Technische Universität München

Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Fachgebiet Biostatistik

Incorporation of external information into clinical prediction models

Sonja Eva Grill

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:		Univ.-Prof. Dr. W. Windisch
Prüfer der Dissertation:	1.	Univ.-Prof. D. P. Ankerst, Ph.D.
	2.	Univ.-Prof. Dr. A. Tellier
	3.	apl. Prof. Dr. K. Herkommer

Die Dissertation wurde am 24.03.2016 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 06.07.2016 angenommen.

Contents

List of Figures

List of Tables

List of Abbreviations

ACS	American Cancer Society
AUC	area under the receiver operating curve
BCSC	Breast Cancer Surveillance Consortium
BI-RADS	Breast Imaging Reporting and Data Systems
CDC	Centers for Disease Control and Prevention
CEPH	Centre d'Etude du Polymorphisme Humain
DRE	digital rectal examination
EDRN	Early Detection Research Network
FDA	U.S. Food and Drug Administration
FDR	first degree relative
GWAS	genome-wide association study
HWE	Hardy-Weinberg-Equilibrium
IARC	International Agency on Research on Cancer
IHGSC	International Human Genome Sequencing Consortium
LR	likelihood ratio
OR	odds ratio
PCPT	Prostate Cancer Prevention Trial
PCPTRC	PCPT Risk Calculator
PSA	prostate specific antigen
RAF	risk allele frequency
RKI	Robert Koch-Institut
ROC	receiver operating characteristics
SDR	second degree relative
SEER	Surveillance, Epidemiology, and End Results Program
SFCD	Swedish Family Cancer Database
SNAP	SNP Annotation and Proxy Search
SNP	single-nucleotide polymorphism
ViraHepC	Study of Viral Resistance to Antiviral Therapy of Chronic Hepatitis C

Publications included in this Thesis

Grill et al. (2015a)

Grill S., Fallah M., Leach R. J., Thompson I. M., Freedland S., Hemminki K., and Ankerst D. P. Incorporation of detailed family history from the Swedish Family Cancer Database into the Prostate Cancer Prevention Trial Risk Calculator. *The Journal of Urology*, 193(2):460-465, 2015

Abstract

Purpose: A detailed family history provides an inexpensive alternative to genetic profiling for individual risk assessment. We updated the PCPT Risk Calculator to include detailed family histories.

Materials and methods: The study included 55,168 prostate cancer cases and 638,218 controls from the Swedish Family Cancer Database who were 55 years old or older in 1999 and had at least 1 male first-degree relative 40 years old or older and 1 female first-degree relative 30 years old or older. Likelihood ratios, calculated as the ratio of risk of observing a specific family history pattern in a prostate cancer case compared to a control, were used to update the PCPT Risk Calculator.

Results: Having at least 1 relative with prostate cancer increased the risk of prostate cancer. The likelihood ratio was 1.63 for 1 first-degree relative 60 years old or older at diagnosis (10.1% of cancer cases vs 6.2% of controls), 2.47 if the relative was younger than 60 years (1.5% vs 0.6%), 3.46 for 2 or more relatives 60 years old or older (1.2% vs 0.3%) and 5.68 for 2 or more relatives younger than 60 years (0.05%vs 0.009%). Among men with no diagnosed first-degree relatives the likelihood ratio was 1.09 for 1 or more second-degree relatives diagnosed with prostate cancer (12.7% vs 11.7%). Additional first-degree relatives with breast cancer, or first-degree or second-degree relatives with prostate cancer compounded these risks.

Conclusions: A detailed family history is an independent predictor of prostate cancer compared to commonly used risk factors. It should be incorporated into decision making for biopsy. Compared with other costly biomarkers it is inexpensive and universally available.

Candidate's contribution: Development of statistical methodology, analysis of the data, discussion of results, creating figures and tables, writing the manuscript, development of the

R-algorithm and online risk tool.

Grill et al. (2015b)

Grill S., Fallah M., Leach R. J., Thompson I. M., Hemminki K., and Ankerst D. P. A simple-to-use method incorporating genomic markers into prostate cancer risk prediction tools facilitates future validation. *Journal of Clinical Epidemiology*, 68(5):563-573, 2015

Abstract

Objectives: To incorporate single-nucleotide polymorphisms (SNPs) into the Prostate Cancer Prevention Trial Risk Calculator (PCPTRC).

Study Design and Setting: A multivariate random-effects meta-analysis of likelihood ratios (LRs) for 30 validated SNPs was performed, allowing the incorporation of linkage disequilibrium. LRs for a SNP were defined as the ratio of the probability of observing the SNP in prostate cancer cases relative to controls and estimated by published allele or genotype frequencies. LRs were multiplied by the PCPTRC prior odds of prostate cancer to provide updated posterior odds.

Results: In the meta-analysis (prostate cancer cases/controls = $386,538/985,968$), all but two of the SNPs had at least one statistically significant allele LR ($P < 0.05$). The two SNPs with the largest LRs were rs16901979 [LR = 1.575 for one risk allele, 2.552 for two risk alleles (homozygous)] and rs1447295 (LR = 1.307 and 1.887, respectively).

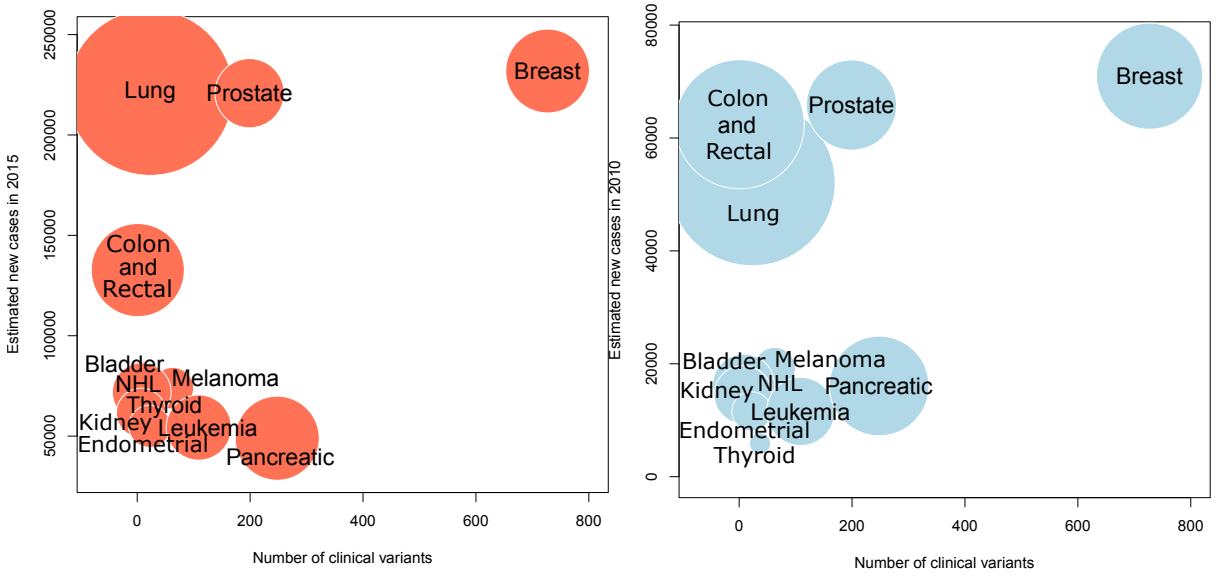
Conclusion: The substantial investment in genome-wide association studies to discover SNPs associated with prostate cancer risk and the ability to integrate these findings into the PCPTRC allows investigators to validate these observations, to determine the clinical impact, and to ultimately improve clinical practice in the early detection of the most common cancer in men.

Candidate's contribution: Development of statistical methodology, analysis of the data, discussion of results, creating figures and tables, writing and revising the manuscript, development of the R-algorithm and online risk tool. Preliminary results of the study were in part included in the candidate's master's thesis.

1 Introduction

With the emergence of targeted therapies and personalized approaches in oncology along with a new technological era of "Big Data", cancer risk prediction has evolved as a premier research area in statistics (Adams, 2015). Individualized risk assessment is especially relevant as cancer is one of the most common diseases. The life time risk of developing cancer for men and women at some point during their life is approximately 39.6 % according to U.S. numbers from the Surveillance, Epidemiology, and End Results Program (SEER, 2015a). In the U.S., for example, cancer is the second most common cause of death. Over half a million Americans were expected to die of cancer in 2015, representing about 24% of the total number of deaths, which is about 2.5 million per year in the U.S. in recent years (American Cancer Society, ACS 2015, and Centers for Disease Control and Prevention, CDC 2013). Using data published by the ACS for 2015, Figure 1a illustrates, in three dimensions, the total number of estimated deaths and new cases in the U.S. for the 12 most common cancers versus the number of published clinical variants associated with each cancer type. For comparison, most recent available numbers for Germany from 2010 for the same cancers are shown in Figure 1b (Robert-Koch-Institut, RKI 2013). The number of clinical variants were obtained from the ClinVar website (Landrum et al., 2014), which is an online archive with information on relationships between medically important variants, such as single-nucleotide polymorphisms (SNPs), and diseases. Lung cancer accounts for most deaths, followed by colon and rectal cancer in both countries. The most apparent difference between the U.S. and German numbers is that the estimated new cases for colon and rectal cancer is relatively higher in Germany, exceeding even that of lung cancer. In fact, colon and rectal cancer incidence is higher in Europe in general than in the U.S. (International Agency on Research on Cancer, IARC 2012). A study by Simko et al. states that one reason for this could be the identification of genetic mutations associated with an increased colon and rectal cancer risk in Ashkenazis in Central Europe (Simko and Ginter, 2016). Breast cancer has by far the most published clinical variants, which could be because it is the most common cancer in women leading to major investments in breast cancer research (RKI, 2013, ACS, 2015).

Figure 1 reflects the current high research effort dedicated to biomarkers, especially in genetics with respect to its role in cancer development and progression. With the life time risk of developing cancer of almost 40%, the need for screening programs, early detection, targeted therapies, and personalized approaches in oncology is becoming increasingly important (U.S.



(a) U.S.A. (b) Germany

Figure 1: Estimated new cases and deaths versus published clinical variants for the 12 most common cancers (a) for the U.S. in 2015 and (b) for Germany in 2010. Area of bubbles indicate the number of estimated deaths, where radius equals square root of corresponding number, after dividing by $\pi = 3.14$. The x-axis shows the number of clinical variants that were reported by multiple submitters on the ClinVar website and is the same in both subfigures. The y-axis shows the number of estimated new cases of the specific cancer type. NHL abbreviation for "Non Hodgkin Lymphoma". (RKI, 2013, Landrum et al., 2014, ACS, 2015)

Food and Drug Administration, FDA 2013, SEER 2015a). Development of statistical methods for calculating individual cancer risks plays a crucial role in translation. Large and complex data sets are continuously under construction and evolution, particularly in the field of genetics. These data sources have to be capitalized upon in order to improve cancer risk prediction, and by that support the fast pace of promising new treatment and early detection research (Early Detection Research Network, EDRN 2016). Risk prediction tools can no longer afford to be static, they must evolve with new data becoming available (Strobl et al., 2015). When novel risk factors are detected, it is desirable to update existing models to take advantage of the information. Information on the new markers may be available from relatively small case-control studies or large population registries, whose focus might differ from that of the existing model (Grill et al., 2015a,b). For example, critical variables in the existing model might not be collected at all.

In this thesis, statistical methods were developed to update the Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) with information on two new risk factors, a) detailed family history and b) SNPs (Grill et al., 2015a,b). Prostate cancer is the second leading cause of death for men in the U.S., after lung cancer, and the third most common cause of death for men

in Germany, after lung and colon and rectal cancer (SEER, 2015b, RKI, 2013). According to SEER statistics, 14% of men will develop prostate cancer at some point during their lifetime (SEER, 2015b), indicating that prostate cancer risk prediction is highly clinically relevant. Most prostate cancer risk prediction tools have been built based on single cohorts incorporating only traditional risk factors or are based on very small cohort sizes, see, for example, Thompson et al. (2006), Roobol et al. (2013). The wide range of prostate cancer risk prediction models alone has been noticed by Vickers (2010), for instance, and a similar updated literature search for publications from 01/01/1985 to 03/15/2016 yielded 894 results for the key words "nomogram prostate cancer" and 882 results for "prostate cancer risk prediction" on PubMed (2016).

The PCPTRC predicts the likelihood of detecting prostate cancer in the case that a biopsy is performed. The original version of the PCPTRC includes six risk factors, prostate-specific antigen (PSA), digital rectal examination (DRE) findings, age, race, family history of prostate cancer (yes/no), and prior biopsy history (Thompson et al., 2006). The PCPTRC online tool at myprostatecancerrisk.com assists clinicians and their patients around the world with clinical decision making, in particular whether or not a patient should have a prostate biopsy. Figure 2 shows a screenshot of the online tool. Aside from assisting patients, the online tool has fostered numerous external validations, rapidly gaining evidence regarding its appropriateness across a range of populations (Parekh et al., 2006, Eyre et al., 2009, Hernandez et al., 2009, Cavadas et al., 2010, Kaplan et al., 2010, Nam et al., 2011, Trottier et al., 2011, Oliveira et al., 2011, Zhu et al., 2012, Ankerst et al., 2012a, Lee et al., 2013). Since its establishment, the PCPTRC has been modified to incorporate newly discovered markers for prostate cancer, including the urine marker PCA3 and the serum marker percent free PSA, using a Bayesian technique called the likelihood ratio (LR) (Ankerst et al., 2008, 2012b). The online updated risk tool has also been validated in 218 patients from a multicenter Italian study by Perdona et al. (2011), 601 patients from a prospective study in Lyon, France by Ruffion et al. (2013) and 100 patients in Catania, Italy by Pepe and Aragona (2013).

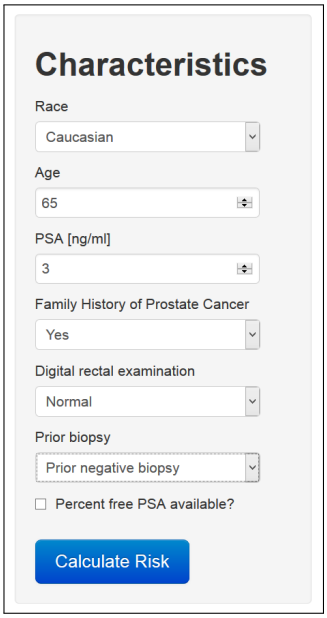


Figure 2: PCPTRC 2.0 entry page with six original risk factors (Ankerst et al., 2015).

The publications in this thesis extend the PCPTRC by two additional genetic risk factors, detailed family history and SNPs. It has been shown by large scale studies that familial back-

ground of prostate cancer and related cancers, often referred to as detailed family history, plays an important role in estimating a man's prior risk of developing prostate cancer (Roudgari et al. (2012) $n = 976,859$ participants; Albright et al. (2015) $n = 635,443$; Randazzo et al. (2015), $n = 4,932$). Commonly, family history is only crudely assessed by asking a simple yes/no question as to whether a patient ever had a first degree relative diagnosed with prostate cancer (Roobol et al., 2013, Ankerst et al., 2014). In Grill et al. (2015a) detailed family history measures with respect to prostate and breast cancer, including first and second degree relatives as well as age at diagnosis, were analyzed using registry data from Sweden contained in the Swedish Family Cancer Database (SFCD). The SFCD is a nationwide connected framework and one of the largest population registries worldwide (Hemminki et al., 2010). The PCPTRC was updated by LRs formed from detailed family history summaries obtained from the SFCD.

The second risk factor, studied in Grill et al. (2015b), was SNPs. A SNP characterizes the smallest possible mutation in the genome: the change of a single base pair of a gene (Campbell, 2008). Multiple confirmatory genome-wide association studies (GWASs) identifying common and rare SNPs for prostate and other cancers have been completed (Gudmundsson et al., 2007a, Yeager et al., 2007, Duggan et al., 2007, Thomas et al., 2008, Eeles et al., 2009, Al Olama et al., 2015). In the analysis to be discussed, SNPs from multiple studies were combined using a meta-analysis providing a statistically more reliable estimate (Grill et al., 2015b). The PCPTRC was updated by information on 30 SNPs, multiply validated as associated with an increased risk of prostate cancer. The purpose of both updates to the PCPTRC was to provide an extended risk prediction tool for prostate cancer that could be posted online.

Aside from the PCPTRC and other prostate cancer risk prediction models, hundreds of clinical risk prediction tools have been developed targeting different disease outcomes. Some of the most commonly used tools are the Framingham risk calculator for cardiovascular events (Wilson et al., 1998) and the Breast Cancer Risk Assessment Tool, often referred to as the Gail model (Gail et al., 1989). These tools have been built on very large cohort and trial populations. Variables included in the risk prediction models are routinely collected factors, such as blood serum markers, blood pressure, and demographic or behavioral measures like tobacco, alcohol consumption and age.

In addition to the established risk factors, novel markers, such as molecular or genetic markers, have been incorporated into many cancer risk models (Gail, 2008, 2009, Wacholder et al., 2010, Raji et al., 2010, Akamatsu et al., 2012, Lindström et al., 2012, Johansson et al.,

2012, Kader et al., 2012, Newcombe et al., 2012) as well as models for other conditions such as heart disease (Goldstein et al., 2014) or type 2 diabetes (Walford et al., 2014). Unfortunately, to the author's knowledge, to date nearly all of these have not been posted online, either because the novel markers failed to improve the model significantly or possibly because of a lack of interest in maintaining an online risk tool. The most recent version of the PCPTRC, called PCPTRC 2.0, was written in the R shiny package (Chang et al., 2015), which automatically connects to a specialized R server (Ankerst et al., 2015). The online PCPTRC has been managed by Prof. Ankerst as helpdesk since 2006. The tasks include answering patient queries and maintaining and financing the server. This illustrates the amount of upkeep necessary to maintain an online calculator.

In breast cancer risk models, new markers other than family history and SNPs have been included that go beyond those contained in the Gail model including breast density only (Tice et al., 2008) and breast density in combination with use of hormone therapy (Barlow et al., 2006). In these models, breast density was reported as part of a mammogram and classified into the four categories in the Breast Imaging Reporting and Data System's (BI-RADS) coding system (Barlow et al., 2006, Tice et al., 2008). However, only the model by Tice et al. is currently available online (BCSC Risk Calculator, 2015). The Farmingham risk score for coronary heart disease has been extended by six additional literature-derived risk factors, using synthesis analysis by Hu et al. (2014). The authors have used partially adjusted relative risks for combining the additional risk factors. Also this extension of the Farmingham risk score could not be found online.

In addition to the two applications for updating a risk prediction tool, an extensive simulation study was performed comparing several different updating methods to those used in Grill et al. (2015a,b), (Grill et al., 2016). The scenario where an existing risk prediction model is updated with information on a new marker from an external cohort or case-control study was investigated under various settings. Varying degrees of dependence between the old and new risk factors, and different types of markers were studied. Synthetic and data-based simulations were performed, with details discussed in Section 3.

2 Materials and Methods

2.1 Data sets

Data sets from different sources were analyzed in this thesis. In the following, brief descriptions of these are provided, details can be found in the respective publications.

Swedish Family Cancer Database (SFCD) In Grill et al. (2015a) data from the SFCD were analyzed. The SFCD includes the entire population of Sweden (those born after 1931 and their biological parents) and is the largest comprehensive family cancer registry in the world (Hemminki et al., 2010). Data contained in the registry are not self-reported, but instead assimilated from a nationwide connected network of multi-generational populations, death and cancer registries. Since the update in 2010, it now contains more than 12.2 million individuals and more than 1.1 million primary cancers (Hemminki et al., 2010). In the analysis in Grill et al. (2015a) men were selected according to the following criteria: alive at the beginning of the study period from 1999 to 2010 and free from prostate cancer, $\geq$ 55 years old in 1999, at least one recorded male FDR $\geq$ 40 years old and at least one reported female FDR $\geq$ 30 years old. Men meeting these requirements were divided into those who developed prostate cancer and those that did not during the subsequent 11 years up until 2010, resulting in 55,168 prostate cancer cases and 638,218 controls to be analyzed.

Genome-wide association studies (GWAS) In Grill et al. (2015b) publicly available results from GWASs in the form of high-risk allele frequencies or genotype counts were extracted for prostate cancer cases and controls. These GWASs were published in a prior meta-analysis on odds ratios (Kim et al., 2010) with the exception of one study that was published since then (Amundadottir et al., 2006, Duggan et al., 2007, Yeager et al., 2007, 2009, Gudmundsson et al., 2007a,b, 2008, 2009, Eeles et al., 2008, 2009, Sun et al., 2008, Al Olama et al., 2009, Hsu et al., 2009, Lindström et al., 2012). Most of the studies were performed on exclusively Caucasian populations including Iceland, Australia, Sweden, and the United States among others. Members of other ethnicities such as African Americans were excluded from the analysis. If a later published GWAS included the same participants as in a prior publication, the prior study was excluded to prevent double counting.

2.2 Updating a risk prediction model

It is assumed that an existing risk model includes p risk factors $\mathbf{X} = \left(X_1, \ldots, X_p\right)^T$ used for predicting a binary outcome Y, with $Y = 0$ denoting non-diseased and $Y = 1$ diseased. A new study or data source is available providing information on a new marker Z. The new study may measure parts or all of $\mathbf{X}$, but not necessarily.

The original risk prediction model is assumed to have been estimated using a logistic regression model, which is the most commonly used model for binary outcomes in medical statistics, yielding the probability of disease as

$$R_{\mathbf{X}} = \hat{P}\left(Y = 1|\mathbf{X}\right) = \frac{\exp(\gamma_0 + \boldsymbol{\gamma_1}^T\mathbf{X})}{1 + \exp(\gamma_0 + \boldsymbol{\gamma_1}^T\mathbf{X})}, \tag{2.1}$$

where $\boldsymbol{\gamma_1} = (\gamma_{11}, \ldots, \gamma_{1p})^T$ are the log odds ratios for $\mathbf{X}$. Extensions to other models for binary outcomes, such as the probit model, are straight forward. Conditioning on Z and $\mathbf{X}$, $P(Y|Z, \mathbf{X})$ in general has the form

$$P\left(Y|Z, \mathbf{X}\right) = \frac{P\left(Y|\mathbf{X}\right) P(Z|Y, \mathbf{X})}{\sum\limits_{Y} P\left(Y|\mathbf{X}\right) P(Z|Y, \mathbf{X})}, \tag{2.2}$$

using Bayes' Theorem. Taking the ratio of (2.2) evaluated at $Y = 1$ to $Y = 0$ and moving to the log odds scale yields

$$\underbrace{\log\left\{\frac{P(Y = 1|Z, \mathbf{X})}{P(Y = 0|Z, \mathbf{X})}\right\}}_{\log(posterior\,odds)} = \underbrace{\log\left\{\frac{P(Y = 1|\mathbf{X})}{P(Y = 0|\mathbf{X})}\right\}}_{\log(prior\,odds)} + \underbrace{\log\left\{\frac{P\left(Z|Y = 1, \mathbf{X}\right)}{P\left(Z|Y = 0, \mathbf{X}\right)}\right\}}_{log\{LR_Y(\mathrm{Z}|\mathbf{X})\}}, \tag{2.3}$$

where $LR_Y\left(\mathrm{Z}|\mathbf{X}\right)$ denotes the LR of Z conditional on $\mathbf{X}$ and according to disease status Y. This expression shows the relationship between the prior model and the desired updated posterior risk model.

Ankerst et al. (2008, 2012b) proposed updating risk models using the LR of $Z|(\mathbf{X}, Y)$. If the old and the new risk factors are independent or independence has to be assumed because of non-overlapping studies, the LR simplifies to

$$LR_Y\left(Z|\mathbf{X}\right) = \frac{P(Z|Y = 1, \mathbf{X})}{P(Z|Y = 0, \mathbf{X})} = \frac{P(Z|Y = 1)}{P(Z|Y = 0)} = LR_Y\left(Z\right), \tag{2.4}$$

which is sometimes referred to as "independence Bayes" (Hand and Yu, 2001). As stated before, the goal is to update an existing risk prediction tool by the information contained in a new factor

Z. This can be done by solving (2.3) for the risk of disease

$$P(Y = 1|Z, \mathbf{X}) = \frac{LR_Y\,(Z|\mathbf{X})\,(prior\,odds)}{LR_Y\,(Z|\mathbf{X})\,(prior\,odds) + 1}, \tag{2.5}$$

and the chance of "no disease"

$$P(Y = 0|Z, \mathbf{X}) = \frac{1}{LR_Y\,(Z|\mathbf{X})\,(prior\,odds) + 1}, \tag{2.6}$$

where the *prior odds* arise from the existing model, in the case of logistic regression: $\exp(\gamma_0 + \boldsymbol{\gamma_1}^T\mathbf{X})$.

The LR depends on the distribution of the new marker and can be calculated in different ways. In Ankerst et al. (2008) the LR for the novel urine marker PCA3 was estimated using multiple regression of log-transformed PCA3 on the predictors PSA, DRE and prior biopsy, fit separately to cancer cases and controls. The predictors were chosen using model selection techniques. Ankerst et al. (2012b) estimated the LR for percent free PSA and [-2]proPSA by fitting two multivariate regressions to cases and controls, using again model selection to chose which components of $\mathbf{X}$ to use as predictors. Another possibility of estimating the LR would be to fit one joint model to cases and controls by including the disease Y as an additional predictor into the model, thereby constraining the variances of the new marker distribution in cases and controls to be equal. A third alternative is to assume independence between the old and the new predictors using $LR_Y(Z)$ instead of $LR_Y(Z|\mathbf{X})$. These alternative ways for estimating the LR are examined via a simulation study in Section 3.

2.3 Detailed family history

For the incorporation of detailed family history in the PCPTRC 2.0, Grill et al. (2015a) used the LR method assuming independence, described in (2.4). Definitions of male FDRs and second degree relatives (SDRs), used in the SFCD, are illustrated in Figure 3. In addition the registry contains age at diagnosis and information on female FDRs and SDRs. Detailed family history patterns comprising the number of male FDRs and SDRs diagnosed with prostate cancer, their ages at diagnosis (<60 years versus ≥ 60 years of age), and number of FDRs diagnosed with breast cancer were computed for cases and controls. These patterns were chosen based on Roudgari et al., who found these summaries to be statistically significantly as-

sociated with prostate cancer risk (Roudgari et al., 2012). The FDR and SDR patterns as well as the combinations of those with age were grouped and analyzed as categorical variables rather than continuous variables due to small sample sizes especially for some extreme patterns. In more detail, FDR prostate cancer history was stratified by whether cancer was diagnosed before versus at or after age 60 as well as by whether zero, one, or two or more FDRs were diagnosed. FDR breast cancer and SDR prostate cancer history were only stratified into two groups, no versus one or more respective relatives diagnosed to avoid small sample sizes. The upper threshold of 60 years is commonly used to distinguish cancer diagnosed at an earlier age, thought to be a stronger genetic risk factor, compared with a later diagnosis (Roudgari et al., 2012). The LRs for FDR family history patterns were calculated for men meeting the requirements mentioned earlier, at least 1 male FDR ≥ 40 years old and at least 1 female FDR ≥ 30 years old.

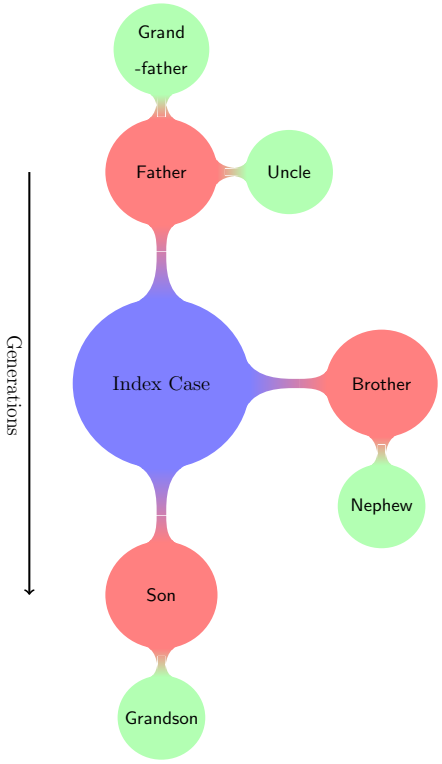


Figure 3: Pedigree for detailed family history measures of prostate cancer, with red indicating FDRs and green SDRs of the index case (Grill et al., 2015a).

The number of relatives falling into the three categories "Prostate Cancer < 60", "Prostate Cancer ≥ 60" and "Breast Cancer" resulted into analyzing 23 different FDR family history patterns (table in Grill et al. 2015a). Therefore, the new marker Z_{FDR} takes the values $Z_{FDR} = 1, ..., 23$, with each number representing one specific family history pattern. The LR is in this case defined as the ratio of two multinomial probability densities in cases versus controls. Let π_i^{ca} be the probability for a new subject to be in a specific detailed family history pattern $Z_{FDR} = i$ for cases ($Y = 1$) and π_i^{co} respectively for controls ($Y = 0$). The LR is then defined as

$$LR_Y(Z_{FDR}) = \frac{P(Z_{FDR} \mid Y = 1)}{P(Z_{FDR} \mid Y = 0)} = \frac{\prod_{i=1}^{23} (\pi_i^{ca})^{I(Z_{FDR}=i)}}{\prod_{i=1}^{23} (\pi_i^{co})^{I(Z_{FDR}=i)}}, \tag{2.7}$$

with $\sum_{i=1}^{23} \pi_i^{ca} = 1$, $\sum_{i=1}^{23} \pi_i^{co} = 1$ and I an indicator function with $I(E) = 1$ if event E is true and $I(E) = 0$ otherwise.

