found in each cellular nucleus. The way from chromosome to single base pair information is illustrated (Figure from: National Institutes of Health, National Human Genome Research Institute, Division of Intramural Research, website:

http://www.genome.gov//Pages/Hyperion//DIR/VIP/Glossary/Illustration/chromosome.cfm).

1.3.2 Single nucleotide polymorphisms

1.3.2.1 Single nucleotide polymorphisms as genetic markers

The focus of this dissertation is set on association analysis with single nucleotide polymorphisms (SNPs), i.e. single nucleotide exchanges that - by definition of a polymorphism - make up at least 1% of the alleles in the population under study. SNPs are estimated to describe about 90% of the genetic inter-individual variability with respect to single nucleotide exchanges. Therefore, SNPs are of high interest for research on complex diseases and health conditions with high prevalence in the general population. SNPs from autosomal chromosomes encode the two *alleles* (i.e. nucleotide bases) from the two chromosomes at a specific position in the genome. The less frequent allele is called *minor allele*, the more frequent allele is the *common allele*. SNPs from the male sex chromosome (Y) are special as long as they do not lie at the ends of the chromosome which are known as pseudoautosomal regions and are similar between X and Y chromosome. The non-pseudoautosomal SNPs only encode information from one haploid chromosome, do not recombine in meiosis, have thus different properties with respect to formal genetics and are not discussed in detail.

A *genotype* describes the alleles present for a certain SNP in a single individual. For an autosomal or pseudoautosomal SNP a genotype consists of two alleles. Genotypes are labeled to be either *homozygous*, i.e. both chromosomes carry the same allele, or *heterozygous*, i.e. different alleles on the two chromosomes. For example, a SNP with the known alleles A and G may yield genotypes AA, AG or GG, where AG is the heterozygous genotype. The vast majority of SNPs is biallelic, i.e. two different alleles are reported to occur in the specific locus. Nevertheless, it is also

possible to have more than two different alleles reported at a specific locus. These are considered special cases and require special modeling.

Large databases with reported SNPs are available. For example, the dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP) to date has registered about 11 million SNPs. Until now, about 3.97 Million of these have been validated in the European descent population in the course of the HapMap Project (http://www.hapmap.org/downloads/index.html.en), which aimed to investigate genetic heterogeneity within and between different ethnic groups [Frazer et al., 2007]. Therefore, it provides additional information e.g. on SNP correlations and haplotype patterns in several populations. These and other databases are utilised for selection of SNPs for genotyping in genetic association studies.

The type of possible genetic markers for association analysis, however, is not restricted to SNPs. Alternative markers for association analysis include sequence repeats in the DNA such as short tandem repeats (e.g. microsatellites) or larger copy number variations (formerly termed deletions and insertions) as well as methylation patterns. The variety of genetic markers is growing due to improved knowledge of molecular genetics.

1.3.2.2 Genotyping

Calling of SNP genotypes (*genotyping*) in a study sample can be performed on different platforms. One possibility mainly used for genetic studies with limited number of SNPs is to apply primer extensions on parts of each of the two chromosomal strands. The length and mass of the extension products then depends on the alleles at the specific SNP. Subsequently, genotypes can be determined through mass spectroscopy. In this case, genotyping is for example carried out

through matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) analysis of the obtained primer extension products. Genotyping calls can then be made with MassArray RT software (Sequenom, San Diego, USA), where the different alleles are visualised through different mass peaks. Details on this technology are given in [Vollmert et al., 2007].

1.3.2.3 Quality control

Quality control is an important issue in genetic analyses. Before any further analysis is performed, the quality of the DNA samples is checked through polymerase chain reaction (PCR). Genotyping itself involves further quality checks like double genotyping of a predefined percentage of SNPs as well as positive and negative controls. Furthermore, the specific structure of sex chromosomes (Y chromosome only in men, heterozygous genotypes in X chromosome only in women) allow for validation of sex in the database. Random doubles allow for determination of random genotyping error which can later be accounted for in analysis [Heid et al., 2008].

Person-wise genotyping success rates, i.e. the percentage of successfully genotyped SNPs per person, gives additional information on DNA quality. Persons with low number of successfully called SNPs are often excluded from further analysis as a general problem with the specific DNA sample is indicated.

The SNP-wise genotyping success rate in a study sample gives information about the quality of the genotyping process. Genotyping success rates above 95% are generally favourable. Low genotyping success rate may pinpoint to problems during the calling process or problematic assay designs. In this case, mainly the number of heterozygotes is affected because here, the signals for the two different alleles have to be measured whereas homozygotes only yield one signal (see section 1.3.2.2).

Therefore, it is also important to control for deviating numbers of heterozygotes. In unselected, randomly mating populations the proportion of heterozygotes should relate to the proportion of homozygous carriers of the different alleles, with an equation called *Hardy-Weinberg equilibrium* (*HWE*). Let *f*(A) be the frequency of allele A and *f*(B)=1-*f*(A) be the frequency of allele B. Then, the expected genotype frequencies are: $f(AA)=f(A)^2$, $f(BB)=f(B)^2$, $f(AB)=2f(A)f(B)$. It is possible to test for *HWE* given the allele frequencies as well as the observed genotype frequencies through Chi^2 test, or, if the expected number of copies of subjects with any genotype falls below 5, Fisher's exact test. However, if a SNP is supposed to be associated to severe diseases that result in early exclusion of specific genotypes, the balance between genotypes may be affected in population-based studies (mainly in higher age-groups) and thus resulting in *HWE* violation due to selection.

Another important measure to control for in association studies is the frequency of the minor allele (*MAF*). Low *MAF* indicates low power to detect genetic effects, which are generally considered to be small for polygenetic diseases. Furthermore, outliers in the phenotype variable may become extremely influential in small groups and general distributional assumptions for regression models may be violated. Therefore, a minimum *MAF* should be available for SNPs that are examined for association to any disease-related outcome. Thresholds for minimum *MAF* are defined per study dependent on sample size and study design. Often a minimum *MAF* of 1% or 5% is required. For example, given a population of 1500 persons, a SNP with *MAF*=1% would yield no more than 10 heterozygotes and 10 homozygous carriers of the minor allele under the assumption of perfect *HWE*.

In case of multi-SNP analysis, it is also important to check the correlation between SNPs. For highly correlated SNPs, the additional information obtained through analysis of all SNPs may be small compared to the analysis of a single SNP. For bins

of highly correlated SNPs, it may be sufficient to choose one or more representative SNPs for analysis of a set of subsequent SNPs. These systematically chosen SNPs are called tagging-SNPs [Stram et al., 2003].

An overview of further considerations with respect to quality issues and study design is given in [Hattersley and McCarthy, 2005].

1.4 *Genetic association studies*

1.4.1 Localisation of phenotype-associated genetic variants

Localisation of disease-causing genetic variants within the genome is generally a stepwise approach. In the first instance, often, general information about heritability of diseases is gathered in order to give a first idea of the impact of a genetic component on disease development. Especially large, unselected populations are more and more in the scope of genetic epidemiological research as they allow best for general validity of possible findings.

Linkage studies may be considered a first traditional step to obtain a rough localisation of disease-causing genetic variants as well as the inheritance mode through investigation of cosegregation between genetic markers and diseases within families [Dawn and Barrett, 2005], [Spielman et al., 1993]. This step, however, recently moves more and more into the background.

Currently, there is increasing interest in *association analysis* in form of *candidate gene studies* or *genome wide screens* which allow for direct identification of disease-related genetic variants [Cordell and Clayton, 2005]. Candidate gene studies focus on gene regions that have already been identified as possible carriers of disease-related genetic variants, e.g. known from the literature, functional analysis or from a

positional indication through linkage studies. Genome wide association studies aim at screening the whole genome in order to find new regions giving a strong signal of association to disease. The possibility to identify completely new areas that play a major role in the pathogenesis of disease makes genome-wide association studies an important tool. In contrast to candidate gene studies, however, there is often low power to find disease-related genetic markers, due to the small expected genetic effect sizes (except for monogenic disorders) and the often high number of statistical tests that are performed. The growing importance of association analysis for identification of genetic variants that indicate increased predisposition to develop disease leads to intensified work on related technical as well as methodological issues.

Once having identified genetic variants that are associated to the phenotype of interest, further refined studies within the respective gene region follow as well as functional studies like gene expression analysis, which validate causality and improve insight into the underlying biological mechanisms. An overview over the field of genetic epidemiology and research focuses can be found e.g. in [Kaprio, 2000] or [Burton et al., 2005].

1.4.2 Genetic effect models

1.4.2.1 Genetic effect model definition

The connection from genotype to the visible phenotype may be closer specified through segregation analysis within families. For definition of a model for association analysis, a priori information about the inheritance mode and the necessary genetic effect model is crucial. Otherwise assumptions have to be made.

Genotypes in SNPs are coded by the number of minor alleles observed on the two homologous chromosomes. Often, a dose-effect relation between the number of copies of the minor allele and the phenotype is assumed. If this relation is linear, the genetic effect model is called *additive* or *codominant*. If the phenotype has to be transformed for statistical analysis, as in loglinear, logistic or survival models, an originally additively modelled effect may become *multiplicative* on the transformed scale of the phenotype variable.

Sometimes, however, no dose effect is visible. In case of a *dominant* effect model, the phenotype is affected as soon as one copy of the minor allele is present but is unchanged even if further copies of the minor allele are observed. In case of *recessive* effect models, the phenotype is only affected in presence of two copies of the minor allele.

Nevertheless, it is also possible to avoid constraints on the genetic effect model and model the two possible genotypes with at least one minor allele as two single indicator variables. The unconstraint model, however, increases the number of degrees of freedom used in the statistical tests for association and, therefore, may be disadvantageous if, for example, correction for multiple testing is required.

In case of low frequency of homozygous carriers of the minor allele, which mainly occurs for SNPs with low *MAF*, separate modelling of this group (within unconstraint or recessive genetic effect models) is problematic and it may be either excluded or similar as in an assumed dominant model, pooled to the group of heterozygotes.

Due to the high effort of segregation analysis and increasing speed in association analysis, a priori knowledge about the genetic effect model is often not available and a dose-effect, hence, an additive effect model is assumed.

1.4.2.2 Coding of SNP variables

Coding of the genotypes as variables for statistical analysis depends on the assumed genetic effect model. For additive or multiplicative models, the variable has a trichotomous design and counts the number of copies of the minor allele. In the association model, it enters as a linear term. Dominant or recessive models need the genotype coded as dummy variable, which is set to 1 for the genotypes assumed to affect the phenotype. The unconstraint model requires two dummy variables, each indicating presence of one of the possible variants in the genotype carrying at least one copy of the minor allele.

An interesting variant is the coding of unconstraint genetic effect models through a count variable, as under the assumption of an additive effect model, and an extra indicator for the group of heterozygotes [Schaid, 2004]. This special coding allows clearer insight into the genetic effect model through statistical testing. In case of strictly additive effects, the additional indicator for the group of heterozygotes is supposed to be not significant. For dominant effects, its estimate is supposed to equal the effect obtained for the additive term, whereas for recessive models, it should equal the additive term's effect estimate with reversed sign.

1.4.3 Methods to quantify the genetic effect

1.4.3.1 Estimation of genetic effect sizes

Estimation of genetic effect sizes depends on the distribution of the phenotype and the necessary regression model for association analysis (see section 1.2.3).

In linear regression, the genetic variants are entered into the model corresponding to the chosen genetic effect model either as single covariates or as a set of covariates. Additional adjustment for environmental covariates and confounders is possible.

However, it is recommended to investigate models with and models without adjustment for additional covariates. In some cases, it may even be favourable to consider different adjusted models in order to better specify side or confounder effects.

Estimated effect sizes are interpreted as relative changes in the phenotype obtained for the variant compared to the reference. In additive, dominant or unconstraint genetic effect models, the reference is the group of homozygous carriers of the common allele. For recessive effect models, the reference also includes the heterozygotes. For additive models, the estimated effect size is interpreted as change in the phenotype per copy of the minor allele.

P-values for judgement of significance of the estimated effect sizes $\hat{\beta}$ are usually obtained through statistical tests like t-tests or Wald tests on the null hypothesis $\hat{\beta} = 0$. Like in general epidemiologic studies, p-values below 5% are usually considered significant for single SNP analysis. In multi-SNP analysis or different models including different adjustment for environmental covariates, correction for multiple testing may be necessary to guarantee the overall significance level of 5%.

1.4.3.2 Quantification of the impact of genetic variants

In genetic association studies, it is of growing interest to report more than estimators of genetic effect size and p-values in order to allow for further specification of the impact of genetic variants on disease outcome. This is particularly an issue in studies based on population representative cohorts without special ascertainment criteria (such as e.g. extreme phenotype selection). For quantification of the impact of

genetic variants, a general focus is set on quantitative criteria that allow for model comparison and model selection.

Likelihood-based criteria like Akaike's information criterion (AIC) or likelihood ratio tests are often the basis of model selection. The values themselves, however, are difficult to interpret and do not allow the quantification of the impact of covariates on the outcome. Measures of discrimination quantify the predictive capability of a model if the association analysis is seen as a classification problem. Therefore, they may be suitable within a logistic regression framework or classification and regression trees (CART). However, measures of discrimination like AUC (area under the receiver operator characteristic curve) tend to ignore the associated effect size [Cook, 2007] which should also contribute to a general measure of impact. In the case of continuous outcomes, measures of discrimination would furthermore lead to a loss of information due to the reduction of the analysis to a classification problem. Therefore, variation-based measures such as the explained variation, obtained as R^2 in linear regression, are generally more attractive to measure the impact of genetic variants.

In the following, possible definitions of R^2 are given for linear and generalised linear regression.

R^2 in linear regression

For linear regression models, the estimation of the explained variation or predictive accuracy is obtained through calculation of the coefficient of determination, R^2 [Rosthoj and Keiding, 2004]. This coefficient may be interpreted as the squared correlation coefficient as well as on basis of sums of squares or variance components. The definition of R^2 based on sums of squares is as follows:

Let $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ with $\bar{y}$ = average over all outcomes y_i be the total sum of squares and $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ be the sum of squared error or residuals, i.e. discrepancy between observed outcomes y_i and their expectation $\hat{y}_i$ given the model is true. Then, R^2 is calculated as:

$$R^2 = 1 - \frac{SSE}{SST}$$

A discussion of possible alternative formulations is given in [Kvalseth, 1985]. It is also possible to correct R^2 for the number of predictor variables through introduction of a shrinkage term [Van Houwelingen and Le Cessie, 1990].

R^2 in generalised linear regression

For models other than the linear model, the appropriate definition of R^2 is less clear, which is often due to missing residual definitions. And the proper choice becomes a challenging task [Kvalseth, 1985]. A widely accepted definition of R^2 is the general coefficient of determination for generalised linear models which is calculated based on the likelihood of a model [Nagelkerke, 1991] and is often implemented in standard software packages. Let $L(0)$ be the log-likelihood of the null model and $L(\hat{\beta})$ be the log-likelihood of the fitted model for the sample of size n. Then R^2 as defined by Nagelkerke is calculated as follows:

$$R_N^2 = 1 - \left(\frac{L(0)}{L(\hat{\beta})} \right)^{\frac{2}{n}}$$

However, this general definition suffers in case of discrete distributions, e.g. in logistic regression models. In these models, the maximum attainable value may be below 0.

The maximum attainable value $R^2_{\max}$ is calculated through setting $L(\hat{\beta}) = 1$. Division of R^2_N by $R^2_{\max}$ leads to the adjusted coefficient $R^2_{adj} = \frac{R^2_N}{R^2_{\max}}$, which is often also returned by standard software.

2 Impact of genetic variants on survival phenotypes

2.1 *Aim of the study*

2.1.1 Genetic association analysis with survival phenotypes

The link between genetics, epidemiology and statistics is a currently fastly developing field. In genetics, studies were initially mainly performed within families. Later, the field of genetic research was extended to include case-control studies. The breakthrough of population-based genetics started with genotyping and evaluation of large cross-sectional studies and prospective cohort studies. There is more and more extensive long-term follow-up data available for large cohorts, which allow analysis of longitudinal development of diseases, incidence and survival. In statistics, new methods for analysis of family data and analysis of large scale genetic data had to be developed. Existing methods for analysis of cohorts (i.e. longitudinal data or survival data) had to be extended for application in genetic epidemiology. Hence, for each new setting, the link between the three involved research areas (genetics, epidemiology and statistics) had to be created and knowledge had to be transferred and combined. Therefore, genetic epidemiology can be seen as a strongly multidisciplinary field of research. A list of currently available statistical methods for genetic association studies with different epidemiologic study designs can be found in [Cordell and Clayton, 2005].

Statistical methods from the survival framework allow analysing the association between early onset or occurrence of diseases, e.g. type 2 diabetes or myocardial infarction, or death and genetic variants. Especially for type 2 diabetes, the current worldwide epidemic [Wild et al., 2004] invoked intensified research in the field of

genetic epidemiology - with the aim to identify genetic variants that may be one reason for the increased spread of the disease. Since, the link between genetic epidemiology and survival analysis has been established, it is possible to not only analyse genetic effects on incidence through methods for case-control studies. Moreover, it is now possible to analyse genetic effects on the time of disease onset through definition of (disease-free) survival phenotypes as combination of:

1) a time variable (time of disease onset, if the disease occurred, or last time of follow-up, if the disease has not occurred until the last follow-up)
2) a status indicator (indicator whether the disease occurred)

The link between survival analysis, as a specialised field of statistics, and genetic research is still young and adaptation of specialised statistical methodology to the new fields is still ongoing.

2.1.2 Measures of the impact of genetic variants on survival phenotypes

It is of increasing interest not only to assess existence of association, but also to quantify the impact of a genetic variant on phenotypes through measures of explained variation. As described in section 1.4.3.2, criteria measuring explained variation exist for quantitative phenotypes that are normally distributed (R^2) or for phenotypes analysed through generalised linear regression, e.g. dichotomous outcome in case-control data, (R^2_N or R^2_{adj}).

Standard analysis through Cox regression yields the usual measures of association between genetic variants and time of diagnosis (or death): Hazard ratios and p-values help quantifying the strength of the association between genetic variant and survival phenotype. Measuring the impact of genetic variants on survival phenotypes

by means of a criterion comparable to R^2, however, is not yet possible, as a generally recommended criterion measuring explained variation for survival data is still missing. Currently, a number of possible solutions addressing this topic exist (a selection is presented in section 2.1.4.1). However, no systematic investigation with a particular focus on possible application for genetic association studies has been conducted, yet.

