Estimation of Latent Familial Risks for Colorectal Cancer

Anna Rieger

München 2017

Estimation of Latent Familial Risks for Colorectal Cancer

Anna Rieger

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

eingereicht von
Anna Rieger
aus München

München, den 12. Juli 2017

Erstgutachter:	Prof. Dr. Ulrich Mansmann
Zweitgutachter:	Prof. Dr. Matthias Schmid
Tag der Einreichung:	12. Juli 2017
Tag der mündlichen Prüfung:	10. November 2017

Summary

In this thesis, a Bayesian approach is provided to estimate the posterior risk of experiencing a familial clustering of colorectal cancer, i. e. of being a "risk family" for colorectal cancer (CRC). The practical relevance is given, as a fifth to a quarter of all CRCs occurs in familial clusters. Only a small part is attributed to known genes causing diseases such as HNPCC or FAP. The cause for the remaining parts of CRC cases with familial clustering is unknown so far.

As no clear genetic model of "risk families" exists, several statistical approaches regarding heredity and penetrance are developed which cover a range of biologically plausible settings. Penetrance describes the duration until the outbreak of a disease caused by specific genes. The Weibull distribution is therefore suitable from a statistical point of view. It serves as likelihood in the Bayesian context. The parameters of the Weibull distribution are chosen such that they fit the natural incidence in Upper Bavaria, Germany (provided by Munich Cancer Registry (MCR)). Simple hereditary mechanisms, which provide the second component of a genetic model, are considered here. Complete risk transmission from parents to children serves as thinking model. Random risk transmission mimics monogenetic or polygenetic inheritance. In that case risk carrier property is inherited by the children with a certain probability. The prior is determined by the prevalence of the risk carrier property (e. g. unknown genes) and the inheritance mechanism. It gives the *a priori* risk of being a "risk family", i. e. the risk without any knowledge about family history and properties of family members like age and sex.

Simulations are used to analyse if the proposed method is working. Estimation is done via grid search as the likelihood and posterior can be written down in closed form. The quality of the prediction of having more than one CRC case in the family, i. e. being a "risk family", is analysed by ROC curves. The Bayesian method described above is compared to a questionnaire developed by "Netzwerk gegen Darmkrebs" that provides a four-point score to evaluate familial CRC risk.

Real data to apply the described methods arise from a study ("Familien schützen und stärken – Umgang mit familiärem Darmkrebs") running from September 2012 until June 2014 in the catchment area of the MCR. The data set consists of 792 families with about 4000 family members. After excluding "families" with only one member, 611 families with 669 patients in total remain for a meaningful analysis. Both estimates arising from grid search and plugging in of epidemiological parameters from literature are used to estimate the posterior risk in the data set of the study.

Zusammenfassung

In dieser Arbeit wird ein Bayesianischer Ansatz verwendet, um die Posteriori-Wahrscheinlichkeit für eine familiäre Häufung von Darmkrebserkrankungen zu berechnen (Darmkrebs-Risiko-Familie). Dies ist praktisch relevant, da etwa eine Fünftel bis ein Viertel aller Darmkrebsfälle eine familiäre Häufung aufweist. Nur ein geringer Teil kann bereits bekannten Genen zugeordnet werden, welche Krankheiten wie HNPCC oder FAP auslösen. Für den übrigen Teil sind die Ursachen bisher unbekannt.

Da kein klares genetisches Modell für Risiko-Familien besteht, wurden verschiedene statistische Ansätze bezüglich Vererbung und Penetranz entwickelt, welche eine Reihe biologisch plausibler Szenarien abdecken. Penetranz beschreibt in der Genetik die Dauer bis zum Ausbruch der jeweiligen Krankheit. Die Weibull-Verteilung ist daher aus statistischer Sicht geeignet. Deren Parameter werden so gewählt, dass sie der natürlichen Inzidenz im Einzugsgebiet des Tumorregisters München (TRM) entsprechen. Es werden einfache Vererbungsmechanismen betrachtet, welche die zweite Komponente eines genetischen Modells darstellen. *Complete risk transmission* dient als Gedanken-Modell. *Random risk transmission* stellt mono- oder polygenetische Vererbung nach, bei der die Risikoträger-Eigenschaft mit einer gewissen Wahrscheinlichkeit vererbt wird. Die Prävalenz der Risikoträger-Eigenschaft (z. B. von unbekannten Genen) und der Vererbungsmechanismus bestimmen die Priori-Wahrscheinlichkeit eine Risiko-Familie zu sein, d. h. die Wahrscheinlichkeit ohne Wissen über die Krebsgeschichte einer Familie oder Eigenschaften wie Alter und Geschlecht ihrer Mitglieder zu haben.