Let n_i^{ca} be the number of men with a specific family history pattern $Z = i$ in cases and n_i^{co} respectively in controls with n^{ca} and n^{co} the total number of cases and controls in the data set. In Grill et al. (2015a) the probabilities in the multinomial densities were estimated as empirical proportions from the data as follows:

$$\pi_i^{ca} = \frac{n_i^{ca}}{n^{ca}} \quad \text{and} \quad \pi_i^{co} = \frac{n_i^{co}}{n^{co}}. \tag{2.8}$$

In other words, the LR was calculated as the ratio of risk of observing a specific family history pattern in a prostate cancer case compared to a control. The resulting LR values for all 23 FDR family history patterns can be found in the table in Grill et al. (2015a).

The LRs for joint FDR and SDR family history were calculated by multiplying the LRs for FDRs by the conditional LRs for SDRs stratified by FDR family history

$$\begin{aligned} LR_Y(Z_{SDR}, Z_{FDR}) &= \frac{P(Z_{SDR}, Z_{FDR} \mid Y = 1)}{P(Z_{SDR}, Z_{FDR} \mid Y = 0)} \\ &= \frac{P(Z_{SDR} \mid Z_{FDR}, Y = 1)}{P(Z_{SDR} \mid Z_{FDR}, Y = 0)} \cdot \frac{P(Z_{FDR} \mid Y = 1)}{P(Z_{FDR} \mid Y = 0)}, \end{aligned} \tag{2.9}$$

where $Y = 1$ indicates prostate cancer and $Y = 0$ no prostate cancer. The conditional LRs were calculated for men fulfilling the same requirements as before and one additional condition: there exists at least 1 male SDR $\geq$ 40 years old for the index case, in order to avoid an underestimated contrast between those with affected SDRs and those without. The data underlying the calculation of the conditional LRs is given in Table 1. The resulting LR values are provided in the Supplementary Table in Grill et al. (2015a).

The confidence intervals of LRs were derived using the delta method and the Bonferroni adjustment accounting for the number of simultaneous intervals was used to obtain 95% overall confidence (Bishop et al., 1975). For updating the PCPTRC 2.0 with the corresponding detailed family history LRs, it was assumed that the original FDR family history question was answered with no in order to avoid double counting.

Family history pattern	Cases (%)	Controls (%)
No FDR family history (13,335 cases, 196,796 controls)		
No SDR	11,635 (87.3)	173,784 (88.3)
≥ 1 SDR	1,700 (12.7)	23,012 (11.7)
≥1 FDR ≥ 60 years, none < 60 years (3,891 cases, 25,380 controls)		
No SDR	3,339 (85.8)	22,042 (86.8)
≥ 1 SDR	552 (14.2)	3,338 (13.2)
≥1 FDR < 60 years, none ≥ 60 years (373 cases, 1,949 controls)		
No SDR	316 (84.7)	1,671 (85.7)
≥ 1 SDR	57 (15.3)	278 (14.3)
≥1 FDR < 60 years, ≥1 FDR ≥ 60 years (218 cases, 567 controls)		
No SDR	182 (83.5)	489 (86.2)
≥ 1 SDR	36 (16.5)	78 (13.8)

Table 1: Total number and percent of men with no or one or more SDRs diagnosed with prostate cancer stratified by four FDR prostate cancer family history patterns in cases and controls.

2.4 Single nucleotide polymorphisms

The aim of Grill et al. (2015b) was to incorporate SNPs into the PCPTRC to predict the risk of prostate cancer in a man considering prostate biopsy based on the original risk factors and additional information on SNPs. De Iorio et al. introduced a method for incorporating external information on linkage disequilibrium (LD) between genetic markers into a SNP-phenotype association analysis using odds ratios (De Iorio et al., 2011). LD is a nonrandom sharing of combinations of alleles/variants at two or more loci and will be covered in detail in the following paragraph (Lewontin and Kojima, 1960). Grill et al. (2015b) adapted this idea and developed two new meta-analysis approaches for LRs, one analyzing each SNP separately and one importing LD information between pairs of SNPs from external sources and by that analyzing multiple SNPs at the same time. In total, 30 SNPs were analyzed based on 22 GWASs, which were reported by Kim et al. with one additional study that has appeared since then (Kim et al., 2010, Lindström et al., 2012). The included SNPs were multiply validated and considered to be strongly linked to prostate cancer. An overview of which studies reporting which SNPs is given in Table 2. More detailed information for each SNP, extracted from the published papers, can

	SNP																											
Study	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
UK/AUS	x		x			x	x			x	x	x			x	x				x									x	
BPC3	x	x	x	x		x	x	x	x		x				x	x		x	x	x	x	x	x	x			x		x	x
Iceland(ICR)		x			x									x											x					
IKO,RUNMC		x			x								x	x			x						x	x	x					
Spain		x			x								x	x			x						x	x	x					
CAPS		x			x										x		x					x	x			x				
Baltimore		x																												
Chicago		x			x								x	x			x						x	x	x					
Nashville		x			x									x											x					
Rochester		x																												
GWAS Iceland				x				x	x				x			x	x	x	x		x	x	x	x			x	x		x
UK				x				x	x									x									x			x
AUS				x				x	x									x									x			x
Finland					x									x											x					
PLCO															x	x	x					x	x		x	x				
CPS-II															x	x	x					x	x		x	x				
ATBC															x	x	x					x	x		x	x				
HPFS															x	x	x					x	x		x	x				
CeRePP															x	x	x					x	x		x	x				
CONOR															x															
JHU															x							x	x			x				
EPIC															x															

Table 2: Overview of all GWASs and the corresponding SNPs that were measured by them. The references of all studies mentioned here as well as the SNP identifier corresponding to SNPs 1-30 are given in Grill et al. (2015b) and the original GWAS publications (Amundadottir et al., 2006, Duggan et al., 2007, Yeager et al., 2007, 2009, Gudmundsson et al., 2007a,b, 2008, 2009, Eeles et al., 2008, 2009, Sun et al., 2008, Al Olama et al., 2009, Hsu et al., 2009, Kim et al., 2010, Lindström et al., 2012).

be found in Grill et al. (2015b).

For analyzing multiple SNPs at the same time, linkage disequilibrium (LD) might need to be taken into account. If high LD is apparent between a pair of SNPs they can not be regarded as independent. LD decays with an increase in distance between two SNPs (Reich et al., 2001), where distance in this context refers to the number of base pairs between two loci. There are two commonly used measures for LD, the statistic D and the correlation coefficient r or its square r^2 (Lewontin, 1964, Reich et al., 2001, Balding et al., 2006). It is assumed that two alleles are coded by a/A and b/B. Let c_a, c_A, c_b, c_B denote the corresponding allele frequencies and $h_{11}, h_{12}, h_{21}, h_{22}$ the haplotype frequencies for the four possible combinations, where "1" stands for the small letters a or b and "2" stands for the capital letters A and B. Haplotypes are combinations of two alleles and their probabilities will be discussed in more detail in Section 2.4.2. Then r^2 is defined as follows:

$$r^2 = \frac{D^2}{c_A \cdot c_a \cdot c_B \cdot c_b} = \frac{(h_{22} - c_A \cdot c_B)^2}{c_A \cdot c_a \cdot c_B \cdot c_b}, \tag{2.10}$$

with $D = h_{22} - c_A \cdot c_B$ (Hill and Robertson, 1968, Balding et al., 2006). The lower the value

of r^2, the lower the LD. In the literature a threshold value of $r^2 < 0.2$ is commonly used as an indicator for SNPs to not be in LD, see, for example, Lindström et al. (2012). In the human genome each locus has its own structure concerning LD between SNPs (Bonnen et al., 2002). The online pairwise LD measurement tool SNP Annotation and Proxy Search (SNAP) was used to identify the LD structure (Johnson et al., 2008). SNAP provides LD information in the form of r^2, which is later transformed to the measure D. In SNAP different SNP reference data sets were available, including the HAPMAP (short for haplotype map, 3 different releases, The International HapMap Consortium 2003) and the 1000 Genome Project (1000 Genome Project, 2012). The Centre d'Etude du Polymorphisme Humain (CEPH) population (Utah Residents with Northern and Western European Ancestry, description in The International HapMap Consortium 2005) was chosen as the reference and both, the 1000 Genome Project as well as the HAPMAP were considered for obtaining estimates for the LD structure. In the following, the two meta-analysis methods are introduced, one taking LD structure into account and one assuming independence between SNPs and analyzing them separately.

2.4.1 Meta-analysis assuming independence between SNPs

GWASs typically report allele frequencies in cancer cases and controls by classifying one of the alleles as the risk allele, see, for example, Eeles et al. (2008) or Gudmundsson et al. (2009). The risk allele is defined as the allele which had the highest frequency among cancer cases in the study. The studies did not always agree on the risk allele classification, thus the one reported by the majority of studies in the analysis was used. The risk allele is denoted by R and the companion allele by r. Since humans are diploid organisms, there are three possible genotypes, rr, Rr and RR, necessitating the estimation of a probability density function for calculating the LRs comprising three probabilities (Campbell, 2008). The multinomial random variable Z is coded in the index as $Z = 0, 1, 2$ indicating the number of risk alleles for genotypes rr, rR, and RR, respectively. Frequencies of the genotypes in cases (ca) and controls (co) are denoted by π_Z^{ca} and π_Z^{co}, whereas c_R^{ca} and c_R^{co} denote the risk allele frequency (RAF) for cases and controls. The genotype frequencies were estimated assuming Hardy-Weinberg-Equilibrium (HWE) in two ways, depending on how they were reported in the study, either using RAFs, $\pi_0^{ca} = (1 - c_R^{ca})^2, \pi_1^{ca} = 2 \cdot c_R^{ca} \cdot (1 - c_R^{ca})$, or genotype counts, $\pi_0^{ca} = n_0^{ca}/N^{ca}, \pi_1^{ca} = n_1^{ca}/N^{ca}$. Hereby, n_Z^{ca} is the number of cases with the corresponding genotype $Z = 0, 1, 2$ in the study considered, N^{ca} is the total number of cases in the same study and respectively for controls. In both

cases $\pi_2^{ca} = 1 - \pi_0^{ca} - \pi_1^{ca}$ and respectively for controls. Since this relationship holds, the three probabilities within cases and controls are constrained by summing up to one. The HEW, named after Hardy and Weinberg, describes a population gene pool that is not undergoing evolution, which implicates that frequencies of alleles and genotypes remain constant over generations (Falconer and Mackay, 1996, Campbell, 2008). The three-dimensional vector of log-transformed LRs, corresponding to the values $Z = 0, 1, 2$, for study d ($d = 1, ..., D$) is:

$$\mathbf{T} = \log(\mathbf{LR})^d = \begin{pmatrix} \log(LR_0) \\ \log(LR_1) \\ \log(LR_2) \end{pmatrix}_d = \begin{pmatrix} \log(\frac{\pi_0^{ca}}{\pi_0^{co}}) \\ \log(\frac{\pi_1^{ca}}{\pi_1^{co}}) \\ \log(\frac{\pi_2^{ca}}{\pi_2^{co}}) \end{pmatrix}_d = \begin{pmatrix} \log(\frac{\pi_0^{ca}}{\pi_0^{co}}) \\ \log(\frac{\pi_1^{ca}}{\pi_1^{co}}) \\ \log(\frac{1-\pi_0^{ca}-\pi_1^{ca}}{1-\pi_0^{co}-\pi_1^{co}}) \end{pmatrix}_d . \tag{2.11}$$

It is assumed that $\log(\mathbf{LR})^d \sim N_3(\boldsymbol{\mu}_d, C_d)$ for $d = 1, ..., D$, a tri-variate normal distribution, with study-specific mean $\boldsymbol{\mu}_d = (\mu_1, \mu_2, \mu_3)_d$, and variance-covariance matrix C_d.

Next, the vector of study-specific means, also referred to as the random effects, is assumed to follow a normal distribution, $\boldsymbol{\mu}_d \sim N_3(\boldsymbol{\mu}, \Sigma)$, with $\boldsymbol{\mu}$ indicating the population mean vector and Σ the between-study variance-covariance matrix. A random-effects meta-analysis framework is used since this allows for a degree of heterogeneity among studies as opposed to a fixed-effects meta-analysis (van Houwelingen et al., 2002). This two-stage formulation implies that marginally the study estimates of the LRs follow a normal distribution with two additive components of variance, within- and between-study: $\log(\mathbf{LR})^d \sim N(\boldsymbol{\mu}, C_d + \Sigma)$. The following shows how the within-study variance matrix C_d is derived by the delta method based on study estimates. However, thereafter C_d for each study is assumed fixed and known in the meta-analysis.

It can be assumed that the genotype counts for cases and controls follow a multinomial distribution:

$$(n_0^{co}, n_1^{co}, n_2^{co})_d \sim \text{Multinomial}(N^{co}, \pi_0^{co}, \pi_1^{co}, \pi_2^{co})_d, \tag{2.12}$$

$$(n_0^{ca}, n_1^{ca}, n_2^{ca})_d \sim \text{Multinomial}(N^{ca}, \pi_0^{ca}, \pi_1^{ca}, \pi_2^{ca})_d. \tag{2.13}$$

In the two-dimensional case the variance-covariance matrix of the multinomial distribution for

the pair n_0^{ca}, n_1^{ca} reads for cases

$$\mathrm{cov}(n_0^{ca}, n_1^{ca})_d = \begin{pmatrix} N^{ca}\pi_0^{ca}(1-\pi_0^{ca}) & -N^{ca}\pi_0^{ca}\pi_1^{ca} \\ -N^{ca}\pi_0^{ca}\pi_1^{ca} & N^{ca}\pi_1^{ca}(1-\pi_1^{ca}) \end{pmatrix}_d \tag{2.14}$$

and respectively for controls. Moving from genotype counts to probabilities, the variance-covariance matrix of π_0^{ca}, π_1^{ca} has the form

$$\mathrm{cov}(\pi_0^{ca}, \pi_1^{ca})_d = \begin{pmatrix} \pi_0^{ca}(1-\pi_0^{ca})/N^{ca} & -\pi_0^{ca}\pi_1^{ca}/N^{ca} \\ -\pi_0^{ca}\pi_1^{ca}/N^{ca} & \pi_1^{ca}(1-\pi_1^{ca})/N^{ca} \end{pmatrix}_d \tag{2.15}$$

and similarly for controls. Only π_0^{ca} and π_1^{ca} are taken into account here, since π_2^{ca} can be calculated by $\pi_2^{ca} = 1 - \pi_0^{ca} - \pi_1^{ca}$ as stated earlier. Since cases and controls are considered independent, the variance covariance matrix of $\mathbf{\Pi}_d = (\pi_0^{co}, \pi_1^{co}, \pi_0^{ca}, \pi_1^{ca})_d^T$ is

$$V_d = \begin{pmatrix} \pi_0^{co}(1-\pi_0^{co})/N^{co} & -\pi_0^{co}\pi_1^{co}/N^{co} & 0 & 0 \\ -\pi_0^{co}\pi_1^{co}/N^{co} & \pi_1^{co}(1-\pi_1^{co})/N^{co} & 0 & 0 \\ 0 & 0 & \pi_0^{ca}(1-\pi_0^{ca})/N^{ca} & -\pi_0^{ca}\pi_1^{ca}/N^{ca} \\ 0 & 0 & -\pi_0^{ca}\pi_1^{ca}/N^{ca} & \pi_1^{ca}(1-\pi_1^{ca})/N^{ca} \end{pmatrix}_d . \tag{2.16}$$

The vector of LRs, $\mathbf{T}$ can be rewritten as

$$\mathbf{T} = \begin{pmatrix} \log(\frac{\pi_0^{ca}}{\pi_0^{co}}) \\ \log(\frac{\pi_1^{ca}}{\pi_1^{co}}) \\ \log(\frac{1-\pi_0^{ca}-\pi_1^{ca}}{1-\pi_0^{co}-\pi_1^{co}}) \end{pmatrix} = f(\pi_0^{co}, \pi_1^{co}, \pi_0^{ca}, \pi_1^{ca}) = f(\mathbf{\Pi}), \tag{2.17}$$

with $f : \mathbb{R}^4 \to \mathbb{R}^3$. Using the delta method, C_d is obtained by

$$C_d = \left(\frac{\partial T_i}{\partial \Pi_j}\right)_{i,j} \cdot V_d \cdot \left(\frac{\partial T_i}{\partial \Pi_j}\right)_{i,j}^T, \tag{2.18}$$

where $\left(\frac{\partial T_i}{\partial \Pi_j}\right)_{i,j}$ has the form

$$\left(\frac{\partial T_i}{\partial \Pi_j}\right)_{i,j} = \begin{pmatrix} -1/\pi_0^{co} & 0 & 1/\pi_0^{ca} & 0 \\ O & -1/\pi_1^{co} & 0 & 1/\pi_1^{ca} \\ 1/(1-\pi_0^{co}-\pi_1^{co}) & 1/(1-\pi_0^{co}-\pi_1^{co}) & -1/(1-\pi_0^{ca}-\pi_1^{ca}) & -1/(1-\pi_0^{ca}-\pi_1^{ca}) \end{pmatrix}. \tag{2.19}$$

The R package mvmeta by Gasparrini et al. (2012) can fit the general 2-stage normal multivariate random-effects meta-analysis (R Core Team, 2013). It was used to fit the meta-analysis model for LRs to one SNP at a time, choosing restricted maximum likelihood (REML) for the estimation. The list of within-study variance-covariance matrices C_d is part of the input to the mvmeta function as well as study-specific LR vectors estimated on each study measuring the corresponding SNP. The meta-analysis model uses these to estimate the overall mean $\boldsymbol{\mu}$ and the between-study variance-covariance matrix Σ. $\boldsymbol{\mu}$ and diagonal elements of $C_d + \Sigma$ yield 95% confidence intervals for the LR vectors that were used to update the PCPTRC.

The incorporation of multiple SNPs together into the PCPTRC is straight forward when independence is assumed between the SNPs. Under independence the LR for a group of n SNPs with respective genotypes Z factorizes

$$\begin{aligned} LR_Y(SNP_Z^1, \cdots, SNP_Z^n|\mathbf{X}) &= \frac{P(SNP_Z^1, \cdots, SNP_Z^n|\mathbf{X}, Y=1)}{P(SNP_Z^1, \cdots, SNP_Z^n|\mathbf{X}, Y=0)} \\ &= \frac{\prod_{i=1}^n P(SNP_Z^i|\mathbf{X}, Y=1)}{\prod_{i=1}^n P(SNP_Z^i|\mathbf{X}, Y=0)} = \prod_{i=1}^n LR_Z^i, \end{aligned} \tag{2.20}$$

where LR_Z^i is the LR of SNP i with genotype Z, which were obtained by separate meta-analyses. Due to this factorization, the joint effect of several SNPs can still be incorporated when assuming independence. Equation (2.20) shows the general formulation of the LR, for specific use the probability densities have an indicator function for the genotype as in (2.7).

2.4.2 Meta-analysis incorporating LD between SNPs

If LD structure was apparent between a group of SNPs, they were analyzed together in one meta-analysis to account for the LD dependency. This had to be done in conjunction with the incorporation of LD information from external sources, since the single studies only report marginal genotype frequencies and therefore no information on LD between pairs of SNPs.

Thus, the traditional meta-analysis framework was modified to allow the import of LD from external sources as described in the following. In contrast to the method in Section 2.4.1, where LD was ignored, the SNPs were not analyzed one at a time, but blocks of SNPs were formed of those in LD. The meta-analysis model for analyzing a block of SNPs is now described in detail.

The $3m$-dimensional vector of LRs for m SNPs in one LD block is assumed to follow a multivariate normal distribution

$$\begin{pmatrix} & \log(LR_0)^1 & \\ SNP^1 & \log(LR_1)^1 & \\ & \log(LR_2)^1 & \\ & \log(LR_0)^2 & \\ SNP^2 & \log(LR_1)^2 & \\ & \log(LR_2)^2 & \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ & \log(LR_0)^m & \\ SNP^m & \log(LR_1)^m & \\ & \log(LR_2)^m & \end{pmatrix}_d \sim \mathrm{N}\left(\begin{pmatrix} \mu_0^1 \\ \mu_1^1 \\ \mu_2^1 \\ \mu_0^2 \\ \mu_1^2 \\ \mu_2^2 \\ \cdot \\ \cdot \\ \cdot \\ \mu_0^m \\ \mu_1^m \\ \mu_2^m \end{pmatrix}_d , C_d \right), \tag{2.21}$$

with study-specific means $\boldsymbol{\mu_d}$, and variance-covariance matrix C_d, here a $3m \times 3m$ matrix. It is again assumed that $\boldsymbol{\mu_d} \sim N(\boldsymbol{\mu}, \Sigma)$ similar to Section 2.4.1 and therefore, marginally the log LR vector follows a $N(\boldsymbol{\mu}, C_d + \Sigma)$ distribution. The matrix C_d is constructed similarly as in the previous section.

How this is done is now explained. First, in contrast to the previous model, two SNP loci and their joint genotype probabilities are considered at the same time, since the LD is measured pairwise. The pairwise analyses are later connected to form a single within-study variance-covariance matrix C_d, comprising one LD block of m SNPs. The vector of joint probabilities of two loci is denoted as

$$\boldsymbol{\pi} = (\pi_{11}^{co}, \pi_{21}^{co}, \pi_{31}^{co}, \pi_{12}^{co}, \pi_{22}^{co}, \pi_{32}^{co}, \pi_{13}^{co}, \pi_{23}^{co}, \pi_{33}^{co}, \pi_{11}^{ca}, \pi_{21}^{ca}, \pi_{31}^{ca}, \pi_{12}^{ca}, \pi_{22}^{ca}, \pi_{32}^{ca}, \pi_{13}^{ca}, \pi_{23}^{ca}, \pi_{33}^{ca})_d^T$$

$= (\boldsymbol{\pi}^{\boldsymbol{co}}, \boldsymbol{\pi}^{\boldsymbol{ca}})_d^T$, separating cases and controls. This is necessary because the LR is a function of all components of this vector. $\boldsymbol{\pi}$ can be derived using haplotype probabilities assuming HWE and LD, shown in detail in Tables 3 and 4 (De Iorio et al., 2011).

Haplotype	a	A	Sum
b	$h_{11} = c_a c_b + D$	$h_{12} = c_A c_b - D$	c_b
B	$h_{21} = c_a c_B - D$	$h_{22} = c_A c_B + D$	$c_B = 1 - c_b$
Sum	c_a	$c_A = 1 - c_a$	1

Table 3: Haplotype probabilities for two loci. The LD is incorporated using the measure D, obtained from a separate source to the frequencies c_a and c_b.

Genotype	aa	aA	AA	Sum/ Marginal prob.
bb	$\pi_{11} = h_{11}^2$	$\pi_{12} = 2h_{11}h_{12}$	$\pi_{13} = h_{12}^2$	$\pi_{1.}$
bB	$\pi_{21} = 2h_{11}h_{21}$	$\pi_{22} = 2h_{11}h_{22} + 2h_{12}h_{21}$	$\pi_{23} = 2h_{12}h_{22}$	$\pi_{2.}$
BB	$\pi_{31} = h_{21}^2$	$\pi_{32} = 2h_{21}h_{22}$	$\pi_{33} = h_{22}^2$	$\pi_{3.}$
Sum/ Marginal prob.	$\pi_{.1}$	$\pi_{.2}$	$\pi_{.3}$	1

Table 4: Genotype probabilities for two loci as constructed from haplotypes. The allele in capital letters stands for the risk allele.

The haplotype probabilities $h_{11}, h_{12}, h_{21}, h_{22}$ and thus also the probabilities $\boldsymbol{\pi}$ are constrained since $c_a + c_A = 1$ and $c_b + c_B = 1$. The LD is incorporated in the form of $D = h_{22} - c_A c_B$, which measures the difference between the expected and the observed haplotype probabilities. In Table 3, D is added and subtracted from the product of allele frequencies, since the simple relationship $D_{AB} = -D_{Ab} = -D_{aB} = D_{ab}$ holds (Balding et al., 2006). Thus, if two loci are independent, $D = 0$, and for example, $h_{22} = c_A c_B$, meaning that the joint probability of two alleles occurring equals the product of the marginal probabilities that each occurs. Table 4 shows how haplotype probabilities are combined to form genotype probabilities. For example, genotype "bbaa" can only arise from a maternal and paternal haplotype of the form "ab" and "ab" and therefore has the probability $\pi_{11} = h_{11}^2$. The genotype "bBaA", however, can arise from multiple maternal and paternal haplotypes, "ab" and "AB" or "Ab" and "aB", and consequently the formula reflecting these combinations reads $\pi_{22} = 2h_{11}h_{22} + 2h_{12}h_{21}$, see Table 4.

It is assumed that $\boldsymbol{\pi}$ follows approximately a multivariate normal distribution with variance-covariance matrix V_d

$$V_d = \begin{pmatrix} \frac{1}{N^{co}}(D_{\boldsymbol{\pi}^{co}} - \boldsymbol{\pi}^{co}\boldsymbol{\pi}^{co^T}) & \mathbf{0} \\ \mathbf{0} & \frac{1}{N^{ca}}(D_{\boldsymbol{\pi}^{ca}} - \boldsymbol{\pi}^{ca}\boldsymbol{\pi}^{ca^T}) \end{pmatrix}_d, \tag{2.22}$$

where $D_{\boldsymbol{\pi}^{co}}$ is a 9×9 matrix that has $\boldsymbol{\pi}^{co}$ as the diagonal elements and zeros otherwise, and $D_{\boldsymbol{\pi}^{ca}}$ is defined similarly for cases. $\mathbf{0}$ is a 9×9 matrix only containing zeros, representing the independence between cases and controls that is assumed within each study. The vector of log LRs, $\mathbf{T}$, for two loci can be written in terms of the genotype probabilities $\boldsymbol{\pi}$

$$\mathbf{T} = \begin{pmatrix} \log(LR_0)^1 \\ \log(LR_1)^1 \\ \log(LR_2)^1 \\ \log(LR_0)^2 \\ \log(LR_1)^2 \\ \log(LR_2)^2 \end{pmatrix}_d = \begin{pmatrix} \log\left(\frac{\pi_{11}^{ca}+\pi_{21}^{ca}+\pi_{31}^{ca}}{\pi_{11}^{co}+\pi_{21}^{co}+\pi_{31}^{co}}\right) \\ \log\left(\frac{\pi_{12}^{ca}+\pi_{22}^{ca}+\pi_{32}^{ca}}{\pi_{12}^{co}+\pi_{22}^{co}+\pi_{32}^{co}}\right) \\ \log\left(\frac{\pi_{13}^{ca}+\pi_{23}^{ca}+\pi_{33}^{ca}}{\pi_{13}^{co}+\pi_{23}^{co}+\pi_{33}^{co}}\right) \\ \log\left(\frac{\pi_{11}^{ca}+\pi_{12}^{ca}+\pi_{13}^{ca}}{\pi_{11}^{co}+\pi_{12}^{co}+\pi_{13}^{co}}\right) \\ \log\left(\frac{\pi_{21}^{ca}+\pi_{22}^{ca}+\pi_{23}^{ca}}{\pi_{21}^{co}+\pi_{22}^{co}+\pi_{23}^{co}}\right) \\ \log\left(\frac{\pi_{31}^{ca}+\pi_{32}^{ca}+\pi_{33}^{ca}}{\pi_{31}^{co}+\pi_{32}^{co}+\pi_{33}^{co}}\right) \end{pmatrix}_d = \begin{pmatrix} \log\left(\frac{\pi_{.1}^{ca}}{\pi_{.1}^{co}}\right) \\ \log\left(\frac{\pi_{.2}^{ca}}{\pi_{.2}^{co}}\right) \\ \log\left(\frac{\pi_{.3}^{ca}}{\pi_{.3}^{co}}\right) \\ \log\left(\frac{\pi_{1.}^{ca}}{\pi_{1.}^{co}}\right) \\ \log\left(\frac{\pi_{2.}^{ca}}{\pi_{2.}^{co}}\right) \\ \log\left(\frac{\pi_{3.}^{ca}}{\pi_{3.}^{co}}\right) \end{pmatrix}_d = f(\boldsymbol{\pi}), \tag{2.23}$$

with $f : \mathbb{R}^{18} \to \mathbb{R}^6$ and the Jacobian of f with respect to $\boldsymbol{\pi}$, $\left(\frac{\partial f_i}{\partial \pi_j}\right)_{i,j}$:

$$\begin{pmatrix} \frac{-1}{\pi_{.1}^{co}} & \frac{-1}{\pi_{.1}^{co}} & \frac{-1}{\pi_{.1}^{co}} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\pi_{.1}^{ca}} & \frac{1}{\pi_{.1}^{ca}} & \frac{1}{\pi_{.1}^{ca}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{-1}{\pi_{.2}^{co}} & \frac{-1}{\pi_{.2}^{co}} & \frac{-1}{\pi_{.2}^{co}} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\pi_{.2}^{ca}} & \frac{1}{\pi_{.2}^{ca}} & \frac{1}{\pi_{.1}^{ca}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{-1}{\pi_{.3}^{co}} & \frac{-1}{\pi_{.3}^{co}} & \frac{-1}{\pi_{.3}^{co}} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\pi_{.3}^{ca}} & \frac{1}{\pi_{.3}^{ca}} & \frac{1}{\pi_{.3}^{ca}} \\ \frac{-1}{\pi_{1.}^{co}} & 0 & 0 & \frac{-1}{\pi_{1.}^{co}} & 0 & 0 & \frac{-1}{\pi_{1.}^{co}} & 0 & 0 & \frac{1}{\pi_{1.}^{ca}} & 0 & 0 & \frac{1}{\pi_{1.}^{ca}} & 0 & 0 & \frac{1}{\pi_{1.}^{ca}} & 0 & 0 \\ 0 & \frac{-1}{\pi_{2.}^{co}} & 0 & 0 & \frac{-1}{\pi_{2.}^{co}} & 0 & 0 & \frac{-1}{\pi_{2.}^{co}} & 0 & 0 & \frac{1}{\pi_{2.}^{ca}} & 0 & 0 & \frac{1}{\pi_{2.}^{ca}} & 0 & 0 & \frac{1}{\pi_{2.}^{ca}} & 0 \\ 0 & 0 & \frac{-1}{\pi_{3.}^{co}} & 0 & 0 & \frac{-1}{\pi_{3.}^{co}} & 0 & 0 & \frac{-1}{\pi_{3.}^{co}} & 0 & 0 & \frac{1}{\pi_{3.}^{ca}} & 0 & 0 & \frac{1}{\pi_{3.}^{ca}} & 0 & 0 & \frac{1}{\pi_{3.}^{ca}} \end{pmatrix}.$$

The variance-covariance matrix of $\mathbf{T}$ is then obtained by

$$\overline{C}_d = \left(\frac{\partial f_i}{\partial \pi_j}\right)_{i,j} \cdot V_d \cdot \left(\frac{\partial f_i}{\partial \pi_j}\right)_{i,j}^T, \qquad \overline{C}_d \in \mathbb{R}^{6\times 6}, \tag{2.24}$$

following the delta method. In a last step, the matrices $\overline{C}_d$, calculated for each pair of SNPs in one LD block, are fused to form the overall within-study variance-covariance matrix C_d. The R package mvmeta by Gasparrini et al. (2012) was again used for estimation. LR vectors and 95% confidence intervals were obtained as described in the previous section.