2.1.3 Aim of this thesis

As survival outcomes are a composition of the two variables *observation time* and *event indicator*, no clear residual definition is available for survival analysis. Therefore, the definition of a criterion of predictive accuracy or of explained variation of a model, similar to R^2 in linear regression, currently leads to a variety of answers. Research within this topic faces a long history and is still ongoing. New methods are developed and existing possible solutions are permanently extended and improved. Therefore, no clear general recommendation is available, yet.

The aim of this thesis was to identify the optimal criterion for judging the impact of genetic variants on survival phenotypes. The literature was searched for criteria suitable to quantify the contribution of a trichotomous or dichotomous variable (as for SNP data) to survival phenotypes (survival or disease-free survival data). Eligible criteria should have an interpretation close to R^2 in linear regression models and fulfil the following characteristics:

(a) reasonable criteria values in the range [0;1] in order to allow for interpretation as percentage of prediction quality

(b) increasing values with increasing effect size

(c) independence of the percentage of censored observations in the data

Identified eligible criteria were tested for their suitability to quantify the impact of SNP genotypes on survival phenotypes through simulation studies. Simulation scenarios were aimed to cover realistic situations (regarding *MAF*, genetic effect size, genetic effect model, censoring percentage and censoring mechanism) as well as extreme settings in order to investigate the range of criteria values and their dependence on the varied parameters.

Furthermore, we aimed to illustrate the performance of the criteria for real data situations. Therefore, the criteria were applied to association analysis for SNPs and survival data from the KORA (Kooperative Gesundheitsforschung in der Region Augsburg) survey S3 from the region of Augsburg, South Germany, including its ten-year follow-up (F3). Genotyped SNPs from genes that are considered as potential candidates for association to outcomes related to severe diseases, such as type 2 diabetes or dyslipidemia were available for analysis. Severe diseases like type 2 diabetes and related markers are known to be associated to comorbidity and mortality [Bell et al., 2005],[Capri et al., 2006]. Therefore, we aimed to analyse these SNPs for association to survival and to elucidate the performance of the criteria to quantify the impact of SNP genotypes on survival phenotypes.

2.1.4 Literature search

2.1.4.1 Overview of available criteria

A general idea about the structure of criteria measuring the impact of covariates has been given in section 1.4.3.2. Criteria measuring the impact of covariates can generally be divided into three main categories:

1) likelihood-based criteria
2) measures of discrimination
3) variance-based criteria

For survival data analysis, a comparison of eight criteria from these three fields, that were at least discussed until 1994, is given in [Schemper and Stare, 1996]. More were invented and improved later. Some criteria investigated during literature search are mentioned in Table I. Some of which, like the variance-based criteria, already face a long history of development and improvement. Two criteria, *D* and the *Brier Score*, are difficult to categorise into one of the upper groups. Their structure is variance-based but originates from a discrimination point of view.

Table I: A selection of criteria proposed for judging the impact of genetic variants.

Characterisation	**Criterion**	**Reference**
Likelihood-based criteria:	*AIC* or *BIC*	e.g. [Harrell, Jr., 2001]
	R^2_{LR}	[Maddala G.S., 1983], [Magee, 1990]
	R^2_N	[Nagelkerke, 1991]
	$\rho^2_{W,A}$ or ρ^2_{PM}	[Kent and O'Quigley, 1988]
Measures of discrimination:	*AUC*	e.g. [Harrell, Jr., 2001]
	Somer's D_{XY}	[Somers, 1962]
	Harrell's C	[Harrell, Jr. et al., 1982]
	D_1 or *SEP*	[Sauerbrei et al., 1997]
Variance-based measure of discrimination:	*D*	[Royston and Sauerbrei, 2004]
	Brier Score	[Graf et al., 1999]
Variance-based criteria:	$k_{d.norm}$	[Stark, 1997]
	V and V_w	[Schemper and Henderson, 2000]
	R^2_{sch}	[O'Quigley and Xu, 2001]

2.1.4.2 Criteria selection

As described in section 1.4.3.2, the general focus was set on measures of predictive accuracy or explained variation, hence, variation-based criteria. Likelihood-based criteria are often difficult in interpretation or need to account for additional correction factors (see e.g. section 1.4.3.2). Measures of discrimination are not eligible for non-categorical phenotypes. The literature search yielded three variance-based criteria: $k_{d.norm}$ [Stark, 1997], V [Schemper and Henderson, 2000], and R^2_{sch}, defined as R^2 based on Schoenfeld residuals [O'Quigley and Xu, 2001]. These three criteria that have been proposed for quantification of the contribution of covariates to survival phenotypes allow for interpretation as percentage. They are described in detail below (section 2.2.1). All three criteria incorporate the estimated effect size $\hat{\beta}$ from the Cox model, without or with adjustment for additional covariates. The two criteria, *D* and the *Brier Score*, are each closely related to one of the criteria chosen for closer investigation in simulation studies: A direct correspondence between the *Brier Score* and *V* has been shown, and *D* is closely related to $k_{d.norm}$.

2.2 *Methods*

2.2.1 The three selected criteria

2.2.1.1 Criterion based on cumulated hazard ($k_{d.norm}$)

The first criterion is based on deviance residuals, which are derived from martingale residuals. Martingale residuals are defined per individual *i* as the difference between the observed survival status, given by the status indicator δ_i=I[*individual i has an event*], and the cumulative hazard at the observed time t_i [Therneau et al., 1990]: $M_i = \delta_i - \Lambda(t_i)$, for i=1,…, n subjects. Martingale residuals may be applied to

investigate the functional form of covariates [Therneau et al., 1990]. However, they have a highly skewed distribution and are not suitable for definition of a performance criterion like a measure of explained variation.

The following transformation of Martingale residuals yields the deviance residuals, which may also be interpreted as contributions to the deviance of the model:

$$dev.res_i = \mathrm{sgn}(M_i)\sqrt{-2(M_i + \delta_i \ln(\delta_i - M_i))}$$

Square root and logarithm, here, result in a more normalised distribution of deviance residuals, compared to Martingale residuals. For the definition of a criterion of prognostic value, it has been proposed to apply an absolute loss function on deviance residuals to measure the difference between null model and covariate model. With $dev.res_i$ being the residual from the null model and $dev.res_{i|X}$ the residual from the covariate model, the criterion $k_{d.norm}$, according to [Stark, 1997], is written as

$$k_{d.norm} = \frac{\sum_{i=1}^{n} \left| dev.res_i - dev.res_{i|X} \right|}{\sum_{i=1}^{n} \left| dev.res_i \right|}$$

As $k_{d.norm}$ incorporates the full vector of estimated effect sizes $\hat{\beta}$ from the Cox model in the above stated cumulative hazard function $\Lambda(t_i)$ when computing M_i, this criterion measures the impact of the full set of covariates in the case of more than one covariate in the model.

2.2.1.2 Criteria based on variation of individual survival curves (V and V_w)

The second criterion is based on measuring the weighted difference in the variation of the individual survival curves [Schemper and Henderson, 2000], for which a direct correspondence to the Brier score was shown [Gerds and Schumacher, 2006]. The variation of individual survival curves dependent on time *t* is defined as *S*(*t*)(1-*S*(*t*)). A

criterion of relative gain is then formulated as the relative difference of the integrated and weighted variance over the complete observation time $[0;\tau]$ between its expectation $E_X\left[S(t|X)\{1-S(t|X)\}\right]$ in the covariate model and the null model. Introduction of weighting function $f(t)$ leads to the definition of criterion V:

$$V(\tau)=\frac{\int_0^\tau S(t)\{1-S(t)\}f(t)dt-\int_0^\tau E_X\left[S(t|X)\{1-S(t|X)\}\right]f(t)dt}{\int_0^\tau S(t)\{1-S(t)\}f(t)dt}.$$

The complete estimation equation can be found in [Schemper and Henderson, 2000]. The weighting function $f(t)$ is introduced to reduce dependence on censoring in the data and incorporates the “reverse Kaplan-Meier estimator” [Schemper and Smith, 1996], [Altman et al., 1995]. Note that estimation is only performed for event time points $t_{\delta_i=1}$.

A slight modification of V, denoted as V_w, is obtained if the order of the operations is exchanged, i.e. if the integration is performed for weighted relative differences in variances:

$$V_w(\tau)=\frac{\int_0^\tau \frac{S(t)\{1-S(t)\}-E_X\left[S(t|X)\{1-S(t|X)\}\right]}{S(t)\{1-S(t)\}}f(t)dt}{\int_0^\tau f(t)dt}$$

Again, as V and V_w involve the full vector of estimated effect sizes $\hat{\beta}$ from the Cox model in the above stated survival function $S(t)$, these criteria measure the impact of the full set of covariates in the case of more than one covariate in the model.

2.2.1.3 Criterion based on variation of Schoenfeld residuals (R^2_{sch})

The third criterion is based on Schoenfeld residuals [Schoenfeld, 1982], which are the summands of the score function derived from the partial likelihood. These residuals measure the difference between the observed covariate values at event times $t_{\delta_i=1}$ and their expectations given the estimated $\hat{\beta}$ from the Cox model. Hence, Schoenfeld residuals are defined per covariate X_k as a vector over event time points $t_{\delta_i=1}$:

$$res\left(\hat{\beta}_k, t_{\delta_i=1}\right) = X_k\left(t_{\delta_i=1}\right) - E\left(X_k \middle| t_{\delta_i=1}, \hat{\beta}_k\right) = X_k\left(t_{\delta_i=1}\right) - \sum_{t_j \geq t_i} X_k\left(t_j\right) \frac{\exp\left(\hat{\beta}_k X_k\left(t_j\right)\right)}{\sum_{t_l \geq t_i} \exp\left(\hat{\beta}_k X_k\left(t_l\right)\right)}$$

As in the context of linear models, a criterion similar to ordinary R^2 can be defined based on squared residuals at event times [O'Quigley and Flandre, 1994]. For the calculation of the necessary null model residuals, covariate values are expected not to be associated with time and are randomly assigned over event time points dependent on the survival distribution. This becomes clear by setting $\hat{\beta} = 0$ in the upper residual definition. In case of tied failure times, residuals can be split randomly between the observations.

For more than one covariate, the Schoenfeld residual definition can be extended to measure differences in the prognostic index $\eta = \hat{\beta}'X$ instead of differences in a single covariate [Andersen et al., 1983]. The residual for the prognostic index is therefore the linear combination $\hat{\beta}' res\left(\hat{\beta}, t_{\delta_i=1}\right)$. Furthermore, introduction of weights $w\left(t_{\delta_i=1}\right)$, defined as the difference in step height of the marginal Kaplan-Meier curve at times $t_{\delta_i=1}$, reduces dependence on censoring. The criterion R^2_{sch} defined by [O'Quigley and Xu, 2001] is then written as:

$$R^2{}_{sch}\left(\hat{\beta}\right) = 1 - \frac{\sum_{\delta_i=1}\left[\hat{\beta}' res\left(\hat{\beta}, t_{\delta_i=1}\right)\right]^2 w\left(t_{\delta_i=1}\right)}{\sum_{\delta_i=1}\left[\hat{\beta}' res\left(0, t_{\delta_i=1}\right)\right]^2 w\left(t_{\delta_i=1}\right)}.$$

Therefore, the Schoenfeld residuals allow for both measuring a single covariate's impact while adjusting for other covariates as well as measuring the impact of a full set of covariates.

2.2.2 Simulation studies

In order to evaluate these criteria with respect to their performance in genetic association studies, simulation studies were performed for biallelic single SNPs. Different scenarios were defined through different genetic effect models, minor allele frequencies (*MAF*) of SNPs under the assumption of Hardy-Weinberg equilibrium (*HWE*) and different effect sizes, as well as failure times created for varying censoring percentage and type of censoring mechanism. For each scenario, 200 datasets each consisting of n=1000 observations were simulated.

2.2.2.1 Simulation of genetic variants

In order to create situations comparable to SNP association studies, SNP genotype data were generated as random variables $X \in \{0, 1, 2\}$ with X representing the number of copies of the minor allele. Probabilities for X, which correspond to the genotype frequencies (π_0, π_1, π_2), were calculated under the assumption of perfect *HWE* for *MAF* of 5%, 10%, 25% and 50% as $\pi_0=(1-MAF)^2$, $\pi_1=2*MAF*(1-MAF)$ and $\pi_2=MAF^2$.

Variance-based criteria, e.g. R^2 in linear regression, also depend on the variance of the covariate. The assumption of perfect *HWE* restricts the variance of the genetic covariate. Therefore, additional simulations were performed for situations with

violation of *HWE* with genotype frequencies (π_0, π_1, π_2) set to (0.6, 0.3, 0.1) (i.e. *MAF*=25%), (0.3, 0.4, 0.3) (i.e. *MAF*=50%) and (0.2, 0.7, 0.1.) (i.e. *MAF*=45%). These genotype frequencies yielded p-values below 10^{-9} for Fisher's exact test for *HWE*.

Simulations were performed for additive, dominant and recessive genetic effect models with corresponding coding of the genetic covariate. The *HR* for the association of the genetic variant was varied as $HR \in \{1.25, 1.5, 2, 4, 8\}$. Especially genetic effect sizes between 1 and 2 are expected to be realistic effect sizes in SNP analysis of complex diseases, whereas higher effect sizes may occur for major genes or in extreme settings. For exploring limitation behaviours of the three criteria, also a hazard ratio of *HR*=128 was simulated, which is a very extreme effect size and most unrealistic.

2.2.2.2 Simulation of survival outcome

Simulation of survival outcomes as combination of failure time and status indicator required generation of two random variables for each individual *i, i*=1,…,*n*: one event time and one for censoring time. Event times T_i were generated as exponentially distributed random variables according to [Bender et al., 2005]. The generation of censoring times C_i depended on the chosen censoring mechanism as described below. The final failure time t_i and the survival status indicator δ_i per individual were then calculated as $t_i=\min(T_i, C_i)$ and $\delta_i=I[T_i \leq C_i]$.

2.2.2.3 Simulation of censoring times

Censoring was chosen to imitate two common study designs: Setting 1 (fixed censoring): All individuals enter the study at the same time and general censoring is

applied at time τ. Hence, time of censoring is constantly $C=\tau$ for all individuals i. This corresponds e.g. to cross-sectional surveys including follow-up. Setting 2 (random censoring): Individuals enter the study continuously over time until general censoring is introduced at time τ. In this setting, the time of censoring is a uniformly distributed random variable $C\sim\text{Unif}[0,\tau]$. This situation with ongoing recruiting is often found in clinical studies. For each scenario, values of τ were varied to obtain censoring percentages of 10%, 25%, 50%, 80%, and 90%.

2.2.2.4 Extended simulation scenarios with continuous covariates

In genetic association analyses, it is often necessary to also include environmental covariates which follow a continuous distribution. Therefore, the performance of the presented criteria was also evaluated for continuous covariate distributions. Chosen covariate distributions were as follows:

1) a standard normal distribution: $X\sim N(0;1)$
2) a standard uniform distribution: $X\sim\text{Unif}[0;1]$
3) a normal distribution with variance 1/12: $X\sim N(0;1/12)$

These distributions were chosen for the following reasons:

1) Represents a standard covariate in epidemiologic studies.
2) May be a covariate that is also included in the study design. Study participants, for example, are often chosen uniformly distributed over age.
3) Has equal variance as 2 and has been added for comparison between normal and uniform distribution.

The assumed hazard ratios in these settings were similar to the simulations with trichotomous SNP covariates with hazard ratios set to 1.25, 1.5, 2, 4, and 8. However, it was necessary to reduce the extreme hazard ratio from the genetic settings (*HR*=128) to *HR*=32 in order to guarantee convergence of the model estimation.

2.2.2.5 Bivariate simulations with genetic variants and a continuous covariate

Finally, to mimic a more realistic situation where a quantitative prognostic factor explains a substantial proportion of the survival outcome, simulation scenarios combining genetic and quantitative variables were created. Here, a standard normally distributed covariate $X \sim N(0;1)$ was simulated with $\log(HR)=1$, thus a *HR* of 2.72, additional to the genetic variant as described before (see section 2.2.2.1).

2.2.2.6 Statistical analysis and simulation summary

For each dataset in the SNP simulations, the effect estimates $\hat{\beta}$ of the SNP from fitting a Cox-model and the criteria $k_{d.norm}$, V and R^2_{sch} were calculated under the assumption of an additive, dominant or a recessive genetic effect model. The criterion V_w, as a possible alternative to V, was also calculated, but is not in the focus of this thesis as it is very similar to V.

The additional settings for SNPs violating the *HWE* assumption were analysed only for additive effects because deviations from *HWE* are mainly visible in trichotomous coding of the SNP covariate.

For each dataset in the simulations with a continuously distributed covariate, the corresponding estimates of log(*HR*) per unit increase were estimated. For the simulation scenarios combining both a genetic and a quantitative variable, the Cox

model included both the SNP genotype and the quantitative variable. The criteria values then depict the combined impact of both. For each simulation scenario, mean and standard deviation of the effect estimates and each criterion were computed across the 200 simulations.

Simulated datasets where the Cox proportional hazards model showed convergence problems were discarded and replaced in order to guarantee an overall number of 200 evaluated datasets. Convergence problems occurred mainly in case of almost monomorphic behaviour of the genetic variable in the event group (δ_i=1). Therefore, mainly recessive models for SNPs with low *MAF* were affected and recessive models were only generated for minor allele frequencies of 25% and 50%.

All simulations and statistical analysis were performed with the R software version 2.4.1[R Development Core Team, 2007].

2.2.3 Real data analysis

2.2.3.1 The KORA data S3/F3 for survival analysis

The KORA (Kooperative Gesundheitsforschung in der Region Augsburg) study divides into four baseline surveys conducted in the years 1984/85 (S1), 1989/90 (S2), 1994/95 (S3) and 1999/2001 (S4). These surveys have been conducted as population-based samples stratified by age and sex drawn from the local registries. The study region comprises Augsburg and its two surrounding counties. Study participants were invited to undergo medical examination in the KORA study centre. Persons who did not want to participate were asked to answer a non-responder questionnaire. Follow-up information in surveys S1-S3 was gathered through two postal questionnaires (GEFU1 and GEFU2) in the years 1997/98 and 2002/03. Indication of disease in these short questionnaires required validation through the

attending physician and written consent of the study participant. In case of death, the exact date of death was confirmed through death certificates from the local health departments. Therefore, exact dates of death or first diagnosis of disease, e.g. myocardial infarction, coronary heart disease or type 2 diabetes, are available for analysis within survival framework.

Furthermore, thorough 10-year-follow-up data are available for participants in KORA S3, who were invited to the KORA study centre for re-examination in the years 2004-2006. Therefore, more phenotypes are available for this follow-up than for GEFU data. More details on KORA as a research platform for health research can be found in [Holle et al., 2005].