Die Nutzbarkeit der vorgeschlagenen Methode wird mit Simulationen überprüft. Die Schätzung der Parameter erfolgt mittels *grid search*, da Likelihood und Posteriori in geschlossener Form dargestellt werden können. Die Trennqualität der Vorhersage, mehr als einen Darmkrebsfall in der Familie aufzuweisen, wird mit Hilfe von ROC-Kurven überprüft. Der Bayesianische Ansatz wird mit dem Fragebogen des „Netzwerk gegen Darmkrebs“ verglichen, der mit vier Fragen das familiäre Darmkrebs-Risiko evaluiert.

Die Methode wird auch bei der Studie „Familien schützen und stärken – Umgang mit familiärem Darmkrebs“ angewendet, welche von September 2012 bis Juni 2014 im Einzugsgebiet des TRM lief und 792 Familien mit etwa 4000 Mitgliedern rekrutierte. Nach Ausschluss von „Familien“ mit nur einem Mitglied verbleiben 611 Familien mit 669 Patienten für eine sinnvolle Analyse. Es werden sowohl geschätzte als auch epidemiologische Parameter aus der Literatur genutzt, um die Posteriori-Wahrscheinlichkeit im Studien-Datensatz zu berechnen.

List of Abbreviations

APC	adenomatous polyposis coli (gene)
CRC	colorectal cancer
CTM	complete risk transmission
DCO	death certificate only
DNA	deoxyribonucleic acid
EM	expectation-maximisation (algorithm)
FAP	familial adenomatous polyposis
GAM	generalized additive model
gFOBT	guaiac fecal occult blood test
HNPCC	hereditary non-polyposis colon cancer, also called Lynch syndrome
MCEM	Monte Carlo EM (algorithm)
MCR	Munich Cancer Registry ("Tumorregister München")
ML	maximum likelihood
MLE	maximum likelihood estimator
MSI	microsatellite instability
NACRC	Network against colorectal cancer (registered society) ("Netzwerk gegen Darmkrebs e.V.")
ROC	receiver operating characteristic
RR	relative risk
RRE	risk as random effects
RTM	random risk transmission
TRM	Tumorregister München

Contents

1
Overview

Motivation The motivation for this thesis is to provide a tool to refine screening for colorectal cancer (CRC). Familial clustering of CRC is not uncommon, but arises only in a small part of cases due to hereditary conditions (see section 3.2). The offer from statutory health insurances in Germany may come too late for members of a family with familial CRC risk. To prevent persons from "over-screening" by means of cumbersome colonscopy and to give the persons at risk an appropriate screening offer, one needs to differentiate between "risk families" and families without familial CRC risk. Unfortunately, risk factors for familial CRC are unknown to date. So, e. g. genetic screening for causing genes is not possible and one needs to go new ways. This thesis may provide a tool to boost tailored screening to the public.

Outline This thesis is constructed as follows: Section 2 shortly introduces the topic. Section 3 illuminates the situation of colorectal cancer (CRC) in Germany and introduces the topic of familial clustering of diagnoses. A specific study called `family study` is introduced in section 4. The two data sets used in this thesis are there described in detail. The Bayesian risk score is proposed in section 5, also presenting methods to estimate it. The *in silico* study is introduced in this section as well. Results of the simulation study and the application to the `family study` are given in section 6. Section 7 contains the discussion of the proposed method of the Bayesian risk score, of the `family study` and of the results gained from the *in silico* and the `family study`. A short outlook is presented in section 8. Further information and details to the results are given in the appendix (section A and section B).

Contributing manuscripts An essence of this work is already written down in the following article that introduces the Bayesian risk score and its estimation via grid search as a precursor of new screening methods for familial CRC risk. It presents the results of the simulation study and the application to the `family study`:

- A. Rieger and U. Mansmann: *Bayesian Prediction of Being a Colorectal Cancer Risk Family*. Submitted at Biometrical Journal. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2017.
  Anna Rieger developed the method of the Bayesian posterior risk score under

supervision of Ulrich Mansmann. Anna Rieger conducted all analyses and simulations as well as the data preparation of the `family study`. Anna Rieger made the R code and the analyses reproducible and drafted the manuscript. Ulrich Mansmann contributed to the manuscript.