Simulations were performed to explore the impact of LD on the LR estimates, with increasing LD values between SNPs by running 1,000 simulations for each setting. Meta-analyses accounting for LD versus assuming independence were compared with respect to resulting LR estimates and corresponding 95% confidence intervals.

3 Discussion

This chapter combines a discussion of the results from the two publications underlying this thesis (Grill et al., 2015a,b) and includes further research now submitted for publication.

3.1 Detailed family history - a quasi-genetic marker for cancer risk prediction

Advantages of the detailed family history update and general discussion of the method

Detailed family history is a convenient marker that is easy to collect compared to other genetic risk factors, such as SNPs, which can be expensive. Together with PSA and DRE, family history is routinely collected in clinical practice in the U.S. (Thompson et al., 2006, Ankerst et al., 2014). The flexibility of the LR approach used in Grill et al. (2015a) implies that the LRs available in analytical form could be applied to update any of the multiple prostate cancer risk prediction tools currently available online, not just the PCPTRC. This is possible since information on the risk factors already included in the PCPTRC is not needed to calculate the LRs. Another advantage of the LR method is that the information on detailed family history from a separate large registry, the SFCD, can be fused with the PCPT cohort, where detailed family history was not collected. The SFCD is currently the largest source of comprehensive detailed family history measures, offering sufficient statistical power for the determination of the independent predictive value of family history on prostate cancer risk (Hemminki et al., 2010).

Common clinical risk factors, such as PSA or DRE, are neither measured in the SFCD nor in most large cohorts specializing in genetic markers, such as the many GWASs used for detection and validation of SNPs. A mathematical merger of large cohorts specializing in the measurement of markers (SFCD for detailed family history versus PCPT for clinical predictors) is needed and this is accomplished through the LRs. Clinical risk prediction tools, that have incorporated family history in addition to the established clinical markers, have commonly relied on self-report limited to a single yes/no question concerning FDR family history of prostate cancer (Thompson et al., 2006, Nam et al., 2006, Macinnis et al., 2011, Roobol et al., 2013). In other diseases such as breast cancer for example this is also typically done (Barlow et al., 2006,

Tice et al., 2008). The study of Grill et al. (2015a) shows that detailed FDR and SDR family history has an added impact on prostate cancer risk and the data is based on exact records and confirmed cases from the SFCD.

In the calculation of the LRs for FDR and SDR family history, the dependence between the new risk factor, detailed family history, and the PCPTRC risk factors, PSA, DRE, age, race, FDR family history (yes/no) and prior biopsy was not taken into account within the strata of prostate cancer cases and controls. The following constraints were assumed for updating the PCPTRC with the detailed family history LRs: the answer to the FDR family history question of prostate cancer, that was already part of the PCPTRC, was set to no in order to not double count for the family history influence on prostate cancer risk. Furthermore, the update was constrained to Caucasians only, since the SFCD mainly included people of Caucasian ethnicity (Hemminki et al., 2010). There was no information on PSA, DRE or prior biopsy given in the SFCD, however, age and race information was available. Age was taken into account as a categorical stratification only: "FDRs with prostate cancer < 60 years" and "FDR with prostate cancer ≥ 60 years" were two variables included in the categorical analysis. Another adjustment could have been done, as an extension of this work, by fitting two separate multinomial models with 23 categories for the 23 detailed family history patterns (table in Grill et al. 2015a) to cases and controls with the overlapping risk factors as covariates similar to Ankerst and colleagues (Ankerst et al., 2008, 2012b). An alternative would be to fit only one multinomial model including the outcome Y as an additional covariate. These methods are discussed in Section 3.4. However, the multinomial models would have been very sparse especially with respect to rare family history patterns. Sample sizes as well as power of the statistical model would have been concerning issues as well. Therefore, these adjustments were not performed in the present study in order to preserve the large sample sizes that the SFCD offered as well as to provide an empirical estimate of the LRs without bringing in the inevitable assumptions of a statistical model.

Previous studies describing the association of family history with prostate cancer

Family history measures have been extensively investigated with respect to prostate cancer risk prediction for years. Albright et al. analyzed a total of 635,443 males, all with ancestral genealogy data (Albright et al., 2015). The prostate cancer diagnosis information was taken from the Utah Cancer Registry. First, second and third-degree relative risks were assessed.

Relative risk for a specific family history constellation was defined as the ratio of the number of observed cases versus the number of expected cases. Relative risks based on age at diagnosis were higher for earlier age at diagnoses, which is similar to the findings in Grill et al. (2015a). The presence of prostate cancer in second- and even third-degree relatives also contributed significantly to the risk of developing prostate cancer (Albright et al., 2015). Pakkanen et al. conducted a Finnish study, including 202 families with 617 prostate cancer cases and a control group of 3,011 individuals. They found an earlier age of onset of prostate cancer and observed a higher PSA level in prostate cancer cases with positive family history (Pakkanen et al., 2012). An earlier meta-analysis of 13 studies by Johns and Houlston reported that the risk of prostate cancer for FDRs of men with prostate cancer could be approximately 2.5-times greater than for men without a family history (Johns and Houlston, 2003). This is a selection of studies reporting the association of family history with prostate cancer risk among many others (Lesko et al., 1996, Ghadirian et al., 1997, Bratt et al., 1999, Hemminki and Czene, 2002, Nam et al., 2006, Hemminki et al., 2006, Xu et al., 2009, Williams et al., 2012).

Although the link of family history with an increased risk of prostate cancer has been shown in many studies as described above and also in the present study, this risk factor still has to be critically evaluated in each situation or study. According to the aforementioned studies family history plays an important role in diagnosis and onset of prostate cancer. However, there are studies that report no association of family history with the clinical endpoints: prostate cancer survival or recurrence free survival. Some examples are Siddiqui et al. (2006), Roehl et al. (2006) and Brath et al. (2015), among others. These findings underscore the importance of specifying the population and clinical outcome when investigating the impact of family history. Nevertheless, predicting prostate cancer diagnosis remains the focus of this thesis.

Next to the SFCD, other comparable population-based data sets or registries exist on which studies with focus on the association between family history and prostate cancer or other diseases were performed. Matikaine et al. did a population-based, cancer registry study in Finland including 1,546 prostate cancer patients and 11,427 FDRs, identified through parish records (Matikaine et al., 2001). Landgren et al. conducted a population-based case-control study with data from Sweden and Denmark to investigate the association of chronic lymphocytic leukemia with family history of autoimmune and other diseases (Landgren et al., 2006). In total approximately 80,000 index subjects and FDRs were analyzed. The Swedish subjects were extracted from the SFCD and the Danish subjects from the Danish cancer registry in

conjunction with the Danish central population registry. Peto et al. analyzed 3,295 breast cancer patients and 11,678 FDRs, both obtained from a register of households as part of a population-based cohort study in the UK (Peto et al., 1996). 32,534 individuals in the Iceland cancer registry were investigated by Amundadottir et al. (2004) with respect to patterns of cancer distributions in families. The authors state that 95% of the cancer cases in the registry are histologically verified. In the U.S. state Utah, a similar population registry to the SFCD exists: the Utah population database. Kerber et al., for example, analyzed 662,515 individuals from this database with respect to familial risk of 40 cancers (Kerber and O'Brien, 2005). Similar to Grill et al. (2015a), family history was not self-reported but registry-based in all five studies. Nevertheless, the SFCD currently remains the largest source of comprehensive detailed family history measures (Hemminki et al., 2010).

Limitations and future directions

In the following the study in Grill et al. (2015a) is discussed with respect to shortcomings, extensions and alternatives under consideration of the current state of research. An important limitation to note is that an internal or external validation set for the detailed family history update to the PCPTRC is lacking. This would be the proof-of-principle whether a model strategy works in practice. Currently, there is no suitable study available measuring both, the PCPTRC risk factors and detailed family history, which would be required in order to perform a validation of the updated risk model that contains both. To comply with the PCPTRC, it would be necessary to have PSA measured less than one year before biopsy as well as prior biopsy, race, age, DRE and detailed family history recorded among participants of one single large cohort. Large sample sizes of the order of those in the SFCD are needed for accurately assessing the effect of rare family history patterns, such as more than one FDR affected with prostate cancer. In order to redress this issue, the extended calculator has been made available online to facilitate external validation (www.myprostatecancerrisk.com, Ankerst et al. 2015).

Previous updates to the PCPTRC for other markers also validated this way (Parekh et al., 2006, Eyre et al., 2009, Hernandez et al., 2009, Scales et al., 2009, Cavadas et al., 2010, Kaplan et al., 2010, Nam et al., 2011, Trottier et al., 2011, Oliveira et al., 2011, Perdona et al., 2011, Zhu et al., 2012, Ankerst et al., 2012a,b, Lee et al., 2013, Pepe and Aragona, 2013). Figure 4 shows an online snapshot of the PCPTRC page including the detailed family history update

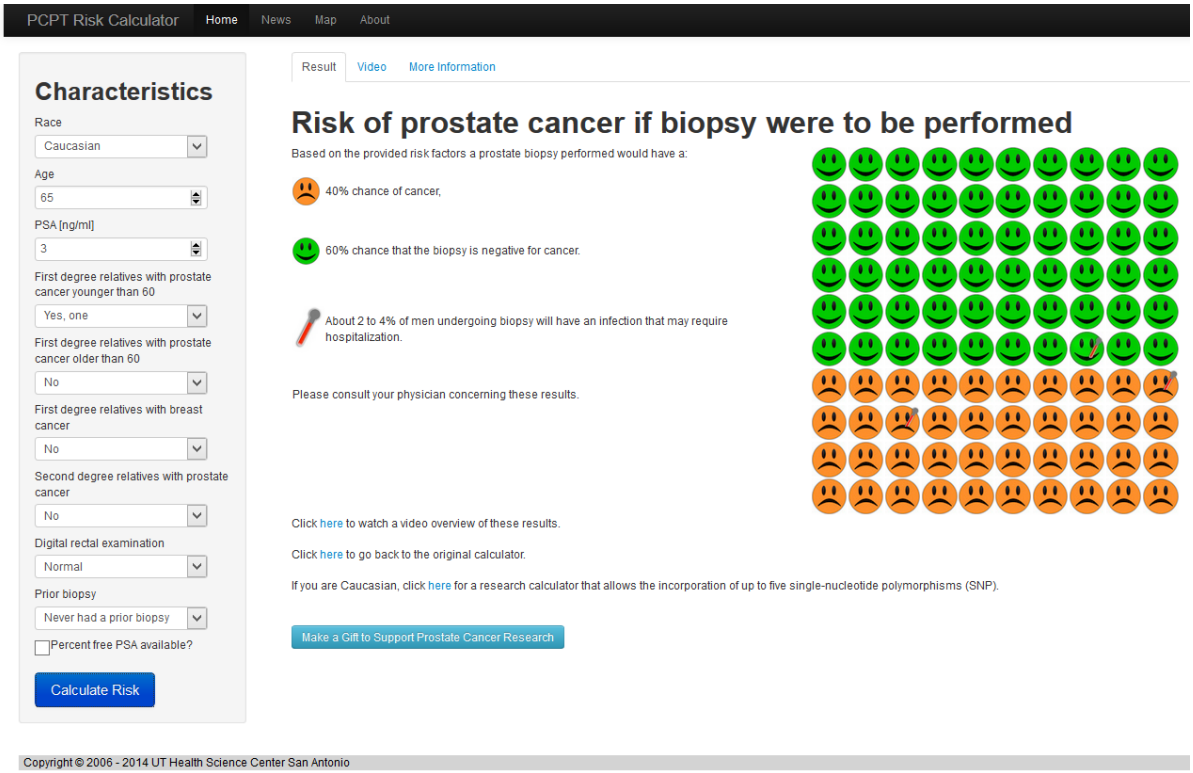


Figure 4: PCPTRC online entry page with detailed family history update. The risk of prostate cancer was calculated for a 65 year old Caucasian man with a PSA of 3 ng/ml, a normal DRE, no prior biopsy, one first degree relative diagnosed with prostate cancer younger than 60 years, no first degree relatives with prostate cancer older than 60 years, no first degree relatives with breast cancer and no second degree relatives with prostate cancer, www.myprostatecancerrisk.com (Ankerst et al., 2015).

with four new options: *1. First degree relatives with prostate cancer younger than 60, 2. First degree relatives with prostate cancer older than 60, 3. First degree relatives with breast cancer and 4. Second degree relatives with prostate cancer.* In this particular example the risk was calculated for a 65 year old Caucasian man with a PSA of 3 ng/ml, a normal DRE, no prior biopsy, one first degree relative with prostate cancer younger than 60, no first degree relative with prostate cancer older than 60, no first degree relative with breast cancer and no second degree relative with prostate cancer. A chance of prostate cancer of 40% is predicted if biopsy were to be performed. This is graphically illustrated by 40 sad orange emoticons and 60 happy green ones. This calculated risk can serve as one additional part of the decision process of the physician and the patient if a biopsy should be performed.

PSA screening leads to overdetection of low-grade prostate cancer resulting in a higher burden of morbidity for those who do not profit from medical treatment or interventions (Ilic et al., 2011, Lawrentschuk et al., 2011). Therefore, Gleason score, a measure for separating low-grade (Gleason grade < 7) and high-grade (Gleason grade ≥ 7) prostate cancer (Ankerst et al.,

2014), would have been a desirable additional information. The PCPTRC 2.0 distinguishes low-grade from high-grade prostate cancer (Ankerst et al., 2014), however, it was not available in the SFCD with the result that only cancer yes/no could be considered. An expectation bias might have increased rates of low-grade prostate cancer in the SFCD. A man who has a brother diagnosed with prostate cancer is more likely to get screened himself. This route increases the chance of cancer detection and association with family history because the person himself is more aware of the disease, which could then introduce expectation bias (Lang and Secic, 2006). Ankerst et al. showed that family history was significantly associated with low-grade but not high-grade prostate cancer, which again stresses the importance of assessing tumor grade in prostate cancer (Ankerst et al., 2014). If Gleason score would have been available, the LRs for updating the PCPTRC 2.0 could have been easily extended to allow 3 outcomes, no cancer, low-grade cancer and high-grade cancer, instead of the binary outcome cancer yes/no. This was done before in a similar fashion by updating the PCPTRC 2.0 with percent free PSA (Ankerst et al., 2014). A problem with tumor grading, however, are current trends in the grading system over time. Pathological grading patterns could have changed dramatically since the time of data acquisition in one of the two sources. Therefore, a dynamical annual updating technique could be considered as in Strobl et al. (2015).

The impact of the PSA screening era on LR estimates and a potentially resulting detection bias was assessed as well. Since much of the information in the SFCD comes prior to the PSA screening era in the late 1980s and prostate cancer screening remains less prevalent in Europe than in the US today (Brasso et al., 1998, Lu-Yao and Greenberg, 1994, Neppl-Huber et al., 2012), the effect of this bias on LR estimates is not expected to be of much magnitude. Nevertheless, this issue was investigated in the SFCD study time from 1999 to 2010 by dividing it into two periods, one from 1999 to 2005 and a later one from 2006 to 2010. This division was performed to investigate if this rather long time frame might have disguised a time period effect. The analysis revealed that the LR estimates of the family history patterns mostly increased in the second period, which could indeed be interpreted as a small but increased detection bias. Yet, in both periods the LR estimates were similar confirming detailed family history as an independent predictor of prostate cancer.

Another limitation of the study in Grill et al. (2015a) is that the LRs were estimated from the Swedish population whereas the original PCPTRC has been built on an American study population (Ankerst et al., 2014). Therefore, two different populations were combined in the

analysis. This was a necessity since a similar U.S. American registry like the SFCD was not available. One aspect that differs widely between both countries is the health care system. In Sweden, health care is universal and every legal resident has coverage (Mossialos et al., 2014). In the U.S., however, before compulsory insurance through Obama-Care was introduced, about 20% of the people aged 19 to 64 were uninsured, according to numbers from June-to-September 2013 (Collins et al., 2014). The question arises if both health care systems in Sweden and the U.S. lead to a potentially differing incidence or diagnosis rates. In the detailed family history update of the PCPTRC, the prevalence is captured in the intercept, γ_0, which was estimated on the original PCPTRC population as the updated probability of prostate cancer is given by

$$P(Y = 1|Z, \mathbf{X}) = \frac{LR_Y(Z)\exp(\gamma_0 + \boldsymbol{\gamma}_1^T\mathbf{X})}{LR_Y(Z)\exp(\gamma_0 + \boldsymbol{\gamma}_1^T\mathbf{X}) + 1}. \tag{3.1}$$

Since the PCPTRC was built on a U.S. population and is primarily intended for a U.S. population, the prevalence is well defined. If, however, the PCPTRC was applied to a different population, the possible impact of differences in the populations would need to be well investigated and adjustments might have to be performed. The Swedish and the PCPT study population could also differ in their age or race distribution. As described before, the detailed family history update of the PCPTRC was limited to Caucasians only to overcome a possible difference in the race distribution between the two populations. However, possibly differing age distributions were not investigated. Thus, the generalizability of this combination of two populations still needs to be validated by an external data set. If there would be a considerable difference between the two populations, methods need to be developed that consider, for example, shrinkage or distance metrics in order to combine two data sources that differ in their covariate distributions. Methods dealing with similar situations were proposed by Wiens et al. (2014) and Debray et al. (2015) and are described in Section 3.4.6.

With the determination of the sequence of the human genome, the field of genetics has made significant progress in the last two decades (IHGSC, 2004). The identification of genetic variants and their association with cancer risk has become a large research field (see also Figure 1). Therefore, SNPs, as an alternative to the quasi-genetic marker, detailed family history, will be discussed in the next section.

3.2 SNPs as biomarkers for cancer risk prediction

Previous studies incorporating SNPs into risk prediction tools

There is a pressing need to detect better markers for the early detection of many diseases, including cancer. This has lead to a massive amount of studies and research in general towards detection of risk-associated genetic loci and other biomarkers as illustrated in Figure 1. Numerous studies have developed a genetic multi-risk score for prostate cancer risk prediction and most of them have shown that the score itself (Machiela et al., 2011, Zheng et al., 2012, Wu et al., 2013, Szulkin et al., 2015) or wrapped in a model with other clinical predictors improves discrimination (Newcombe et al., 2012, Xu et al., 2009, Hsu et al., 2010, Macinnis et al., 2011, Kader et al., 2012, Ribeiro et al., 2012, Agalliu et al., 2013, Jiang et al., 2013). A risk score is a weighted combination of several SNPs.

Zheng et al. built a genetic score comprised of 33 SNPs, which were found to be associated with prostate cancer and validated in European studies (Zheng et al., 2012). These 33 SNPs include all 30 SNPs that were analyzed in Grill et al. (2015b) as reported by Kim et al. (2010) and three additional ones. They re-estimated the odds ratios (ORs) to calculate the genetic score in a Chinese case-control study with 1,108 cases and 1,525 controls. The genetic score with all 33 SNPs, using 10-fold cross validation reached an area under the receiver operating curve (AUC) of 0.63 and 0.62 when only using 11 significant SNPs (Zheng et al., 2012). The AUC equals the probability that the model classifies a case higher than a non-case and ranges between 0.5 and 1. A value of 0.5 is only as good as random guessing and values higher than 0.5 are desired (Fawcett, 2006).

Machiela et al. evaluated genetic scores for prostate cancer in 1,164 cases and 1,113 controls from the Prostate Lung Colorectal and Ovarian Cancer Screening Trial and breast cancer in 1,145 cases and 1,142 controls from the Nurses Health Study (Machiela et al., 2011). The genetic score comprised 30 SNPs and overlaps in 20 SNPs with the analysis in Grill et al. (2015b). Using 10-fold cross validation, the genetic score with only published risk alleles achieved the highest, but rather low AUC of 0.57 for prostate cancer and 0.53 for breast cancer (Machiela et al., 2011).

Szulkin et al. showed that adding new genetic variants to established variants in a polygenetic risk score increases the predictive accuracy for prostate cancer (Szulkin et al., 2015).

The risk scores were built on a training set with more than 25,000 individuals from different populations: Denmark, U.S., Germany, UK, Sweden, Australia, EU, Bulgaria and Poland. The scores were tested on an independent test set from the UK with 1,370 cases and 1,239 controls. The final score contained 65 SNPs including 23 of those analyzed in Grill et al. (2015b). However, no clinical variables were included and the detected increase in AUC from established to expanded SNPs was very low, from 0.67 to 0.68 (Szulkin et al., 2015).

The studies, that wrapped a genetic risk-score in a model with other clinical predictors (Newcombe et al., 2012, Xu et al., 2009, Hsu et al., 2010, Macinnis et al., 2011, Kader et al., 2012, Ribeiro et al., 2012, Agalliu et al., 2013, Jiang et al., 2013), in general, have the advantage that both types of predictors are measured on the same patients in contrast to the study in Grill et al. (2015b) where the predictors come from two different sources. Kader et al. evaluated 1,654 men, of which 410 had a positive biopsy and 1,244 a negative biopsy, in the placebo arm of the Reduction by Dutasteride of Prostate Cancer Events (REDUCE) trial (Kader et al., 2012). The clinical variables included in the model were age, DRE, prostate volume, PSA, PSA density, free-to-total PSA ratio, and number of cores sampled at baseline biopsy. Further, a genetic score of 33 selected SNPs, containing all 30 SNPs that were studied in Grill et al. (2015b), was assessed. The genetic score itself reached the highest AUC (0.59) compared to the remaining clinical variables, whereas both sources together obtained an AUC of 0.66. However, these AUCs were estimated using the entire cohort. When splitting the cohort into a fitting and a test set and using four-fold cross-validation, the AUC for the combined clinical and genetic model was only 0.64. This study is an example that has the advantage of having both types of predictors measured on the same patients. The tradeoff, however, is the small sample size of only 1,654 patients (Kader et al., 2012).

A commentary by Chatterjee et al. stated that including susceptibility SNPs into a risk prediction model that do not meet stringent genome-wide significance thresholds may also improve discriminatory performance (Chatterjee et al., 2011). The authors commented on a finding by van Zitteren et al. who reported an AUC of 0.67 for 41 genetic variants with regard to breast cancer risk prediction in the general population (van Zitteren et al., 2011). According to Chatterjee et al. this rather high AUC result could originate from genetic variants that are not detectable by standard GWAS protocols, such as the Illumina Infinium 660K array, which means that the 41 SNPs could include some rare, uncommon SNPs, deletions or copy number variations (Chatterjee et al., 2011).

Another important point when comparing different studies are the underlying patient profiles, for example, ethnicities, and the validation set that is used for calculating the AUCs. Validation results can differ strongly depending on the data set considered, see for example, Ankerst et al. (2012a). Therefore, the AUC reported by the different studies above have to be interpreted and compared with care also because some authors used internal and some external validation data sets.

Studies reporting conflicting results

There are conflicting reports suggesting that SNP information is not able to add much accuracy to a risk prediction tool that already includes the established risk factors for a certain disease. Sullivan et al. investigated the association of 61 SNPs with prostate-cancer specific endpoints as well as with PSA values at diagnosis in 1,354 individuals treated for localized prostate cancer (Sullivan et al., 2015). 10 of the SNPs analyzed in Grill et al. (2015b) were part of the 61 SNPs used by Sullivan et al. After correcting for multiple testing they found a significant link between one SNP (rs17632542) and PSA values. However, they did not find a significant association between any loci and disease-specific endpoints (Sullivan et al., 2015).

Little et al. confirmed this finding in a review (Little et al., 2015). The appraisal of 21 studies with multi-gene panels of 2 to 35 SNPs in prostate cancer risk assessment revealed that the improvement, with respect to clinical validity of SNP panels, is at best very small. A similar study by Park et al. came to the same conclusion for prostate and other types of cancer, arguing through simulations that SNPs to be discovered in the future by even larger GWASs than now will not add much discriminative accuracy to existing models (Park et al., 2012).

An earlier study by Park et al. estimated that on average 67 susceptibility loci exist, based on effect size calculations from published GWASs, for each of the three types of cancer: breast, prostate and colon cancer (Park et al., 2010). The authors included 20 already discovered SNPs into the effect size calculations, of which 5 were associated with prostate cancer and were all part of the analysis in Grill et al. (2015b). Park et al. stated that this group of 67 hypothetical and projected SNPs was estimated to only explain 17% of the genetic variation for each of the three cancer types and would only reach an AUC of 0.635 (Park et al., 2010).

These findings contradict the ones by van Zitteren et al. (2011) and Szulkin et al. (2015) mentioned above. Therefore, one can conclude that SNPs are not able to replace the established risk factors in cancer risk prediction, but rather have to be seen as an additional source of infor-

mation that can improve predictive accuracy. Accordingly, research still has to find additional strong risk factors in order to further significantly improve risk prediction models (Park et al., 2012).

Advantages of SNP update to the PCPTRC and general discussion of the method

Single SNPs have been validated by additional studies since the meta-analysis performed in Grill et al. (2015b) as for example the recently reported SNP rs2735839 by Helfand et al. (2015). Under the assumption of no LD, new results can be simply added to the PCPTRC by performing a meta-analysis on the LR for the new SNP. This is an advantage as any group of SNPs can be incorporated in a flexible and easy fashion since no strict pre-manufactured risk score comprising specific SNPs was used. Furthermore, the existing meta-analysis can be easily extended by a new study as the one by Helfand et al. (2015).

In addition, the information incorporated in Grill et al. (2015b) does not originate from a single study but was acquired by pooling information from several GWASs using a meta-analysis. Therefore, the estimated results are statistically more reliable. Especially in a data-intense era where "Big Data" has become a keyword (Adams, 2015), techniques for combining data from different sources have become more relevant. In the analysis in Grill et al. (2015b) data from multiple GWAS populations in the U.S. and Europe are fused with that from the U.S. PCPT population.

When comparing both meta-analysis approaches introduced in Sections 2.4.1 and 2.4.2, the two methods showed very similar results, probably due to very small LD values (see Table 2 in Grill et al. 2015b, all values of $r^2 < 0.1$). In most cases the meta-analysis taking LD into account had slightly more confined confidence intervals for the LR estimates. This occurs because more data is pooled when analyzing a group of SNPs together and thereby effectively increasing the sample size for that SNP by borrowing from neighbors, compared to analyzing each SNP in isolation. However, the approach incorporating LD also showed numerical instability for the largest LD block of 5 SNPs. The simulations that were performed to explore the impact of LD on the meta-analysis results further showed that the magnitude of LD between two SNPs did not have a high impact on the LR estimates or the 95% confidence intervals.

Difficulties with data from published GWASs

In some of the GWASs the controls also included women and children, as for example in Gudmundsson et al. (2008). Gudmundsson et al. used 21,372 controls in the Icelandic study population of which 12,060 were female. The age range of the controls was from 4 years to 102 years which implies that children were included. The inclusion of females and children as controls in a study on prostate cancer, which mainly affects men older than 50 years (ACS, 2015), is a major point and needs to be stated. However, this information was not clearly given in the main paper but was only found in a long supplementary appendix.

Overlap or reuse of populations is another issue with data from GWASs. Eeles et al., for example, analyzed a population from the UK and Australia in two stages (Eeles et al., 2008). The first stage included only data from studies in the UK, whereas in stage 2, cases and controls were selected partly from the same studies as in stage 1 with similar criteria. This indicates that data were probably reused in the second stage. Thus, studies have to be selected with care in order to only include independent results. In this particular case, only stage 2 was included in the meta-analysis in Grill et al. (2015b).

Limitations and future directions

A limitation of the study in Grill et al. (2015b) is that, with the SNP information originating from GWASs, no individual risk factor information, even age, is available. Thus, a possible dependence structure between the old and new risk factors cannot be taken into account and independence has to be assumed in constructing the LRs. In the updating method in Grill et al. (2015b), the LR is calculated by estimating two separate probability densities, one for cases and one for controls. This might reduce the severity of the independence assumption. Some studies, however, reported a dependence between SNPs and PSA (Gudmundsson et al., 2010, Sävblom et al., 2014, Chang et al., 2014). SNPs and race are also correlated so that GWASs are performed on separate ethnic groups. Gudmundsson et al. and Amundadottir et al. for example reported GWAS results for African Americans separately from European populations (Gudmundsson et al., 2007a, Amundadottir et al., 2006). As the PCPTRC population was predominantly Caucasian (> 95%, Ankerst et al. 2014) only GWASs based on Caucasian populations were selected for the meta-analysis in Grill et al. (2015b) and the update of the PCPTRC is only made possible for Caucasians. In the future, when more data might become available, a similar

update could be feasible with GWAS results for African American populations. The PCPTRC population was restricted to patients with 55 years of age or older. Information on individual patient age was not available from the GWASs. Therefore, the patients used for calculating the LRs could not be restricted to that age as it was done for the detailed family history update in Grill et al. (2015a). The two remaining original predictors in the PCPTRC, prior biopsy and DRE, are unlikely to be directly correlated with genetic variants, especially when cases and controls are treated separately.