Mortality data to apply the selected criteria were obtained from the KORA study S3. Mortality follow-up after 10 years and DNA samples were available for 4420 study participants as well as information on sex and age. A short description of the data available for analysis is given in Table II.

Table II: Description of the KORA S3-F3 data available for mortality analysis

Sample size	4420
Men	50.4%
Average age at baseline (S3)	49.5 years
Mortality	7.6%
Median follow-up of all survivors	3549 days

In total, 51 autosomal SNPs from different genes with possible or known associations with severe diseases or related parameters such as diabetes, myocardial infarction or dyslipidemia were chosen. As these diseases and related conditions are supposed to be related to mortality, it was hypothesised that some of these SNPs would be

associated with mortality. Genotyping had been performed using MALDI-TOF MS technology. Because association with mortality could affect the genotype distribution in the higher age group violation of *HWE* is a possible indicator for association to mortality. Therefore, deviations from *HWE* were tested by the exact *HWE* test described by [Emigh, 1980], but did not result in exclusion from analysis of the particular SNP.

2.2.3.2 Adding simulation of SNPs associated with mortality

As no strong association with mortality was expected for the investigated real SNPs, artificial SNPs with minor allele frequency of 25%, perfect *HWE* and three different magnitudes of additive genetic effect on the β level (i.e. multiplicative effect on hazard ratio scale) on mortality were created as follows:

First, genotypes of an initial artificial SNP were assigned to individuals based on the percentiles of ordered failure times from the KORA data. Percentiles were chosen corresponding to the genotype frequencies (i.e. for *MAF*=25%: π_0=56.25%, π_1=37.5%, π_2=6.25%). Homozygous carriers of the minor allele were chosen to be at the highest risk for death. Therefore, genotype value 2 was assigned to the 6.25% shortest survival times and genotype value 0 was assigned to the 56.25% longest survival times. No association of genotypes with other covariates was assumed. Given the high censoring percentage in the KORA data, this design, would have lead to an almost complete separation of cases from the group of genotype 0. This would have resulted in convergence problems for estimation of the association, especially for a dominant model. Therefore, a small random error was added on each genotype group. Variation of the percentage of randomly assigned genotypes from the original artificial SNP yielded SNPs with different degrees of association to mortality:

- To obtain strongest association to mortality, an “extreme SNP” was generated through random assignment of genotypes with probability of 2% to the genotype carrying one minor allele more or less than the original genotype.
- A “strong SNP” with lower but still strong association to mortality was obtained through random assignment of 25% of genotypes to any genotype with probabilities corresponding to the original genotype frequencies.
- A “moderate SNP” with moderate association to mortality was created the same way as the “strong SNP” but through random assignment of 50% of the original genotypes.

Note that the chosen simulation approach keeps *MAF*=25% for all artificial SNPs.

2.2.3.3 Statistical analysis and the impact of the genetic variants

In real data analysis, the Cox model was fitted for time since baseline survey S3. Each of the SNPs (51 real SNPs and three artificial lethal SNPs) was analysed unadjusted as well as adjusted for age and sex and values for $k_{d.norm}$, V and R^2_{sch} were calculated. Note that criteria values in the adjusted models depict the combined impact of SNP, age and sex.

As genotype distributions of SNPs with association to mortality are supposed to affect longevity, it is possible that dependence on age is already visible at baseline. Therefore, further investigation was conducted for real data SNPs:

1. In order to investigate whether any dependence of genotype distributions on age exists in real SNPs, linear models adjusted for sex were fitted against age at baseline.

2. Another Cox model was added for the real data SNPs with the time variable defined as *age at death* and adjustment for sex. This model was supposed to better account for this possible situation.

The significance level corrected for multiple testing was obtained according to Bonferroni as the significance level of 5% divided by the number of models multiplied by the number of independent marker loci. The latter was calculated through spectral decomposition of the SNP correlation matrix as described by [Li and Ji, 2005].

2.3 *Results*

2.3.1 Results from SNP simulation studies

2.3.1.1 Overview

As described in section 2.1.3, comparison of the performance of the selected criteria in genetic association analysis with survival phenotype was conducted for simulated data with different scenarios (varying *MAF*, genetic effect size, genetic effect model, censoring percentage and censoring mechanism). Evaluation of the three investigated criteria focused on the following items:

(a) reasonable values in the range [0;1]

(b) increasing values with increasing effect size.

(c) independence of censoring

Each of these items required inspections from several points of view and direct as well as indirect links were discovered. In the following, results are presented corresponding to the upper main items a, b and c.

Throughout the investigation, criterion V_w yielded values very close to those of V - with a tendency to be slightly smaller. Therefore, results for V_w are not presented in detail in the following.

2.3.1.2 Reasonable values in the range [0;1]

Observed range of criteria values

As a demand of interpretability as percentage of maximum possible impact on the phenotype, criteria values should range between 0 and 1, with 0 denoting no impact and 1 denoting 100% impact on the survival phenotype. In all SNP simulation scenarios, the mean values of all investigated criteria were limited to the interval [0;1]. V, however, did not cover the full range. Simulation results are presented in Figure 4 for the case of a SNP with *MAF*=25% under the assumption of fixed censoring.

V yielded generally low values and did not reach unity even for the most extreme scenario. For example, for the data presented in Figure 4, even for the highest effect size of *HR*=128, V hardly exceeded 60-65%, while R^2 and $k_{d.norm}$ approached 100%. For the lowest effect size *HR*=1.25, values close to zero were obtained for V and R^2_{sch} and values up to 10-15% for $k_{d.norm}$.

For single simulated datasets with low effect size, V and R^2_{sch} yielded slightly negative values (>-0.01). This cannot be seen from Figure 4 as the mean values over the 200 simulated datasets were positive. Values below 0 indicate a better fit of the null model than the covariate model, which may occur in cases of low association. As, for both criteria, mean values are positive, single values slightly below 0 may be interpreted as “no impact”.

Dependence on genotype variance

Criteria like R^2 in linear regression measure the reduction of variance in the phenotype due to the covariates in the model. Therefore, they also depend on the variance of the covariates.

For genetic variants, the variance of the genotype covariate X depends on the genotype frequencies and the assumed genetic effect model. The variance for qualitative covariates X is generally defined as:

$\mathrm{var}(X) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$, with n=sample size and $\overline{X}$ denoting the mean value of X

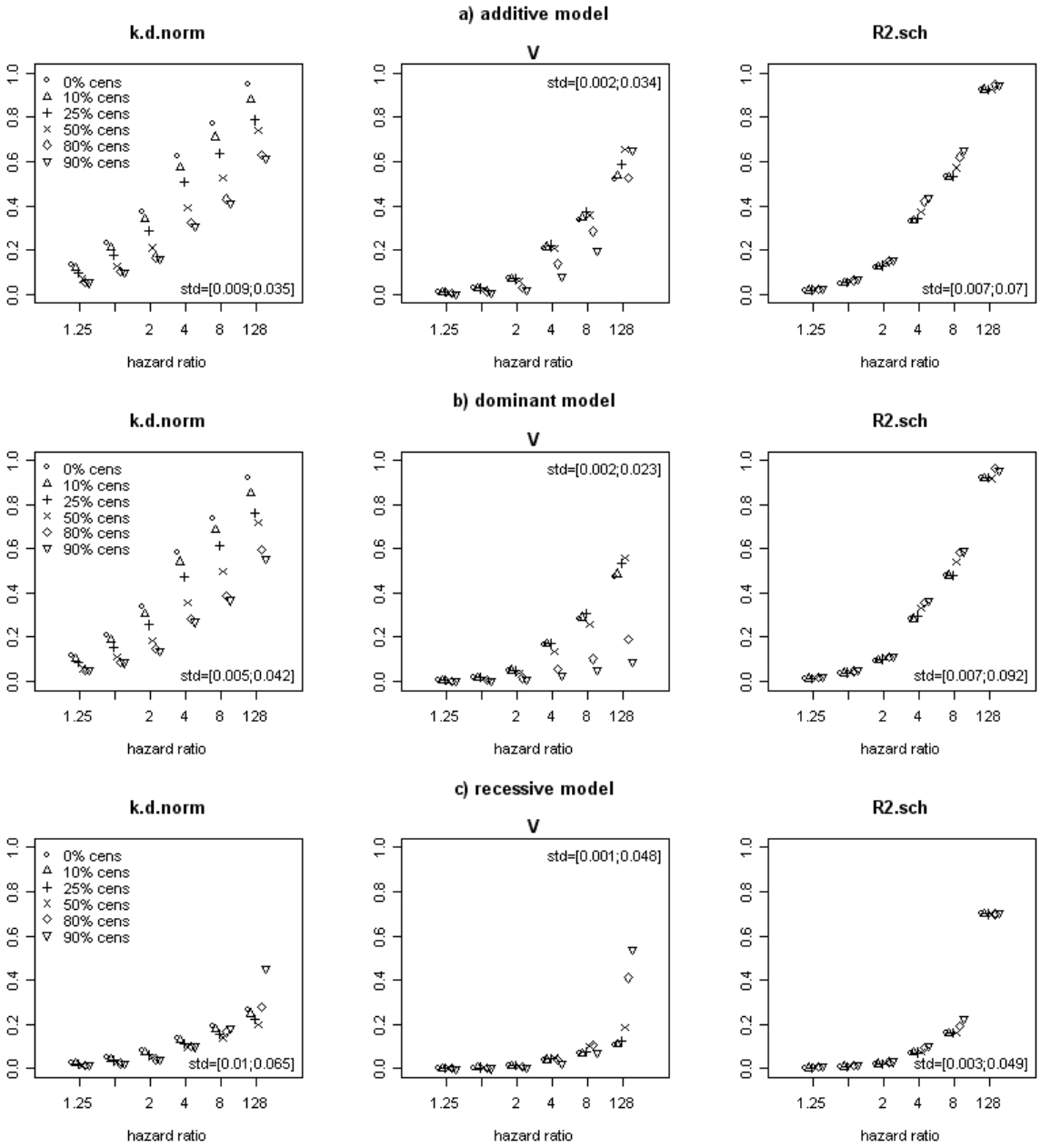

Figure 4: Survival data with fixed censoring were simulated *for a single covariate* $X \in \{0;1;2\}$ with sampling probabilities calculated from *MAF*=25%. Mean values of the *investigated criteria judging the impact of the single genetic covariate* (a) for additive and (b) dominant and (c) recessive effect models are plotted against hazard ratios (*HR*). For each value of *HR*, results for different censoring percentages (cens) are presented as different point characters (adding small scatter on the x-axis for differentiation). The range of standard deviations (std) averaged over the 200 simulated data sets per scenario is given in each panel.

For genotype data $\overline{X}$ = 2*MAF*, and, for trichotomous coding of the genotypes, the sum is calculated over genotypes. With n_0, n_1 and n_2 denoting the genotype counts, the upper formula, therefore, reduces as follows:

$$\mathrm{var}(X) = \frac{n_0}{n}(2*MAF)^2 + \frac{n_1}{n}(1-2*MAF)^2 + \frac{n_2}{n}(2-2*MAF)^2$$

The fractions, then, represent the genotype frequencies.

In case of perfect *HWE* (as in the majority of the chosen simulation scenarios), all genotype frequencies can directly be calculated from the *MAF* and the formula for the genotype variance can be directly calculated from this single parameter:

$$\mathrm{var}(X) = (1-MAF)*(2*MAF)^2 + 2*MAF*(1-MAF)(1-2*MAF)^2 + MAF*(2-2*MAF)^2$$

For dichotomous coding of SNPs, as in recessive or dominant genetic effect models, calculation of the variance is based on the formula for binomial data:

$\mathrm{var}(X) = n_1 * \left(1 - \frac{n_1}{n}\right)$, with n_1 denoting the number of observations with X=1

As n_1 either equals the sum of observed numbers of genotypes 1 and 2 (dominant model) or the number of genotypes 2 (recessive model), the derivation of the variance, again, is directly related to the *MAF* and therefore straightforward.

The variances calculated for the genetic covariates in the chosen simulation settings are shown in Table III.

The simulations revealed that all criteria increase with increasing *MAF* (assuming *HWE*) or, more generally speaking, increasing genotypic variance (independent of *HWE* assumption).

Table III: Genotype frequencies, *MAF* and variance of the genetic covariate in the chosen genetic simulation settings are shown.

Genotype frequencies for genotypes (0;1;2)	*MAF*	Variance of genetic covariate		
		Additive effect model	Dominant effect model	Recessive effect model
90.25%, 9.5%, 0.25%	5%	9.68%	8.85%	---
81%, 18%, 1%	10%	18.02%	15.41%	---
56.25%, 37.5%, 6.25%	25%	37.45%	24.64%	5.82%
25%, 50%, 25%	50%	50.05%	18.77%	18.77%
60%, 30%, 10%*	25%	45.05%	---	---
20%, 40%, 30%*	50%	60.06%	---	---
20%, 70%, 10%*	45%	29.03%	---	---

* SNP violating *HWE* with $p<10^{-9}$; simulations were only performed for the additive effect model

In Figure 5, this dependence is shown for the additive effect models. Dominant and recessive effect models have been excluded from Figure 5 due to the different scaling of the genetic covariate. In order to exclude any effects resulting from dependence on censoring, Figure 5 is only displayed for settings without censoring.

Highest criteria values were observed for SNPs with genotype frequencies set to (0.3; 0.4; 0.3), which is the setting that is close to an equally balanced design and therefore highest variance in the trichotomous SNP covariate. The relation between the criteria values and the variance of the genotypic covariate was found as an almost linear trend. The general dependence on the covariate's variance is also verified in the additional simulations for continuous covariate distributions, as described later.

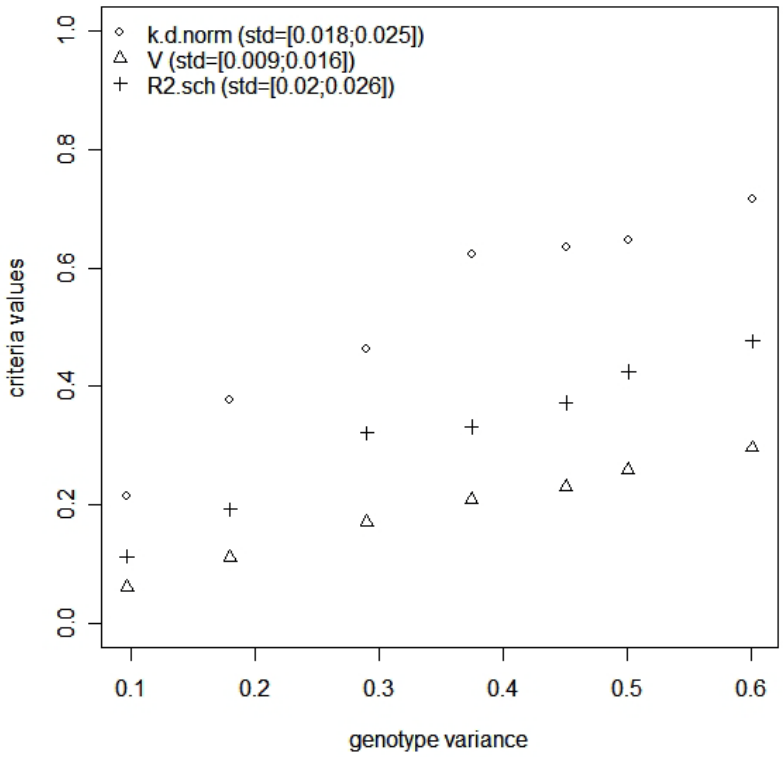

Figure 5: Average values of criteria are plotted for all additive effect models included in simulations with *HR*=4 and without censoring against the calculated variance of the genotypic covariate.

Summary of results: limitation behaviour

For all of the investigated criteria, the limitation to the range [0; 1], which is necessary for interpretation as percentage of explained variation, is fulfilled in the SNP simulation scenarios. A general increase with increasing genotype variance has been observed, which is in line with the expected characteristics of variance-based criteria. In case of very small effects, $K_{d.norm}$ and R^2_{sch} yield values close to zero and approach unity for extremely large effects. Values of V, however, are generally low and coverage of the full range is not verified. Slightly negative values of V and R^2_{sch} in may occur in the case of low effect size and indicate poor impact of the genotype on the survival phenotype.

2.3.1.3 Dependence on the genetic effect size

General dependence on the genetic effect size

As can be seen from Figure 4, all criteria generally increase with the genetic effect size. The increase, however, seems also to be related to the underlying genetic effect model. Whereas it is clearly visible for additive and dominant genetic effect models, it is less clear for recessive effect models. This finding has been observed throughout the simulations.

Dependence on variance of effect size

The link between the criteria and effect size was further investigated through analysis of standard deviations of both, criteria and effect size, over the 200 simulated datasets within each scenario. Averaged standard deviations of the criteria in each scenario were generally low (Figure 4). However, comparison of results over all simulation scenarios showed that R^2_{sch} is most sensitive to the variance of the estimated effect size, which also increases with increasing censoring percentage, i.e. increasing uncertainty in the data. For example, for models with *HR*=8 in combination with 80% censoring, a high standard deviation of the estimated $\hat{\beta}$ in the 200 datasets yielded up to 12% standard deviation for R^2_{sch}. Standard deviations of the criteria V (and V_w) and $k_{d.norm}$ also showed dependence on censoring percentage and the estimated effect size but hardly exceeded 4%.

Summary of results: dependence on effect size

For all criteria, the values increase for increasing effect size as postulated. The underlying genetic effect model, however, seems to play an important role.

Particularly the recessive effect model exhibits lower values in all criteria than the other genetic effect models. This is obviously due to the low genotype variance in recessive effect models (see section 2.3.1.2). Furthermore, R^2_{sch} is most sensitive to variations of the estimated effect size. Simulation settings with high variation of the estimated genetic effect size were observed to result in increased variation of R^2_{sch}. Variation of the estimated genetic effect size was found to be directly related to the censoring percentage in the data. Therefore, indirect influence of censoring on values of R^2_{sch} is possible, as described in the next section.

2.3.1.4 Dependence on censoring

Dependence on censoring mechanism

In first instance, the two censoring mechanisms (fixed censoring and random censoring) were compared. The comparison revealed no major differences between the two scenarios in the general settings with censoring percentages below 80%. The criteria results from settings shown in Figure 4 for fixed censoring are displayed for random censoring on the same data in Figure 6. Only slight differences were observed: for censoring ≥80, the decrease of V in the dominant model with high effect size was less and R^2_{sch} was strictly robust against censoring below 80%. The discrepancies, however, were small. Hence, all further general examination was limited to simulations with fixed censoring a time $C=\tau$, which is the more common situation in genetic association studies within general populations and regular follow-ups.