A short section in this thesis is devoted to the estimation of the parameters used in the Bayesian risk score by means of an expectation-maximisation (EM) algorithm. An accelerated version of the EM algorithm needed for the estimation of the Bayesian risk score is introduced in the following article:

- A. Engelhardt, A. Rieger, A. Tresch and U. Mansmann: *Efficient Maximum Likelihood Estimation for Pedigree Data with the Sum-Product Algorithm*. Accepted at Human Heredity. Karger Publishers, Basel, Switzerland, 2017. [18]
  Alexander Engelhardt developed the accelerated EM algorithm and drafted the manuscript. Anna Rieger conducted the data preparation and contributed marginally to the method. Anna Rieger, Achim Tresch and Ulrich Mansmann contributed to the manuscript.

Software The code and data to reproduce the results of this thesis can be found on LRZ Gitlab: https://gitlab.lrz.de/AnnaRieger/BayesianPosteriorScore.git
It has been written using R version 3.3.3 [64] on platform `x86_64-pc-linux-gnu` (64-bit). The versions of the attached packages were: ROCR version 1.0-7 [68], gplots version 3.0.1 [77], survival version 2.40-1 [70], mgcv version 1.8-16 [82], nlme version 3.1-131 [59] and e1071 version 1.6-7 [46].

2
Introduction

Colorectal cancer (CRC) is one of the most prevalent cancers in Germany with about 60 000 new cases each year [35]. People are getting older and that leads to an increased cumulative risk for cancer [27]. CRC is a big challenge for the health care system [66]. So, studying the prevalence of cancer and especially the prevention of cancer is a research field highly relevant for society.

Familial clustering of CRC diagnoses is not uncommon (about a quarter of cases, see section 3.2), but only a small proportion is hereditary, i. e. caused by genes. The remaining part has to date unknown causes and is called "familial CRC" instead of "hereditary CRC" like syndromes such as "hereditary nonpolyposis colorectal cancer" (HNPCC) or "familial adenomatous polyposis" (FAP). It has been already shown with data from Sweden that offspring of a CRC case have an approximately doubled to tripled CRC risk [5]. Several other studies and meta-analyses came to the same result (see section 3.2). Persons at familial risk get CRC approximately ten years earlier than persons without familial burden (e. g. [10], see section 3.2). Persons with risk for familial CRC should be identified to give them a tailored CRC screening. Screening can prevent cancer diagnoses and therefore decrease prevalence, as precancerous lesions (adenoma) are usually removed during screening colonoscopy [23, 41]. On the other hand, persons at *no* risk should not be "overscreened". Colonoscopy, the method of choice for CRC screening, is perceived as cumbersome and stressful. So, redundant application should be prevented.

Hints for familial CRC risk are very unspecific. They include early age of onset or familial clustering of CRC cases as already mentioned. Furthermore, it is not possible to do genetic testing or something similar to detect "risk families", as the causing genes are still unknown. Familial risk may also be caused by the family's common lifestyle, but current research found no association [84]. So far, questionnaires are used to identify persons and families with risk for familial clustering of CRC. But they are mainly constructed to identify genetic burden [52, 65], which does not coincide with familial burden as pointed out before. In this thesis, a new tool to identify "risk families" is developed. For that purpose, a Bayesian risk score is developed by means of a Bayesian posterior approach that is based on the family's history of CRC and the family tree. It has several advantages, e. g. it takes into account the size of the family,

which is not always given in the questionnaires used so far. E. g., if a certain number of diagnoses with the relatives needs to be gained to get a score point, small families are discriminated against. It needs the full information on the family tree and the family's CRC history. The Bayesian risk score gives quite stable results in terms of discrimination between "risk families" and "normal" families (see section 6).

A hereditary disease is often described by a so called genetic model. It has two components: penetrance and hereditary mechanism. Penetrance describes how fast the disease associated to the gene breaks out. The hereditary mechanism describes how and therefore how probably the disease is inherited by offspring. The genetic model is not known for familial CRC. If it would be known, it would be possible, together with the family's cancer history and the family tree, to calculate a so called posterior risk of being a "risk family" in a Bayesian way. The unknown genetic model is mimicked by different simple statistical approaches within the proposed method of Bayesian risk score that cover a range of plausible biological settings (see section 5).