An extension of the prediction of cancer versus no cancer would be to differentiate between high- and low-grade prostate cancer as implemented in the PCPTRC 2.0 (Ankerst et al., 2014). However, the GWASs considered here did not assess tumor grade. With regard to overdetection of low-grade, possibly non-life-threatening tumors, this is an important issue as discussed in detail in Section 3.1. GWAS results with the additional information on tumor grade would be very useful in the future to address this concern.

As for the detailed family history update, a suitable validation data set for the SNP update is lacking. For this reason, the PCPTRC including 30 SNPs, that have been found by several GWASs to be associated with prostate cancer, has been made freely available online and is open for validation. The study by Kader et al. described earlier would be suitable for validating the PCPTRC SNP update since it contains both, the clinical risk factors of the PCPTRC and information on SNPs (Kader et al., 2012). The genetic score of 33 SNPs even contains all 30 SNPs that were studied in Grill et al. (2015b). Negotiations between Prof. Dr. Ankerst and Dr. Kader are in progress. A limitation of the study by Kader et al., however, is the small sample size and that only simple self-reported and not detailed family history is measured.

Figure 5 shows the webpage of the updated PCPTRC with two selected SNPs. Up to five SNPs can be selected at the same time together with the genotype of zero, one or two risk alleles. In this particular example the risk was calculated for a 65 year old Caucasian man with a PSA of 3 ng/ml, no family history of prostate cancer, a normal DRE, no prior biopsy and the two SNPs, rs1465618 (1 risk allele) and rs12621278 (2 risk alleles). A chance of prostate cancer of 23 % is predicted if a biopsy were to be performed. This is only a 1% point higher chance than without the specification of the SNPs, where the risk of cancer would be 22%.

Furthermore, family history and SNPs next to other biomarkers and clinical variables seem to be just one part of the big picture regarding cancer research. It is now thought that not only

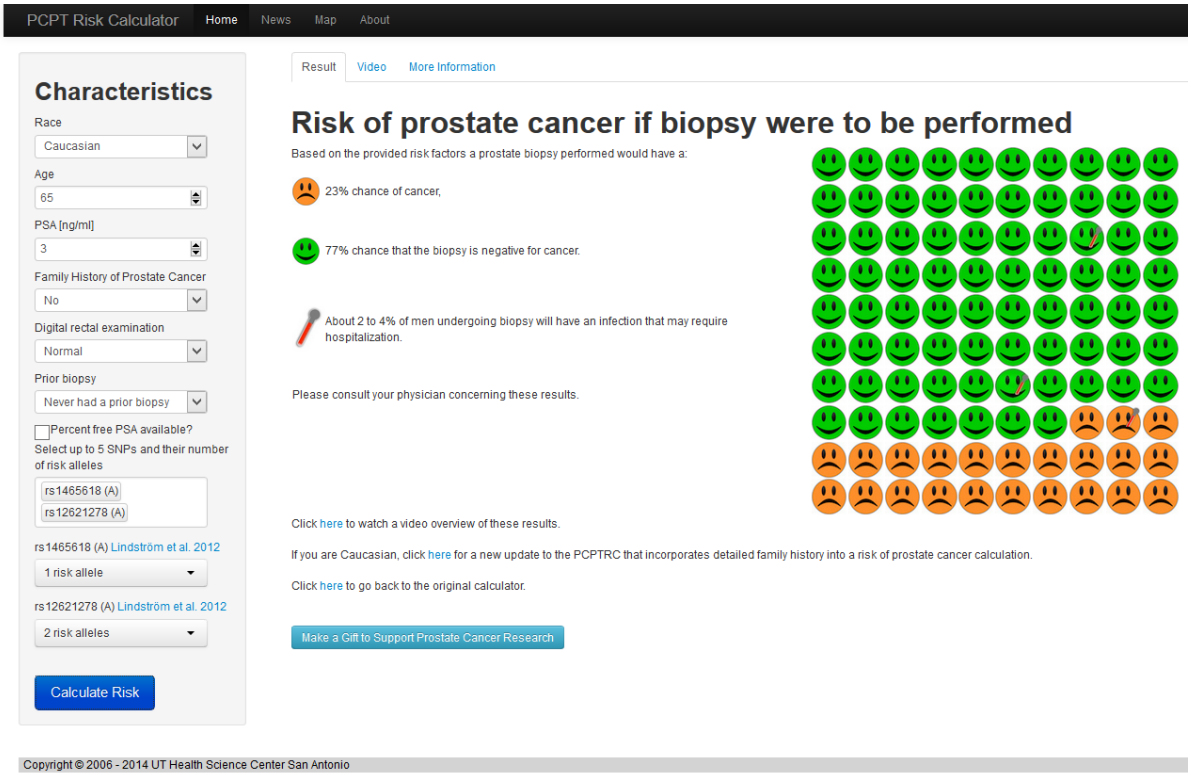


Figure 5: PCPTRC online entry page with SNP update. The risk was calculated for a 65 year old Caucasian man with a PSA of 3 ng/ml, no family history of prostate cancer, a normal DRE, no prior biopsy and the two SNPs, rs1465618 (1 risk allele) and rs12621278 (2 risk alleles), www.myprostatecancerrisk.com (Ankerst et al., 2015).

the genetic makeup itself including family history and genetic variants, but also the interplay between genetics and environment as well as lifestyle play an important role (ACS, 2015). Gene-environment interactions form a research field themselves in which also the establishment of new statistical methodology is needed (Hunter, 2005, Mukherjee et al., 2012, Wu et al., 2015). Inclusion of gene-gene or gene-environment interactions may improve discriminative accuracy (Purcell et al., 2009). Only very little cancer types are strongly hereditary, meaning that inherited genetic variants transfer a very high risk by themselves (ACS, 2015). Therefore, it might be worth taking gene-environmental interactions into account when updating an existing risk prediction tool as an extension of this work.

3.3 Comparison of both risk factors: detailed family history and SNPs

Not only the magnitude of the LR of a certain risk factor is of importance, but also its population prevalence, estimated here by the control prevalence in the studies (Kooter et al., 2011).

This means that the usefulness of a risk factor is also influenced by how many people in the population actually carry the risk factor. A rare allele, which only appears in less than 1% of the population, for example, implies that 99% are not affected by it and hence the risk remains unaffected in nearly all of the validation sets, making it difficult to note an improvement to predictions. In order to compare both risk factors, detailed family history and SNPs, considering this circumstance, a graph of LR values versus control prevalence is shown in Figure 6. The LR values for twelve different FDR and SDR family history patterns are taken from a condensed version of the table in Grill et al. (2015a) conditioning on at least ten people representing each family history pattern. Concerning SNPs, the genotype with two risk alleles was chosen. The detailed family history patterns show higher LR values than SNPs and in return SNPs show higher prevalences. Combinations of SNPs would have higher LRs but their corresponding population frequency would also decrease. An ideal marker, from a predictive point of view, would be one with a high population prevalence and high LR values, because this would affect a high fraction of the population and would have high impact. Both markers illustrate the trade-off between prevalence and effect in this context.

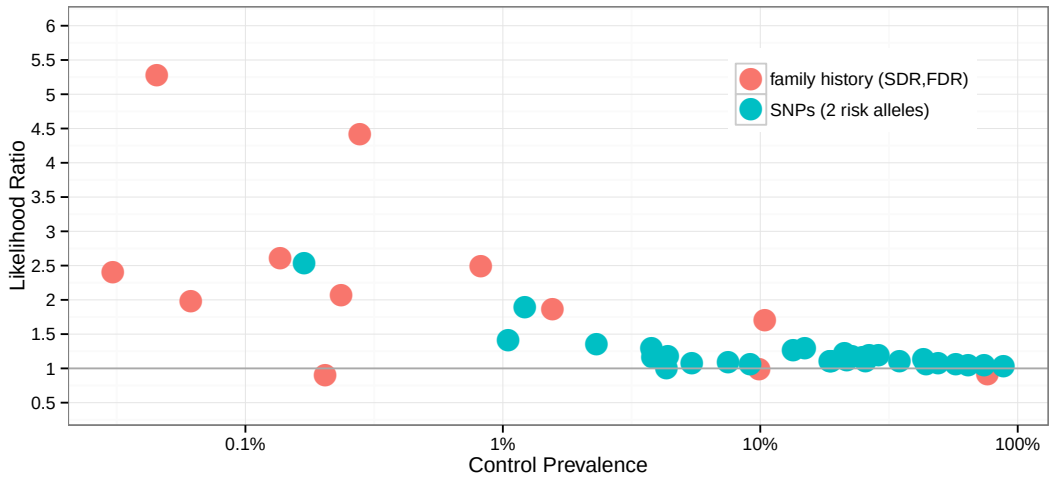


Figure 6: Magnitudes of LRs versus control prevalence for SFCD FDR and SDR detailed family history patterns and GWAS meta-analysis estimates for SNPs. For the LR values of the SNPs the genotype with 2 risk alleles was chosen. The control prevalence on the x-axis is on a $\log_{10}$ scale.

A series of risk models developed on 7,509 prostate cancer cases and 7,652 controls from the National Cancer Institute Breast and Prostate Cancer Cohort Consortium found that the best risk model included both, genetic markers and self-reported family history of prostate cancer (Lindström et al., 2012). 23 of the 25 SNPs studied by Lindström et al. were also analyzed

in Grill et al. (2015b). As described in the previous section, Kader et al analyzed 1,654 men following an initial negative biopsy and considered PSA and percent free PSA along with family history, other established clinical risk factors for prostate cancer and a 33-SNP genetic score. They found the simple yes/no FDR family history question to remain independently statistically significant. In the best clinical model for high-grade prostate cancer including the genetic score, family history had an odds ratio similar to that for a unit-increase of the genetic SNP score and both were significant (OR=2.20, p=0.002 versus OR=1.61, p=0.003, respectively, Kader et al. 2012). These findings suggest that the inclusion of both risk factors might lead to even better performance than just including one of the two markers into a prediction model.

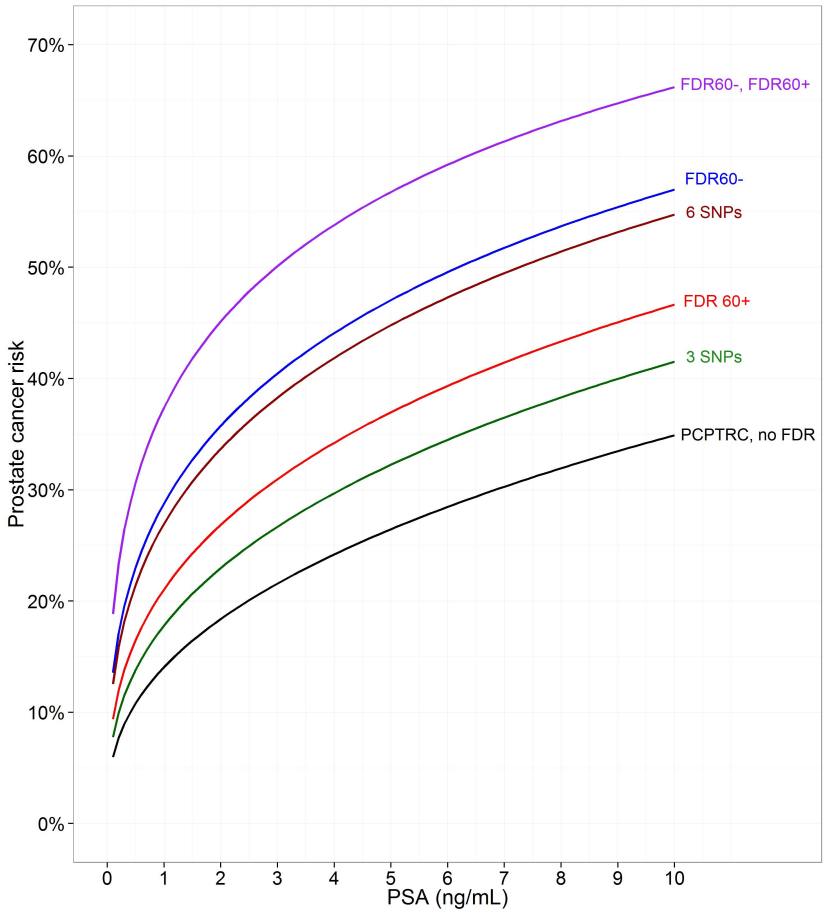


Figure 7: Risk curves according to FDR family history, SNPs and PSA level for a 65-year old Caucasian man with a normal digital rectal exam finding and no history of a prior biopsy. The bottom curve is the risk from the non-updated PCPTRC with no FDR, which corresponds to the risk for a man without history of a FDR with prostate cancer at any age. The upper 5 risk curves represent the updated PCPTRC for detailed family history from likelihoods computed from the SFCD for 3 different scenarios and 2 SNP risk curves for the following SNP profiles, 6 SNPs (rs620861, rs6983267, rs11649743, rs4430796, rs721048, rs1859962) and 3 SNPs (rs620861, rs6983267, rs11649743), each contributing two risk alleles.

Figure 7 compares risk curves for selected FDR family history patterns: a FDR with prostate cancer <60 years of age "FDR60-", a FDR with prostate cancer ≥ 60 years of age "FDR60+" and a combination of both scenarios "FDR60-, FDR60+", and 6 SNPs (rs620861, rs6983267, rs11649743, rs4430796, rs721048, rs1859962), each contributing two risk alleles. These SNPs were chosen because they were reported by a large number of studies, thereby producing accurate LR results. Risk curves are plotted as a function of PSA, with other PCPTRC variables remaining fixed. Two SNP risk curves are plotted, one with a combination of three SNPs (rs620861, rs6983267, rs11649743) and one with all six SNPs. The figure shows that the family history LRs and combinations of SNPs have similar net effects on the estimated risk curves.

3.4 A comparison of statistical methods for updating risk prediction models

In Sections 3.1 and 3.2 and the corresponding publications Grill et al. (2015a,b), a LR based method assuming independence between the old and the new risk factors was used, primarily because it was the only option due to lack of data. It is now investigated how the independence assumption impacts the LR methods to update a risk tool using simulations. Also other approaches proposed by several authors are discussed (Albert, 1982, Spiegelhalter and Knill-Jones, 1984, Janssens et al., 2005, Ankerst et al., 2008, 2012b) and their robustness to the independence assumption as well as other model assumptions is determined for comparison. This work is submitted for publication and in the following referred to as Grill et al. (2016). It is assumed that a risk prediction model using logistic regression was fit to a large cohort with original predictors, as for example the PCPTRC and as described in Section 2.2. This original model is then updated by information from a new cohort or case-control study which contains the original risk factors as well as a new marker. Various simulation settings with a range of dependence between the old and new risk factors are considered. The multiple methods are also applied to a real data example based on the Viral Resistance to Antiviral Therapy of Chronic Hepatitis C (ViraHepC) study, a multicenter clinical trial for testing the response to antiviral therapy for hepatitis C patients in African Americans and Caucasians (ViraHepC, 2002-2006).

3.4.1 Description of methods

The general setup for updating a risk prediction model has been described in Section 2.2. In the following several approaches, including the one used in Grill et al. (2015a,b), for incorporating a new marker from a new study into an existing risk model are investigated. Some methods assume independence between the new marker Z and the old predictors $\mathbf{X} = (X_1, ..., X_p)^T$ and some incorporate dependence in various ways. It is assumed that the true outcome-predictor relationship is given by the logistic model

$$P(Y = 1|Z, \mathbf{X}) = \frac{\exp\left(\mu_y + \beta_{yz}Z + \boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X}\right)}{1 + \exp\left(\mu_y + \beta_{yz}Z + \boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X}\right)}, \tag{3.2}$$

where $\boldsymbol{\beta}_{y\mathbf{x}} = (\beta_{y1}, \ldots, \beta_{yp})^T$ are the log odds ratios for $\mathbf{X}$ and β_{yz} is the log odds ratio for Z.

Approaches for estimating the LR

Joint estimation of the LR As mentioned in Section 2.2 there are several possibilities for estimating $LR_Y(Z|\mathbf{X})$. One way is to estimate both densities in the LR, $P(Z|Y = 1, \mathbf{X})$ and $P(Z|Y = 0, \mathbf{X})$, in a joint manner by including the outcome Y additionally to the old predictors $\mathbf{X}$ in a single regression model of Z. For a binary marker the single logistic model is $\text{logit}\{P(Z = 1|Y, \mathbf{X})\} = \alpha_0 + \boldsymbol{\alpha}_1^T\mathbf{X} + \alpha_2 Y$ and the LR model introduced in (2.3) can be written as

$$LR_Y(Z|\mathbf{X}) = \frac{\left\{1 + \exp\left(-\alpha_0 - \boldsymbol{\alpha}_1^T\mathbf{X}\right)\right\}^Z \left\{1 + \exp(\alpha_0 + \boldsymbol{\alpha}_1^T\mathbf{X})\right\}^{1-Z}}{\left\{1 + \exp\left(-\alpha_0 - \boldsymbol{\alpha}_1^T\mathbf{X} - \alpha_2\right)\right\}^Z \left\{1 + \exp\left(\alpha_0 + \boldsymbol{\alpha}_1^T\mathbf{X} + \alpha_2\right)\right\}^{1-Z}}. \tag{3.3}$$

And for a continuous marker the joint log LR model is

$$\begin{aligned} \log\{\text{LR}_Y(Z|\mathbf{X})\} &= -\left(Z - \alpha_0 - \alpha_2 - \boldsymbol{\alpha}_1^T\mathbf{X}\right)^2/(2\sigma^2) + \left(Z - \alpha_0 - \boldsymbol{\alpha}_1^T\mathbf{X}\right)^2/(2\sigma^2) \\ &= \alpha_2(Z - \alpha_0 - \boldsymbol{\alpha}_1^T\mathbf{X} - \alpha_2/2)/\sigma^2, \end{aligned} \tag{3.4}$$

where α_2 is the coefficient for Y in the linear model for the combined data and σ the standard deviation. In Equation (3.4) only the means differ but not the variances between cases and controls and therefore the log LR in the LR-joint model is linear in Z, which makes it equivalent to linear discrimination, see, for example, Chapter 6, Anderson (1984).

Separate estimation of the LR for cases and controls Here the numerator and the denominator of the LR are estimated by fitting two different models to the new data set, one

to cases ($Y = 1$) and one to controls ($Y = 0$). For example, when Z is a binary marker, fit by separate logistic regressions for cases and controls

$$\mathrm{LR}_Y(Z|\mathbf{X}) = \frac{\left\{1+\exp\left(-\alpha_{00}-\boldsymbol{\alpha}_{10}^T\mathbf{X}\right)\right\}^Z\left\{1+\exp(\alpha_{00}+\boldsymbol{\alpha}_{10}^T\mathbf{X})\right\}^{1-Z}}{\left\{1+\exp\left(-\alpha_{01}-\boldsymbol{\alpha}_{11}^T\mathbf{X}\right)\right\}^Z\left\{1+\exp\left(\alpha_{01}+\boldsymbol{\alpha}_{11}^T\mathbf{X}\right)\right\}^{1-Z}}, \tag{3.5}$$

where $\alpha_{1Y} = \left(\alpha_{11}^Y,\ldots,\alpha_{1p}^Y\right)^T$ are the parameters for $\mathbf{X}$ in the two separate models for cases ($Y = 1$) and controls ($Y = 0$). For a continuous marker Z that follows a normal distribution with $Z|(\mathbf{X},Y) \sim N(\alpha_{0Y}+\boldsymbol{\alpha}_{1Y}^T\mathbf{X},\sigma_Y^2)$, the fitting of two separate linear regressions to cases and controls yields

$$\log\left\{\mathrm{LR}_Y(Z|\mathbf{X})\right\} = \log(\sigma_0/\sigma_1) - \left(Z-\alpha_{01}-\boldsymbol{\alpha}_{11}^T\mathbf{X}\right)^2/(2\sigma_1^2) + \left(Z-\alpha_{00}-\boldsymbol{\alpha}_{10}^T\mathbf{X}\right)^2/(2\sigma_0^2), \tag{3.6}$$

where σ_1 and σ_0 are the standard deviations of the normal distribution in cases and controls respectively, and $\alpha_{0Y}, \boldsymbol{\alpha}_{1Y}$ are the parameters estimated in the two linear models. In Ankerst et al. (2008, 2012b), variable selection was used to reduce the complexity of the dependence of Z on $\mathbf{X}$. Here, the LR remains quadratic in Z thus this approach corresponds to quadratic discrimination (Chapter 6, Anderson 1984). Further theoretical results for estimating the LR-joint and LR-separate will be shown in Section 3.4.2.

LR estimation under independence of Z and $\mathbf{X}$ The LR method assuming independence between Z and $\mathbf{X}$ has been introduced in Section 2 and was used in Grill et al. (2015a,b). Under independence $\mathrm{LR}_Y(Z|\mathbf{X})$ simplifies to $\mathrm{LR}_Y(Z)$

$$LR_Y(Z|\mathbf{X}) = \frac{P(Z|Y=1,\mathbf{X})}{P(Z|Y=0,\mathbf{X})} = \frac{P(Z|Y=1)}{P(Z|Y=0)} = LR_Y(Z). \tag{3.7}$$

The LR in (3.7) was estimated by fitting two separate models in cases and controls and is termed "LR-ind" in the following in order to distinguish between the different methods.

Joint LR estimation under independence, logistic model with offset Analogous to Equation (3.3) for a binary marker Z with $\mathrm{logit}\{P(Z=1|Y)\} = \alpha_0 + \alpha_2 Y$ the following holds

under independence when fitting the two densities in the LR jointly:

$$\begin{aligned}\log\{LR_Y(Z)\} &= \log\left[\frac{\{1+\exp(-\alpha_0)\}^Z\{1+\exp(\alpha_0)\}^{1-Z}}{\{1+\exp(-\alpha_0-\alpha_2)\}^Z\{1+\exp(\alpha_0+\alpha_2)\}^{1-Z}}\right] \\ &= Z\log\left(\frac{1+\exp(-\alpha_0)}{1+\exp(-\alpha_0-\alpha_2)}\right)+(1-Z)\log\left(\frac{1+\exp(\alpha_0)}{1+\exp(\alpha_0+\alpha_2)}\right) \\ &= \delta_0+\delta_1 Z,\end{aligned} \tag{3.8}$$

with δ_0 and δ_1 defined accordingly as functions of (α_0, α_2). A corresponding result holds for a normally distributed marker and for the exponential family in general as shown in Grill et al. (2016). This linear dependency between $\log\{LR_Y(Z)\}$ and Z was already noted by Albert (1982). He proposed an updating method for a new binary or normal marker that included the prior odds $\gamma_0 + \boldsymbol{\gamma}_1^T\mathbf{X}$ with the parameters from the original model in (2.1) as an offset in a logistic regression of the outcome Y. The new marker Z was added as the only covariate (Albert, 1982). This results in fitting two new parameters, an intercept δ_0 and a parameter for Z, δ_1. The model $R_{\mathbf{X},Z}$ one wishes to estimate can then be written as

$$R_{\mathbf{X},Z} = \hat{P}(Y=1|Z,\mathbf{X},R_{\mathbf{X}}) = \frac{\exp(\gamma_0+\boldsymbol{\gamma}_1^T\mathbf{X}+\delta_0+\delta_1 Z)}{1+\exp(\gamma_0+\boldsymbol{\gamma}_1^T\mathbf{X}+\delta_0+\delta_1 Z)}, \tag{3.9}$$

and will be called "LR-offset" in the following.

LR estimation with shrinkage Spiegelhalter and Knill-Jones extended the approach in (3.7) with the addition of a shrinkage parameter (Spiegelhalter and Knill-Jones, 1984). They first estimated the $\mathrm{LR}_Y(Z)$ and then included this as an independent variable into a logistic regression for the outcome Y with the prior odds $\gamma_0 + \boldsymbol{\gamma}_1^T\mathbf{X}$ as an offset as performed by Albert (1982). The shrinkage factor θ is estimated in this second step. Hence, the posterior odds are

$$\log(\textit{posterior odds}) = \gamma_0 + \boldsymbol{\gamma}_1^T\mathbf{X} + \theta\log\{\mathrm{LR}_Y(Z)\}. \tag{3.10}$$

The additional parameter θ is included in order to adjust or allow for a certain dependence between the effect of Z and $\mathbf{X}$ on the posterior risk. A value of $\theta = 0$ would mean complete correlation between Z and $\mathbf{X}$ and therefore, the new marker does not add any information to the prediction and has no effect. Values between $\theta = 0$ and $\theta = 1$ include a certain dependence structure of different degrees. $\theta = 1$ corresponds to the LR-ind method. This approach will be called "LR-shrink" in the following.

Fitting logistic regression model to new data only

Another simple option to estimate $R_{\mathbf{X},Z}$ is to fit a logistic regression to the new data only and not use the old model at all. This clean-slate approach is investigated in the simulations in order to measure the gain in predictive performance of the other methods that actually include information on the old model $R_{\mathbf{X}}$. $R_{\mathbf{X},Z}$ estimated from this "Logistic new" approach can be written as

$$R_{\mathbf{X},Z} = \hat{P}\left(Y = 1|Z, \mathbf{X}\right) = \frac{\exp(\mu + \boldsymbol{\beta}_{\mathbf{X}}^T \mathbf{X} + \beta_Z Z)}{1 + \exp(\mu + \boldsymbol{\beta}_{\mathbf{X}}^T \mathbf{X} + \beta_Z Z)}. \tag{3.11}$$

3.4.2 Further results and adjustments under rare disease

In general all six methods described above can be applied to non-rare disease cases as well as rare disease cases. There is no strict definition of when a disease is considered rare in a statistical or mathematical context. In the simulations presented here, a prevalence of $P(Y = 1) = 10\%$ or smaller is assumed to characterize rare disease. The main difference between the LR-joint, LR-separate, LR-ind and the remaining three methods is, that the first three ones do not estimate an intercept from the new study used for updating. Therefore, the disease prevalence is specified by the original cohort. This does not require adjustment unless the model would be applied to a population with a different disease prevalence. The remaining methods LR-offset, LR-shrink and Logistic new in Equations (3.9)-(3.11), however, require an adjustment of the intercept if the new study is a case-control study. In a case-control study the disease prevalence does not reflect the one of the general population and thus it is not estimated correctly by these three methods. The method used in Grill et al. (2015a) and Grill et al. (2015b) was the LR-ind and therefore, did not need adjustment although the SNP data came from case-control GWASs, since, for the LR-ind, the disease prevalence is estimated on the prior study. An adjustment is now proposed for the rare disease case for the three methods LR-offset, LR-shrink and Logistic new and it is stated in the following at which step rare disease has to be assumed. For the non-rare disease case a corresponding method still needs to be developed and was not covered in Grill et al. (2016). Prostate cancer is not a rare disease, the prevalence in the PCPTRC cohort, for example, was 18% (Ankerst et al., 2014). However, rare disease scenarios are applicable to many diseases such as thyroid cancer or stomach cancer, which are a lot more rare as they only have a lifetime risk of 1.1% and 0.9% (SEER, 2016b,a). When the new data set is a cohort this problem does not occur for any of the six methods. The disease prevalence is in this sense well

defined in a cohort and therefore the intercept is estimated correctly for the rare disease as well as the non-rare disease case.

Intercept adjustment for LR-offset, LR-shrink and Logistic new for updating with case-control data

When the disease prevalence is known from an external source without error, then the adjusted intercept can be obtained by solving the following equation for μ^*

$$P(Y=1) - \int_Z \int_{\mathbf{X}} \frac{\exp(\mu^* + \hat{g}(\mathbf{X}, Z))}{1 + \exp(\mu^* + \hat{g}(\mathbf{X}, Z))} d\hat{F}(\mathbf{X}, Z) = 0. \tag{3.12}$$

For model (3.9), $\hat{g}(\mathbf{X}, Z) = \boldsymbol{\gamma}_1^T \mathbf{X} + \hat{\delta}_1 Z$, and for model (3.11), $\hat{g}(\mathbf{X}, Z) = \hat{\boldsymbol{\beta}}_{\mathbf{X}}^T \mathbf{X} + \hat{\beta}_Z Z$. The intercept adjustment for LR-shrink is performed in two steps. First an additional intercept θ_0 is included in the model such that

$$\log(\textit{posterior odds}) = \gamma_0 + \boldsymbol{\gamma}_1^T \mathbf{X} + \theta_0 + \theta_1 \log\{\mathrm{LR}_Y(Z)\}. \tag{3.13}$$

In contrast to Equation (3.10), θ_0 here absorbs the case-control sampling ratio. In a second step this intercept needs adjustment by solving Equation (3.12) with $\hat{g}(\mathbf{X}, Z) = \boldsymbol{\gamma}_1^T \mathbf{X} + \hat{\theta}_1 \log\{\mathrm{LR}_Y(Z)\}$. The empirical distribution function, $\hat{F}(\mathbf{X}, Z)$, is estimated in the simulations by splitting it further into $\hat{F}(\mathbf{X}, Z) = \hat{F}(Z|\mathbf{X})\hat{F}(\mathbf{X})$, where $\hat{F}(Z|\mathbf{X})$ is estimated from the controls in the new study. This can be performed when the disease is rare and the controls constitute a random sample of the general population. Then the empirical distribution function $\hat{F}(Z|\mathbf{X})$ estimated on the controls approximates the one of the general population. This is the point where the rare disease assumption was necessary for performing the adjustment in (3.12). If the information is available $\hat{F}(\mathbf{X})$ can be estimated from an external big data source as for example the one where the original model was built on in this simulation setup.

Theoretical results

Furthermore, a theoretical result will be proofed in the following regarding the estimation of the LR in the LR-joint and LR-separate method. This result requires, however, the rare disease assumption. Nevertheless, the simulations revealed that the LR-joint showed very good calibration for all rare and non-rare disease settings, the LR-separate in most scenarios and both methods seem to be very robust in this regard.