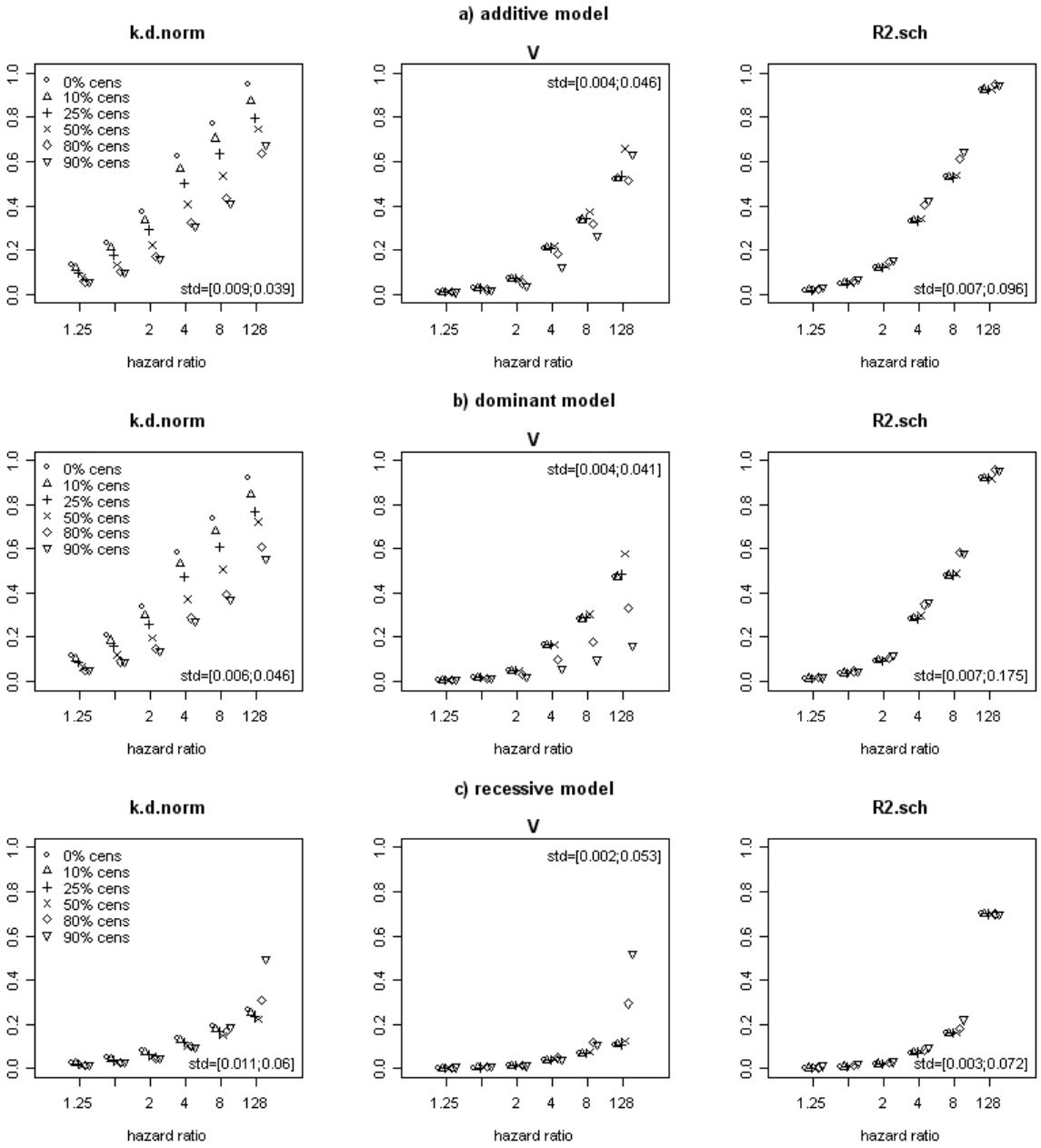

Figure 6: Survival data with *random censoring* were simulated *for a single covariate* $X \in \{0;1;2\}$ with sampling probabilities calculated from *MAF*=25%. Mean values of the *investigated criteria judging the impact of the single genetic covariate* (a) for additive and (b) dominant and (c) recessive effect models are plotted against hazard ratios (*HR*). For each value of *HR*, results for different censoring percentages (cens) are presented as different point characters (adding small scatter on the x-axis for differentiation). The range of standard deviations (std) averaged over the 200 simulated data sets per scenario is given in each panel.

Dependence on censoring percentage

When investigating the dependence of $k_{d.norm}$, V and R^2_{sch} on the censoring percentage, a distinction between the genetic effect models was necessary, again, due to the different range of criteria values.

Values were high for all three criteria under the assumption of an additive or dominant effect model as long as censoring percentage was lower than 80%. The additive model generally yielded slightly higher values. Only for data with high censoring (≥80%), an eminent drop in V was observed in the dominant model. R^2_{sch}, in contrast, was even slightly higher for the dominant model than for the additive model in the extreme settings with HR=128 when combined with censoring ≥80%.

For recessive effect models, all criteria had low values although a general increase was observed for increasing effect size (Figure 4). In contrast to the dominant model, V showed an eminent increase for effect size HR=128 in combination with high censoring percentage (≥80%).

Especially for the additive (as well as for the dominant) effect model, a strong dependence of $k_{d.norm}$ on the censoring percentage could be seen, while the other criteria were more robust against censoring, at least up to 50% (Figure 4). In the recessive model, this dependence is not obvious.

For R^2_{sch}, a minor dependence on censoring may be noticed for high effect sizes in combination with high censoring.

Summary of results: dependence on censoring (c)

A strong dependence on censoring is generally observed for $k_{d.norm}$, while V and R^2_{sch} are more robust. With high censoring (≥80%), however, values of V and R^2_{sch} are also affected. In case of high censoring percentage in the data, the structure of the

underlying censoring mechanism influences this degree of change marginally. Both criteria are slightly less affected, if censoring is random and not conducted at a fixed point on the time scale.

2.3.2 Results from simulations for a single continuous covariate

Simulations for continuous covariates were conducted in order to observe the criteria's behaviour in case of non-genetic covariates as they in a Cox model adjusted for environmental covariates. Chosen covariate distributions were:

1. a standard normally distributed covariate X~N(0;1)
2. a standard uniformly distributed covariate X~unif[0;1]
3. a normally distributed covariate X~N(0;1/12)

In case of a single standard normally distributed covariate, the criterion $k_{d.norm}$ exceeded the desired boundary of 1 in most situations with HR=32 and low censoring (Figure 7). The other criteria, however, hold the desired limit. In none of the genetic simulation scenarios, this exceeding of the limit was observed.

For the standard normal covariate distribution, again, V yielded lowest values among the three criteria with a maximum of 71%, while R^2_{sch} reached up to 95%. Although a general increase with increasing effect size was observed, all criteria in this setting, to some extend, depended on the censoring percentage. Therefore, the precision of the estimation of effect sizes was also examined. While the mean values of the estimated hazard ratios per scenario were independent of censoring, their standard deviations increased with increasing censoring percentage (Table IV). In addition, a general dependence of standard deviations on the underlying effect size was

observed. Highest standard deviations were obtained for maximum censoring percentage in combination with maximum effect size.

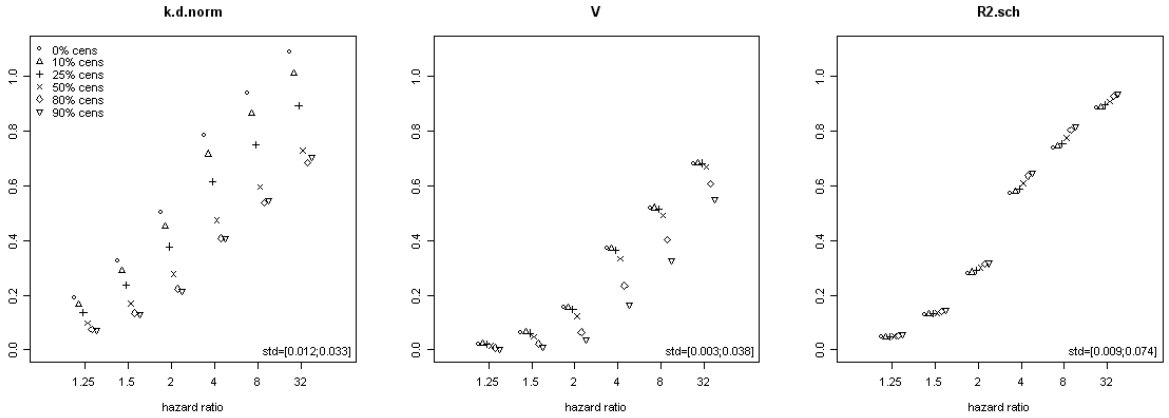

Figure 7: Criteria values for a non-genetic univariate model with one *single standard normal covariate* are plotted as average over 200 simulation scenarios against the *HR* from the corresponding setting with fixed censoring. For each value of *HR*, results for different censoring percentages (cens) are presented as different point characters (adding small scatter on the x-axis for differentiation). The range of standard deviations (std) averaged over the 200 simulated data sets per scenario is given in each panel.

Table IV: Standard deviations of estimated effect sizes β for a standard normally distributed covariate over 200 simulations dependent on censoring percentage (fixed censoring) and effect size.

	cens=0%	**cens=10%**	**cens=25%**	**cens=50%**	**cens=80%**	**cens=90%**
β=log(1.25)	0.033	0.036	0.040	0.047	0.074	0.097
β=log(1.5)	0.034	0.036	0.039	0.048	0.077	0.109
β=log(2)	0.038	0.039	0.043	0.048	0.075	0.101
β=log(4)	0.045	0.045	0.048	0.057	0.084	0.116
β=log(8)	0.065	0.065	0.065	0.072	0.105	0.138
β=log(32)	0.087	0.091	0.097	0.114	0.162	0.231

For the settings with (a) a uniformly and (b) a normally distributed covariate but both with equal variance (var=1/12), criteria values were similar but generally lower than for the setting with a single standard normally distributed covariate (Figure 8). In these settings too, a slight dependence of the criteria on censoring was observed. Like for the standard normally distributed covariate, standard deviations of effect estimates within each setting were found to be dependent on censoring percentage, while mean values were precise (data not shown).

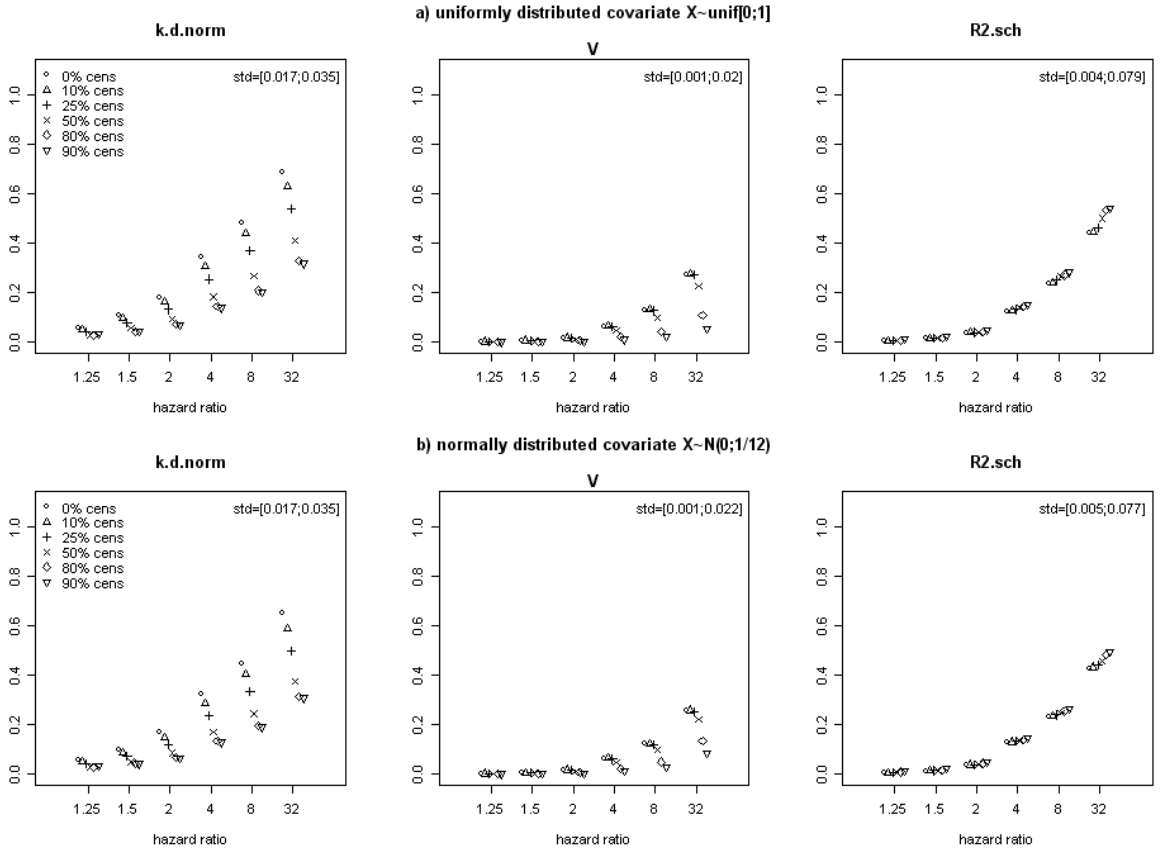

Figure 8: Criteria values for a non-genetic univariate model with a) a *single standard uniform covariate* X~unif[0;1] and b) a *single normally distributed covariate* $X \sim N(0;1/12)$ are plotted as average over 200 simulation scenarios against the *HR* from the corresponding setting. For each value of *HR*, results for different censoring percentages (cens) are presented as different point characters (adding small scatter on the x-axis for differentiation). The range of standard deviations (std) averaged over the 200 simulated data sets per scenario is given in each panel.

In conclusion, the settings with a single continuous covariate refined the insight into the characteristics of the investigated criteria:

(a) Limitation to the range [0;1] is not generally guaranteed for $k_{d.norm}$. In presence of very strong effects and high variance of the covariate, it systematically exceeds the desired maximum of 1. Coverage of the whole range [0;1] is doubted for V, as it hardly exceeds 70% even in extreme situations. Only R^2_{sch} seems to fully exploit the whole range while keeping the limitation.

(b) All criteria increase with increasing effect size and increasing variance of the covariate.

(c) Dependence on censoring, to some extend, is observed for all criteria. Especially for V and R^2_{sch}, this dependence may be due to increasing standard deviation of the estimated effect size with increasing censoring percentage. $k_{d.norm}$, however, already decreases in case of small censoring percentage in the data, where standard deviations of the estimated effect size are only slightly increased.

2.3.3 Results from combining a SNP with a strong continuous predictor

In the case that adjustment for environmental covariates is required in genetic association studies, it is also important to know the criteria's behaviour in this situation. Therefore, simulations with both, a SNP and continuous environmental covariate, were conducted. Here, a strong effect of the environmental covariate on the survival phenotype was assumed, as it is often the case when age or blood markers are included into a model. Adding the presence of a strong continuous predictor to the SNP simulation scenarios was established through incorporation of a

standard normally distributed covariate with effect size *HR*=exp(1)=2.72 to the previously investigated single SNP scenarios.

It should be noted that all criteria, in this setting, measure the impact of the combination of both covariates on the phenotype. The effect of the continuous covariate was kept fixed in order to investigate the criteria's behaviour in case of changes in the genetic effect size alone.

Due to the high impact of the continuous covariate, all criteria values rose to a higher level (Figure 9). The increase due to increasing genetic effect size, however, was low, but still observable in the additive or dominant genetic effect model, at least in case of strong effects. Within recessive effect models, however, none of the criteria seemed to change with increasing genetic effect size.

As in the setting with a single standard normally distributed covariate, $k_{d.norm}$ tended to exceed the desired maximum value of 1 with increasing genetic effect size in combination with low censoring percentages.

Dependence on the censoring percentage was mainly visible for $k_{d.norm}$ under the assumption of an additive effect model. However, none of the criteria was unaffected by censoring. As in the simulations for single continuous covariates, the standard deviation of the estimated effect sizes, calculated over the 200 simulations within each setting, was dependent on the censoring percentage. Standard deviations of the estimated effect sizes were higher for dominant and recessive effect models than for additive effect models and also resulted in increased standard deviations of the criteria in each setting (Figure 9).

Furthermore, average values of estimated genetic effect sizes within each setting were observed to not be robust against the censoring percentage in presence of a strong continuous predictor. Biases in averaged effect estimates of the genetic covariate were mainly observed for the recessive model.