A simulation study ("*in silico*") is done to test the proposed method in a simulated general population and in a simulated selected population of CRC patients and their families (see section 5.4). An application to a real data set is done using data of the so called "`family study`" (see section 4). In all settings, the newly proposed Bayesian risk score is compared to a questionnaire of the "Netzwerk gegen Darmkrebs e.V." (NACRC), where coarse knowledge about the family's CRC history is sufficient as for other questionnaires regarding genetic burden [52, 65] (see section 3.2.3). With both instruments, "risk families" can be identified in a general population and specific screening of the family members could be offered (see section 6).

3
Colorectal Cancer

This chapter provides informations about the situation of colorectal cancer in Germany, the screening guidelines as well as familial clustering of colorectal cancer and its possible causes.

3.1 Colorectal Cancer in Germany

Colorectal cancer (CRC) is one of the most prevalent cancer diseases in Germany [35]. About 60 000 new CRC cases are registered each year by Robert-Koch-Institut [35]. About 25 000 patients with CRC die each year in Germany [35]. The risk for CRC increases with age, considerably beginning with the age of 50 [35, 41, 42]. Men are more often and earlier affected than women [35, 42]. The mean age of diagnosis is 71 years for men and 75 years for women in Germany [35]. Most CRC cases are adenocarcinoma [35]. The progress from adenoma to cancer lasts about ten years [41], enough time to prevent colorectal cancer by screening.

3.1.1 Screening for Colorectal Cancer

In Germany, the population in statutory health insurance (about 87 % in 2011 [54, 55]) can use screening programmes for several cancers. The screening programme for CRC

starts at the age of 50 with a yearly guaiac fecal occult blood test (gFOBT) [41]. Persons aged 55 and older can use a screening colonoscopy which can be repeated after ten years, if the first examination was unsuspicious [41]. During colonoscopy, discovered precursors of CRC like adenoma can be removed. The alternative for colonoscopy is a biannual gFOBT [6, 41]. The colonoscopy participation rate is low in Bavaria (about 15 % to 20 % [42]).

Guideline recommendations are a little bit different [23, 41]: An upper age bound for screening cannot be given since people are getting older and older. The colonoscopy should be the standard procedure for CRC screening, i. e. the gFOBT is not recommended by the guideline. A sigmoidoscopy in combination with a gFOBT is recommended as alternative instead. A yearly gFOBT should be applied to persons with average risk. A colonoscopy should follow a positive gFOBT result. It is recommended to remove the adenomas found during colonoscopy [23, 41] to interrupt the adenoma-carcinoma sequence [23, 81].

A bunch of questionnaires exists to identify patients at raised risk for CRC. However, the predictive value of those questionnaires is low and should not be used in practice for screening for CRC [41]. Instead colonoscopy or alternatives should be chosen as recommended by the guideline [23]. The use of questionnaires for screening for *familial burden* of CRC is untouched by this recommendation.

3.1.2 Comparison with Global Data

CRC is the third most common cancer in Germany as well as worldwide [19, 35]. It is the second most common cancer in women after breast cancer and the third most common cancer in men after prostate and lung cancer [35]. Other sources [9] see it on rank four for men adding stomach cancer and rank three for women adding cancer of the cervix uteri. In Germany, about 13 % of all new cancer cases are CRC. Worldwide, about 9 % to 10 % of all cancers are CRC [9, 19]. The percentages are both in Germany and globally higher in men than in women [19, 35]. The rates are much higher in developed countries than in developing countries [9, 19], indicating that lifestyle has an impact on disease outbreak.

3.1.3 Risk Factors for Colorectal Cancer

Some risk factors, i. e. the lifestyle, can be influenced and changed. These include tobacco consumption [35, 41, 80], physical inactivity [9, 35, 41], obesity [9, 35] and alcohol intake [9, 35]. Unhealthy eating habits like red meat and low fibre content are considered as risk factors, too [9, 35, 41]. These influencable risk factors are in parts correlated to each other. A study among spouses unveiled a correlation between shared environment, i. e. lifestyle, and developing cancer, although spouses have usually different genetic background [75]. A study among adopted persons supports the opposite suspicion, that genetics have more impact than lifestyle [84]. Some chronic inflammatory bowel diseases like Colitis ulcerosa or Crohn's disease increase the risk for CRC as well [9, 35, 41].