If, for a binary marker Z, $P(Z|\mathbf{X})$ is logistic, i.e. $\text{logit}\{P(Z|\mathbf{X})\} = \mu_{zx} + \beta_{z\mathbf{x}}^T\mathbf{X}$, then, under the assumption of rare disease, also $P(Z|\mathbf{X},Y)$ is logistic as shown below. It is assumed that also the distribution of $Y|(X,Z)$ is logistic as stated in Equation (3.2). In general $P(Z|\mathbf{X},Y)$ can be rewritten using Bayes' theorem as follows:

$$P\left(Z|Y,\mathbf{X}\right) = \frac{P\left(Y|Z,\mathbf{X}\right)P\left(Z|\mathbf{X}\right)}{P\left(Y|\mathbf{X}\right)}. \tag{3.14}$$

For controls $(Y = 0)$ this yields

$$\begin{aligned} P\left(Z|Y=0,\mathbf{X}\right) &= \frac{P\left(Y=0|Z,\mathbf{X}\right)P\left(Z|\mathbf{X}\right)}{P\left(Y=0|\mathbf{X}\right)} = \frac{(1-\epsilon_1)P(Z|\mathbf{X})}{(1-\epsilon_2)} \\ &= \frac{1}{(1-\epsilon_2)}P(Z|\mathbf{X}) - \frac{\epsilon_1}{(1-\epsilon_2)}P(Z|\mathbf{X}) \end{aligned}$$

with parameters $\epsilon_1, \epsilon_2 \in [0,1]$. For the rare disease case, ϵ_1 and ϵ_2 are small and this can be further approximated by

$$P(Z|\mathbf{X}) = \frac{\exp\left\{\mathrm{Z}\left(\mu_{zx} + \boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X}\right)\right\}}{1+\exp\left(\mu_{zx} + \boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X}\right)}.$$

For cases using the law of total probability (Meintrup and Schäffler, 2005) in the denominator notice that

$$\begin{aligned} P\left(Z|Y=1,\mathbf{X}\right) &= \frac{P\left(Y=1|Z,\mathbf{X}\right)P\left(Z|\mathbf{X}\right)}{P\left(Y=1|\mathbf{X}\right)} \\ &= \frac{P\left(Y=1|Z,\mathbf{X}\right)P\left(Z|\mathbf{X}\right)}{P\left(Y=1|\mathbf{X},Z=1\right)P\left(Z=1|\mathbf{X}\right) + P\left(Y=1|\mathbf{X},Z=0\right)P\left(Z=0|\mathbf{X}\right)} \\ &= \frac{\frac{\exp\left(\mu_y+\beta_{yz}Z+\boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X}\right)}{1+\exp\left(\mu_y+\beta_{yz}Z+\boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X}\right)}\frac{\exp\left\{Z\left(\mu_{zx}+\boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X}\right)\right\}}{1+\exp\left(\mu_{zx}+\boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X}\right)}}{\frac{\exp\left(\mu_y+\beta_{yz}+\boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X}\right)}{1+\exp\left(\mu_y+\beta_{yz}+\boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X}\right)}\frac{\exp\left(\mu_{zx}+\boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X}\right)}{1+\exp\left(\mu_{zx}+\boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X}\right)} + \frac{\exp\left(\mu_y+\boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X}\right)}{1+\exp\left(\mu_y+\boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X}\right)}\frac{1}{1+\exp\left(\mu_{zx}+\boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X}\right)}}. \end{aligned}$$

Assuming rare disease, the logistic function can be approximated by the exponential function since the following holds for a constant $a \in \mathbb{R}$ and $\exp(ax) < 1$ using the geometric series: $f(x) = \frac{\exp(ax)}{1+\exp(ax)} = \exp(ax)(\sum_{k=0}^{\infty}(-\exp(ax))^k) = \exp(ax)(1 - \exp(ax) + ...) \approx \exp(ax)$. All conditional probabilities for $Y = 1$ can thus be rewritten using this approximation, so that $P(Z|Y = 1, \mathbf{X})$ becomes

$$\begin{aligned}
&\frac{\exp\left(\mu_y + \beta_{yz}Z + \boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X}\right)\frac{\exp\{Z(\mu_{zx}+\boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X})\}}{1+\exp(\mu_{zx}+\boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X})}}{\exp\left(\mu_y + \beta_{yz} + \boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X}\right)\frac{\exp(\mu_{zx}+\boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X})}{1+\exp(\mu_{zx}+\boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X})} + \exp\left(\mu_y + \boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X}\right)\frac{1}{1+\exp(\mu_{zx}+\boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X})}} \\
&= \frac{\exp\left(\beta_{yz}Z\right)\frac{\exp\{Z(\mu_{zx}+\boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X})\}}{1+\exp(\mu_{zx}+\boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X})}}{\exp\left(\beta_{y,z}\right)\frac{\exp(\mu_{zx}+\boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X})}{1+\exp(\mu_{zx}+\boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X})} + \frac{1}{1+\exp(\mu_{zx}+\boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X})}} \\
&= \frac{\exp\left\{Z\left(\mu_{zx} + \boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X}+\beta_{yz}\right)\right\}}{\exp\left(\mu_{zx} + \boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X}+\beta_{yz}\right)+1} = \frac{\exp\left\{Z\left(\mu_{zx}^* + \boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X}\right)\right\}}{\exp\left(\mu_{zx}^* + \boldsymbol{\beta}_{z\mathbf{x}}^T\mathbf{X}\right)+1},
\end{aligned}$$

which is a logistic regression model with new intercept $\mu_{zx}^* = \mu_{zx} + \beta_{yz}$ and otherwise the same logistic coefficients as for the controls. For a normally distributed marker Z a similar argument holds. Assuming that $Z|\mathbf{X} \sim N(\mu_x, \sigma^2)$, then $Z|(\mathbf{X}, Y)$ also follows a normal distribution under the assumption of a rare disease. For controls the result follows immediately by the same argument as above. For cases it holds that

$$P(Z|Y = 1, \mathbf{X}) = \frac{P(Y = 1|Z, \mathbf{X})P(Z|\mathbf{X})}{P(Y = 1|\mathbf{X})}.$$

The denominator can be rewritten using the law of total probability for a continuous random variable with an additional second condition on $\mathbf{X}$ yielding

$$\begin{aligned}
&\frac{P(Y = 1|Z, \mathbf{X})P(Z|\mathbf{X})}{\int P(Y = 1|Z, \mathbf{X})dF(Z|\mathbf{X})} \\
&= \frac{\frac{\exp(\mu_y+\beta_{yz}Z+\boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X})}{1+\exp(\mu_y+\beta_{yz}Z+\boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X})}\exp\left\{-\frac{1}{2\sigma^2}(Z-\mu_x)^2\right\}}{\sqrt{2\pi\sigma^2}\int\frac{\exp(\mu_y+\beta_{yz}Z+\boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X})}{1+\exp(\mu_y+\beta_{yz}Z+\boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X})}dF(Z|\mathbf{X})},
\end{aligned}$$

see, for example, Duda et al. (2001). Using that the logistic function can be approximated by the exponential function under rare disease, as described earlier, this simplifies to

$$\begin{aligned}
&\frac{\exp(\mu_y + \beta_{yz}Z + \boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X})\exp\left\{-\frac{1}{2\sigma^2}(Z-\mu_x)^2\right\}}{\sqrt{2\pi\sigma^2}\int \exp(\mu_y + \beta_{yz}Z + \boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X})dF(Z|\mathbf{X})} \\
&= \frac{\exp\left\{-\frac{1}{2\sigma^2}(Z-\mu_x-\sigma^2\beta_{yz})^2\right\}\exp\left\{\frac{1}{2\sigma^2}(\mu_x+\sigma^2\beta_{yz})^2 - \frac{1}{2\sigma^2}\mu_x^2 + \mu_y + \boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X}\right\}}{\int \exp(\mu_y + \beta_{yz}Z + \boldsymbol{\beta}_{y\mathbf{x}}^T\mathbf{X})\exp\left\{-\frac{1}{2\sigma^2}(Z-\mu_x)^2\right\}dF(Z)} \\
&= \frac{\exp\left\{-\frac{1}{2\sigma^2}(Z-\mu_x-\sigma^2\beta_{yz})^2\right\}\exp\left\{\frac{1}{2\sigma^2}(\mu_x+\sigma^2\beta_{yz})^2 - \frac{1}{2\sigma^2}\mu_x^2\right\}}{\int \exp(\beta_{yz}Z)\exp\left\{-\frac{1}{2\sigma^2}(Z-\mu_x)^2\right\}dF(Z)}.
\end{aligned}$$

The moment-generating-function of the normal distribution, see, for example, Chapter 3 in Roussas (2015), is now applied to the denominator yielding

$$\begin{aligned}
&\frac{\exp\left\{-\frac{1}{2\sigma^2}(Z-\mu_x-\sigma^2\beta_{yz})^2\right\}\exp\left(\mu_x\beta_{yz} + \frac{\sigma^2}{2}\beta_{yz}^2\right)}{\sqrt{2\pi\sigma^2}\exp(\mu_x\beta_{yz} + \frac{\sigma^2}{2}\beta_{yz}^2)} \\
&= \frac{\exp\left\{-\frac{1}{2\sigma^2}(Z-\mu^*)^2\right\}}{\sqrt{2\pi\sigma^2}},
\end{aligned}$$

which is a normal distribution with a new mean $\mu^* = \mu_x + \sigma^2\beta_{yz}$. This result holds in general for the exponential family: if $P(Z|\mathbf{X})$ is in the exponential family then $P(Z|\mathbf{X}, Y)$ has the same general exponential form (Grill et al., 2016). Thus including Y into the regression model for Z in addition to $\mathbf{X}$ and fitting a single model to the combined data accommodates the different intercept terms in the models for $P(Z|\mathbf{X})$ in the exponential family. Alternatively, as performed in the LR-separate, one can separately estimate the numerator and the denominator of the LR by fitting two different models to the new data set, one to cases and one to controls. In this proof the assumption of a rare disease was used, however, as mentioned above, the simulations in Grill et al. (2016) revealed that the LR-joint and LR-separate still show very good calibration for non-rare disease cases, such as $P(Y = 1) = 0.3$, 0.4 or 0.6.

3.4.3 Synthetic simulations

For the synthetic simulations the existing model contained $p = 4$ independent binary covariates $\mathbf{X} = (X_1, X_2, X_3, X_4)^T$ given in model $R_{\mathbf{X}}$ in Equation (2.1), with $P(X_i = 1) = 0.2$, for $i = 1, \ldots 4$. The new marker Z was assumed to follow a mixture of two normals as part of a robustness study for the methods. Additional simulation settings that consider a binary and a normally distributed marker are described in Grill et al. (2016). The mixture of normals was set as

$Z \sim 0.9N(\alpha_{10} + \boldsymbol{\alpha}_1^T\mathbf{X}, 1) + 0.1N(\alpha_{20} + \boldsymbol{\alpha}_2^T\mathbf{X}, 1)$ with $\sigma^2 = 1$. The relationship between $\mathbf{X}$, Z and the outcome Y was specified by a logistic regression model,

$$P(Y = 1|Z, \mathbf{X}) = \frac{\exp(\beta_0 + \boldsymbol{\beta}_\mathbf{X}^T\mathbf{X} + \beta_Z Z)}{1 + \exp(\beta_0 + \boldsymbol{\beta}_\mathbf{X}^T\mathbf{X} + \beta_Z Z)}. \tag{3.15}$$

The intercept β_0 was chosen such that the disease prevalence $P(Y = 1) = 5\%$.

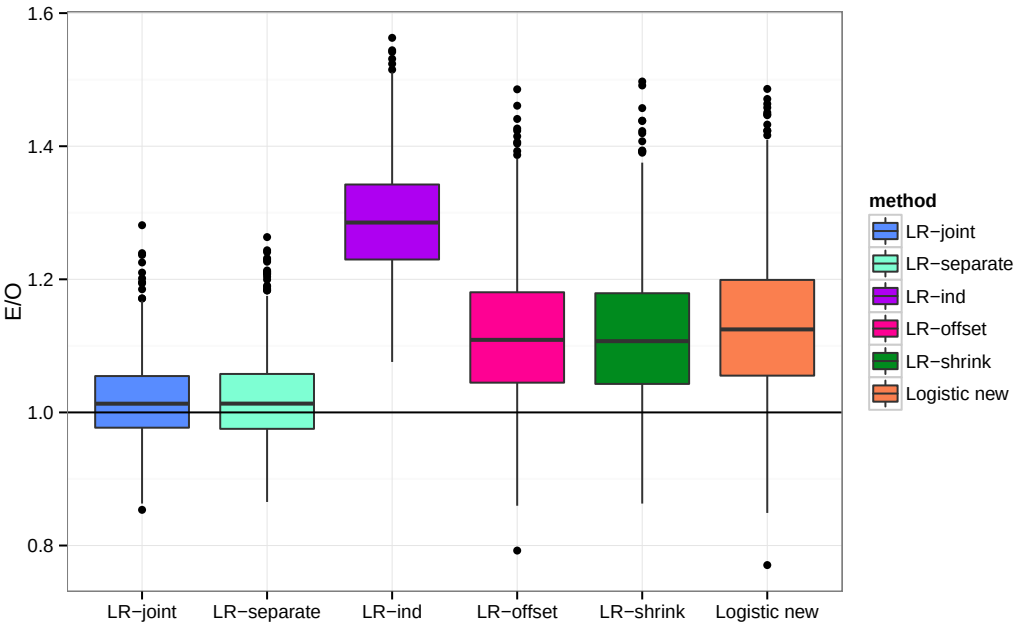
For each scenario $n = 1,000$ simulations were performed by first generating $\mathbf{X}$, then generating Z from the mixture equation and then using $\mathbf{X}$ and Z to obtain the outcome Y from Equation (3.15). One big data set was generated and then split into three non-overlapping sub data sets: data set A with $n_A = 1,000,000$ samples, which was used for estimating model $R_\mathbf{X}$ only based on $\mathbf{X}$, data set B with $n_B = 500$ samples used for updating with the new marker Z and data set C with $n_C = 100,000$ samples used for validation purposes. For the case-control adjustment of the intercept the disease prevalence $P(Y = 1)$ as well as the empirical distribution of $\mathbf{X}$, $\hat{F}(\mathbf{X})$, was estimated on data set A. The conditional distribution $\hat{F}(Z|\mathbf{X})$ was estimated from the controls in data set B and therefore the joint distribution was composed of $\hat{F}(Z, \mathbf{X}) = \hat{F}_{control,B}(Z|\mathbf{X})\hat{F}_A(\mathbf{X})$.

Model calibration was calculated as the number of expected cases by each method divided by the number of observed cases in data set C, E/O= $\sum \hat{p}_i / \sum Y_i$. Overall E/O ratios in data set C as well as ratios in risk groups were evaluated for specific X_1 or Z values, where X_1 is chosen exemplarily for one of the four covariates $\mathbf{X}$ in the existing model. Further, E/O ratios in risk deciles were assessed. As a second measure, the variability of the predictions was calculated by the standard deviation of the means of the predicted probabilities $\sum \hat{p}_i/n$ over the 1,000 simulation runs. The simulations were performed within the R environment (R Core Team, 2013) in combination with SAS 9.4 (SAS Institute Inc.).

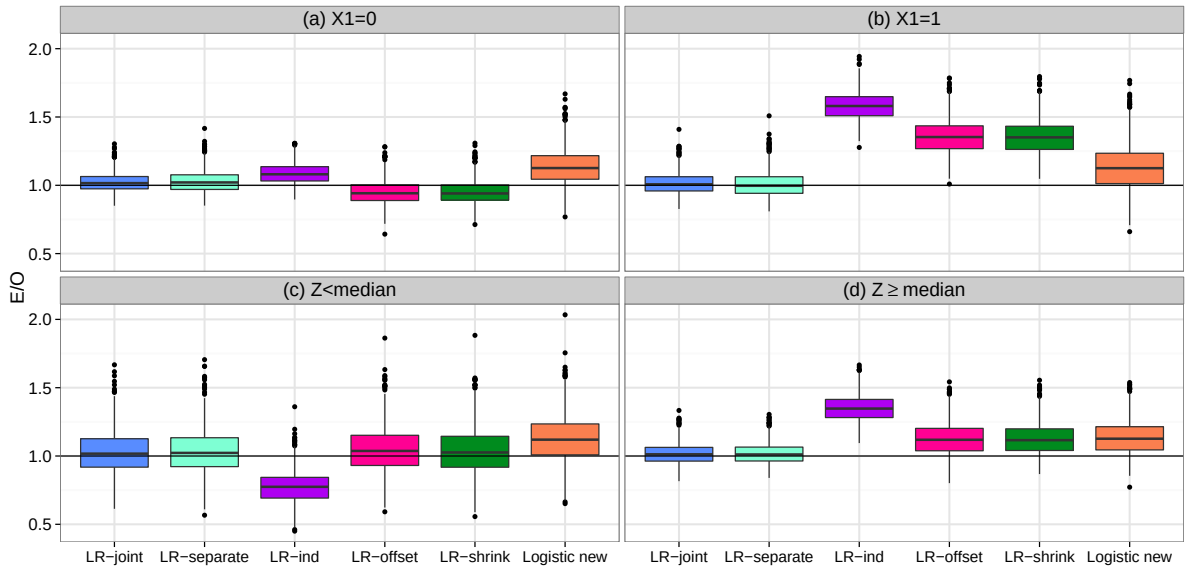
Results of synthetic simulations

The new study data set B was a case-control study, which comprised 250 cases and 250 controls and Z followed the mixture distribution $0.9N(\alpha_{10} + \boldsymbol{\alpha}_1^T\mathbf{X}, 1) + 0.1N(\alpha_{20} + \boldsymbol{\alpha}_2^T\mathbf{X}, 1)$, with $\alpha_{10} = 0, \boldsymbol{\alpha}_1 = (0.7, 0.7, -0.7, -0.7)^T, \alpha_{20} = 0.5, \boldsymbol{\alpha}_2 = (1, 1, -1, -1)^T$. Remaining parameters are given in the captions of Figures 8 and 9. Additional simulation scenarios, where, for example, data set B is a cohort were considered in Grill et al. (2016).

Figure 8 shows boxplots of E/O ratios based on 1,000 simulations. Corresponding numbers



(a) Overall E/O ratios



(b) E/O ratios in risk groups

Figure 8: E/O ratios for case-control setting where Z followed a mixture of normals with $Z \sim 0.9N(\alpha_{10}+\boldsymbol{\alpha}_1^T\mathbf{X},1)+0.1N(\alpha_{20}+\boldsymbol{\alpha}_2^T\mathbf{X},1), \alpha_{10}=0, \boldsymbol{\alpha}_1=(0.7,0.7,-0.7,-0.7)^T, \alpha_{20}=0.5, \boldsymbol{\alpha}_2=(1,1,-1,-1)^T, P(X)=0.2, P(Y)=0.05, \boldsymbol{\beta}_{\mathbf{X}}=(0.5,0.5,-0.5,-0.5)^T, \beta_Z=1$.

can be found in Supplementary Table S1. The two methods LR-separate and LR-joint were nearly unbiased overall (Figure 8a) and also in risk groups defined by Z and X_1, exemplary for one of the four covariates $\mathbf{X}$ in the existing model (Figure 8b). LR-ind showed the largest overall bias of about 29% followed by Logistic new with 14% and LR-offset and LR-shrink with 12% each. All four methods overestimated the true risk. Figure 8b shows that the bias in these

four methods was even more pronounced in some of the risk groups. In cells defined by $X_1 = 1$ (Figure 8b, panel (b)), LR-ind, Logistic new, LR-offset and LR-shrink showed overestimation of 58%, 13%, 36% and 36%, respectively, and of 35%, 14%, 13% and 13%, respectively, in the risk group $Z \geq median$ (Figure 8b, panel (d)). In groups defined by $Z < median$ (Figure 8b, panel (c)), LR-ind underestimated risk by about 23% and LR-offset and LR-shrink underestimated by about 5% in groups defined by $X_1 = 0$ (Figure 8b, panel (a)).

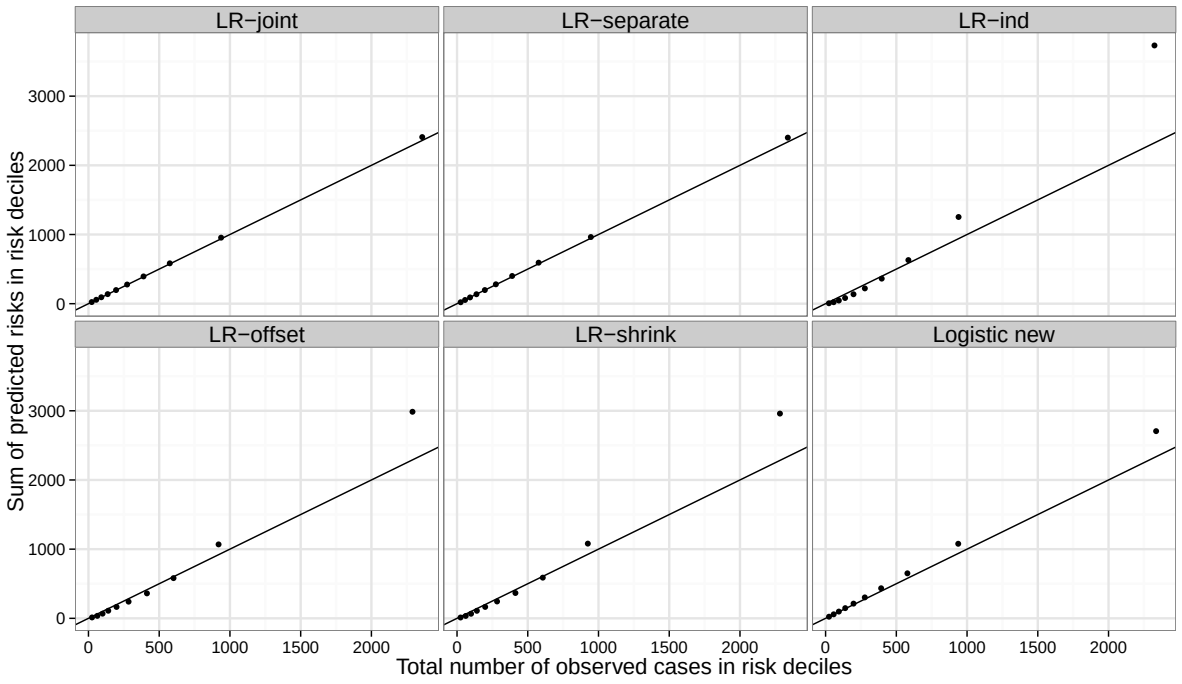


Figure 9: Calibration plot of risk deciles for the case-control setting where Z followed a mixture of normals with $Z \sim 0.9N(\alpha_{10} + \boldsymbol{\alpha}_1^T\mathbf{X}, 1) + 0.1N(\alpha_{20} + \boldsymbol{\alpha}_2^T\mathbf{X}, 1), \alpha_{10} = 0, \boldsymbol{\alpha}_1 = (0.7, 0.7, -0.7, -0.7)^T, \alpha_{20} = 0.5, \boldsymbol{\alpha}_2 = (1, 1, -1, -1)^T, P(X) = 0.2, P(Y) = 0.05, \boldsymbol{\beta}_{\mathbf{X}} = (0.5, 0.5, -0.5, -0.5)^T, \beta_Z = 1$.

Calibration of the six methods is shown in Figure 9. LR-separate and LR-joint showed very good calibration with all ten points of expected over observed cases lying almost exactly on the bisectrix. LR-offset, LR-shrink and Logistic new overestimated slightly in the second highest risk decile and overestimated conspicuously in the highest risk decile. LR-ind, however, already showed high overestimation in the second highest risk deciles, which was even more pronounced in the highest risk decile and therefore showed considerable lack of fit.

3.4.4 ViraHepC simulations

Simulations based on data from the ViraHepC study were performed additionally in order to obtain a realistic set of correlations between predictors (ViraHepC, 2002-2006). The ViraHepC study was conducted from 2002-2006 with the aim to examine differences between Caucasians and African Americans with respect to response to antiviral treatment for hepatitis C (HCV). Sustained virological response (SVR) was the binary outcome Y of interest to predict. Four variables were chosen as the original predictors $\mathbf{X}$, namely race (Caucasian vs. African American), sex (male vs. female), AST/ALT ratio (in quartiles) and Ishak fibrosis score (liver fibrosis stages from normal to cirrhosis in four ordinal categories). Two variables were used as new markers Z_1 and Z_2: interferon lambda 4 (IFNL4, ss469415590) genotype (two categories, $\Delta G/\Delta G$ or $\Delta G/TT$ vs. TT/TT) and pre-treatment HCV-RNA level ($\log_{10}$(IU/ml), continuous), respectively. IFNL4 is a novel gene associated with impaired clearance of hepatitis C virus (Prokunina-Olsson et al., 2013). An existing model based on the covariates $\mathbf{X}$ was updated first with either Z_1 or Z_2 and then jointly with both markers. Correlations between predictors were assessed using Spearman's rank correlation coefficient, ρ, and was highest between the genetic marker IFNL4 and race ($\rho = -0.40$). The correlations between the second marker HVC-RNA and the original covariates were rather weak with the highest value of $\rho = -0.10$ with the variable sex. Therefore, IFNL4 served here as an example of higher correlations between the old and the new predictors, whereas HCV-RNA constituted an example of low correlation.

For the simulations, covariate vectors $(\mathbf{X}, Z_1, Z_2)$ were sampled with replacement from the 350 ViraHepC patients. The outcome Y was not sampled from the data together with the covariates, but generated from a logistic regression model

$$P(Y = 1|\mathbf{X}, Z_1, Z_2) = \frac{\exp\{\beta_0 + \boldsymbol{\beta}_{\mathbf{X}}^T\mathbf{X} + \beta_{Z_1}Z_1 + \beta_{Z_2}Z_2\}}{1 + \exp\{\beta_0 + \boldsymbol{\beta}_{\mathbf{X}}^T\mathbf{X} + \beta_{Z_1}Z_1 + \beta_{Z_2}Z_2\}}, \tag{3.16}$$

where β_0 was chosen such that the outcome prevalence equals $P(Y = 1) = 0.1$. The remaining coefficients were chosen as the values that were obtained by fitting this model to the real data set. As described for the previous simulations the generated data set was split into three non-overlapping parts. A logistic regression was fit to the old covariates $\mathbf{X}$ to estimate the original model $R_{\mathbf{X}}$ in data set A. For updating with both markers jointly to obtain $R_{\mathbf{X},Z_1,Z_2}$, the LR

was calculated on data set B as follows:

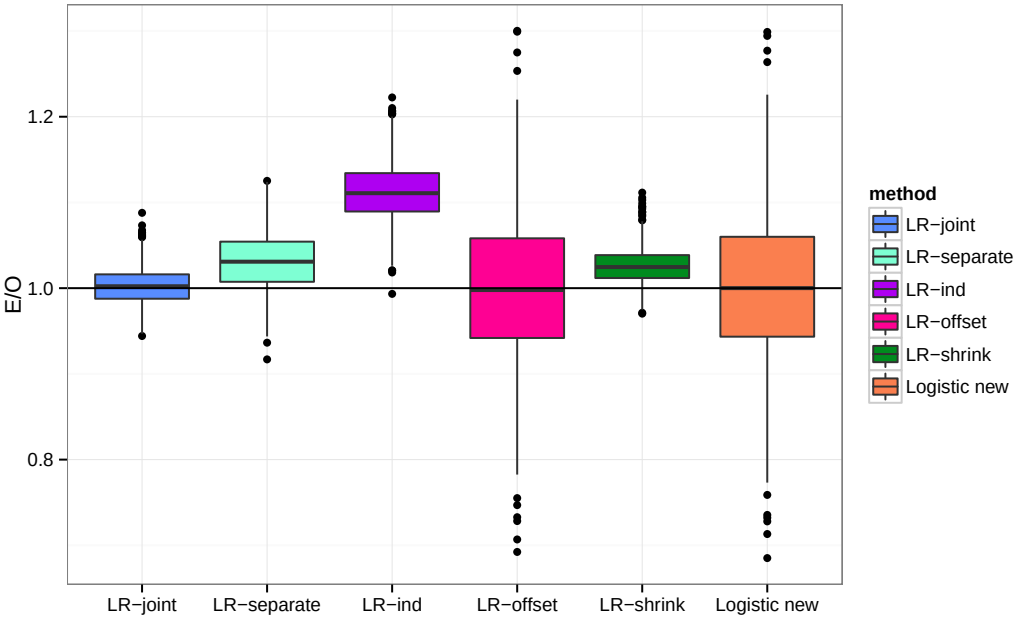
$$\log\{\mathrm{LR}_Y(Z_1, Z_2|\mathbf{X})\} = \log\left\{\frac{P(Z_2|Y=1, Z_1, \mathbf{X})P(Z_1|Y=1, \mathbf{X})}{P(Z_2|Y=0, Z_1, \mathbf{X})P(Z_1|Y=0, \mathbf{X})}\right\} = \log\left\{\frac{P(Z_2|Y=1, Z_1, \mathbf{X})}{P(Z_2|Y=0, Z_1, \mathbf{X})}\right\} + \log\left\{\frac{P(Z_1|Y=1, \mathbf{X})}{P(Z_1|Y=0, \mathbf{X})}\right\} = \log\{\mathrm{LR}_Y(Z_2|Z_1, \mathbf{X})\} + \log\{\mathrm{LR}_Y(Z_1|\mathbf{X})\}. \quad (3.17)$$

For estimating $\log\{\mathrm{LR}_Y(Z_2|Z_1, \mathbf{X})\}$, the marker Z_1 was included as a predictor among the other variables $\mathbf{X}$. The models built on data sets A and B, with $n_A = 1,000,000$ and $n_B = 1,000$, were validated on data set C with $n_C = 100,000$. Additional scenarios were covered in Grill et al. (2016).