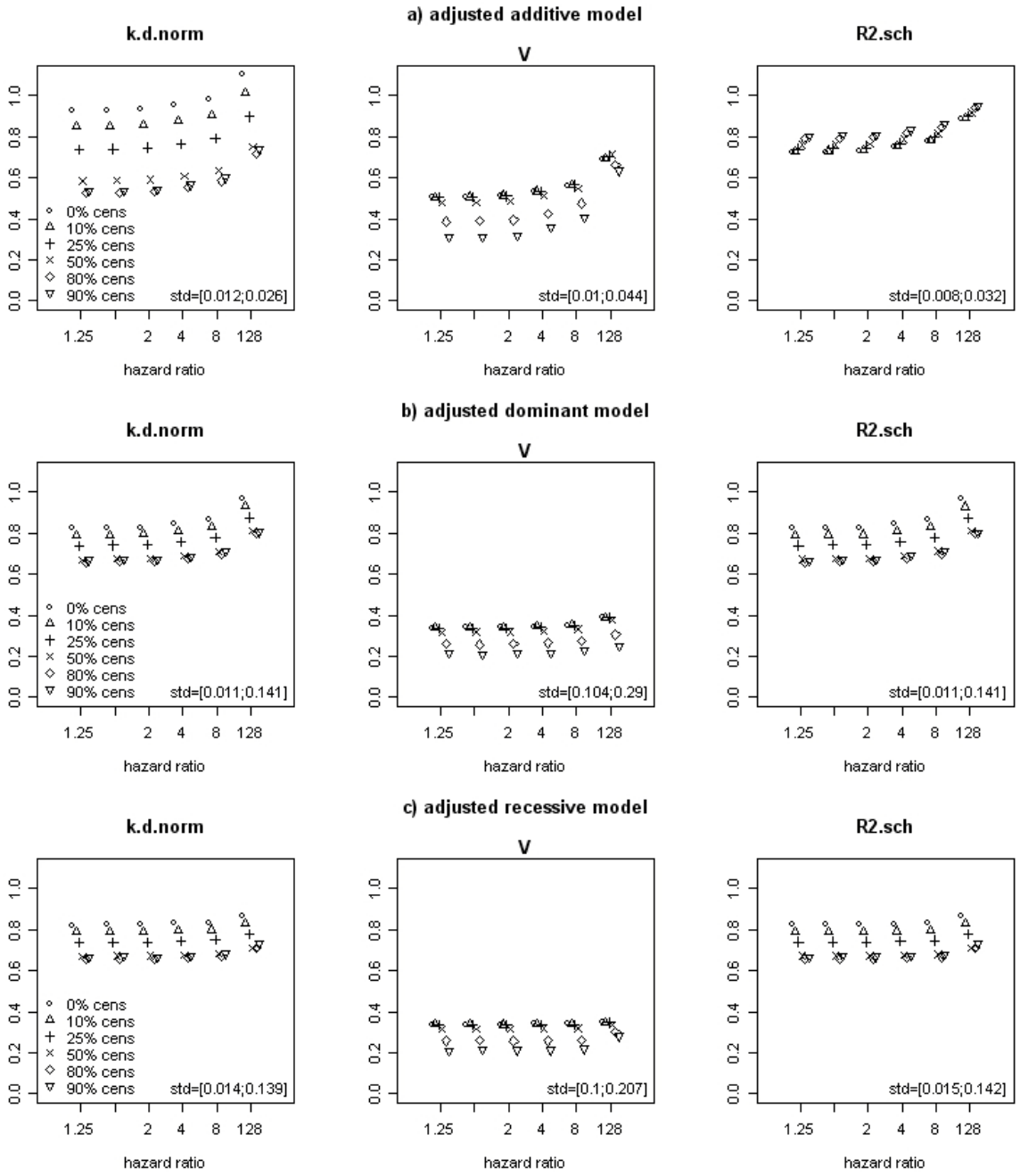

Figure 9: Survival data were simulated for *a genetic covariate* $X \in \{0;1;2\}$ with sampling probabilities calculated from *MAF*=25% *and a standard normally distributed covariate* with fixed *HR*=exp(1). Mean values of the investigated criteria judging the *impact of the genetic and the quantitative covariate combined* (a) for additive, (b) dominant and (c) recessive effect models are plotted against hazard ratios (*HR*) of the genetic covariate. For each value of HR, results for different censoring percentages under the assumption of fixed censoring (cens) are presented as different point characters (adding small scatter on the x-axis for differentiation). The range of averaged standard deviations (std) over the 200 simulated data sets per scenario is given in each panel.

From the setting combining SNPs with a strong continuous predictor, it is concluded that the impact of single SNPs may be masked by the continuous covariate, especially when the genetic effect is small. In part, this may be explained by biased estimates of the genetic effect size. Biases in the genetic effect estimates are also related to the censoring percentage. Therefore, indirect dependence on the censoring percentage is observed.

2.3.4 Results from real data analysis

2.3.4.1 KORA, real SNP analysis

From the 51 SNPs that were initially chosen for analysis, five SNPs show violation of *HWE*, 11 SNPs had *MAF*<10%, four SNPs show no homozygous carriers of the minor allele (rs17366743, rs2066860, rs707922 and ADA22) and gave rise to unreliable estimates for the recessive model (Table V). Only ADA22 and rs2225995 had a genotyping success rate below 90%. No SNPs, however, were excluded from analysis.

Four SNPs (rs174570, rs2225995, rs8065316, ADA22) showed association to age at baseline. However, none of these SNPs was significantly associated to mortality in any of the investigated Cox models.

Further investigation is restricted to the eight SNPs that showed significant results on significance level of 5% in any of the association models. With the significance level corrected to 0.2% for six models and 40 independent SNPs derived from the correlation matrix, none of these detected significances in these SNPs, however, remained after correction for multiple testing.

Only rs6808, rs3834458 and the artificially generated SNPs yielded significant association to a 5% significance level in SNP analysis unadjusted for age and sex

Table V: Description of selected SNPs investigated for this mortality study in KORA S3 (MONICA Augsburg Cohort Study S3, 1984-2005) including position, genotyping success rate (SR), minor allele frequency (*MAF*), p-value resulting from exact test for violation of the Hardy-Weinberg-Equilibrium (*HWE*) and disease relevant association or indication as possible candidate from genome-wide analysis screen (GWA).

SNP_ID	Chromosome / Position	Gene	SR	*MAF*	*HWE*	Association to
rs1234313	1 / 171432870	*TNFSF4*	96.6	30.7	0.8855	Atherosclerosis
rs3850641	1 / 171442455	*TNFSF4*	96.2	15.1	0.6756	Atherosclerosis
rs1234315	1 / 171445086	*TNFSF4*	96.7	45.4	0.5370	Atherosclerosis
rs7566605	2 / 118552495	*INSIG2*	91.3	33.4	0.1284	Obesity
rs17300539	3 / 188042154	*APM1*	98.2	8.6	0.7001	Plasma adiponectin
rs17366743	3 / 188054783	*APM1*	98.7	2.6	0.7652	Plasma adiponectin
rs1800791	4 / 155702759	*FGB*	97.6	13.5	0.4735	Fibrinogen, myocardial infarction
rs1800788	4 / 155703364	*FGB*	90.9	17.9	0.5205	Fibrinogen, myocardial infarction
rs4463047	4 / 155714983	*intergenic*	91.7	13.0	0.2676	Fibrinogen, myocardial infarction
rs6825454	4 / 155720638	*intergenic*	98.1	25.5	0.2154	Fibrinogen, myocardial infarction
rs2070022	4 / 155724398	*FGA*	99.0	16.3	0.2025	Fibrinogen, myocardial infarction
rs2070016	4 / 155729764	*FGA*	91.8	14.9	0.2651	Fibrinogen, myocardial infarction
rs2066861	4 / 155746886	*FGA*	98.9	24.4	1	Fibrinogen, myocardial infarction
rs2066860	4 / 155748924	*FGG*	93.8	3.8	0.1936	Fibrinogen, myocardial infarction
rs1800792	4 / 155753858	*FGG*	93.1	46.6	0.0450	Fibrinogen, myocardial infarction
rs10012555	4 / 155870881	*intergenic*	94.3	11.4	0.5903	Fibrinogen, myocardial infarction
rs10520818	5 / 15642917	*FBXL7*	98.8	12.2	0.8873	GWA
rs707922	6 / 31733486	*APOM*	99.5	6.6	0.5398	HDL cholesterol
rs1800796	7 / 22732771	*IL6*	99.2	5.7	0.4005	Diabetes
rs1800795	7 / 22733170	*IL6*	99.5	43.7	0.7827	Diabetes
rs1105218	8 / 12700580	*intergenic*	97.4	15.4	0.7251	GWA
rs1248696	10 / 79286611	*DGL5*	99.3	9.6	0.7934	inflammatory bowel disease
rs1528133	11 / 8106329	*RIC3*	99.3	6.3	0.0195	nicotinic acetylcholine receptors
rs2071212	11 / 61287413	*C11orf9*	94.7	32.3	0.9437	Fatty acids, atopy
rs174528	11 / 61300075	*C11orf9*	97.3	33.6	0.4519	Fatty acids, atopy
rs174538	11 / 61316656	*C11orf9*	97.3	26.6	0.8760	Fatty acids, atopy
rs174544	11 / 61324329	*FADS1*	94.0	26.6	0.7207	Fatty acids, atopy
rs174553	11 / 61331734	*FADS1*	94.6	30.8	0.0885	Fatty acids, atopy
rs174556	11 / 61337211	*FADS1*	94.2	26.7	0.9369	Fatty acids, atopy
rs174561	11 / 61339284	*FADS1*	97.1	25.5	0.0220	Fatty acids, atopy
rs3834458	11 / 61351497	*FADS2*	93.4	30.1	0.8244	Fatty acids, atopy
rs99780	11 / 61353209	*FADS2*	97.6	30.2	0.1699	Fatty acids, atopy
rs174570	11 / 61353788	*FADS2*	94.7	10.5	0.0316	Fatty acids, atopy
rs2072114	11 / 61361791	*FADS2*	93.5	10.2	0.3064	Fatty acids, atopy
rs174583	11 / 61366326	*FADS2*	93.1	31.6	0.6654	Fatty acids, atopy
rs174602	11 / 61380990	*FADS2*	94.7	17.9	0.5629	Fatty acids, atopy
rs482548	11 / 61389758	*FADS2*	95.4	9.4	0.7862	Fatty acids, atopy
rs174454	11 / 61407323	*FADS3*	93.1	23.9	0.5482	Fatty acids, atopy
rs528285	11 / 61417280	*FADS3*	97.3	33.0	0.1048	Fatty acids, atopy
rs10507197	12 / 104585833	*intergenic*	95.0	40.8	0.5022	GWA
rs1543480	14 / 45533589	*intergenic*	95.2	49.8	0.7344	GWA
rs2225995	14 / 48124872	*intergenic*	84.5	8.1	0.3789	GWA
rs2400464	14 / 98433397	*intergenic*	97.7	37.3	0.1823	GWA
rs293004	15 / 58058615	*RAB3C*	99.0	42.0	0.6418	GWA
rs1588085	15 / 96117717	*intergenic*	97.6	3.3.0	0.1484	GWA
rs6808	17 / 59754307	*PECAM1*	98.8	48.1	0.6713	CAD
rs8065316	17 / 59816347	*PECAM1*	90.6	46.6	0.0057	CAD
rs1390428	18 / 32457205	*FHOD3*	99.1	12.6	0.1310	GWA
ADA22	20 / 42713641	*ADA*	84.6	5.7	0.3641	duration and intensity of deep sleep
rs2038526	20 / 48619056	*PTPN1*	99.3	36.4	1	Diabetes
rs5751876	22 / 23167301	*ADORA2A*	94.3	39.6	0.6504	duration and intensity of deep sleep

(Table VIa). After adjustment for sex and age, five more SNPs showed significant association in any of the proposed genetic effect models (Table VIb). A last SNP (rs1543480) chosen for closer investigation was found to be associated with age at death under the assumption of a recessive genetic effect model (Table VIc). None of these altogether eleven SNPs violated the *HWE* assumption. The significance obtained for rs3834458 under a recessive effect model disappeared after adjusting for sex and age.

Lowest p-values were observed for two SNPs located within the fibrinogen gene (rs2070016 and rs10012555) in the age and sex adjusted model under the assumption of a recessive effect. An effect of these SNPs could not be seen for any of the other models.

Investigation of the criteria $k_{d.norm}$, V and R^2_{sch} in the three genetic effect models showed that none of the real SNPs has a high contribution to the fit of the model (Table VIa). Like in the simulation studies, values of V_w were very close to V with a tendency to be slightly smaller. Therefore, results for V_w are not discussed in detail.

The criteria $k_{d.norm}$ and R^2_{sch} reach up to 4%, whereas V yields values below 0.1%. These rather low values, however, are in the generally reported range of R^2 from SNP association analysis with quantitative phenotypes. For the adjusted models as well as for the models with age at death as outcome, no obvious deviation from the criteria values for the non-genetic models is obvious for any of the SNPs (Table VIb and c). It should be noted that criteria values for non-genetic models varied slightly, if observations with missing genotypes were excluded. Therefore, values in the genetic models may be slightly lower than in the non-genetic model without exclusion of these observations.

Table VI: Real data: For the three different genetic effect models (additive, dominant, recessive), estimated hazard ratios (HR) and corresponding p-values (p) for the seven SNPs from the KORA data that showed p< 0.05 in any of these models as well as the three artificial lethal SNPs. Estimates were derived by Cox regression (a) without and (b) with adjusting for age and sex† for analysis of time since baseline survey S3 and (c) age at death adjusted for sex. Also stated are criteria values $K_{d.norm}$, V, V_w and R^2_{sch}.

(a)

	additive effect model						dominant effect model						recessive effect model					
SNP	HR	p	$K_{d.norm}$	V	V_w	R^2_{sch}	HR	p	$K_{d.norm}$	V	V_w	R^2_{sch}	HR	p	$K_{d.norm}$	V	V_w	R^2_{sch}
Real SNPs :																		
rs17300539	0.806	0.150	0.0236	$4*10^{-4}$	$3*10^{-4}$	0.0067	0.774	0.110	0.0267	$5*10^{-4}$	$4*10^{-4}$	0.0089	1.136	0.830	9*10-4	0	0	$2*10^{-4}$
rs17366743*	0.566	0.074	0.0186	$4*10^{-4}$	$3*10^{-4}$	0.0195	0.567	0.077	0.0184	$4*10^{-4}$	$3*10^{-4}$	0.0185	---	---	---	---	---	---
rs2070016	1.124	0.280	0.0210	$1*10^{-4}$	$1*10^{-4}$	0.0051	1.079	0.550	0.0124	0	0	0.0020	1.697	0.073	0.0124	$6*10^{-4}$	$5*10^{-4}$	0.0096
rs10012555	1.057	0.640	0.0081	$1*10^{-4}$	$1*10^{-4}$	$2*10^{-4}$	1	1	0	0	0	0	1.895	0.074	0.0088	$7*10^{-4}$	$5*10^{-4}$	0.0050
rs10520818	1.125	0.300	0.0184	$2*10^{-4}$	$1*10^{-4}$	0.0038	1.231	0.096	0.0309	$4*10^{-4}$	$3*10^{-4}$	0.0079	0.198	0.110	0.011	$5*10^{-4}$	$4*10^{-4}$	0.0405
rs3834458	0.915	0.310	0.0211	$1*10^{-4}$	$1*10^{-4}$	0.0039	0.987	0.900	0.0027	0	0	0	0.612	0.038	0.0283	$7*10^{-4}$	$5*10^{-4}$	0.0212
rs1543480	0.897	0.170	0.0223	$3*10^{-4}$	$3*10^{-4}$	0.0083	0.958	0.730	0.0066	0	0	0.001	0.767	0.056	0.0375	$8*10^{-4}$	$6*10^{-4}$	0.0133
rs6808	1.164	0.050	0.0309	$6*10^{-4}$	$5*10^{-4}$	0.0133	1.315	0.038	0.0416	$6*10^{-4}$	$5*10^{-4}$	0.0162	1.148	0.270	0.0205	$2*10^{-4}$	$2*10^{-4}$	0.0038
Artificial lethal SNPs for three different degrees lethality:																		
Moderate	1.577	$<10^{-10}$	0.3526	0.1022	0.0803	0.3724	1.167	$<10^{-10}$	0.2276	0.0132	0.0106	0.2623	2.832	$<10^{-10}$	0.2853	0.1579	0.1239	0.3953
Strong	2.472	$<10^{-10}$	0.4760	0.2834	0.2242	0.6002	1.956	$<10^{-10}$	0.3512	0.0293	0.0235	0.5456	3.802	$<10^{-10}$	0.4144	0.3542	0.2797	0.6189
Extreme	56.120	$<10^{-10}$	0.6133	0.5484	0.4315	0.9345	78.218	$<10^{-10}$	0.5385	0.0395	0.0354	0.9405	155.400	$<10^{-10}$	0.5545	0.6750	0.5201	0.8601

(b) †

	additive effect model						dominant effect model						recessive effect model					
SNP	HR	p	$K_{d.norm}$	V	V_w	R^2_{sch}	HR	p	$K_{d.norm}$	V	V_w	R^2_{sch}	HR	p	$K_{d.norm}$	V	V_w	R^2_{sch}
Real SNPs :																		
rs17300539	0.758	0.069	0.4491	0.1044	0.0808	0.7635	0.721	0.044	0.4493	0.1046	0.0810	0.7647	1.227	0.720	0.4481	0.1028	0.0795	0.7651
rs17366743*	0.498	0.028	0.4485	0.1036	0.0802	0.7563	0.498	0.030	0.4484	0.1035	0.0801	0.7562	---	---	---	---	---	---
rs2070016	1.168	0.160	0.4399	0.0970	0.0750	0.7550	1.088	0.500	0.4393	0.0964	0.0744	0.7547	2.332	0.004	0.4417	0.0986	0.0764	0.7567
rs10012555	1.110	0.390	0.4422	0.1048	0.0810	0.7569	1.043	0.750	0.4419	0.1042	0.0805	0.7581	2.327	0.018	0.4430	0.1065	0.0826	0.7570
rs10520818	1.166	0.180	0.4442	0.1008	0.0778	0.7648	1.279	0.048	0.4445	0.1017	0.0784	0.7652	0.209	0.120	0.4460	0.1015	0.0784	0.7611
rs3834458	0.928	0.390	0.4367	0.1003	0.0774	0.7536	0.999	0.990	0.4365	0.0998	0.0770	0.7550	0.638	0.057	0.4381	0.1017	0.0785	0.7534
rs1543480	0.957	0.570	0.4513	0.1057	0.0816	0.7663	1.100	0.46	0.4515	0.1062	0.0821	0.7647	0.785	0.082	0.4525	0.1065	0.0823	0.7692
rs6808	1.165	0.049	0.4445	0.1018	0.0788	0.7645	1.306	0.043	0.4447	0.1023	0.0792	0.7617	1.158	0.240	0.4437	0.1007	0.0779	0.7646
Artificial lethal SNPs for three different degrees lethality:																		
Moderate	1.403	$<10^{-10}$	0.5073	0.2356	0.1892	0.7716	1.033	$<10^{-10}$	0.4610	0.1254	0.0975	0.9204	2.555	$<10^{-10}$	0.4878	0.2907	0.2369	0.7744
Strong	2.191	$<10^{-10}$	0.5835	0.4046	0.3292	0.7937	1.796	$<10^{-10}$	0.4966	0.1585	0.1229	0.9353	3.429	$<10^{-10}$	0.5018	0.4595	0.3757	0.8130
Extreme	39.770	$<10^{-10}$	0.6550	0.6006	0.4852	0.9592	75.790	$<10^{-10}$	0.6281	0.2057	0.1608	0.9453	103.340	$<10^{-10}$	0.5747	0.6074	0.4858	0.9188

* no homozygous carriers of the minor allele observed in the group of cases

† The model for age and sex alone yields hazard ratios of HR=0.485 for age and HR=1.115 for sex=woman. Both p-values are below 10^{-9} and the following values are obtained for the criteria: $K_{d.norm}$=0.4458, V =0.1027, V_w=0.0794 and R^2_{sch}=0.7658.