Non-influencable risk factors are age as for most cancers [9, 41] and male sex [9, 35, 41]. Males seem to develop colorectal neoplasia (i. e. both adenomas and cancer) at an earlier age than women [42]. A retrospective study [39] approves male sex and increasing age as risk factor for detecting adenomas during colonoscopy. Another risk factor is familial burden [9, 35, 41]. Based on the colonoscopy screening programme in Germany, a higher impact of non-influencable male sex and influencable tobacco consumption on the prevalence of colorectal neoplasia like adenoma is revealed compared to the impact of non-influencable family history [29].

3.2 Familial Clustering of Colorectal Cancer

There are several studies about cancer as "familial disease" [27, 76]. The variety of rates of familial clusters is high among the specific organs and sites affected by cancer [27]. Colorectal cancer is found to have the third-highest familial clustering proportion in Sweden [27], with breast cancer having a similar one of about 13 %. Prostate cancer has the highest familial proportion of around 20 % [27]. The main constellation of affected family members found in that study is "parent and offspring" (about 80 % of familial clusters) [27]. A study conducted in general practices in Scotland in persons aged 30 up to 65 reveals a proportion of about 20 % with a family history of colorectal, breast or ovarian cancer [76]. Out of the persons reporting CRC family history, about 5 % meet the national guidelines for genetic counselling [76].

The familial burden partly arises from known genes. Approximately 5 % of all CRC cases are caused by hereditary cancers like hereditary non-polyposis colorectal cancer (HNPCC or Lynch syndrome) or familial adenomatous polyposis (FAP) [41]. The mutations and genetic mechanisms are already known for these syndromes. Concerned persons get usually special screening and genetic counselling [23, 41]. Some other 10 % up to 20 % of all CRC case occur in familial clusters [9, 21, 27, 45, 67, 72]. Familial risk is unspecific, because no genetic background is known so far. It could be caused by another, yet undiscovered single gene or a combination of several genes each with low penetrance or with only recessive inheritance. Lifestyle can be another cause (see section 3.1.3). Of course, clustering of sporadic cases is possible, too. Persons in familial clusters are usually defined in the literature as CRC cases with at least one first-degree relative diagnosed with CRC, too. It is assumed that persons in a familial cluster have an increased risk of disease and/or an earlier mean age of onset. Several studies show evidence for these assumptions:

Offspring of a CRC case have an approximately doubled CRC risk as shown in various population based studies as well as case-control studies [1, 5, 21, 26, 37, 69]. A doubled risk for all relatives of sporadic cancer was shown in various case-control studies already back in the 1980s and 1990s [4, 8, 32]. Some meta-analyses show that the relative risk of persons with a first-degree relative with CRC is about 2 [14, 33]. The risk for offspring is higher, the more relatives are affected [5, 37]. Also persons with relatives with adenoma, i. e. precursors of CRC, have a higher risk of getting CRC [31]. CRC cases

with a family history of CRC have an earlier mean age of onset [38]. The risk is higher, the lower the age of diagnosis in the ancestors is [21, 37, 38, 67, 69]. The other way round, the risk of having an ancestor with CRC is higher, if the patient is younger than 45 years [32]. Persons at familial risk get CRC approximately ten years earlier than persons without familial burden [10]. There are also data regarding clustering of CRC diagnoses from Germany published from two large studies [60, 79]. They use data sources from different regional parts of Germany. The enrolment of the study participants is similar to that of the `family study` (see section 4.1). One study is designed as a case-control study [79]. Both studies rely on the statements of the participating patients regarding family history of CRC. There are 13.5 % respectively 10.2 % among the cases respectively controls with a family history in the case-control study [79]. There, 7.2 % report at least one first-degree relative with CRC and 1.2 % report diagnoses before the age of 50 [60].

German cancer guidelines [23] take up the increased risk for familial clusters. For persons with first-degree relatives with CRC it is recommended to be screened by colonoscopy ten years before the age at diagnosis of the youngest CRC case in the family, at latest reaching 45 years. However, this is not yet implemented in the statutory health care screening programme in Germany. Starting with age 50, screening onset comes too late for persons with familial risk. Instead, earlier screening onset is needed to have the same chance to prevent cancer as in people without familial burden.