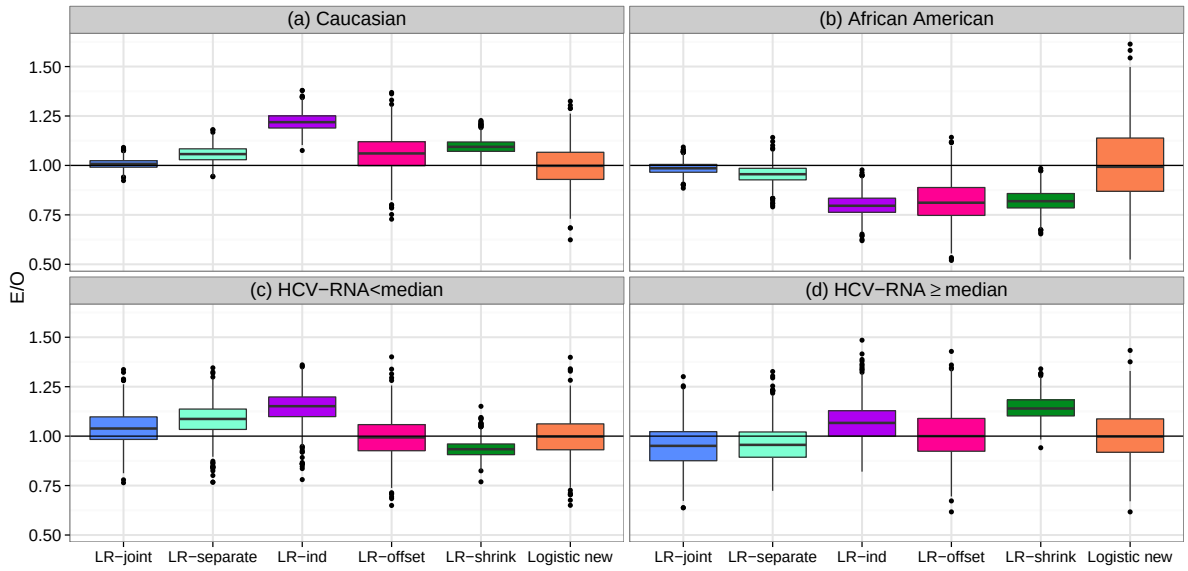
Results of data-based simulations

In the following results are presented when the original model was updated with information from a cohort on both markers IFNL4 (Z_1) and HCV-RNA (Z_2). Figure 10 shows E/O ratios over 1,000 simulations. Corresponding numbers can be found in Supplementary Table S2.

LR-joint, LR-offset and Logistic new showed overall unbiased results (Figure 10a), however, the last two methods showed the largest variability in predictions. LR-ind showed overall the highest overestimation with 11% followed by LR-separate with 3% and LR-shrink with 3%. LR-joint and Logistic new stayed more or less unbiased in all risk groups, whereas LR-offset showed overestimation for Caucasians of 6% and underestimation in African Americans of 18%. LR-ind showed even higher overestimation for Caucasians of 22% (Figure 10b, panel(a)) and for IFNL4 genotype TT/TT of 28% (Supplementary Table S2). Generally speaking, those methods that showed biased results, overestimated for Caucasians and underestimated for African Americans (Figure 10b, panel(a) and (b)).



(a) Overall E/O ratios



(b) E/O ratios in risk groups

Figure 10: E/O ratios for the cohort setting, where models were updated with both markers, IFNL4 and HCV-RNA, with $P(Y) = 0.1$.

When assessing calibration in Figure 11, LR-joint, LR-offset, LR-shrink and Logistic new showed hardly any lack of fit. LR-separate, however, showed slight underestimation in the lower risk deciles and clear overestimation in the highest risk decile, whereas LR-ind displayed a similar but even more pronounced pattern with the overestimation starting in the second highest risk decile.

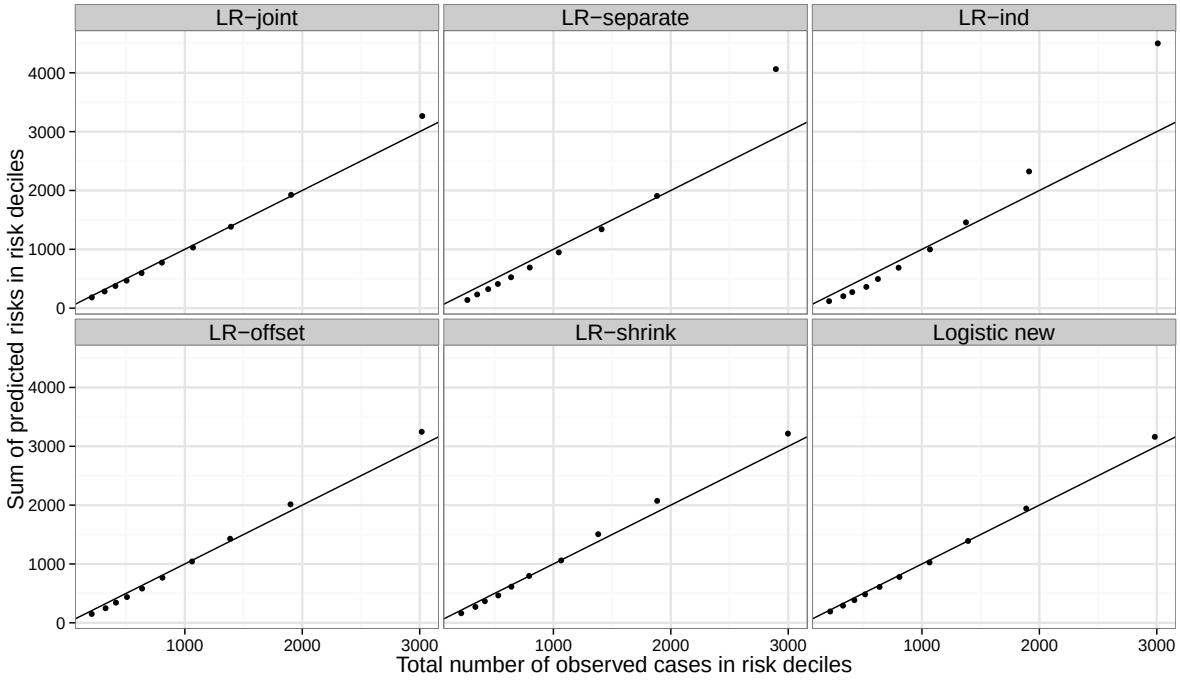


Figure 11: Calibration plot of risk deciles for the cohort setting where models were updated with both markers, IFNL4 and HCV-RNA, with $P(Y) = 0.1$.

3.4.5 Discussion of simulation methods and results

In this section the methods and results of the simulations in Sections 3.4.3 and 3.4.4 are discussed with respect to existing literature. In summary, the simulations revealed that Logistic new and LR-offset showed the largest variability of predictions in most of the settings. The largest bias was witnessed for the LR-ind method for settings with high dependence between the old and new predictors, followed by LR-shrink, LR-offset and Logistic new. However, further simulations of non-rare disease settings with $P(Y = 1) = 0.3, 0.4$ and 0.6 revealed a less pronounced bias for LR-ind than for the rare disease case. In the PCPTRC population the disease prevalence was $P(Y = 1) = 18\%$ and therefore falls in the non-rare disease setting (Ankerst et al., 2014). The intercept adjustment for the case-control setting appeared to be very sensitive to the disease prevalence even for the independence case. The bias appeared due to the fact that the distribution of the risk factors was estimated on the controls of the updating data set and this distribution was used as an approximation for the population. However, the simulations showed that the bias in the intercept adjustment decreased as expected as the disease prevalence decreased. LR-joint and LR-separate were largely unbiased in most settings and were not sensitive to violations of the normality of the new marker as the robustness study revealed. LR-separate, however, showed in some scenarios, for example, in Section 3.4.4, bad calibration

in the highest risk decile and also a small overall bias revealed by the E/O ratios. In contrast, the LR-joint remained unbiased in all settings. Another advantage of the LR-joint is that it requires fitting less parameters than the LR-separate. For a continuous marker $p+3$ parameters were fit for the LR-joint and $2(p+1)+2$ parameters for the LR-separate, where p is the number of predictors in the original model. Overall, the LR-joint showed the best performance and is recommended for updating under scenarios covered in this simulation study when information on the original predictors $\mathbf{X}$ is available. In addition, it is important to note that both methods are very easy to implement in standard software.

Chan et al. compared several methods to update pretest risk with information on a new test, including the three methods: LR-ind, LR-offset and LR-shrink (Chan et al., 2008). They recommended the LR-offset methods for updating. However, the simulations presented here did not confirm this finding, since LR-offset showed the largest variance and also considerable bias, especially for some settings for updating with case-control data. Furthermore, Chan et al. only built the models in one real data set comprising 309 patients with chronic obstructive airway disease, and assessed performance in a second, independent validation study comprising 161 individuals. They assessed performance in quintiles of risk rather than deciles as presented here (Chan et al., 2008).

One of the most important findings of these simulations was that the assumption of independence is crucial and has to be well assessed, since the LR-ind method could have large bias for a strong dependence between the old and new predictors. However, the simulations also revealed that this bias is a lot less pronounced for non-rare disease settings. If possible, meaning that required data are available, dependence structures should be taken into account. Based on the simulations performed, the LR-joint method, which assumes equal variances between cases and controls and corresponds to linear discriminant analysis, should be recommended for updating among the methods that were investigated here. However, this was only investigated for a binary outcome and a single marker Z, which is binary, normally distributed or following a mixture of normals or a combination of a continuous and a binary marker as part of the data-based simulations. Extended simulations would need to be examined for generalizing this to a multinomial outcome Y and additional combinations of new markers.

The LR has not only been studied when updating a risk prediction tool but also with respect to evaluating the diagnostic performance of a new marker as well as to compare the predictive information of two markers (Gu and Pepe, 2009, 2011). Gu and Pepe (2009) evaluated the

diagnostic performance of a single continuous marker and investigated the diagnostic LR as a covariate specific function estimated by logistic regression. However, the authors did not study its properties or make comparisons with other methods. Gu and Pepe (2011) extended this idea and compared six methods for estimating the diagnostic LR. They extended the logistic regression approach from Janssens et al. (2005) to continuous markers, and for simplicity, they only considered scenarios without covariates. Further, the diagnostic LR functions were used to compare the predictive information of two markers by simulation studies (Gu and Pepe, 2011).

Huang et al. combined logistic regression with receiver operating characteristics (ROC) curves using the covariate specific diagnostic LR to evaluate classification performance as well as model the probability of a disease (Huang et al., 2013). The authors considered the scenario of combining biomarker data sets from different sources, however, only when ROC curves are similar across the different sources. Covariate adjustment was covered and three estimators were proposed and compared in simulations as well as in a real data example: a pseudo-likelihood estimator, a constrained maximum likelihood estimator and an estimated empirical likelihood estimator (Huang et al., 2013).

The focus of this thesis was to incorporate a new marker in an existing risk prediction tool. Several different methods were examined with respect to their performance under various settings. It was assumed that the predictive performance of the new marker had been investigated in previous studies justifying its use. Consequently, the added value of the new marker had been shown and so this aspect is not covered here.

Liu et al. incorporated longitudinal sequences of biomarkers in a risk prediction model using a likelihood ratio statistic similar to Ankerst et al. (2008) and Gu and Pepe (2009), (Liu and Albert, 2014). A pattern mixture model framework to predict a dichotomous disease outcome from longitudinal biomarker data was proposed. The authors used Bayes' theorem to estimate individual risk scores including a LR statistic as the combination rule. Covariate-specific combinations of biomarkers were covered, however, the purpose of Liu et al. was not to update an existing risk prediciton tool with new information, but rather to combine multiple longitudinal biomarkers for improving predictive accuracy (Liu and Albert, 2014). Longitudinal marker levels had been previously analyzed similarly in LRs by Skates et al. using Markov chain Monte Carlo methods to obtain posterior probabilities within a Bayesian setting (Skates et al., 2001).

One aspect that could be investigated further is the flexibility of the methods that were

considered. Within the LR for LR-joint or LR-separate, for example, the covariates were included in a linear fashion in the logistic or linear regressions for the simulations. In practice, however, new continuous markers might not be normally distributed and the distribution of the new marker needs to be carefully evaluated. The new marker has to be modeled in each case separately by looking at the distribution and applying transformations when necessary. This is performed in Ankerst et al. (2008, 2012b, 2014) as well as model selection. In addition, more flexible ways could be considered, as for example splines, which were investigated in Nieboer et al. (2015). Nieboer et al. assessed the performance of different functions included in a regression model with respect to model performance as well as internal and external validation. They found that nonlinear models, which are more flexible, showed better performance in the scenario of internal validation. Yet, when comparing models with regard to external validation, the less flexible functions led to better performance (Nieboer et al., 2015). However, their investigations did not consider an updating setting, which was the focus of this thesis. Furthermore, more flexible approaches need to be investigated with great care. The simulations showed that letting the variances differ between cases and controls in the LR-separate, which makes this approach a bit more flexible than the LR-joint, already lead to overestimation in the highest risk decile in some settings, which did not appear for the LR-joint method.

3.4.6 Future directions

In the simulations presented in the previous sections, it was assumed that the new data as well as the validation data set come from the same underlying population as the original data set. This is a strong assumption and might not always hold in real world scenarios. Therefore, future investigations are needed referring to which method to use for a differing updating data set. Practically speaking, the new data set could come from a different country (as was the case in Grill et al. (2015a), Sweden and U.S.) or originate from a different ethnic group. The age structure in the SFCD and the PCPTRC population could be compared and if necessary methods using, for example, shrinkage or distance metrics could be developed and applied. Two approaches by Wiens et al. and Debray et al. investigated similar problems (Wiens et al., 2014, Debray et al., 2015). Wiens et al. investigated how to combine data from different hospitals with partly overlapping risk factors for prediction (Wiens et al., 2014). Debray et al. discussed the relatedness of a development and validation data set with respect to case-mix differences. They incorporated these differences into the interpretation of external validation results (Debray

et al., 2015). Approaches in these directions could be investigated to study scenarios where data sets originating from different populations are fused.

As the simulations revealed that conditioning on the predictors in the original model lead to less biased results, the detailed family history update could be extended by using multinomial models with 23 categories conditioning on age and race for estimating the LR. However, this has to be evaluated with care, since the large number of categories could lead to severe sparsity which could cause problems in estimating the model parameters. For the SNP update there are unfortunately no other covariates available. In addition, the simulations revealed that constraining the variances to be equal between cases and controls when estimating the LR lead to better model performance. Consequently the percent free PSA update in Ankerst et al. (2014) could be revisited. The LRs for the percent free update were obtained by fitting three normal linear regressions of log-base-2-transformed percent free PSA on the PCPT risk factors separately in the three outcome groups of low-grade, high-grade and no cancer. The simulation results suggest that fitting a single model and by that constraining the variances between the groups to be equal could lead to even better performance. The LR-joint method could also be extended by adding interaction terms between the disease outcome Y and the predictors $\mathbf{X}$ since this could possibly even better accommodate the dependence structure between the old and new predictors. Simulation studies would need to be performed to investigate these two extensions further.

The methods discussed here could be applied for updating prediction models for other cancer types, such as genetic variants or mammographic density in breast cancer. This is ongoing work in collaboration with the University of Texas Health Science Center at San Antonio.

3.5 Conclusions

The main conclusions of this thesis are summarized in the following.

- An update of detailed family history measures to the PCPTRC was provided. Measures of detailed family history have been shown to be independent predictors of prostate cancer also in combination with other common markers. Therefore, risk assessment can be improved by including this risk factor into medical decision making as part of the widely used PCPTRC. An advantage of detailed family history is that it is easily available and not cost intensive to collect compared to some molecular biomarkers, such as PSA or

genetic factors.

- An easy-to use method for incorporating any group of SNPs into an existing risk prediction model was developed and is now available as an update to the PCPTRC. The meta-analysis of several GWASs yielded independent predictive effects of most of the 30 SNPs considered with respect to prostate cancer risk, one of the most common cancer types in men. Therefore, the immense investment in the discovery and validation of SNPs by GWASs can be translated to clinical practice and further validated by incorporation into existing risk tools.

- Validation of both updates in Grill et al. (2015a,b) is needed and hopefully achieved soon with online accessibility at www.myprostatecancerrisk.com.

- Simulation studies of six updating methods, including the one used in Grill et al. (2015a,b), revealed that it is desirable to account for dependence structures between the old and new predictors in the estimation of the LR if this is possible. A joint estimation of the probability densities in the LR in cases and controls is recommended. LR estimation under independence showed a large bias for strong dependence between old and new predictors for rare disease scenarios, which was, however, a lot less pronounced for the non-rare disease case.

4 Summary

Novel genetic markers as well as biomarkers for cancer risk prediction are discovered continuously, which leads to an increasing need for incorporating these new markers into existing risk prediction tools. Risk calculators are often built on very large cohorts, whereas new biomarkers are usually measured on different populations, such as small studies or population registries. The predictors in these populations might overlap only party or not at all. This thesis investigated the fusion of different data sources using two applications that update the Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) with new information. The PCPTRC was built on a large prevention trial and calculates the risk of detecting prostate cancer in case a biopsy were to be performed using six established risk factors.

The first application studied detailed family history of prostate cancer or breast cancer as predictors of prostate cancer from the Swedish Family Cancer Database, one of the world's largest population registries. The study investigated different family history patterns distinguishing between age at diagnosis (younger or older than 60 years of age), first and second degree relatives as well as prostate and breast cancer history. The PCPTRC was then updated with detailed family history variables. In the second application, a meta-analysis was performed of 30 single-nucleotide polymorphisms (SNPs) that were multiply validated by genome wide association studies. Both, a meta-analysis technique assuming independence between SNPs and one incorporating linkage disequilibrium were developed.

In the case of both markers, detailed family history and SNPs, a Bayesian technique called the likelihood ratio (LR) was used to update the existing risk calculator with information on the two new markers. This technique allows the incorporation of any group of SNPs or detailed family history patterns. Both updates were made freely available under www.myprostatecancerrisk.com for physician and patient use as well as for research and validation purposes.

In addition to the two applications on prostate cancer risk, the LR-based method was investigated and compared in detail to other updating techniques by extensive synthetic and data-based simulations. Six different methods were studied: three different LR-based methods, one assuming independence and two incorporating dependence structures between the new and old predictors; one approach using an offset term for the original model; one including an additional shrinkage factor and one fitting a logistic regression to the new data only. Updating

settings were studied when the new data arose from a cohort as well as a case-control study. Calibration of the new updated models and the variability of the predictions were assessed under various settings, including some that violate the underlying assumption of independence of new to old markers or of the distribution of the new marker. The synthetic simulations examined three different scenarios for the new marker: a binary marker, a normally distributed marker and a marker following a mixture of normal distributions. For the data-based simulations, two new markers for updating were considered: a binary (genotype) and a continuous (RNA-levels) variable. Moreover, a combination of both markers was investigated as well. Overall, the LR-based method that constrained the variances between cases and controls to be equal and incorporated a dependence structure between the new and old predictors showed the best performance.

5 Zusammenfassung

Die laufende Entdeckung von neuen genetischen Markern sowie Biomarkern im Allgemeinen zur Krebsrisikovorhersage macht es notwendig die Information dieser neuen Marker in bestehende Risikovorhersagemodelle einzubinden. Risikorechner entstehen oft auf der Datengrundlage von sehr großen Kohorten, wohingegen neue Biomarker in der Regel in anderen Populationen, wie zum Beispiel kleine Studien oder Populationsregister, gemessen werden. Die Variablen der verschiedenen Studien sind dabei oft nicht deckungsgleich. Ziel dieser Arbeit ist es, die Fusion von verschiedenen Datenquellen anhand von zwei Anwendungen zu untersuchen, welche den Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) mit neuen Informationen updaten. Der PCPTRC basiert auf einer großen Präventionsstudie und berechnet das Prostatakrebsrisiko im Falle einer Biopsie anhand von sechs etablierten Risikofaktoren.

Die erste Fallstudie beschäftigte sich mit detaillierter Familienanamnese von Prostata- und Brustkrebs als Prädiktoren für Prostatekrebs. Die Daten stammen aus der Swedish Family Cancer Database, eines der größten Populationsregister weltweit. Es wurden verschiedene Muster der Familienanamnese im Bezug auf das Alter bei Diagnose (jünger oder älter als 60 Jahre), Anzahl an Verwandten ersten und zweiten Grades sowie Prostata- und Brustkrebs untersucht. Der PCPTRC wurde durch Variablen zur detaillierten Familienanamese erweitert. In der zweiten Fallstudie wurde eine Metaanalyse von 30 Einzelnukleotid-Polymorphismen (SNPs) durchgeführt, welche von mehreren genomweiten Assoziationsstudien validiert wurden. Zwei Metanalyse-Techniken wurden entwickelt, die erste nimmt Unabhängigkeit zwischen den einzelnen SNPs an und die zweite berücksichtigt Linkage Disequilibrium zwischen den SNPs.

Für die beiden Marker, detailierte Familienanamnese und SNPs, wurde eine Bayessche Methode verwendet, welche mithilfe des Likelihood Ratios (LR) ein bestehendes Risikomodell um die Information über die neuen Marker erweitert. Diese Technik ermöglicht die Erweiterung des Risikomodells um jede beliebige Gruppe von SNPs oder Familienanamnese Variablen. Beide Updates sind unter www.myprostatecancerrisk.com frei zugänglich für behandelnde Ärzte, Patienten oder zu Forschungs- und Validierungszwecken.

Zusätzlich zu den beiden Fallstudien für Prostatakrebs, wurde die LR-basierte Methode durch umfangreiche synthetische und datenbasierte Simulationen im Detail untersucht und mit weiteren Methoden verglichen. Sechs verschiedene Methoden wurden untersucht: drei verschiedene LR-basierte Methoden, davon eine, die Unabhängigkeit annimmt und zwei, die

Abhängigkeiten zwischen den neuen und alten Prädiktoren berücksichtigen; ein Ansatz, der einen Offset-Term für das ursprüngliche Modell verwendet; eine Methode, die einen zusätzlichen Shrinkage-Faktor integriert und ein Modell, dass eine logistische Regression nur auf die neuen Daten fittet. Es wurden Updating-Szenarien analysiert, bei denen die neuen Daten sowohl von einer Kohorte als auch von einer Fall-Kontroll-Studie stammen können. Die Kalibrierung der erweiterten Modelle sowie die Variabilität in den Vorhersagen wurden unter verschiedenen Voraussetzungen und Konstellationen eruiert, welche auch Verletzungen der Modellannahmen beinhalten, wie zum Beispiel die Unabhängigkeit der neuen und alten Variablen oder die Verteilung des neuen Markers. Im Rahmen der synthetischen Simulationen wurden drei verschiedene Markertypen untersucht: ein binärer Marker, ein normalverteilter Marker und ein Marker, welcher mit einer Mischung von zwei Normalverteilungen generiert wurde. Bei den datenbasierten Simulationen wurden zwei neue Marker berücksichtigt: eine binäre Variable (Genotyp) und eine kontinuierliche Variable (RNA-Level). Darüber hinaus wurde auch die Kombination beider Marker untersucht. Insgesamt hat die LR-basierte Methode, welche die Varianzen in Fällen und Kontrollen gleichsetzt und Abhängigkeiten zwischen neuen und alten Variablen berücksichtigt, am besten abgeschnitten.

6 References

1000 Genome Project. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65, 2012.

ACS. *American Cancer Society: Cancer Facts and Figures 2015.* Atlanta, 2015. URL `http://www.cancer.org/acs/groups/content/@editorial/documents/document/acspc-044552.pdf`. Accessed Jan 23, 2016.

Adams J. U. Genetics: Big hopes for big data. *Nature*, 527(7578):S108–S109, 2015.

Agalliu I., Wang Z., Wang T., Dunn A., Parikh H., Myers T., Burk R., and Amundadottir L. Characterization of snps associated with prostate cancer in men of ashkenazic descent from the set of gwas identified snps: Impact of cancer family history and cumulative snp risk prediction. *PLoS ONE*, 8(4), 2013.

Akamatsu S., Takahashi A., Takata R., Kubo M., Inoue T., Morizono T., Tsunoda T., Kamatani N., Haiman C. A., Wan P., Chen G. K., Le Marchand L., Kolonel L. N., Henderson B. E., Fujioka T., Habuchi T., Nakamura Y., Ogawa O., and Nakagawa H. Reproducibility, performance, and clinical utility of a genetic risk prediction model for prostate cancer in japanese. *PLoS ONE*, 7(10):e46454, 2012.

Al Olama A. A., Kote-Jarai Z., Giles G. G., Guy M., Morrison J., Severi G., Leongamornlert D. A., Tymrakiewicz M., Jhavar S., Saunders E., Hopper J. L., Southey M. C., Muir K. R., English D. R., Dearnaley D. P., Ardern-Jones A. T., Hall A. L., O'Brien L. T., Wilkinson R. A., Sawyer E., Lophatananon A., Horwich A., Huddart R. A., Khoo V. S., Parker C. C., Woodhouse C. J., Thompson A., Christmas T., Ogden C., Cooper C., Donovan J. L., Hamdy F. C., Neal D. E., Eeles R. A., and Easton D. F. Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nature Genetics*, 41(10):1058–1060, 2009.

Al Olama A. A., Dadaev T., Hazelett D. J., Li Q., Leongamornlert D., Saunders E. J., Stephens S., Cieza-Borrella C., Whitmore I., Benlloch Garcia S., Giles G. G., Southey M. C., Fitzgerald L., Gronberg H., Wiklund F., Aly M., Henderson B. E., Schumacher F., Haiman C. A., Schleutker J., Wahlfors T., Tammela T. L., Nordestgaard B. G., Key T. J., Travis R. C., Neal D. E., Donovan J. L., Hamdy F. C., Pharoah P., Pashayan N., Khaw K.-T., Stanford J. L., Thibodeau S. N., Mcdonnell S. K., Schaid D. J., Maier C., Vogel W., Luedeke M., Herkommer

K., Kibel A. S., Cybulski C., Wokoorczyk D., Kluzniak W., Cannon-Albright L., Brenner H., Butterbach K., Arndt V., Park J. Y., Sellers T., Lin H.-Y., Slavov C., Kaneva R., Mitev V., Batra J., Clements J. A., Spurdle A., Teixeira M. R., Paulo P., Maia S., Pandha H., Michael A., Kierzek A., Govindasami K., Guy M., Lophatonanon A., Muir K., Vinuela A., Brown A. A., The PRACTICAL Consortium, COGS-CRUK GWAS-ELLIPSE (Part of GAME-ON) Initiative, The Australian Prostate Cancer BioResource, The UK Genetic Prostate Cancer Study Collaborators, The UK ProtecT Study Collaborators, Freedman M., Conti D. V., Easton D., Coetzee G. A., Eeles R. A., and Kote-Jarai Z. Multiple novel prostate cancer susceptibility signals identified by fine-mapping of known risk loci among europeans. *Human Molecular Genetics*, 24(19):5589–602, 2015.

Albert A. On the use and computation of likelihood ratios in clinical chemistry. *Clinical Chemistry*, 28(5):1113–1119, 1982.

Albright F., Stephenson R., Agarwal N., Teerlink C., Lowrance W., Farnham J., and Albright L. Prostate cancer risk prediction based on complete prostate cancer family history. *Prostate*, 75(4):390–398, 2015.

Amundadottir L. T., Thorvaldsson S., Gudbjartsson D. F., Sulem P., Kristjansson K., Arnason S., Gulcher J. R., Bjornsson J., Kong A., Thorsteinsdottir U., and Stefansson K. Cancer as a complex phenotype: pattern of cancer distribution within and beyond the nuclear family. *PLoS Medicine*, 1(3):e65, 2004.

Amundadottir L. T., Sulem P., Gudmundsson J., Helgason A., Baker A., Agnarsson B. A., Sigurdsson A., Benediktsdottir K. R., Cazier J.-B., Sainz J., Jakobsdottir M., Kostic J., Magnusdottir D. N., Ghosh S., Agnarsson K., Birgisdottir B., Le Roux L., Olafsdottir A., Blondal T., Andresdottir M., Gretarsdottir O. S., Bergthorsson J. T., Gudbjartsson D., Gylfason A., Thorleifsson G., Manolescu A., Kristjansson K., Geirsson G., Isaksson H., Douglas J., Johansson J.-E., Balter K., Wiklund F., Montie J. E., Yu X., Suarez B. K., Ober C., Cooney K. A., Gronberg H., Catalona W. J., Einarsson G. V., Barkardottir R. B., Gulcher J. R., Kong A., Thorsteinsdottir U., and Stefansson K. A common variant associated with prostate cancer in european and african populations. *Nature Genetics*, 38(6):652–658, 2006.

Anderson T. *An Introduction to Multivariate Statistical Analysis, 2nd Edition.* Wiley, New York, 1984.

Ankerst D. P., Groskopf J., Day J. R., Blase A., Rittenhouse H., Pollock B. H., Tangen C., Parekh D., Leach R. J., and Thompson I. Predicting prostate cancer risk through incorporation of prostate cancer gene 3. *The Journal of Urology*, 180(4):1303 – 1308, 2008.

Ankerst D. P., Böck A., Freedland S., Thompson I., Cronin A. M., Roobol M., Hugosson J., Stephen Jones J., Kattan M., Klein E., Hamdy F., Neal D., Donovan J., Parekh D., Klocker H., Horninger W., Benchikh A., Salama G., Villers A., Moreira D., Schröder F. H., Lilja H., and Vickers A. Evaluating the pcpt risk calculator in ten international biopsy cohorts: results from the prostate biopsy collaborative group. *World Journal of Urology*, 30(2):181–187, 2012a.

Ankerst D. P., Koniarski T., Liang Y., Leach R. J., Feng Z., Sanda M. G., Partin A. W., Chan D. W., Kagan J., Sokoll L., Wei J. T., and Thompson I. M. Updating risk prediction tools: A case study in prostate cancer. *Biometrical Journal*, 54(1):127–142, 2012b.

Ankerst D. P., Hoefler J., Bock S., Goodman P. J., Vickers A., Hernandez J., Sokoll L. J., Sanda M. G., Wei J. T., Leach R. J., and Thompson I. M. Prostate cancer prevention trial risk calculator 2.0 for the prediction of low- vs high-grade prostate cancer. *Urology*, 83(6):1362–7, 2014.