(c) †

	additive effect model						dominant effect model						recessive effect model					
SNP	HR	p	$K_{d.norm}$	V	V_w	R^2_{sch}	HR	p	$K_{d.norm}$	V	V_w	R^2_{sch}	HR	p	$K_{d.norm}$	V	V_w	R^2_{sch}
Real SNPs :																		
rs17300539	0.805	0.150	0.1503	0.0080	0.0062	0.1310	0.774	0.110	0.1499	0.0080	0.0063	0.1330	1.059	0.920	0.1502	0.0076	0.0059	0.1236
rs17366743	0.566	0.074	0.1511	0.0077	0.0060	0.1304	0.567	0.078	0.1509	0.0077	0.0060	0.1298	---	---	---	---	---	---
rs2070016	1.120	0.300	0.1323	0.0061	0.0049	0.1000	1.069	0.600	0.1324	0.0060	0.0047	0.0972	1.759	0.055	0.1343	0.0066	0.0053	0.1006
rs10012555	1.075	0.550	0.1496	0.0081	0.0063	0.1210	1.018	0.890	0.1495	0.0080	0.0062	0.1213	1.968	0.059	0.1510	0.0088	0.0069	0.1229
rs10520818	1.146	0.230	0.1464	0.0075	0.0058	0.1210	1.263	0.061	0.1465	0.0079	0.0061	0.1247	0.192	0.100	0.1519	0.0079	0.0061	0.1439
rs3834458	0.904	0.250	0.1486	0.0080	0.0062	0.1242	0.975	0.820	0.1486	0.0078	0.0060	0.1198	0.598	0.030	0.1545	0.0086	0.0067	0.1390
rs1543480	0.888	0.130	0.1412	0.0074	0.0058	0.1171	0.940	0.630	0.1413	0.0070	0.0055	0.1102	0.761	0.049	0.1409	0.0078	0.0061	0.1243
rs6808	1.164	0.049	0.1489	0.0082	0.0064	0.1370	1.327	0.032	0.1492	0.0083	0.0065	0.1348	1.139	0.300	0.1489	0.0077	0.0060	0.1290

* no homozygous carriers of the minor allele observed in the group of cases

† The model for sex alone yields HR=0.458 for sex=woman with p-value below 10^{-10} and the following values are obtained for the criteria: $K_{d.norm}$=0.1523, V =0.0079, V_w=0.0062 and R^2_{sch}=0.1272.

2.3.4.2 Analysis of artificial SNPs in KORA

For the artificially generated SNPs, all criteria values were substantially higher with an increase with increasing lethality, as postulated. Highest values throughout all the models were obtained for R^2_{sch} for the "extreme SNP" (with R^2_{sch}>0.86), which was generated to yield almost perfect association to mortality and therefore values close to 1 in all criteria. All criteria's connection to the estimated hazard ratio and the p-value was apparent as they increased either with the distance of the estimated hazard ratio from 1 (null effect) or with decrease in the p-value.

For the initially generated artificial SNP, the true effect model was known. For generation of SNPs with different degrees of association to mortality, different percentages of the genotypes from this original SNP were randomly assigned. Therefore, it cannot be guaranteed that with increasing random assignment of genotypes, the original genetic effect model is kept for these SNPs. The effect of misspecification of the assumed effect model on the criteria could be seen best for the "extreme SNP", which is closest to the original SNP. The two derived SNPs with lower association to mortality, however, were not excluded from this analysis.

V obviously suffered from a substantial drop in value under the false assumption of a dominant effect for all artificially generated SNPs, while the other criteria were less affected. In addition, *V* would clearly have favoured the recessive over the additive model, while only $k_{d.norm}$ yields highest values for the additive model. For R^2_{sch}, a decision between the additive effect model and one of the dichotomous effect models was difficult in the unadjusted model. In the adjusted model, the correct model would have been favoured only for the "extremely lethal" SNP. For the two other SNPs, the dominant effect model yielded highest values. The Kaplan Meier survival curves obtained for the "extreme SNP" is displayed in Figure 10. The major discrepancy

between the curves was found for homozygous carriers of the minor allele and censored times were highly clustered at the end of the study period.

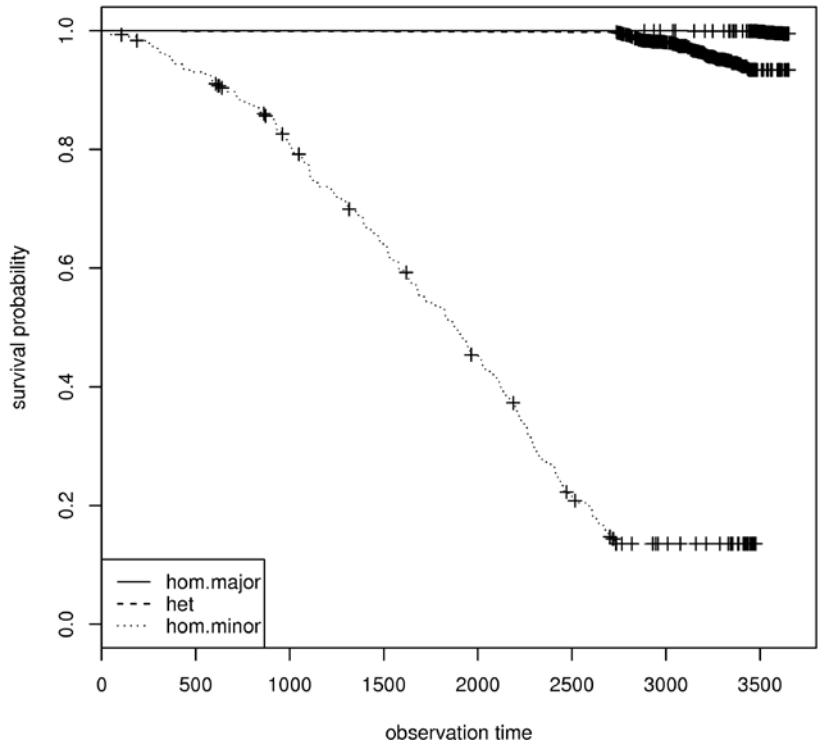

Figure 10: Kaplan-Meier curves obtained for genotypes of the "extreme SNPl" artificially generated SNP (homozygous carriers of the major allele, heterozygotes, homozygous carriers of the minor allele).

After adjustment for sex and age, the criteria's values increased substantially for all artificial SNPs (Table VIb). However, differences between criteria values of the models were low, which indicates that the major increase of the predictive capability discovered by all criteria was due to adjustment alone. Compared to real SNPs, the gain in value obtained for the artificially generated SNPs after adjustment for sex and age was smaller for all criteria. V and V_w, again, were substantially smaller under the assumption of a dominant effect compared to the other genetic effect models.

3 Discussion

3.1 *Overview*

It is of increasing interest to quantify the impact of genetics on health related parameters or disease development. A generally comprehensible measure apart, from effect size and p-value, is the percentage of the impact on the phenotype which can be explained through genetic variation alone. In case of continuously measured health parameters (e.g. blood parameters), it is easy to judge the percentage of variation in the health parameter that can be explained by variation in genetic variants. The appropriate measure R^2 is included as standard output in association analysis software. For case-control studies a derivation of R^2 for logistic regression models is also available.

Meanwhile, more and more population-based studies provide follow-up data and allow for analysing mortality or time of occurrence of disease. These analyses, however, require application of methods from the statistical field of survival analysis. Here, the phenotype is a composition of two variables: the indicator of the health state at the end of the observation time and the length of the observation time. This special situation, without a clear definition of residuals, makes it difficult to define a measure comparable to R^2. If, for example, the impact of a genetic variant on early occurrence of type 2 diabetes is to be judged, the investigator is usually confronted with a variety of available criteria. Without deeper knowledge about the structure of the available criteria and their interpretation, the proper choice becomes a challenging task.

The aim of this thesis was to identify a criterion which is eligible for analysis of survival data while being close to classical R^2 and its interpretation as percentage of

explained variation in the phenotype. Furthermore, the identified criterion was required to be applicable to genetic data, which is often categorical with low number of categories. Appropriate requirements were defined and eligible criteria were selected after a thorough literature review focusing on established as well as less known criteria. Simulation studies with a broad variety of settings for genetic data (in form of SNPs) and real data analysis were then conducted and gave insight into strengths and limitations of the selected criteria.

As expected, no perfect solution could be found. Limitation behaviour (0%-100%) and dependence on censoring (percentage of observations that are still under observation at the end of the study period) turned out to be major or minor nuisances for all criteria. Some of which, however, can be explained through indirect dependence on the estimation under the standard Cox proportional hazards regression model. Therefore, as a side effect of the study, further insight into the special characteristics of the Cox proportional hazards model was gained, which is of importance for genetic as well as non-genetic association studies. The detailed discussion of results is given in the following chapters. After this and accounting for additional properties leads to the conclusion that a clear recommendation of a criterion comparable to classical R^2 can be established for application for genetic association with survival phenotypes.

3.2 *Main results*

Judging the impact of covariates in survival analysis through measures of explained variation, like the R^2 measure in linear regression, touches a general problem of statistics with a variety of answers. It was the aim of this investigation to identify the criterion that suits best to quantify the impact of genetic variants (i.e. binary or

trichotomous covariates for SNP genotypes) on survival phenotypes. A suitable criterion was expected to fulfil the following three requirements:

(a) reasonable criteria values in the range [0;1] in order to allow for interpretation as percentage of prediction quality
(b) increasing values with increasing effect size
(c) independence of the percentage of censored observations in the data

Literature review revealed three different criteria that were considered potentially eligible for this purpose. The first criterion, $k_{d.norm}$, is based on absolute differences in deviance residuals between null model and covariate model. The next criterion, V, aims to measure the difference in variation of the individual survival curves between the two models. For criterion V, a variant V_w exists, which is characterised by a different scheme of weighting and integrating but yields values generally close to V. The last criterion selected for investigation, R^2_{sch}, is close to traditional R^2 in linear regression, and is defined based on the weighted sums of squared Schoenfeld residuals.

Through extensive simulation studies, it was observed that none of the investigated criteria completely fulfils the predefined requirements.

The main findings with respect to the three predefined requirements are discussed in the following:

a) The requirement of limitation to the desired range [0;1] – which is important for interpretation as percentage – was fulfilled by all criteria for mean values across the 200 simulation runs per scenario, at least in all single SNP simulations. Note that in the rare instances with low effect size, where V (and V_w) and R^2_{sch} yielded slightly negative values (> -1%), the fit of the null model is slightly better than the

fit of the covariate model and an improvement of the model cannot be seen. These values, therefore, may be interpreted as zeros and are not considered to be major problems. Simulations with continuous covariates, however, revealed that the required limitation is not generally guaranteed for $k_{d.norm}$. Furthermore, *V* seems not to cover the full range [0;1] and yields generally low values. The range of criteria values could also be linked to the genotype variance. High *MAF*, and therefore high genotype variance, resulted in higher criteria values, which is in line with the general properties of R^2-like criteria. Generally, it could be seen that the impact of SNPs with recessive effects on survival time in the overall population is mostly low and only visible in the presence of high *MAF* (or genotypic variance) or strong effects. This is most likely due to the lower power and generally less explained variation in the phenotype for recessive effect models. This general problem, on the other hand, is not only observed for survival data.

b) All investigated criteria generally increased with increasing effect size. The estimated effect sizes, however, were potentially biased in presence of strong continuous predictors. Furthermore, the variance of the estimated effect sizes was found to be dependent on the censoring percentage in the data. Increased imprecision of the effect estimate in case of high censoring, therefore, resulted in indirect dependence on censoring of the criteria.

c) The two chosen censoring mechanisms representing the two most common study designs in cohort studies hardly affected the criteria values. Minor discrepancies for high censoring percentages can be explained through slightly more robust estimation of effect sizes under the assumption of random censoring. $k_{d.norm}$ generally strongly depended on the percentage of censored observations in the

data. For the other two criteria, observed dependencies can mainly be explained through increased variation of the estimated effect size as mentioned above.

3.3 *Criteria selection*

After a thorough literature search, the investigation was restricted to the criteria $k_{d.norm}$, V and R^2_{sch}, although a broad variety is available, due to several reasons:

- Likelihood-based criteria like *AIC* were excluded due to their lack of interpretation although they are still recommended for additional source of information in model selection.
- Harrell's C-index is calculated from the receiver operator characteristic (ROC) [Harrell, Jr. et al., 1982], [Harrell, Jr., 2001] based on risk ranks and is therefore independent of the underlying effect size [Cook, 2007], which violates one of the predefined general requirements.
- The Brier score was originally developed for logistic regression and also found adaptation to survival data [Graf et al., 1999], [Schumacher et al., 2003], [Gerds and Schumacher, 2006]. As a time-dependent measure of mean squared error based on the considerations of [Korn and Simon, 1990], integration of the Brier score over time was proposed as a criterion measuring the predictive capability of the model. A direct correspondence to the investigated criterion V has been shown and comparable performance was concluded [Gerds and Schumacher, 2006]. The criterion V_w was described as interesting alternative to V [Schemper and Henderson, 2000], and thus it was decided to include V as well as V_w, but with a focus on V.
- Two criteria measuring the separation of survival curves were proposed [Sauerbrei et al., 1997], [Royston and Sauerbrei, 2004]. The first of which is

based on mean absolute differences between the estimated effect sizes $\hat{\beta}$ and their estimated mean. Therefore, this definition is close to the investigated criterion $k_{d.norm}$. The second proposed criterion is based on the variation of the prognostic index. Hence, it is more appropriate for association analysis with continuous covariates and not for categorical data such as tri- or dichotomous SNP data.

Therefore, the criteria $k_{d.norm}$, V and R^2_{sch} were identified as the most promising and representative measures and were thoroughly investigated for application in genetic association studies with survival phenotypes.

3.4 *Criteria characterisation*

3.4.1 Characteristics of $k_{d.norm}$

$k_{d.norm}$ is the only criterion that is calculated per subject, independent of its event status indicator. Apart from the advantage of easy calculation of $k_{d.norm}$, this criterion emerged as the only criterion that clearly identifies the correct genetic effect model in the presented real data setting with the artificially generated SNPs for crude as well as for age- and sex-adjusted analysis, which could be of advantage in application for model selection. But more detailed evaluation is needed to clarify whether this observation may be generalised.

Drawbacks of $k_{d.norm}$ are: 1) its strong dependence on censoring and 2) the violation of the limitation to the range [0;1] in presence of strong continuous predictors. These drawbacks can be explained as follows:

1) In contrast to the other criteria considered here, $k_{d.norm}$ is not corrected for censoring, which explains the observed strong dependence on the censoring percentage in the analysed data.
2) For some settings with a normally distributed covariate, $k_{d.norm}$ exceeded the desired maximum value of 1. Generally, this is considered to be an indication for some misspecification [Kvalseth, 1985]. In the present study, it is likely to result from the strongly skewed distribution of martingale residuals in the range $[-\infty,1]$. Deviance residuals are normalised transformations of martingale residuals. In case of high cumulative hazards, as they may occur for highly varying covariate values, normalisation fails and the deviance residuals also tend to extreme values in the covariate model. Therefore, their expected value can deviate from 0, as exemplified in Figure 11. In case of deviation of the deviance residuals' mean value from 0, limitation of $k_{d.norm}$ to the range of [0;1] is not guaranteed – especially in association analysis with adjustment for environmental covariates. In real data single SNP analysis without adjustment for other covariates, though, the situation with $k_{d.norm}$ exceeding 1 is very unlikely to occur due to the limited variation of the di- or trichotomous SNP covariate under the assumption of *HWE* and the limited range of realistic genetic effect sizes.

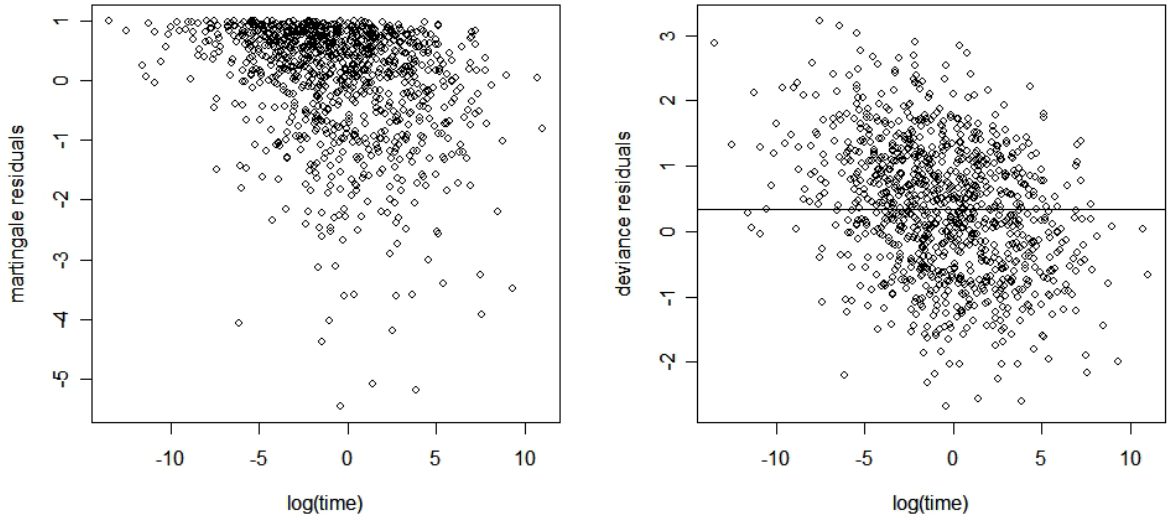

Figure 11: Martingale and deviance residuals are displayed for a model with uncensored exponentially distributed survival times associated with a standard normally distributed covariate with *HR*=32. For deviance residuals a horizontal line indicates the residual mean value which deviates from zero.

3.4.2 Characteristics of *V*

The criterion *V*, in contrast to $k_{d.norm}$, was limited to the range [0;1] and was robust against moderate censoring in the data. The variant V_w yielded similar, but slightly lower values than *V* throughout the simulations. Therefore, it is not discussed in detail. Major drawbacks of *V* are: 1) its generally low values and 2) its sensitivity to high censoring in the data. These drawbacks and related problems are discussed as follows:

1) In all simulation settings as well as for the real data examples with the artificially generated lethal SNPs, these values hardly exceeded 60%. This may indicate a generally different boundedness with the maximum attainable value <1. Especially recessive effect models with low power can hardly be differentiated.

2) *V* emerged to be sensitive to high censoring percentages, especially if the censoring accumulates at the end, which is often the case for real data with a predefined follow-up period. In these cases, values are very low, which may be due to the extremely low variation of the survival curves. The proper choice of the genetic effect model, here, also plays an important role - as could be seen for the presented mortality data analysis with the artificially generated SNPs. *V* suffered from a substantial drop under the false assumption of a dominant effect and for crude as well as for adjusted analysis, the false recessive effect model yielded clearly higher values than the true additive model. If specification of the genetic effect model would have been based on *V*, the wrong model would therefore be chosen. As *V* is calculated based on survival curves, this circumstance, however, is likely to be due to the fact that the discrepancy of the survival curves is lowest under the assumption of a dominant effect model and highest under the assumption of a recessive effect model (see Figure 10).

3.4.3 Characteristics of R^2_{sch}

R^2_{sch}, apart from some rare cases with slightly negative values, generally held the limitation to the range [0;1], which is, obviously in contrast to *V*, fully exploited. The investigations revealed that, among all investigated criteria, only R^2_{sch} was able to approach (and not exceed) the maximum value 1 in realistic situations, where the main influence on disease emerges from a strong quantitative prognostic factor, while the genetic component only gives small additional rise in risk.

Among all criteria, R^2_{sch} showed strongest dependence on the effect size and its estimation. As estimation of $\hat{\beta}$ may be affected by the censoring percentage in the

data, R^2_{sch} also shows indirect dependence on censoring (see section 3.1). Mean values of the criterion, however, are not affected in the presented simulation studies and no general strong dependence on censoring, as for $k_{d.norm}$, is obvious.

Minor drawbacks of R^2_{sch}:

1) As R^2_{sch} is dependent on the variance of the genetic covariate, its applicability to differentiate between different genetic effect models may be questionable.
2) Its interpretation as a measure of variation in the covariates given failure time may be confusing. But this apparently reverse way of conditioning is justified after closer inspection of the Cox model (see definition of the score function) and therefore straightforward [Xu, 1996], [Xu and O'Quigley, 1999].

3.5 *Outlook*

3.5.1 Strengths and limitations

The major strength of this study was the high variety of simulation settings covering realistic as well as extreme settings with respect to:

- genetic effect sizes
- genotype distributions (varying *MAF* with and without *HWE*)
- genetic effect models
- censoring percentage in the data as well as censoring mechanism
- continuous covariate distributions (as single prognostic factor or as additional adjustment covariate)

Variation of all these parameters allowed extensive comparison of the investigated criteria and evaluation of their sensitivity to a variety of components in combination. The availability of mortality data from KORA S3/F3 allowed for judgement of the

criteria's performance in realistic situations, where only small impact of genetic variants on survival phenotypes is expected to occur. The additionally simulated artificial SNPs gave insight on how the criteria would perform in case a SNP of the defined degree of association to mortality existed and how they react if adjustment for environmental factors with true effects on the survival phenotype is included.

Although the simulation study already involved lots of tuning parameters, there could be further items of interest, e.g.:

- situations with more than one single SNP
- varying effect sizes for the environmental covariate in the bivariate simulations
- other genetic variants like haplotypes that are possibly not coded as dichotomous or trichotomous covariates
- inclusion of a set of environmental covariates, confounders or gene-environment interactions

The KORA mortality data example was added in order to investigate the criteria's performance in real data situations. However, no strong associations were detected. This gap was filled through simulation of artificial SNPs based on the mortality data. On the other hand, it could also be interesting to investigate more complex models, e.g. include more environmental factors for adjustment, to analyse cause-specific mortality or other survival phenotypes like incident type 2 diabetes or myocardial infarction.

3.5.2 Possible applications and extensions of R^2_{sch}

Major advantages of R^2_{sch} are its ease of computation and its high flexibility. Lots of additional properties and extensions have already been described in the literature. Comparable information is not yet available for the other criteria investigated in this study. Some of the additional properties and possible extensions should be

mentioned here, as they can easily be realised for R^2_{sch} due to its direct connection to the score function of the partial likelihood:

Similar to R^2 for linear regression, R^2_{sch} allows for decomposition into sums of squares and interpretation as proportion of variation explained by the model at least asymptotically.

In addition, it is possible to calculate confidence intervals by recalculating R^2_{sch} for estimated confidence limits of $\hat{\beta}$, which is realised through simple replacement of the point estimate of $\hat{\beta}$ by it's a) upper and b) lower confidence limit. For the other criteria the calculation would be more complex as the effect estimate is only incorporated indirectly in their definition.

Furthermore, R^2_{sch} can easily be extended to situations where the assumption of proportional hazards is violated. In case of non-proportional hazards, a possible solution is to let at least one effect estimate vary over time as a time-varying coefficient $\beta(t)$. Another solution could be stratified analysis. These and other extended settings and possible tests are discussed and presented in [Therneau and Grambsch, 2000]. For time-varying models, it is also possible to define Schoenfeld residuals and R^2_{sch} due to the direct connection to the score function. This has been exemplified by [Xu and Adak, 2002] through introduction of a time-varying effect in form of a step function. Here, R^2_{sch} is applied as model-selection criterion to derive the number of necessary steps of the time-varying effect.

It is also possible to extend the definition of R^2_{sch} to weighted settings, such as for case-cohort data. This setting offers the possibility to reduce genotyping costs within large cohorts through sampling of a population-representative subcohort and inclusion of all cases, i.e. all non-censored individuals. Hence, sampling weights are assigned to each subject, for which it is necessary to account in estimation.

Different weighting schemes and robust variance estimators for this kind of study have been proposed [Therneau and Li, 1999]. Again, the direct connection of R^2_{sch} to the score function allows for computation of this criterion. For the other criteria, adaptation to the weighted setting with possible cluster definitions for cases that are entered twice (as random sample from the cohort and as part of the case sample) is less clear and needs more research. A description of a KORA case-cohort, which is defined for incident cases of coronary heart disease and type 2 diabetes surveys S1, S2 and S3 is available [Thorand et al., 2005]. Application of R^2_{sch} to genetic association studies in the KORA case-cohort is planned.

Another major advantage is the possibility to calculate partial coefficients, which also suggests its application for model selection. The partial coefficient calculated for the genetic variants draws a possible connection to the estimation of heritability in family studies. Details and an overview of possible extensions and applications can be found in [Xu, 1996], [O'Quigley and Xu, 2001], and [O'Quigley and Xu, 2006].

Its high flexibility and especially the possibility to calculate partial coefficients is of major interest for genetic association studies, where the genetic effect often has to be judged in presence of non-genetic covariates. Adopting the general definition of partial coefficients for R^2_{sch} [O'Quigley and Xu, 2006], the impact of the genetic variants on the outcome's variation in presence of environmental covariates can be established as follows: Let $X_{g1},\ldots, X_{gm}$ be the subset of m genetic covariates and $X_{e1},\ldots X_{ek}$ be k environmental covariates. With $\hat{\beta}_X$ being the vector of estimations $\hat{\beta}$ from the model including covariates X, the partial R^2_{sch} for the genetic covariates can then be calculated as:

$$R^2_{sch}\left(\hat{\beta}_{X_{g1},\ldots,X_{gm}|X_{e1},\ldots,X_{ek}}\right)=1-\frac{1-R^2_{sch}\left(\hat{\beta}_{X_{g1},\ldots,X_{gm},X_{e1},\ldots,X_{ek}}\right)}{1-R^2_{sch}\left(\hat{\beta}_{X_{e1},\ldots,X_{ek}}\right)}$$

Partial coefficients allow for closer investigation of model building procedures which include, for example, stepwise addition of sets of components, i.e. genetic covariates in step one, environmental covariates in step two and gene-environment interactions in the third step. In this case, the additional information obtained from the set of gene-environment interactions can be judged. Unfortunately, no definition of partial coefficients is currently available for any of the other criteria investigated in this study. Due to the lack of association of the SNPs investigated in the real data example, no contribution of the candidates was detected through calculation of these partial coefficients in adjusted mortality analysis except for the artificially generated SNPs. Highest values were obtained for the ”extremely lethal” SNP, where partial coefficients of R^2_{sch} for the additive, dominant and recessive effect model (adjusted for age and sex) were calculated as 82.58%, 76.63% and 65.33%, respectively. Hence, most of the variation is explained by the genetic covariate alone. Comparison of the partial coefficients for the different genetic effect models as in a model building situation would here clearly favour the true additive effect coding over the two dichotomous variants.

3.6 *Conclusion*

The present study showed that none of the investigated criteria proposed for judgement of the impact of covariates on survival phenotypes perfectly fulfilled our requirements, which also shows why no general recommendation is available, yet. The limitation behaviour of $k_{d.norm}$ and V were found to be major problems for interpretation as percentage. Altogether, our requirements were best fulfilled by R^2_{sch}

which is also closest related to estimation in the Cox model and the definition of classical R^2 from linear regression.

Therefore, R^2_{sch} is recommended as a powerful and highly flexible tool for quantification of the impact of genetic variants on survival phenotypes. The extensive simulation settings also indicate that this recommendation may not only be restricted to genetic association studies but also account for general epidemiologic studies.

Summary

Reporting the impact of genetic variants on diseases by means of a percentage of impact has become a standard question in genetic epidemiological studies. In case of cross-sectional studies with continuous phenotypes or case-control studies, measures like R^2 or derivations are already available. They allow quantifying the impact of genetic variants by a measure of percentage of explained variation in the phenotype. For survival phenotypes (e.g. mortality or incidence), however, the definition of a comparable criterion is still unclear. Therefore, genetic variants are usually only judged through effect size estimates and p-values when they are analysed in their association to survival phenotypes.

The aim of this thesis was to identify the criterion which suits best for quantification of the impact of genetic variants on survival phenotypes, similar to classical R^2. For none of the investigated criteria, investigations focusing on applicability in genetic association analysis, with the special character of genetic variants as statistical covariates, are available, yet.

In first instance, a thorough literature search was performed. It revealed three criteria that were generally considered eligible for measuring of the impact of genetic variants as percentage – comparable to a measure of explained variation.

The three identified criteria measure:

(1) difference between deviance residuals ($k_{d.norm}$)

(2) variation of survival curves (V)

(3) variation of Schoenfeld residuals (R^2_{sch}).

These were subsequently compared in their performance for SNP data through thorough simulation studies with a variety of scenarios (with respect to phenotype and genetic variants) and application to KORA mortality data.

The focus of the evaluation was set on the following predefined requirements:

(a) reasonable criteria values in the range [0;1] in order to allow for interpretation as percentage of prediction quality

(b) increasing values with increasing genetic effect size

(c) independence of the percentage of censored observations in the data

However, none of the investigated criteria perfectly fulfilled these requirements. In the simulation studies, the deviance residuals' criterion showed high dependence on the censoring percentage and is not generally limited to the range [0; 1]. The second criterion (variation of survival curves) hardly reached values above 60%. The requirements were best fulfilled by the criterion based on Schoenfeld residuals. Additionally to the good performance in genetic simulation studies, and application to mortality data, a variety of possible extensions and applications are available for this criterion. Therefore, it is recommended as a powerful and highly flexible tool for judgement of the impact of genetic variants in genetic association studies with survival outcome, which, in addition, is relatively easy to calculate.

Therefore, it is now possible to fill the gap of a missing criterion like R^2 for judgment of the impact of genetic variants in analysis of survival phenotypes. Furthermore, a deeper insight into the Cox proportional hazards model was gained. Therefore, some general problems which may occur in genetic association analysis with survival phenotypes could be identified.

Zusammenfassung

Die Frage nach der Bedeutung genetischer Varianten für Erkrankungen im Sinne einer Prozentzahl des Einflusses ist inzwischen zu einer Standardfrage genetische epidemiologischer Studien geworden. Bei Querschnittsstudien mit kontinuierlichen Phänotypen oder Fallkontrollstudien stehen bereits Maße wie R^2 o.ä. zur Verfügung. Diese erlauben die Beurteilung des Einflusses der genetischen Varianten im Sinne der erklärten Varianz. Für Survivalphänotypen (wie z.B. Mortalität oder Inzidenz) ist die Definition eines vergleichbaren Kriteriums allerdings noch unklar. Somit beschränkt sich die Beurteilung genetischer Einflüsse bei der Analyse von Survivalphänotypen häufig auf Effektstärken und p-Werte.

Ziel dieser Arbeit war es, ein Kriterium zu identifizieren, das am ehesten dazu geeignet ist, genetische Varianten im Sinne eines klassischen R^2-Kriteriums in ihrem Einfluss auf Survivalphänotypen zu beurteilen. Für keines der betrachteten Kriterien wurden bisher Untersuchungen hinsichtlich der Eignung für genetische Assoziationsanalysen durchgeführt, die sich durch Besonderheiten der genetischen Varianten als Kovariablen im statistischen Sinne auszeichnen.

Zunächst wurde dazu eine umfangreiche Literatursuche durchgeführt, über die drei Kriterien identifiziert wurden, die prinzipiell geeignet schienen, eine Interpretation des genetischen Einflusses im Sinne eines Maßes erklärter Varianz zu ermöglichen.

Die drei identifizierten Kriterien beruhen auf:

(1) Differenz zwischen Devianzresiduen ($k_{d.norm}$)

(2) Variation individueller Survivalkurven (V)

(3) Variation von Schoenfeld-Residuen (R^2_{sch}).

Diese wurden anschließend anhand von umfangreichen Simulationsstudien mit einer Vielfalt an Szenarien (bzgl. Phänotyp und genetischer Varianten) sowie einer Anwendung auf KORA-Mortalitätsdaten hinsichtlich ihrer Eignung für SNP-Assoziationsstudien untersucht. Bei der Beurteilung standen die folgenden im Vorfeld definierten Anforderungen im Vordergrund:

(a) sinnvolle Werte im Bereich [0;1] um eine Interpretation als Prozent erklärter Variation zu gewährleisten
(b) größere Werte mit wachsender Effektstärke
(c) Unabhängigkeit vom Zensierungsanteil in den Daten

Die Untersuchung zeigte, dass keines der verwendeten Kriterien gänzlich diese Anforderungen erfüllte. Das auf Devianzresiduen basierende Kriterium zeigte in den Simulationsstudien eine starke Abhängigkeit vom Zensierungsanteil in den Daten und hielt keine generelle Limitierung des Wertebereichs auf [0;1] ein. Das zweite Kriterium (Variation individueller Survivalkurven) erreichte selten Werte über 60%. Die gestellten Anforderungen wurden am besten durch das auf Schoenfeld-Residuen basierende Kriterium erfüllt. Zusätzlich zu der positiven Beurteilung im Rahmen der genetischen Simulationsstudien und der Anwendung auf Mortalitätsdaten, stehen für dieses Kriterium eine Vielzahl an Anwendungsmöglichkeiten und möglicher Erweiterungen zur Verfügung. Daher wird es als starkes und hoch flexibles Maß zur Beurteilung des Einflusses genetischer Varianten in Assoziationstudien mit Survival-Phänotypen empfohlen, das zudem noch relativ einfach zu berechnen ist.

Hiermit ist es nun möglich, die Lücke eines fehlenden Kriteriums zur Beurteilung des Einflusses genetischer Varianten im Sinne eines R^2 zu füllen. Zusätzlich konnte ein tieferer Einblick in das Cox proportional hazards Modell gewonnen werden. Daher

konnten einige generelle Probleme, die bei genetischen Assoziationsstudien mit Survivalphänotypen auftreten können, identifiziert werden.

References

Kaplan EL, Meier P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**:457-481.

Cox DR. 1972. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society: Series B* **34**:187-220.

Cox DR. 1975. Partial likelihood. *Biometrika* **62**:269-276.

Grambsch PM, Therneau TM. 1994. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**:515-526.

Frazer KA, Ballinger DG, Cox DR et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**(7164):851-861.

Vollmert C, Hahn S, Lamina C et al. 2007. Calpain-10 variants and haplotypes are associated with polycystic ovary syndrome in Caucasians. *Am.J.Physiol Endocrinol.Metab* **292**(3):E836-E844.

Heid IM, Lamina C, Küchenhoff H et al. 2008. Estimating the SNP Genotype Misclassification from Routine Double Measurements in a Large Epidemiological Sample. *American Journal of Epidemiology* .

Stram DO, Haiman CA, Hirschhorn JN et al. 2003. Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of

unrelated subjects with an example from the Multiethnic Cohort Study. *Hum.Hered.* **55**(1):27-36.

Hattersley AT, McCarthy MI. 2005. What makes a good genetic association study? *Lancet* **366**(9493):1315-1323.

Dawn TM, Barrett JH. 2005. Genetic linkage studies. *Lancet* **366**(9490):1036-1044.

Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am.J.Hum.Genet.* **52**(3):506-516.

Cordell HJ, Clayton DG. 2005. Genetic association studies. *Lancet* **366**(9491):1121-1131.

Kaprio J. 2000. Science, medicine, and the future. Genetic epidemiology. *BMJ* **320**(7244):1257-1259.

Burton PR, Tobin MD, Hopper JL. 2005. Key concepts in genetic epidemiology. *Lancet* **366**(9489):941-951.

Schaid DJ. 2004. Evaluating associations of haplotypes with traits. *Genet.Epidemiol.* **27**(4):348-364.

Cook NR. 2007. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* **115**(7):928-935.

Rosthoj S, Keiding N. 2004. Explained variation and predictive accuracy in general parametric statistical models: the role of model misspecification. *Lifetime Data Analysis* **10**(4):461-472.

Kvalseth TO. 1985. Cautionary Note About R^2. *The American Statistician* **39**(4):279-285.

Van Houwelingen JC, Le Cessie S. 1990. Predictive value of statistical models. *Stat.Med.* **9**(11):1303-1325.

Nagelkerke NJD. 1991. A note on a general definition of the coefficient of determination. *Biometrika* **78**(3):691-692.

Wild S, Roglic G, Green A, Sicree R, King H. 2004. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care* **27**(5):1047-1053.

Bell CG, Walley AJ, Froguel P. 2005. The genetics of human obesity. *Nat.Rev.Genet.* **6**(3):221-234.

Capri M, Salvioli S, Sevini F et al. 2006. The genetics of human longevity. *Ann.N.Y.Acad.Sci.* **1067**:252-263.

Schemper M, Stare J. 1996. Explained variation in survival analysis. *Stat.Med.* **15**(19):1999-2012.

Harrell FE, Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer: New York, 2001.

Maddala G.S. *Limited-Dependent and Qualitative Variables in Econometriocs*. Cambridge University Press: Cambridge, UK, 1983.

Magee L. 1990. R^2 meaasures based on Wald and likelihood ratio joint significance tests. *The American Statistician* **44**:250-253.

Kent TJ, O'Quigley J. 1988. Measures of dependence for censored survival data. *Biometrika* **75**:525-534.

Somers RH. 1962. A new asymmetric measure of association for ordinal variables. *American Sociological Review* **27**:799-811.

Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. 1982. Evaluating the yield of medical tests. *JAMA* **247**(18):2543-2546.

Sauerbrei W, Hubner K, Schmoor C, Schumacher M. 1997. Validation of existing and development of new prognostic classification schemes in node negative breast cancer. German Breast Cancer Study Group. *Breast Cancer Res.Treat.* **42**(2):149-163.

Royston P, Sauerbrei W. 2004. A new measure of prognostic separation in survival data. *Stat.Med.* **23**(5):723-748.

Graf E, Schmoor C, Sauerbrei W, Schumacher M. 1999. Assessment and comparison of prognostic classification schemes for survival data. *Stat.Med.* **18**(17-18):2529-2545.