Ankerst D. P., Schepsmeier U., and Hoefler J. *PCPT Risk Calculator*. UT Health Science Center San Antonio, 2015. URL `www.myprostatecancerrisk.com`. Accessed Dec 29, 2015.

Balding D. J., Bishop M., and Cannings C. *Handbook of Statistical Genetics*, volume 2. Wiley, 3 edition, 2006.

Barlow W., White E., Ballard-Barbash R., Vacek P., Titus-Ernstoff L., Carney P., Tice J., Buist D., Geller B., Rosenberg R., Yankaskas B., and Kerlikowske K. Prospective breast cancer risk prediction model for women undergoing screening mammography. *Journal of the National Cancer Institute*, 98(17):1204–1214, 2006.

BCSC Risk Calculator. *Breast Cancer Surveillance Consortium Risk Calculator*. 2015. URL `http://tools.bcsc-scc.org/BC5yearRisk/`. Accessed Feb 10, 2016.

Bishop Y. M. M., Fienberg S. E., and Holland P. W. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: MIT Press, 1975.

Bonnen P. E., Wang P. J., Kimmel M., Chakraborty R., and Nelson D. L. Haplotype and linkage disequilibrium architecture for human cancer-associated genes. *Genome Research*, 12 (12):1846–1853, 2002.

Brasso K., Friis S., Kjaer S. K., Jorgensen T., and Iversen P. Prostate cancer in denmark: a 50-year population-based study. *Urology*, 51(4):590–4, 1998.

Brath J. M., Grill S., Ankerst D. P., Thompson J. I. M., Gschwend J. E., and Herkommer K. No detrimental effect of a positive family history on long-term outcomes following radical prostatectomy. *The Journal of Urology*, 2015.

Bratt O., Kristoffersson U., Lundgren R., and Olsson H. Familial and hereditary prostate cancer in southern sweden. a population-based case-control study. *European Journal of Cancer*, 35 (2):272–7, 1999.

Campbell N. A. *Biology (Eighth Edition)*. Pearson Benjamin Cummings, San Francisco, 2008. ISBN 978-0-321-53616-7.

Cavadas V., Osorio L., Sabell F., Teves F., Branco F., and Silva-Ramos M. Prostate cancer prevention trial and european randomized study of screening for prostate cancer risk calculators: A performance comparison in a contemporary screened cohort. *European Urology*, 58 (4):551 – 558, 2010.

CDC. *Deaths: Final Data from 2013*. Centers for Disease Control and Prevention, U.S. Department of Health and Human Sevices, 2013. URL `http://www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64_02.pdf`. Table 1, Accessed Dec 14, 2015.

Chan S. F., Deeks J. J., Macaskill P., and Irwig L. Three methods to construct predictive models using logistic regression and likelihood ratios to facilitate adjustment for pretest probability give similar results. *Journal of Clinical Epidemiology*, 61(1):52 – 63, 2008.

Chang B.-L., Hughes L., Chen D. Y., Gross L., Ruth K., and Giri V. N. Validation of association of genetic variants at 10q with prostate-specific antigen (psa) levels in men at high risk for prostate cancer. *BJU International*, 113(5b):E150–E156, 2014.

Chang W., Cheng J., Allaire J., Xie Y., and McPherson J. *shiny: Web Application Framework for R*, 2015. URL `http://CRAN.R-project.org/package=shiny`.

Chatterjee N., Park J.-H., Caporaso N., and Gail M. H. Predicting the future of genetic risk prediction. *Cancer Epidemiology, Biomarkers & Prevention*, 20(1):3–8, 2011.

Collins S. R., Rasmussen P. W., and Doty M. M. Gaining ground: Americans health insurance coverage and access to care after the affordable care acts first open enrollment period. *Commonwealth Fund*, 16, 2014. New York, pub. no. 1760.

De Iorio M., Newcombe P. J., Tachmazidou I., Verzilli C. J., and Whittaker J. C. Bayesian semiparametric meta-analysis for genetic association studies. *Genetic Epidemiology*, 35(5): 333–340, 2011.

Debray T. P., Vergouwe Y., Koffijberg H., Nieboer D., Steyerberg E. W., and Moons K. G. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology*, 68(3):279–89, 2015.

Duda R. O., Hart P. E., and Stork D. G. *Pattern classification*. Wiley, 2001. Appendix 4, equation 84.

Duggan D., Zheng S. L., Knowlton M., Benitez D., Dimitrov L., Wiklund F., Robbins C., Isaacs S. D., Cheng Y., Li G., Sun J., Chang B.-L., Marovich L., Wiley K. E., Blter K., Stattin P., Adami H.-O., Gielzak M., Yan G., Sauvageot J., Liu W., Kim J. W., Bleecker E. R., Meyers D. A., Trock B. J., Partin A. W., Walsh P. C., Isaacs W. B., Grönberg H., Xu J., and Carpten J. D. Two Genome-wide Association Studies of Aggressive Prostate Cancer Implicate Putative Prostate Tumor Suppressor Gene DAB2IP. *Journal of the National Cancer Institute*, 99(24): 1836–1844, 2007.

EDRN. *Highlights of the accomplishments of the Early Detection Research Network, National Cancer Institute*. 2016. URL `http://edrn.nci.nih.gov/resources/highlights`. Accessed Feb 2, 2016.

Eeles R. A., Kote-Jarai Z., Giles G. G., Olama A. A. A., Guy M., Jugurnauth S. K., Mulholland S., Leongamornlert D. A., Edwards S. M., Morrison J., Field H. I., Southey M. C., Severi G., Donovan J. L., Hamdy F. C., Dearnaley D. P., Muir K. R., Smith C., Bagnato M., Ardern-Jones A. T., Hall A. L., O'Brien L. T., Gehr-Swain B. N., Wilkinson R. A., Cox A., Lewis S., Brown P. M., Jhavar S. G., Tymrakiewicz M., Lophatananon A., Bryant S. L., Horwich A., Huddart R. A., Khoo V. S., Parker C. C., Woodhouse C. J., Thompson A., Christmas

T., Ogden C., Fisher C., Jamieson C., Cooper C. S., English D. R., Hopper J. L., Neal D. E., and Easton D. F. Multiple newly identified loci associated with prostate cancer susceptibility. *Nature Genetics*, 40(3):316–321, 2008.

Eeles R. A., Kote-Jarai Z., Al Olama A. A., Giles G. G., Guy M., Severi G., Muir K., Hopper J. L., Henderson B. E., Haiman C. A., Schleutker J., Hamdy F. C., Neal D. E., Donovan J. L., Stanford J. L., Ostrander E. A., Ingles S. A., John E. M., Thibodeau S. N., Schaid D., Park J. Y., Spurdle A., Clements J., Dickinson J. L., Maier C., Vogel W., Dork T., Rebbeck T. R., Cooney K. A., Cannon-Albright L., Chappuis P. O., Hutter P., Zeegers M., Kaneva R., Zhang H.-W., Lu Y.-J., Foulkes W. D., English D. R., Leongamornlert D. A., Tymrakiewicz M., Morrison J., Ardern-Jones A. T., Hall A. L., O'Brien L. T., Wilkinson R. A., Saunders E. J., Page E. C., Sawyer E. J., Edwards S. M., Dearnaley D. P., Horwich A., Huddart R. A., Khoo V. S., Parker C. C., Van As N., Woodhouse C. J., Thompson A., Christmas T., Ogden C., Cooper C. S., Southey M. C., Lophatananon A., Liu J.-F., Kolonel L. N., Le Marchand L., Wahlfors T., Tammela T. L., Auvinen A., Lewis S. J., Cox A., FitzGerald L. M., Koopmeiners J. S., Karyadi D. M., Kwon E. M., Stern M. C., Corral R., Joshi A. D., Shahabi A., McDonnell S. K., Sellers T. A., Pow-Sang J., Chambers S., Aitken J., Gardiner R. A. F., Batra J., Kedda M. A., Lose F., Polanowski A., Patterson B., Serth J., Meyer A., Luedeke M., Stefflova K., Ray A. M., Lange E. M., Farnham J., Khan H., Slavov C., Mitkova A., Cao G., and Easton D. F. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nature Genetics*, 41(10):1116–1121, 2009.

Eyre S. J., Ankerst D. P., Wei J. T., Nair P. V., Regan M. M., Bueti G., Tang J., Rubin M. A., Kearney M., Thompson I. M., and Sanda M. G. Validation in a multiple urology practice cohort of the prostate cancer prevention trial calculator for predicting prostate cancer detection. *The Journal of Urology*, 182(6):2653 – 2658, 2009.

Falconer D. S. and Mackay T. F. C. *Introduction to Quantitative Genetics*. Pearson, London, 1996.

Fawcett T. An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

FDA. US Food and Drug Administration. Report: Paving the Way for Personalized Medicine, FDAs Role in a New Era of Medical Product Development, 2013. URL `http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/PersonalizedMedicine/UCM372421.pdf`.

Gail M. H. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *Journal of the National Cancer Institute*, 100(14):1037–1041, 2008.

Gail M. H. Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. *Journal of the National Cancer Institute*, 101(13):959–963, 2009.

Gail M. H., Brinton L. A., Byar D. P., Corle D. K., Green S. B., Schairer C., and Mulvihill J. J. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute*, 81(24):1879–1886, 1989.

Gasparrini A., Armstrong B., and Kenward M. G. Multivariate meta-analysis for non-linear and other multi-parameter associations. *Statistics in Medicine*, 31(29):3821–3839, 2012.

Ghadirian P., Howe G. R., Hislop T. G., and Maisonneuve P. Family history of prostate cancer: a multi-center case-control study in canada. *International Journal of Cancer*, 70(6):679–81, 1997.

Goldstein B. A., Knowles J. W., Salfati E., Ioannidis J. P., and Assimes T. L. Simple, standardized incorporation of genetic risk into non-genetic risk prediction tools for complex traits: coronary heart disease as an example. *Frontiers in Genetics*, 5(254), 2014.

Grill S., Fallah M., Leach R. J., Thompson I. M., Freedland S., Hemminki K., and Ankerst D. P. Incorporation of detailed family history from the swedish family cancer database into the pcpt risk calculator. *The Journal of Urology*, 193(2):460–465, 2015a.

Grill S., Fallah M., Leach R. J., Thompson I. M., Hemminki K., and Ankerst D. P. A simple-to-use method incorporating genomic markers into prostate cancer risk prediction tools facilitated future validation. *Journal of Clinical Epidemiology*, 68(5):563–573, 2015b.

Grill S., Ankerst D. P., Gail M. H., Chatterjee N., and Pfeiffer R. M. Comparison of approaches for incorporating new information into existing risk prediction models. 2016. Manuscript submitted for publication, under revision, Departments of Life Sciences and Mathematics of the Technical University Munich, Germany.

Gu W. and Pepe M. S. Estimating the capacity for improvement in risk prediction with a marker. *Biostatistics*, 10(1):172–186, 2009.

Gu W. and Pepe M. S. Estimating the diagnostic likelihood ratio of a continuous marker. *Biostatistics*, 12(1):87101, 2011.

Gudmundsson J., Sulem P., Manolescu A., Amundadottir L. T., Gudbjartsson D., Helgason A., Rafnar T., Bergthorsson J. T., Agnarsson B. A., Baker A., Sigurdsson A., Benediktsdottir K. R., Jakobsdottir M., Xu J., Blondal T., Kostic J., Sun J., Ghosh S., Stacey S. N., Mouy M., Saemundsdottir J., Backman V. M., Kristjansson K., Tres A., Partin A. W., Albers-Akkers M. T., Godino-Ivan Marcos J., Walsh P. C., Swinkels D. W., Navarrete S., Isaacs S. D., Aben K. K., Graif T., Cashy J., Ruiz-Echarri M., Wiley K. E., Suarez B. K., Witjes J. A., Frigge M., Ober C., Jonsson E., Einarsson G. V., Mayordomo J. I., Kiemeney L. A., Isaacs W. B., Catalona W. J., Barkardottir R. B., Gulcher J. R., Thorsteinsdottir U., Kong A., and Stefansson K. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nature Genetics*, 39(5):631–637, 2007a.

Gudmundsson J., Sulem P., Steinthorsdottir V., Bergthorsson J. T., Thorleifsson G., Manolescu A., Rafnar T., Gudbjartsson D., Agnarsson B. A., Baker A., Sigurdsson A., Benediktsdottir K. R., Jakobsdottir M., Blondal T., Stacey S. N., Helgason A., Gunnarsdottir S., Olafsdottir A., Kristinsson K. T., Birgisdottir B., Ghosh S., Thorlacius S., Magnusdottir D., Stefansdottir G., Kristjansson K., Bagger Y., Wilensky R. L., Reilly M. P., Morris A. D., Kimber C. H., Adeyemo A., Chen Y., Zhou J., So W.-Y., Tong P. C. Y., Ng M. C. Y., Hansen T., Andersen G., Borch-Johnsen K., Jorgensen T., Tres A., Fuertes F., Ruiz-Echarri M., Asin L., Saez B., van Boven E., Klaver S., Swinkels D. W., Aben K. K., Graif T., Cashy J., Suarez B. K., van Vierssen Trip O., Frigge M. L., Ober C., Hofker M. H., Wijmenga C., Christiansen C., Rader D. J., Palmer C. N. A., Rotimi C., Chan J. C. N., Pedersen O., Sigurdsson G., Benediktsson R., Jonsson E., Einarsson G. V., Mayordomo J. I., Catalona W. J., Kiemeney L. A., Barkardottir R. B., Gulcher J. R., Thorsteinsdottir U., Kong A., and Stefansson K. Two variants on chromosome 17 confer prostate cancer risk, and the one in tcf2 protects against type 2 diabetes. *Nature Genetics*, 39(8):977–983, 2007b.

Gudmundsson J., Sulem P., Rafnar T., Bergthorsson J. T., Manolescu A., Gudbjartsson D., Agnarsson B. A., Sigurdsson A., Benediktsdottir K. R., Blondal T., Jakobsdottir M., Stacey S. N., Kostic J., Kristinsson K. T., Birgisdottir B., Ghosh S., Magnusdottir D. N., Thorlacius S., Thorleifsson G., Zheng S. L., Sun J., Chang B.-L., Elmore J. B., Breyer J. P., McReynolds K. M., Bradley K. M., Yaspan B. L., Wiklund F., Stattin P., Lindstrom S., Adami H.-O., McDonnell S. K., Schaid D. J., Cunningham J. M., Wang L., Cerhan J. R., St Sauver J. L., Isaacs S. D., Wiley K. E., Partin A. W., Walsh P. C., Polo S., Ruiz-Echarri M., Navarrete

S., Fuertes F., Saez B., Godino J., Weijerman P. C., Swinkels D. W., Aben K. K., Witjes J. A., Suarez B. K., Helfand B. T., Frigge M. L., Kristjansson K., Ober C., Jonsson E., Einarsson G. V., Xu J., Gronberg H., Smith J. R., Thibodeau S. N., Isaacs W. B., Catalona W. J., Mayordomo J. I., Kiemeney L. A., Barkardottir R. B., Gulcher J. R., Thorsteinsdottir U., Kong A., and Stefansson K. Common sequence variants on 2p15 and xp11.22 confer susceptibility to prostate cancer. *Nature Genetics*, 40(3):281–283, 2008.

Gudmundsson J., Sulem P., Gudbjartsson D. F., Blondal T., Gylfason A., Agnarsson B. A., Benediktsdottir K. R., Magnusdottir D. N., Orlygsdottir G., Jakobsdottir M., Stacey S. N., Sigurdsson A., Wahlfors T., Tammela T., Breyer J. P., McReynolds K. M., Bradley K. M., Saez B., Godino J., Navarrete S., Fuertes F., Murillo L., Polo E., Aben K. K., van Oort I. M., Suarez B. K., Helfand B. T., Kan D., Zanon C., Frigge M. L., Kristjansson K., Gulcher J. R., Einarsson G. V., Jonsson E., Catalona W. J., Mayordomo J. I., Kiemeney L. A., Smith J. R., Schleutker J., Barkardottir R. B., Kong A., Thorsteinsdottir U., Rafnar T., and Stefansson K. Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nature Genetics*, 41(10):1122–1126, 2009.

Gudmundsson J., Besenbacher S., Sulem P., Gudbjartsson D. F., Olafsson I., Arinbjarnarson S., Agnarsson B. A., Benediktsdottir K. R., Isaksson H. J., Kostic J. P., Gudjonsson S. A., Stacey S. N., Gylfason A., Sigurdsson A., Holm H., Bjornsdottir U. S., Eyjolfsson G. I., Navarrete S., Fuertes F., Garcia-Prats M. D., Polo E., Checherita I. A., Jinga M., Badea P., Aben K. K., Schalken J. A., van Oort I. M., Sweep F. C., Helfand B. T., Davis M., Donovan J. L., Hamdy F. C., Kristjansson K., Gulcher J. R., Masson G., Kong A., Catalona W. J., Mayordomo J. I., Geirsson G., Einarsson G. V., Barkardottir R. B., Jonsson E., Jinga V., Mates D., Kiemeney L. A., Neal D. E., Thorsteinsdottir U., Rafnar T., and Stefansson K. Genetic correction of psa values using sequence variants associated with psa levels. *Science Translational Medicine*, 2(62):62ra92, 2010.

Hand D. and Yu K. Idiot's bayes - not so stupid after all? *International Statistical Review*, 69 (3):385–398, 2001.

Helfand B., Roehl K., Cooper P., McGuire B., Fitzgerald L., Cancel-Tassin G., Cornu J.-N., Bauer S., Van Blarigan E., Chen X., Duggan D., Ostrander E., Gwo-Shu M., Zhang Z.-F., Chang S.-C., Jeong S., Fontham E., Smith G., Mohler J., Berndt S., McDonnell S., Kittles R., Rybicki B., Freedman M., Kantoff P., Pomerantz M., Breyer J., Smith J., Rebbeck T.,

Mercola D., Isaacs W., Wiklund F., Cussenot O., Thibodeau S., Schaid D., Cannon-Albright L., Cooney K., Chanock S., Stanford J., Chan J., Witte J., Xu J., Bensen J., Taylor J., and Catalona W. Associations of prostate cancer risk variants with disease aggressiveness: results of the nci-spore genetics working group analysis of 18,343 cases. *Human Genetics*, 134(4): 439–450, 2015.

Hemminki K. and Czene K. Age specific and attributable risks of familial prostate carcinoma from the family-cancer database. *Cancer*, 95(6):1346–53, 2002.

Hemminki K., Granström C., Sundquist J., and Bermejo J. L. The updated swedish family-cancer database used to assess familial risks of prostate cancer during rapidly increasing incidence. *Hereditary Cancer in Clinical Practice*, 4(4):186–192, 2006.

Hemminki K., Ji J. G., Brandt A., Mousavi S. M., and Sundquist J. The swedish family-cancer database 2009: prospects for histology-specific and immigrant studies. *International Journal of Cancer*, 126:2259–2267, 2010.

Hernandez D. J., Han M., Humphreys E. B., Mangold L. A., Taneja S. S., Childs S. J., Bartsch G., and Partin A. W. Predicting the outcome of prostate biopsy: comparison of a novel logistic regression-based model, the prostate cancer risk calculator, and prostate-specific antigen level alone. *BJU International*, 103(5):609–614, 2009.

Hill W. and Robertson A. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38(6):226–231, 1968.

Hsu F.-C., Sun J., Wiklund F., Isaacs S. D., Wiley K. E., Purcell L. D., Gao Z., Stattin P., Zhu Y., Kim S.-T., Zhang Z., Liu W., Chang B.-L., Walsh P. C., Duggan D., Carpten J. D., Isaacs W. B., Grönberg H., Xu J., and Zheng S. L. A novel prostate cancer susceptibility locus at 19q13. *Cancer Research*, 69(7):2720–2723, 2009.

Hsu F.-C., Sun J., Zhu Y., Kim S.-T., Jin T., Zhang Z., Wiklund F., Kader A. K., Zheng S. L., Isaacs W., Grönberg H., and Xu J. Comparison of two methods for estimating absolute risk of prostate cancer based on single nucleotide polymorphisms and family history. *Cancer Epidemiology, Biomarkers & Prevention*, 19(4):1083–1088, 2010.

Hu G., Root M., and Duncan A. Adding multiple risk factors improves framingham coronary heart disease risk scores. *Vascular health and risk management*, 10:557–562, 2014.

Huang Y., Pepe M., and Feng Z. Logistic regression analysis with standardized markers. *Annals of Applied Statistics*, 7(3):1640–1662, 2013.

Hunter D. Gene-environment interactions in human diseases. *Nature Reviews Genetics*, 6(4): 287–298, 2005.

IARC. *Colorectal Cancer, Estimated Incidence, Mortality and Prevalence Worldwide in 2012.* International Agency on Research on Cancer, World Health Organization, Section of Cancer Surveillance, Globocan 2012, 2012. URL `http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx`. Accessed Dec 14, 2015.

IHGSC. International human genome sequencing consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431:931945, 2004.

Ilic D., O'Connor D., Green S., and Wilt T. J. Screening for prostate cancer: an updated cochrane systematic review. *BJU International*, 107(6):882–91, 2011.

Janssens A. C. J. W., Deng Y., Borsboom G. J. J. M., Eijkemans M. J. C., Habbema J. D. F., and Steyerberg E. W. A new logistic regression approach for the evaluation of diagnostic test results. *Medical Decision Making*, 25(2):168–177, 2005.

Jiang H., Liu F., Wang Z., Na R., Zhang L., Wu Y., Zheng J., Lin X., Jiang D., Sun J., Zheng S., Ding Q., and Xu J. Prediction of prostate cancer from prostate biopsy in chinese men using a genetic score derived from 24 prostate cancer risk-associated snps. *Prostate*, 73(15): 1651–1659, 2013.

Johansson M., Holmström B., Hinchliffe S. R., Bergh A., Stenman U.-H., Hallmans G., Wiklund F., and Stattin P. Combining 33 genetic variants with prostate-specific antigen for prediction of prostate cancer: Longitudinal study. *International Journal of Cancer*, 130(1):129–137, 2012.

Johns L. E. and Houlston R. S. A systematic review and meta-analysis of familial prostate cancer risk. *BJU International*, 91(9):789–94, 2003.

Johnson A. D., Handsaker R. E., Pulit S. L., Nizzari M. M., O'Donnell C. J., and de Bakker P. I. W. Snap: a web-based tool for identification and annotation of proxy snps using hapmap. *Bioinformatics*, 24(24):2938–2939, 2008. URL `https://www.broadinstitute.org/mpg/snap/`. Accessed Feb 17, 2014.

Kader A. K., Sun J., Reck B. H., Newcombe P. J., Kim S.-T., Hsu F.-C., Jr. R. B. D., Tao S., Zhang Z., Turner A. R., Platek G. T., Spraggs C. F., Whittaker J. C., Lane B. R., Isaacs W. B., Meyers D. A., Bleecker E. R., Torti F. M., Trent J. M., McConnell J. D., Zheng S. L., Condreay L. D., Rittmaster R. S., and Xu J. Potential impact of adding genetic markers to clinical parameters in predicting prostate biopsy outcomes in men following an initial negative biopsy: Findings from the reduce trial. *European Urology*, 62(6):953 – 961, 2012.

Kaplan D. J., Boorjian S. A., Ruth K., Egleston B. L., Chen D. Y. T., Viterbo R., Uzzo R. G., Buyyounouski M. K., Raysor S., and Giri V. N. Evaluation of the prostate cancer prevention trial risk calculator in a high-risk screening population. *BJU International*, 105(3):334–337, 2010.

Kerber R. A. and O'Brien E. A cohort study of cancer risk in relation to family histories of cancer in the utah population database. *Cancer*, 103(9):1906–15, 2005.

Kim S. T., Cheng Y., Hsu F.-C., Jin T., Kader A. K., Zheng S., Isaacs W., and Sun J. Prostate cancer risk-associated variants reported from genomewide association studies: meta-analysis and their contribution to genetic variation. *Prostate*, 70:1729–1738, 2010.

Kooter A. J., Kostense P. J., Groenewold J., Thijs A., Sattar N., and Smulders Y. M. Integrating information from novel risk factors with calculated risks: The critical impact of risk factor prevalence. *Circulation*, 124(6):741–745, 2011.

Landgren O., Engels E. A., Caporaso N. E., Gridley G., Mellemkjaer L., Hemminki K., Linet M. S., and Goldin L. R. Patterns of autoimmunity and subsequent chronic lymphocytic leukemia in nordic countries. *Blood*, 108(1):292–6, 2006.

Landrum M., Lee J., Riley G., Jang W., Rubinstein W., Church D., and Maglott D. Clinvar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1):D980–D985, 2014. Accessed Sept 7, 2015.

Lang T. A. and Secic M. *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers*. American College of Physicians, 2 edition, 2006. ISBN: 1-930513-69-0.

Lawrentschuk N., Daljeet N., Trottier G., Crawley P., and Fleshner N. E. An analysis of world

media reporting of two recent large randomized prospective trials investigating screening for prostate cancer. *BJU International*, 108(8 Pt 2):E190–5, 2011.

Lee D. H., Jung H. B., Park J. W., Kim K. H., Kim J., Lee S. H., and Chung B. H. Can western based online prostate cancer risk calculators be used to predict prostate cancer after prostate biopsy for the korean population? *Yonsei Medical Journal*, 54(3):665–671, 2013.

Lesko S. M., Rosenberg L., and Shapiro S. Family history and prostate cancer risk. *American Journal of Epidemiology*, 144(11):1041–7, 1996.

Lewontin R. C. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics*, 49(1):49–67, 1964.

Lewontin R. C. and Kojima K. The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4):458–472, 1960.

Lindström S., Schumacher F., and Cox D. G. Common genetic variants in prostate cancer risk prediction - Results from the NCI Breast and Prostate Cancer Cohort Consortium (BPC3). *Cancer Epidemiology, Biomarkers & Prevention*, 21:437–44, 2012.

Little J., Wilson B., Carter R., Walker K., Santaguida P., Tomiak E., Beyene J., Usman Ali M., and Raina P. Multigene panels in prostate cancer risk assessment: a systematic review. *Genetics in Medicine*, 2015.

Liu D. and Albert P. Combination of longitudinal biomarkers in predicting binary events. *Biostatistics*, 15(4):706–718, 2014.

Lu-Yao G. and Greenberg E. Changes in prostate cancer incidence and treatment in USA. *The Lancet*, 343(8892):251 – 254, 1994. Originally published as Volume 1, Issue 8892.

Machiela M., Chen C.-Y., Chen C., Chanock S., Hunter D., and Kraft P. Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. *Genetic Epidemiology*, 35(6):506–514, 2011.

Macinnis R., Antoniou A., Eeles R., Severi G., Al Olama A., Mcguffog L., Kote-Jarai Z., Guy M., O'Brien L., Hall A., Wilkinson R., Sawyer E., Ardern-Jones A., Dearnaley D., Horwich A., Khoo V., Parker C., Huddart R., Van As N., Mccredie M., English D., Giles G., Hopper J., and Easton D. A risk prediction algorithm based on family history and common genetic

variants: Application to prostate cancer with potential clinical impact. *Genetic Epidemiology*, 35(6):549–556, 2011.

Matikaine M. P., Pukkala E., Schleutker J., Tammela T. L., Koivisto P., Sankila R., and Kallioniemi O. P. Relatives of prostate cancer patients have an increased risk of prostate and stomach cancers: a population-based, cancer registry study in finland. *Cancer Causes Control*, 12(3):223–30, 2001.

Meintrup D. and Schäffler S. *Stochastik*. Springer, Berlin, 2005.

Mossialos E., Wenzl M., Osborn R., and Anderson C. 2014 international profiles of health care systems. *Commonwealth Fund*, January 2014. URL `http://www.commonwealthfund.org/~/media/files/publications/fund-report/2015/jan/1802_mossialos_intl_profiles_2014_v7.pdf?la=en`. pub. no. 1802.

Mukherjee B., Ahn J., Gruber S., and Chatterjee N. Testing gene-environment interaction in large-scale case-control association studies: Possible choices and comparisons. *American Journal of Epidemiology*, 175(3):177–190, 2012.

Nam R. K., Toi A., Trachtenberg J., Klotz L. H., Jewett M. A., Emami M., Sugar L., Sweet J., Pond G. R., and Narod S. A. Making sense of prostate specific antigen: Improving its predictive value in patients undergoing prostate biopsy. *The Journal of Urology*, 175(2):489 – 494, 2006.

Nam R. K., Kattan M. W., Chin J. L., Trachtenberg J., Singal R., Rendon R., Klotz L. H., Sugar L., Sherman C., Izawa J., Bell D., Stanimirovic A., Venkateswaran V., Diamandis E. P., Yu C., Loblaw D. A., and Narod S. A. Prospective multi-institutional study evaluating the performance of prostate cancer risk calculators. *Journal of Clinical Oncology*, 29(22): 2959–2964, 2011.

Neppl-Huber C., Zappa M., Coebergh J. W., Rapiti E., Rachtan J., Holleczek B., Rosso S., Aareleid T., Brenner H., and Gondos A. Changes in incidence, survival and mortality of prostate cancer in europe and the united states in the psa era: additional diagnoses and avoided deaths. *Annals of Oncology*, 23(5):1325–34, 2012.

Newcombe P. J., Reck B. H., Sun J., Platek G. T., Verzilli C., Kader A. K., Kim S.-T., Hsu F.-C., Zhang Z., Zheng S. L., Mooser V. E., Condreay L. D., Spraggs C. F., Whittaker

J. C., Rittmaster R. S., and Xu J. A comparison of bayesian and frequentist approaches to incorporating external information for the prediction of prostate cancer risk. *Genetic Epidemiology*, 36(1):71–83, 2012.

Nieboer D., Vergouwe Y., Roobol M. J., Ankerst D. P., Kattan M. W., Vickers A. J., and Steyerberg E. W. Nonlinear modeling was applied thoughtfully for risk prediction: the prostate biopsy collaborative group. *Journal of Clinical Epidemiology*, 68(4):426–34, 2015.

Oliveira M., Marques V., Carvalho A. P., and Santos A. Head-to-head comparison of two online nomograms for prostate biopsy outcome prediction. *BJU International*, 107(11):1780–1783, 2011.

Pakkanen S., Kujala P. M., Ha N., Matikainen M. P., Schleutker J., and Tammela T. L. Clinical and histopathological characteristics of familial prostate cancer in finland. *BJU International*, 109(4):557–63, 2012.

Parekh D. J., Ankerst D. P., Higgins B. A., Hernandez J., Canby-Hagino E., Brand T., Troyer D. A., Leach R. J., and Thompson I. M. External validation of the prostate cancer prevention trial risk calculator in a screened population. *Urology*, 68(6):1152 – 1155, 2006.

Park J.-H., Wacholder S., Gail M. H., Peters U., Jacobs K. B., Chanock S. J., and Chatterjee N. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics*, 42(7):570–575, 2010.

Park J.-H., Gail M., Greene M., and Chatterjee N. Potential usefulness of single nucleotide polymorphisms to identify persons at high cancer risk: An evaluation of seven common cancers. *Journal of Clinical Oncology*, 30(17):2157–2162, 2012.

Pepe P. and Aragona F. Prostate cancer detection rate at repeat saturation biopsy: Pcpt risk calculator versus pca3 score versus case-finding protocol. *The Canadian Journal of Urology*, 20:6620–6624, 2013.

Perdona S., Cavadas V., Lorenzo G. D., Damiano R., Chiappetta G., Prete P. D., Franco R., Azzarito G., Scala S., Arra C., Sio M. D., and Autorino R. Prostate cancer detection in the "grey area" of prostate-specific antigen below 10 ng/ml: Head-to-head comparison of the updated pcpt calculator and chun's nomogram, two risk estimators incorporating prostate cancer antigen 3. *European Urology*, 59(1):81 – 87, 2011.

Peto J., Easton D. F., Matthews F. E., Ford D., and Swerdlow A. J. Cancer mortality in relatives of women with breast cancer: the opcs study. office of population censuses and surveys. *International Journal of Cancer*, 65(3):275–83, 1996.

Prokunina-Olsson L., Muchmore B., Tang W., Pfeiffer R., Park H., Dickensheets H., Hergott D., Porter-Gill P., Mumy A., Kohaar I., Chen S., Brand N., Tarway M., Liu L., Sheikh F., Astemborski J., Bonkovsky H., Edlin B., Howell C., Morgan T., Thomas D., Rehermann B., Donnelly R., and O'Brien T. A variant upstream of ifnl3 (il28b) creating a new interferon gene ifnl4 is associated with impaired clearance of hepatitis c virus. *Nature Genetics*, 45(2): 164–171, 2013.

PubMed. *PubMed database: Currently indexed journals. U.S. National Library of Medicine, National Institutes of Health.* 2016. URL `http://www.ncbi.nlm.nih.gov/pubmed`. Accessed Mar 15, 2016.

Purcell S., Wray N., Stone J., Visscher P., O'Donovan M., Sullivan P., and Sklar P. Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder. *Nature*, 460(7256):748–752, 2009.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`.

Raji O. Y., Agbaje O. F., Duffy S. W., Cassidy A., and Field J. K. Incorporation of a genetic factor into an epidemiologic model for prediction of individual risk of lung cancer: The liverpool lung project. *Cancer Prevention Research*, 3(5):664–669, 2010.

Randazzo M., Müller A., Carlsson S., Eberli D., Huber A., Grobholz R., Manka L., Mortezavi A., Sulser T., Recker F., and Kwiatkowski M. A positive family history as a risk factor for prostate cancer in a population-based study with organised prostate-specific antigen screening: results of the swiss european randomised study of screening for prostate cancer (ERSPC, Aarau). *BJU International*, 2015.

Reich D. E., Cargill M., Bolk S., Ireland J., Sabeti P. C., Richter D. J., Lavery T., Kouyoumjian R., Farhadian S. F., Ward R., and Lander E. S. Linkage disequilibrium in the human genome. *Nature*, 411:199–204, 2001.

Ribeiro R., Monteiro C., Azevedo A., Cunha V., Ramanakumar A., Fraga A., Pina F., Lopes C., Medeiros R., and Franco E. Performance of an adipokine pathway-based multilocus genetic risk score for prostate cancer risk prediction. *PLoS ONE*, 7(6), 2012.

RKI. *Krebs in Deutschland 2009/2010*, volume 9. Robert Koch-Institut (Hrsg) und die Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. (Hrsg), Berlin, 2013. ISBN 978-3-89606-221-5.

Roehl K. A., Loeb S., Antenor J. A., Corbin N., and Catalona W. J. Characteristics of patients with familial versus sporadic prostate cancer. *The Journal of Urology*, 176(6 Pt 1):2438–42, 2006.

Roobol M. J., Zhu X., Schroder F. H., van Leenders G. J., van Schaik R. H., Bangma C. H., and Steyerberg E. W. A calculator for prostate cancer risk 4 years after an initially negative screen: Findings from erspc rotterdam. *European Urology*, 63(4):627–33, 2013.

Roudgari H., Hemminki K., Brandt A., Sundquist J., and Fallah M. Prostate cancer risk assessment model: a scoring model based on the swedish family-cancer database. *Journal of Medical Genetics*, 49:345–352, 2012.

Roussas G. G. *An introduction to probability and statistical inference.* Academic Press, London, U.K., 2 edition, 2015. ISBN: 978-0-12-800437-1.

Ruffion A., Devonec M., Champetier D., Decaussin-Petrucci M., Rodriguez-Lafrasse C., Paparel P., Perrin P., and Vlaeminck-Guillem V. Pca3 and pca3-based nomograms improve diagnostic accuracy in patients undergoing first prostate biopsy. *International Journal of Molecular Sciences*, 14(9):17767–17780, 2013.

SAS Institute Inc. *SAS 9.4.* Copyright 2015, SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Sävblom C., Hallden C., Cronin A. M., Sall T., Savage C., Vertosick E. A., Klein R. J., Giwercman A., and Lilja H. Genetic variation in klk2 and klk3 is associated with concentrations of hk2 and psa in serum and seminal plasma in young men. *Clinical Chemistry*, 60(3):490–9, 2014.

Scales J. C. D., Zarei M., and Dahm P. Evidence-based urology in practice: likelihood ratios. *BJU International*, 104(7):892–4, 2009.

SEER. *SEER Stat Fact Sheets: All Cancer Sites*. Surveillance, Epidemiology, and End Results (SEER), 2015a. URL `http://seer.cancer.gov/statfacts/html/all.html`. Accessed Jan 25, 2016, Sub (2010-2012).

SEER. *SEER Stat Fact Sheets: Prostate Cancer*. Surveillance, Epidemiology, and End Results (SEER), 2015b. URL `http://seer.cancer.gov/statfacts/html/prost.html`. Accessed Dec 11, 2015, Sub (2010-2012).

SEER. *SEER Stat Fact Sheets: Stomach Cancer*. Surveillance, Epidemiology, and End Results (SEER), 2016a. URL `http://seer.cancer.gov/statfacts/html/stomach.html`. Accessed Feb 18, 2016, Sub (2010-2012).

SEER. *SEER Stat Fact Sheets: Thyroid Cancer*. Surveillance, Epidemiology, and End Results (SEER), 2016b. URL `http://seer.cancer.gov/statfacts/html/thyro.html`. Accessed Feb 18, 2016, Sub (2010-2012).

Siddiqui S. A., Sengupta S., Slezak J. M., Bergstralh E. J., Zincke H., and Blute M. L. Impact of familial and hereditary prostate cancer on cancer specific survival after radical retropubic prostatectomy. *The Journal of Urology*, 176(3):1118–21, 2006.

Simko V. and Ginter E. Region-specific differences in colorectal cancer: Slovakia and hungary have highest incidence in europe. *Bratisl Lek Listy*, 117(2):66–71, 2016.

Skates S., Pauler D., and Jacobs I. Screening based on the risk of cancer calculation from bayesian hierarchical changepoint and mixture models of longitudinal markers. *Journal of the American Statistical Association*, 96(454):429–439, 2001.

Spiegelhalter D. J. and Knill-Jones R. P. Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *Journal of the Royal Statistical Society. Series A (General)*, 147(1):pp. 35–77, 1984.

Strobl A. N., Thompson I. M., Vickers A. J., and Ankerst D. P. The next generation of clinical decision making tools: Development of a real-time prediction tool for outcome of prostate biopsy in response to a continuously evolving prostate cancer landscape. *The Journal of Urology*, 194(1):58 – 64, 2015.

Sullivan J., Kopp R., Stratton K., Manschreck C., Corines M., Rau-Murthy R., Hayes J., Lincon A., Ashraf A., Thomas T., Schrader K., Gallagher D., Hamilton R., Scher H., Lilja H.,

Scardino P., Eastham J., Offit K., Vijai J., and Klein R. J. An analysis of the association between prostate cancer risk loci, psa levels, disease aggressiveness and disease-specific mortality. *British Journal of Cancer*, 113(1):166–72, 2015.

Sun J., Zheng S. L., Wiklund F., Isaacs S. D., Purcell L. D., Gao Z., Hsu F.-C., Kim S.-T., Liu W., Zhu Y., Stattin P., Adami H.-O., Wiley K. E., Dimitrov L., Sun J., Li T., Turner A. R., Adams T. S., Adolfsson J., Johansson J.-E., Lowey J., Trock B. J., Partin A. W., Walsh P. C., Trent J. M., Duggan D., Carpten J., Chang B.-L., Gronberg H., Isaacs W. B., and Xu J. Evidence for two independent prostate cancer risk-associated loci in the hnf1b gene at 17q12. *Nature Genetics*, 40(10):1153–1155, Oct. 2008.

Szulkin R., Whitington T., Eklund M., Aly M., Eeles R., Easton D., Kote-Jarai Z., Amin Al Olama A., Benlloch S., Muir K., Giles G., Southey M., FitzGerald L., Henderson B., Schumacher F., Haiman C., Schleutker J., Wahlfors T., Tammela T., Nordestgaard B., Key T., Travis R., Neal D., Donovan J., Hamdy F., Pharoah P., Pashayan N., Khaw K.-T., Stanford J., Thibodeau S., McDonnell S., Schaid D., Maier C., Vogel W., Luedeke M., Herkommer K., Kibel A., Cybulski C., Lubinski J., Kluzniakniak W., Cannon-Albright L., Brenner H., Butterbach K., Stegmaier C., Park J., Sellers T., Lim H.-Y., Slavov C., Kaneva R., Mitev V., Batra J., Clements J., Spurdle A., Teixeira M., Paulo P., Maia S., Pandha H., Michael A., Kierzek A., Gronberg H., and Wiklund F. Prediction of individual genetic risk to prostate cancer using a polygenic score. *Prostate*, 75(13):1467–1474, 2015.

The International HapMap Consortium. The international hapmap project. *Nature*, 426:789–796, 2003.

The International HapMap Consortium. Hapmap populations. June 2005. URL `http://hapmap.ncbi.nlm.nih.gov/citinghapmap.html`. Accessed Mar 17, 2014.

Thomas et al. D. Multiple loci identified in a genome-wide association study of prostate cancer. *Nature Genetics*, 40:310–315, 2008.

Thompson I., Ankerst D. P., Chi C., Goodman P. J., Tangen C. M., Lucia M. S., Feng Z., Parnes H. L., and Coltman C. A. J. Assessing prostate cancer risk: Results from the prostate cancer prevention trial. *Journal of the National Cancer Institute*, 98:529 – 534, 2006.

Tice J., Cummings S., Smith-Bindman R., Ichikawa L., Barlow W., and Kerlikowske K. Using

clinical factors and mammographic breast density to estimate breast cancer risk: Development and validation of a new predictive model. *Annals of Internal Medicine*, 148(5):337–347, 2008.

Trottier G., Roobol M. J., Lawrentschuk N., Boström P. J., Fernandes K. A., Finelli A., Chadwick K., Evans A., van der Kwast T. H., Toi A., Zlotta A. R., and Fleshner N. E. Comparison of risk calculators from the prostate cancer prevention trial and the european randomized study of screening for prostate cancer in a contemporary canadian cohort. *BJU International*, 108(8b):E237–E244, 2011.

van Houwelingen H. C., Arends L. R., and Stijnen T. Tutorial in biostatistics, advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine*, 21:589–624, 2002.

van Zitteren M., van der Net J. B., Kundu S., Freedman A. N., van Duijn C. M., and Janssens A. C. J. Genome-based prediction of breast cancer risk in the general population: A modeling study based on meta-analyses of genetic associations. *Cancer Epidemiology, Biomarkers & Prevention*, 20(1):9–22, 2011.

Vickers A. Prediction models in urology: Are they any good, and how would we know anyway? *European Urology*, 57(4):571–573, 2010.

ViraHepC. *Study of Viral Resistance to Antiviral Therapy of Chronic Hepatitis C (Virahep-C)*. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), 2002-2006. ClinicalTrials.gov identifier: NCT00038974.

Wacholder S., Hartge P., Prentice R., Garcia-Closas M., Feigelson H. S., Diver W. R., Thun M. J., Cox D. G., Hankinson S. E., Kraft P., Rosner B., Berg C. D., Brinton L. A., Lissowska J., Sherman M. E., Chlebowski R., Kooperberg C., Jackson R. D., Buckman D. W., Hui P., Pfeiffer R., Jacobs K. B., Thomas G. D., Hoover R. N., Gail M. H., Chanock S. J., and Hunter D. J. Performance of common genetic variants in breast-cancer risk models. *New England Journal of Medicine*, 362(11):986–993, 2010.

Walford G. A., Porneala B. C., Dauriz M., Vassy J. L., Cheng S., Rhee E., Wang T. J., Meigs J. B., Gerszten R. E., and Florez J. C. Metabolite traits and genetic risk provide complementary information for the prediction of future type 2 diabetes. *Diabetes Care*, 37 (9):2508–2514, 2014.

Wiens J., Guttag J., and Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 21(4):699–706, 2014.

Williams S. B., Salami S., Regan M. M., Ankerst D. P., Wei J. T., Rubin M. A., Thompson I. M., and Sanda M. G. Selective detection of histologically aggressive prostate cancer: an early detection research network prediction model to reduce unnecessary prostate biopsies with validation in the prostate cancer prevention trial. *Cancer*, 118(10):2651–8, 2012.

Wilson P., D'Agostino R., Levy D., Belanger A., Silbershatz H., and Kannel W. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.

Wu C., Shi X., Cui Y., and Ma S. A penalized robust semiparametric approach for gene-environment interactions. *Statistics in Medicine*, 34(30):4016–30, 2015.

Wu J., Pfeiffer R. M., and Gail M. H. Strategies for developing prediction models from genome-wide association studies. *Genetic Epidemiology*, 37(8):768–777, 2013.

Xu J., Sun J., Kader A. K., Lindström S., Wiklund F., Hsu F.-C., Johansson J.-E., Zheng S. L., Thomas G., Hayes R. B., Kraft P., Hunter D. J., Chanock S. J., Isaacs W. B., and Grönberg H. Estimation of absolute risk for prostate cancer using genetic markers and family history. *Prostate*, 69(14):1565–1572, 2009.

Yeager M., Orr N., Hayes R. B., Jacobs K. B., Kraft P., Wacholder S., Minichiello M. J., Fearnhead P., Yu K., Chatterjee N., Wang Z., Welch R., Staats B. J., Calle E. E., Feigelson H. S., Thun M. J., Rodriguez C., Albanes D., Virtamo J., Weinstein S., Schumacher F. R., Giovannucci E., Willett W. C., Cancel-Tassin G., Cussenot O., Valeri A., Andriole G. L., Gelmann E. P., Tucker M., Gerhard D. S., Fraumeni J. F., Hoover R., Hunter D. J., Chanock S. J., and Thomas G. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genetics*, 39(5):645–649, 2007.

Yeager M., Chatterjee N., Ciampa J., Jacobs K. B., Gonzalez-Bosquet J., Hayes R. B., Kraft P., Wacholder S., Orr N., Berndt S., Yu K., Hutchinson A., Wang Z., Amundadottir L., Feigelson H. S., Thun M. J., Diver W. R., Albanes D., Virtamo J., Weinstein S., Schumacher F. R., Cancel-Tassin G., Cussenot O., Valeri A., Andriole G. L., Crawford E. D., Haiman C. A., Henderson B., Kolonel L., Le Marchand L., Siddiq A., Riboli E., Key T. J., Kaaks R., Isaacs

W., Isaacs S., Wiley K. E., Gronberg H., Wiklund F., Stattin P., Xu J., Zheng S. L., Sun J., Vatten L. J., Hveem K., Kumle M., Tucker M., Gerhard D. S., Hoover R. N., Fraumeni J. F., Hunter D. J., Thomas G., and Chanock S. J. Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nature Genetics*, 41(10):1055–1057, Oct. 2009.

Zheng J., Liu F., Lin X., Wang X., Ding Q., Jiang H., Chen H., Lu D., Jin G., Hsing A., Shao Q., Qi J., Ye Y., Wang Z., Gao X., Wang G., Chu L., Ouyang J., Huang Y., Chen Y., Gao Y., Shi R., Wu Q., Wang M., Zhang Z., Hu Y., Sun J., Zheng S., Gao X., Xu C., Mo Z., Sun Y., and Xu J. Predictive performance of prostate cancer risk in chinese men using 33 reported prostate cancer risk-associated snps. *Prostate*, 72(5):577–583, 2012.

Zhu Y., Wang J.-Y., Shen Y.-J., Dai B., Ma C.-G., Xiao W.-J., Lin G.-W., Yao X.-D., Zhang S.-L., and Ye D.-W. External validation of the prostate cancer prevention trial and the european randomized study of screening for prostate cancer risk calculators in a chinese cohort. *Asian Journal of Andrology*, 14:738–744, 2012.

7 Appendix

7.1 Supporting tables

Methods	Overall	$X_1 = 0$	$X_1 = 1$	$Z < median$	$Z \geq median$
LR-joint	1.018	1.022	1.014	1.028	1.017
	(0.002)	(0.002)	(0.002)	(0.005)	(0.002)
LR-separate	1.020	1.028	1.010	1.035	1.019
	(0.002)	(0.003)	(0.003)	(0.005)	(0.002)
LR-ind	1.290	1.086	1.583	0.774	1.351
	(0.003)	(0.002)	(0.003)	(0.004)	(0.003)
LR-offset	1.118	0.950	1.360	1.046	1.127
	(0.003)	(0.003)	(0.004)	(0.005)	(0.004)
LR-shrink	1.117	0.951	1.357	1.035	1.127
	(0.003)	(0.003)	(0.004)	(0.005)	(0.004)
Logistic new	1.135	1.136	1.134	1.127	1.136
	(0.003)	(0.004)	(0.006)	(0.005)	(0.004)

(a) E/O ratios and standard errors in brackets

Methods	Overall	$X_1 = 0$	$X_1 = 1$	$Z < median$	$Z \geq median$
LR-joint	0.003	0.002	0.008	0.002	0.007
LR-separate	0.003	0.003	0.009	0.002	0.007
LR-ind	0.004	0.004	0.010	0.001	0.009
LR-offset	0.005	0.003	0.012	0.002	0.010
LR-shrink	0.005	0.003	0.013	0.002	0.010
Logistic new	0.005	0.005	0.018	0.002	0.011

(b) Standard deviations of $\bar{E}$

Table S1: Simulation results for updating based on case-control data for a continuous marker Z following a **mixture distribution** of normals with $p = 4, P(X_i = 1) = 0.2, Z \sim 0.9N(\alpha_{10} + \boldsymbol{\alpha}_1^T\mathbf{X}, 1) + 0.1N(\alpha_{20} + \boldsymbol{\alpha}_2^T\mathbf{X}, 1), \alpha_{10} = 0, \boldsymbol{\alpha}_1 = (0.7, 0.7, -0.7, -0.7)^T, \alpha_{20} = 0.5, \boldsymbol{\alpha}_2 = (1, 1, -1, -1)^T, \boldsymbol{\beta}_{\mathbf{X}} = (0.5, 0.5, -0.5, -0.5)^T, \beta_Z = 1$ in Equation (3.15), $P(Y) = 0.05$.

	Overall	**Race**		**IFNL4**		**HCV-RNA**	
Methods		Caucasian	African American	$\Delta G/\Delta G$ or $\Delta G/TT$	TT/TT	<median	≥median
LR-joint	1.002 (0.001)	1.008 (0.001)	0.986 (0.001)	0.994 (0.003)	1.010 (0.003)	1.042 (0.003)	0.952 (0.003)
LR-separate	1.031 (0.001)	1.057 (0.001)	0.957 (0.001)	0.999 (0.003)	1.060 (0.003)	1.086 (0.003)	0.961 (0.003)
LR-ind	1.113 (0.001)	1.220 (0.001)	0.798 (0.002)	0.929 (0.003)	1.275 (0.003)	1.146 (0.002)	1.070 (0.003)
LR-offset	0.999 (0.003)	1.061 (0.003)	0.818 (0.003)	0.995 (0.004)	1.003 (0.004)	0.994 (0.003)	1.006 (0.004)
LR-shrink	1.026 (0.001)	1.096 (0.001)	0.821 (0.002)	1.030 (0.003)	1.023 (0.003)	0.935 (0.001)	1.143 (0.002)
Logistic new	1.000 (0.003)	0.999 (0.003)	1.003 (0.006)	0.996 (0.004	1.003 (0.004)	0.997 (0.003)	1.003 (0.004)

(a) E/O ratios and standard errors in brackets

	Overall	**Race**		**IFNL4**		**HCV-RNA**	
Methods		Caucasian	African American	$\Delta G/\Delta G$ or $\Delta G/TT$	TT/TT	<median	≥median
LR-joint	0.002	0.004	0.001	0.006	0.015	0.010	0.009
LR-separate	0.003	0.006	0.002	0.006	0.017	0.009	0.006
LR-ind	0.003	0.006	0.003	0.006	0.019	0.009	0.009
LR-offset	0.009	0.014	0.005	0.009	0.023	0.012	0.011
LR-shrink	0.002	0.005	0.003	0.006	0.019	0.005	0.005
Logistic new	0.009	0.005	0.010	0.009	0.023	0.012	0.011

(b) Standard deviations of $\bar{E}$

Table S2: Simulation results for updating based on ViraHepC data for the cohort setting with both markers IFNL4 and HCV-RNA, $P(Y) = 0.1$.

7.2 Publications

The publications underlying this thesis, including supporting material, can be accessed via following links:

Grill et al. 2015a doi: `http://dx.doi.org/10.1016/j.juro.2014.09.018`

Grill et al. 2015b doi: `http://dx.doi.org/10.1016/j.jclinepi.2015.01.006`

8 Acknowledgements

First of all, I would like to thank my supervisor, Prof. Dr. Donna Ankerst, for dedicated supervision, endless patience and very productive discussions. Thank you for great professional and personal support for my work and my plans throughout the last years and for making my PhD possible to begin with.

Great thanks go to Prof. Dr. Wilhelm Windisch, Prof. Dr. Aurélien Tellier and Prof. Dr. Kathleen Herkommer for serving on my PhD committee. I want to express my sincere thanks to Dr. Ruth Pfeiffer for making my stay at the NIH possible and for greatly supporting me with our joint research project, as well as Dr. Mitchell Gail for inspiring discussions. I want to thank the urology group at the university hospital Klinikum rechts der Isar for productive collaborations. I would also like to greatly thank all co-authors and project partners, especially Prof. Dr. Ian Thompson and Dr. Mahdi Fallah for their great support and scientific contributions. Many thanks are extended to Dr. Maarit Laaksonen and Prof. Dr. Jake Olivier at the University of New South Wales and Dr. Armando Teixeira-Pinto at the University of Sydney for an inspiring time and for giving me various opportunities to present and discuss my work.

I wish to express my gratitude for my graduate school mentor, Dr. Florian Lipsmeier, for his dedication, great advice, encouragement during the last years and support for my future plans. Sincere thanks go to Dr. Hannes Petermeier for thoughtful comments, encouraging conversations, professional and personal support and to Josef Höfler for great advice, patience and help with the website. I want to thank all my colleagues at the Technische Universität München, Lehrstuhl for Mathematical Modeling of Biological Systems of Prof. Dr. Dr. Fabian Theis for a friendly and cooperative working environment, many discussions, various help and support, my office mates and other PhD students Andreas, Anna, Julie, Michael, Bendix, Lisa and Norbert. I always liked working here. Thanks a lot to Silke for helping me overcome every bureaucratic hurdle and finding answers to all my questions.

Abschließend möchte ich meiner Familie danken, zu allererst meinen Eltern, die mir das Studium und damit auch die Promotion ermöglichten und stets unterstützend hinter mir standen. Herzlichsten Dank auch meinen Brüdern, Michael und Andreas, sowie Annette und Elsa für ihre Ermutigung und Geduld während der letzen Jahre und vor allem meinem Partner Florian für ein stets offenes Ohr, die Motivation und Geduld in allen Lebenslagen und die Unterstützung all meiner Vorhaben von Anfang an.

9 Curriculum Vitae

Personal Information

Sonja Eva Grill

Date of birth: September 23rd, 1988

Place of birth: Nabburg, Germany

Education

Since 10/2013: PhD Student at the Technical University of Munich (TUM), Division of Mathematical Modelling of biological Systems

06/2015 - 07/2015: Predoctoral Fellowship, Biostatistics Branch, National Cancer Institute, Washington D.C., U.S.A.; Supervisor: Dr. Ruth Pfeiffer

07/2014 - 09/2014: Visiting Scientist at the University of New South Wales, Sydney, Australia

12/2011 - 08/2013: Master of Science in Mathematics in Bioscience at TUM

08/2012 - 09/2012: Research Project at the ETH, Division of Theoretical Biology, Zurich, Switzerland

10/2008 - 12/2011: Bachelor of Science in Mathematics at TUM, Minor: Physics

09/1999 - 07/2008: Secondary School, Johann-Andreas-Schmeller Gymnasium, Nabburg, Germany

Publications

- Brath J. M. S., **Grill S.**, Ankerst D. P., Thompson I. M. Jr., Gschwend J. E., and Herkommer K. No detrimental effect of a positive family history on long-term outcomes following radical prostatectomy. *The Journal of Urology*, 195(2):343-348, 2016.
- Morales E. E., **Grill S.**, Svatek R. S., Kaushik D., Thompson I. M. Jr., Ankerst D. P., and Liss M. A. Finasteride reduces risk of bladder cancer in large, prospective screening study. *European Urology*, 69(3):407-410, 2016.

- Kühn A., **Grill S.**, Baumgartner M., Ankerst D. P., and Matyssek R. Daily growth of European beech (Fagus sylvatica L.) on moist sites is affected by short-term drought rather than ozone uptake. *Trees*, 1-19, 2015.

- **Grill S.**, Fallah M., Leach R. J., Thompson I. M., Hemminki K., and Ankerst D. P. A simple-to-use method incorporating genomic markers into prostate cancer risk prediction tools facilitates future validation. *Journal of Clinical Epidemiology*, 68(5):563-73, 2015.

- **Grill S.**, Fallah M., Leach R. J., Thompson I. M., Freedland S., Hemminki K., and Ankerst D. P. Incorporation of detailed family history from the Swedish Family Cancer Database into the Prostate Cancer Prevention Trial Risk Calculator. *The Journal of Urology*, 193(2):460-5, 2015.

Contributions to international conferences

- Baade N., Laenger N., Klorek T., **Grill S.**, Schulwitz H., Albers P., Arsov C., Hadaschik B., Hohenfellner M., Imkamp F., Kuczyk M., Gschwend J., and Herkommer K. Assoziation zwischen PSA-Wert und Familienanamnese im 45-jährigen Kollektiv der deutschen Prostatakarzinom Screening Studie (PROBASE). Bayerisch-Österreichischer Urologenkongress, Augsburg, 2016

- Herkommer K., Laenger N., Klorek T., Ankerst D. P., **Grill S.**, Schulwitz H., Albers P., Arsov C., Hadaschik B., Hohenfellner M., Kuczyk M., Imkamp F., and Gschwend J. The association between family history and prostate-specific antigen from a large group of 45-year old men embarking on prostate cancer screening: Results from the PROBASE trial. AUA Meeting, San Diego and EAU Meeting, Munich, 2016.

- **Grill S.**, Ankerst D. P., and Pfeiffer R. M. Comparison of methods for updating risk prediction models. ENAR Spring Meeting in Austin, Texas, 2016 (presenter).

- Morales E. E., **Grill S.**, Freidberg N. A., Thompson I. M., Svatek R. S., Kaushik D., Ankerst D. P., and Liss M. A. Self-Reported Finasteride Use is Associated with Decreased Incidence of Bladder Cancer: Data from the Prostate, Lung, Colorectal, & Ovarian Cancer Study. AUA Meeting, New Orleans, 2015.

- Goetz J., **Grill S.**, Ankerst D. P., and Tseng T. Acute Urinary Stone Incidence as a Function of Temperature and Lag Length in a Subtropical Climate. AUA Meeting, New

Orleans, 2015.

- Brath J. M. S., **Grill S.**, Ankerst D. P., Gschwend J. E., and Herkommer K. Einfluss der Familienanamnese auf das karzinomspezifische berleben nach radikaler Prostatovesikulektomie bei jungen Prostatakarzinompatienten. Bayerisch-Österreichischer Urologenkongress, Linz, 2015 und Kongress der Deutschen Gesellschaft für Urologie, Hamburg, 2015.

- **Grill S.** and Ankerst D. P. Synthesizing Genetic Markers for Incorporation into Clinical Risk Prediction Tools. ENAR Spring Meeting, Miami, 2015 (presenter).

- **Grill S.** and Ankerst D. P. Synthesizing Genetic Markers for Incorporation into Clinical Risk Prediction Tools. IMS-ASC Konferenz in Sydney 2014 (presenter).