Stark, M. 1997. Beurteilungskriterien für die Güte von Modellen zur Analyse von Überlebenszeiten. *Dissertation*. Berlin. Logos Verlag.

Schemper M, Henderson R. 2000. Predictive accuracy and explained variation in Cox regression. *Biometrics* **56**(1):249-255.

O'Quigley J, Xu R. 2001. Explained variation in Cox regression. In: Crowley J. (ed) *Handbook of Statistics in Clinical Oncology.* Marcel Dekker, Inc.: pp 397-410.

Therneau TM, Grambsch PM, Fleming TR. 1990. Martingale-based residuals for survival models. *Biometrika* **77**(1):147-160.

Gerds TA, Schumacher M. 2006. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biom.J.* **48**(6):1029-1040.

Schemper M, Smith TL. 1996. A note on quantifying follow-up in studies of failure time. *Control Clin.Trials* **17**(4):343-346.

Altman DG, De Stavola BL, Love SB, Stepniewska KA. 1995. Review of survival analyses published in cancer journals. *Br.J.Cancer* **72**(2):511-518.

Schoenfeld DA. 1982. Partial residuals for the proportional hazards regression model. *Biometrika* **69**(1):239-241.

O'Quigley J, Flandre P. 1994. Predictive capability of proportional hazards regression. *Proc.Natl.Acad.Sci.U.S.A* **91**(6):2310-2314.

Andersen PK, Christensen E, Fauerholdt L, Schlichting P. 1983. Measuring prognosis using the proportional hazards model. *Scand.J.Statist.* **10**:49-52.

Bender R, Augustin T, Blettner M. 2005. Generating survival times to simulate Cox proportional hazards models. *Stat.Med.* **24**(11):1713-1723.

R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing.

Holle R, Happich M, Lowel H, Wichmann HE. 2005. KORA--a research platform for population based health research. *Gesundheitswesen* **67 Suppl 1**:S19-S25.

Emigh T. 1980. Comparison of tests for Hardy-Weinberg Equilibrium. *Biometrics* **36**:627-642.

Li J, Ji L. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95**(3):221-227.

Schumacher M, Graf E, Gerds T. 2003. How to assess prognostic models for survival data: a case study in oncology. *Methods Inf.Med.* **42**(5):564-571.

Korn EL, Simon R. 1990. Measures of explained variation for survival data. *Stat.Med.* **9**(5):487-503.

Xu, R. 1996. Inference for the proportional hazards model. *Dissertation*. University of California, San Diego.

Xu R, O'Quigley J. 1999. A R-2 type measure of dependence for proportional hazards models. *Journal of Nonparametric Statistics* **12**(1):83-107.

Therneau TM, Grambsch PM. *Modeling Survival Data. Extending the Cox Model.* Springer-Verlag: New York Berlin Heidelberg, 2000.

Xu R, Adak S. 2002. Survival analysis with time-varying regression effects using a tree-based approach. *Biometrics* **58**(2):305-315.

Therneau TM, Li H. 1999. Computing the Cox model for case cohort designs. *Lifetime.Data Anal.* **5**(2):99-112.

Thorand B, Schneider A, Baumert J et al. 2005. Case-cohort studies: an effective design for the investigation of biomarkers as risk factors for chronic diseases--demonstrated by the example of the MONICA/KORA Augsburg Case-Cohort Study 1984-2002. *Gesundheitswesen* **67 Suppl 1**:S98-102.

O'Quigley J, Xu R. 2006. Explained Variation in Proportional Hazards Regression. In: Crowley J., Ankerst D (eds) *Handbook of Statistics in Clinical Oncology.* Boca Raton: Chapman & Hall/CRC.

Appendix

A1. List of publications and presentations

List of oral presentations

Seminar, Charité, Institut für Sozialmedizin, Epidemiologie und Gesundheitsökonomie, Nov. 2007, Berlin: „Einführung in die genetische Epidemiologie"

Kongress "Medizin und Gesellschaft", Sept. 2007, Augsburg: "Quantification of the contribution of genetic variants in association analysis with survival outcome: three methods in comparison"

Seminar, Georg-August-Universität, Göttingen, Department of Genetic Epidemiology, Jan. 2006: "Judging explained variation in survival models"

Airgene Workshop, Oct. 2005, GSF, Neuherberg: "Haplotypes and the CRP gene"

Seminar, TU München, Institute for Medical Statistics and Epidemiology, Apr. 2004, "Gütekriterien in der Survivalanalyse"

SFB 368, Workshop Höhenried, Jun. 2003, "Modellierung komplexer Interaktionen in der Survivalanalyse"

Poster presentations

48th Cardiovascular Disease Epidemiology and Prevention Conference, and Nutrition, Physical Activity and Metabolism Conference. 13.03-15.03.2008, Colorado Springs, Colorado, USA. Circulation 2008;117: e256 (P221): “Effect of macrophage MIF gene variants and serum concentrations on the risk for coronary heart disease: results from the MONICA/KORA Augsburg Study, 1984-2002”, *authors*: Herder C, Klopp N, Baumert J, Müller M, Khuseyinova N, Meisinger C, Martin S, Illig T, Koenig W, Thorand B

48th Cardiovascular Disease Epidemiology and Prevention Conference, and Nutrition, Physical Activity and Metabolism Conference. 13.03-15.03.2008, Colorado Springs, Colorado, USA. Circulation 2008;117: e255 (P215): “Interleukin-18 gene polymorphisms, interleukin-18 serum concentrations and risk of coronary heart disease: results from the MONICA/KORA Augsburg case-cohort study, 1984-2002”, *authors*: Thorand B, Baumert J, Herder C, Klopp N, Kolz M, Khuseyinova N, Müller M, Loewel H, Illig T, Koenig W

Jahrestagung der deutschen Gesellschaft für Kardiologie, Mannheim 27.03.-29.03.2008: “Interleukin-18 gene polymorphisms and incident coronary heart disease in middle-aged men and women: results from the MONICA/KORA Augsburg case-cohort study, 1984-2002”, *authors*: Thorand B, Baumert J, Herder C, Klopp N, Kolz M, Khuseyinova N, Müller M, Meisinger C, Illig T, Koenig W

XXVII Congress of the European Society of Cardiology, Vienna, 01.09.-05.09.2007. Eur Heart J 2007;28(Abstract Supplement):691 (P4112): “Effect of macrophage MIF

gene variants and serum concentrations on the risk for type 2 diabetes and coronary heart disease: results from the MONICA/KORA Augsburg Study, 1984-2002.", *authors*: Herder C, Klopp N, Baumert J, Müller M, Khuseyinova N, Meisinger C, Martin S, Illig T, Koenig W, Thorand B.

.XXVII Congress of the European Society of Cardiology, Vienna, 01.09.-05.09.2007. Eur Heart J 2007;28(Abstract Supplement):692 (P4115): "Interleukin-18 gene polymorphisms, interleukin-18 serum concentrations and risk of coronary heart disease: results from the MONICA/KORA Augsburg case-cohort study, 1984-2002", *authors*: Thorand B, Baumert J, Herder C, Klopp N, Kolz M, Khuseyinova N, Müller M, Löwel H, Illig T, Koenig W.

Congress „Statistik unter einem Dach", Bielefeld, Mar. 2007: „Quantifying the contribution of genetic variants in association analysis with survival outcome: three methods in comparison", *authors:* Martina Müller, Helmut Küchenhoff, Dörthe Malzahn, Heike Bickeböller(4), Thomas Illig, H.-Erich Wichmann, Iris M. Heid

47th Annual Conference on Cardiovascular Disease Epidemiology and Prevention. 28.02-03.03.07, Orlando, USA. Circulation 2007;115 (8): e299 (P367): "No association between C-reactive protein (CRP) gene polymorphisms, CRP haplotypes and incident type 2 diabetes mellitus (T2DM) in middle-aged men and women: results from the MONICA/KORA Augsburg case-cohort study, 1984-2002", *authors*: Khuseyinova N, Baumert J, Müller M, Klopp N, Kolz M, Meisinger C, Illig T, Thorand B, Koenig W.

Symposium of the German National Genome Research Network (NGFN) 2007, Heidelberg, Germany, 10.11-11.11.2007: "Interleukin-18 gene polymorphisms,

interleukin-18 serum concentrations and risk of coronary heart disease: results from the MONICA/KORA Augsburg case-cohort study, 1984-2002", *authors*: Thorand B, Baumert J, Herder C, Klopp N, Kolz M, Khuseyinova N, Müller M, Loewel H, Illig T, Koenig W

Symposium of the German National Genome Research Network (NGFN) 2007, Heidelberg, Germany, 10.11-11.11.2007: "RANTES/CCL5 gene polymorphisms, serum concentrations and incident type 2 diabetes: results from the MONICA/KORA Augsburg case-cohort study, 1984-2002", *authors*: Herder C, Illig T, Baumert J, Müller M, Klopp N, Khuseyinova N, Meisinger C, Poschen U, Martin S, Koenig W, Thorand B

Symposium of the German National Genome Research Network (NGFN) 2007, Heidelberg, Germany, 10.11-11.11.2007: "Genetic variants in the Upstream Stimulatory Factor 1 (USF1) gene are associated with lipid parameters and T2DM in German Caucasians: Results from the MONICA/KORA Augsburg case-cohort study, 1984-2002", *authors*: Holzapfel C, Baumert J, Grallert H, Müller M, Khuseyinova N, Herder C, Thorand B, Hauner H, Wichmann HE, Koenig W, Illig T, Klopp N

Symposium of the German National Genome Research Network (NGFN) 2006, Heidelberg, Germany, 25.11-26.11.2006: „Quantifying the contribution of genetic variants in association analysis with survival outcome: three methods in comparison", *authors:* Martina Müller, Helmut Küchenhoff, Dörthe Malzahn, Heike Bickeböller(4), Thomas Illig, H.-Erich Wichmann, Iris M. Heid

Symposium of the German National Genome Research Network (NGFN) 2006, Heidelberg, Germany, 25.11-26.11.2006: “Effect of macrophage migration inhibitory factor (MIF) gene variants and serum concentrations on the risk for type 2 diabetes and coronary heart disease: Results from the MONICA/KORA Augsburg Study, 1984-2002”, *authors*: Herder C, Klopp N, Baumert J, Müller M, Khuseyinova N, Meisinger Ch, Martin S, Illig T, Koenig W Thorand B.

Publications

Petter Lennart Seve Ljungman, M.D.; Tom Bellander, PhD; Fredrik Nyberg, MPH MD PhD; Erik Lampa, MSc; Bénédicte Jacquemin, MD PhD; Melanie Kolz, MSc; Timo Lanki, PhD; John Mitropoulos, MD; Martina Müller, MSc; Sally Picciotto, PhD; Riccardo Pistelli, MD; Regina Rückerl, MSc; Wolfgang Koenig, MD Prof; Annette Peters, PhD MPH: *DNA variants, plasma levels and variability of Interleukin-6 in myocardial infarction survivors: Results from the AIRGENE study.* Thrombosis Research (accepted)

Holzapfel C, Baumert J, Grallert H, Mueller AM, Thorand B, Khuseyinova N, Herder C, Meisinger C, Hauner H, Wichmann H, Koenig W, Illig T, Klopp N: *Genetic variants in the USF1 gene are associated with LDL cholesterol levels and incident T2DM in women: results from the MONICA/KORA Augsburg case-cohort study, 1984-2002*. Eur J Endocrinol, 2008 Oct; 159(4): 407-16.

Jacquemin B, Antoniades C, Nyberg F, Plana E, Müller M, Greven S, Salomaa V, Sunyer J, Bellander T, Chalamandaris AG, Pistelli R, Koenig W, Peters A: *Common Genetic Polymorphisms And Haplotypes Of Fibrinogen -a, -ß And -γ Chains Affect*

Fibrinogen Levels And The Response To Proinflammatory Stimulation In Myocardial Infarction Survivors: The AIRGENE Study. Journal of the American College of Cardiology. 2008 Sep 9;52(11):941-52.

Heid IM, Boes E, Müller AM, Kollerits B, Lamina C, Coassin S, Gieger C, Döring A, Klopp N, Frikke-Schmidt R, Tybjærg-Hansen A, Brandstätter A, Luchner A, Meitinger T, Wichmann HE, Kronenberg F: *A Genome-Wide Association Analysis of HDL-Cholesterol in the Population-Based KORA Study Sheds New Light on Intergenic Regions*. Circulation: Cardiovascular Genetics. 2008 Oct;1:10-20.

Herder C, Illig T, Baumert J, Müller M, Klopp N, Khuseyinova N, Meisinger C, Poschen U, Martin S, Koenig W, Thorand B: *RANTES/CCL5 gene polymorphisms, serum concentrations, and incident type 2 diabetes: results from the MONICA/KORA Augsburg case-cohort study, 1984-2002.* Eur J Endocrinol. 2008 May;158(5):R1-5.

Heid I, Lamina C, Küchenhoff H, Fischer G, Klopp N, Kolz M, Grallert H, Vollmert C, Wagner S, Huth C, Müller J, Müller M, Hunt S, Peters A, Paulweber B, Wichmann H, Kronenberg F, Illig T.: *Estimating the Single Nucleotide Polymorphism Genotype Misclassification From Routine Double Measurements in a Large Epidemiologic Sample*. American Journal of Epidemiology . 2008 Sep 12. Epub ahead of print.

Müller, M., Döring, A., Küchenhoff, H., Lamina, C., Malzahn, D., Bickeböller, H., Vollmert, C., Klopp, N., Meisinger, C., Heinrich, J., Kronenberg, F., Wichmann, H.-E., Heid, I.M.: *Quantifying the contribution of genetic variants for survival phenotypes.* Genetic Epidemiology. 32(6), 574-585, 2008

Kolz, M., Baumert, J., Müller, M., Khuseyinova, N., Klopp, N., Thorand, B., Meisinger, C., Herder, C., Koenig, W. and Illig, T.: *Association between variations in the TLR4*

gene and incident type 2 diabetes is modified by the ratio of total cholesterol to HDL-cholesterol. BMC Medical Genetics. 9(1), 9, 2008

Herder, C., Illig, T., Baumert, J., Müller, M., Klopp, N., Khuseyinova, N., Meisinger, C., Martin, S., Thorand, B., Koenig, W.: *Macrophage migration inhibitory factor (MIF) and risk for coronary heart disease: Results from the MONICA/KORA Augsburg Case-Cohort Study, 1984-2002*. Atherosclerosis. 2008 Jan 31. Epub ahead of print.

Herder, C., Klopp, N., Baumert, J., Müller, M., Khuseyinova, N., Meisinger, C., Martin, S., Illig, T., Koenig, W., Thorand, B.: *Effect of macrophage migration inhibitory factor (MIF) gene variants and MIF serum concentrations on the risk of type 2 diabetes: results from the MONICA/KORA Augsburg Case Cohort Study, 1984-2002*. Diabetologia, 2008 Feb;51(2):276-84.

Kolz, M., Koenig, W., Müller, M., Andreani, M., Greven, S., Illig, T., Khuseyinova, N., Panagiotakos, D., Pershagen, G., Salomaa, V., Sunyer, J., Peters, A., for the AIRGENE Study Group: *DNA variants, plasma levels and variability of C-reactive protein in Myocardial infarction survivors: results from the AIRGENE study*. Eur Heart J. 2008 May;29(10):1250-8

Schaeffer L, Gohlke H, Müller M, Heid IM, Palmer LJ, Kompauer I, Demmelmair H, Illig T, Koletzko B, Heinrich J.: *Common genetic variants of the FADS1 FADS2 gene cluster and their reconstructed haplotypes are associated with the fatty acid composition in phospholipids*. Hum Mol Genet. 2006 Jun 1;15(11):1745-56.

Napieralski R, Ott K, Kremer M, Specht K, Vogelsang H, Becker K, Müller M, Lordick F, Fink U, Rüdiger Siewert J, Höfler H, Keller G.: *Combined GADD45A and thymidine*

phosphorylase expression levels predict response and survival of neoadjuvant-treated gastric cancer patients. Clin Cancer Res. 2005 Apr 15;11(8):3025-31.

Mueller, M.: *Goodness-of-fit criteria for survival data.* Discussion Paper 382. Sonderforschungsbereich 368, Ludwig-Maximilians-Universität, München, 2004.

Ott K, Vogelsang H, Mueller J, Becker K, Müller M, Fink U, Siewert JR, Höfler H, Keller G.: *Chromosomal instability rather than p53 mutation is associated with response to neoadjuvant cisplatin-based chemotherapy in gastric carcinoma*. Clin Cancer Res. 2003 Jun;9(6):2307-15.

Müller, M., Ulm, K. (2003): *Implementation of complex interactions in a Cox regression framework*. Discussion Paper 363. Sonderforschungsbereich 368, Ludwig-Maximilians-Universität, München, 2003.

A2. Curriculum vitae

Surname:	Müller
First name:	Andrea Martina
Date / place of birth:	20.08.1974, München
Marital status:	single

Academic Education:

Since Sept. 2004	PhD student at Institute of Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-Universität, Chair of Epidemiology, research unit 'Genetic Epidemiology', Neuherberg, Germany.
2001	Diploma in statistics, Ludwig-Maximilians-Universität München, Germany.
1999/2000	Course of practical work in biology “Botanisches Großpraktikum”
1999	Course of practical work in biology “Tropenbotanik”
1997/1998	Course of practical work in statistics “Immobilienmarktanalyse”
1997	Course of practical work in biology “Artenvielfalt in der Botanik”
1996	Course of practical work in biology “Botanisches Grundpraktikum”
1994/1995	Course of practical work in biology “Genetisches Grundpraktikum”

1994	Begin of studies in statistics with a focus of applications in biology, Ludwig-Maximilians-Universität München, Germany.
1994	Qualification for admission to university

Work Experience:

Since July 2008	Statistical researcher and consultant at the Department of Internal Medicine I (Cardiology), Klinikum Grosshadern, Ludwig-Maximilians-Universität, Munich, Germany
2004-2008	Statistical researcher and consultant at GSF-Institute of Epidemiology (renamed 2008 Helmholtz Zentrum München) and Institute of Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-Universität, Chair of Epidemiology, research unit 'Genetic Epidemiology', Neuherberg, Germany.
2001 – 2004	Statistical researcher and consultant with a focus on cancer research, Institute for Medical Statistics and Epidemiology (IMSE), research unit "Prognostic Factors", TU München, Germany.
1999 – 2001	Student assistant and statistical consultant, Institute for Medical Statistics and Epidemiology (IMSE), research unit “Prognostic Factors”, TU München, Germany.

1999 – 2001	Involvement into fieldwork and study design as interviewer for several research projects on infrastructure by the company Socialdata, Munich.