Families with a higher risk because of expected clustering of CRC cases should be identified for specific screening to prohibit further CRC cases. Persons with a family history of CRC wish to be informed by their general practitioner or their health insurance about their familial risk and disease prevention [61].

Genes causing *hereditary* colorectal cancer are known. No causative genes are known so far for the familial cases and therefore it is really difficult to detect familial risk. To make things worse, hints for familial risk are given quite indirectly. In both the hereditary and the familial case, familial clustering of CRC cases and earlier age of onset are used to define some sort of increased risk.

3.2.1 *Excursus*: Genetics

Heredity is the passing of properties to the descendants [28]. The information about the properties is located on chromosomes, which are situated in the cell nucleus [66]. Chromosomes are build from deoxyribonucleic acid (DNA) [49, 66]. Human cells contain every chromosome two times except the sex chromosome of the males [66]. Father and mother bequeath each every chromosome once [49]. Genes are locations on the DNA strand [66]. Those gene pairs are homologous with respect to form, structure, and sequence of genes, but not mandatory exactly the same. Several variants of genes exist, the so called alleles [28]. Every human can have at most two different alleles of every gene: one from the father and one from the mother. Those build the genotype of a human. The phenotype is then the realisation of genes in life. It is also influenced by environment [28, 49, 66].

An organism is called heterozygous, if it has two alleles of a gene [49]. If both alleles

are identical, the organism is called homozygous [49]. It will develop the property associated to the respective gene with certainty.

A gene variant is called dominant, if it will be developed also in heterozygous organisms [66]. One single (dominant) allele then defines the phenotype, i. e. the development of the related property. Regarding family trees, the related property or disease is pronounced in every generation, if the penetrance is high. If a gene variant is a recessive one, the organism needs two identical alleles to develop the related property [66]. Heterozygous organisms will not develop the property that is coded in this gene. However, they can carry and therefore inherit the allele to offspring. In a family tree regarding recessive inherited diseases, only the few homozygous persons are ill. A mixture form of dominant and recessive are codominant alleles that will lead to developing both properties [28, 66]. An example for a codominant property is the blood group AB.

Penetrance gives the percentage, how strong the characteristic of a property is [66]. It can reach up to 100 % and can vary with age. If the penetrance is not complete, i. e. less than 100 %, some generations in a family tree can be bypassed.

3.2.2 Hereditary Cases

Regarding hereditary CRC, a bunch of diseases is known to be caused by genetic defects with dominant inheritance [37]. Familial adenomatous polyposis (FAP) and hereditary non-polyposis colon cancer (HNPCC) are examples. The former is caused by a mutated copy of the adenomatous polyposis gene (APC) [37]. The latter can have several different causes, but it is always a damage on DNA mismatch repair genes (MLH1, MSH2, PMS2, and MSH6) [37].

HNPCC has a high penetrance of about 80 % [74]. HNPCC patients develop CRC already in early years [37, 41]. It is the most common hereditary CRC disease [9]. Patients may develop also cancer at other typical locations than colon/rectum like ovary, stomach, small bowel and urinary tract [37]. It is possible to test for HNPCC by means of microsatellite instability (MSI) [7]. Amsterdam criteria can be used to identify persons at risk for HNPCC. They were introduced in the year 1990 and revised 1998 [74]. An alternative to the Amsterdam criteria are the Bethesda criteria from 1997 [65], they were revised in 2004 [73].

FAP patients develop hundreds to thousands of adenoma in early years of life, which progress almost certainly to cancer, if they are not removed [37]. The responsible gene APC is a tumour suppressor. FAP patients have a mutation [15]. The disease breaks out if another lesion arises there by spontaneous mutations during “normal” cell division on the not-damaged allele. This phenomenon is called “two-hit theory” [15]. It is possible to test for lesions in APC and to use those test results for genetic counselling [57]. There exist also several weaker forms of FAP [37]. One APC variant is over-present in Ashkenazi Jews [37, 44].

3.2.3 Identification of Familial Clusters

Testing for genetic damages and gene variants is possible for hereditary CRC as the causing genes are known. They are unknown for unspecific familial CRC. The gene or genes are not discovered yet, but it is assumed, that penetrance is low to moderate [37]. Nevertheless, it is necessary to have a tool to identify the families at risk for unspecific familial CRC. Then, screening recommendations and rules of statutory health insurances can be adjusted to provide risk families specific screening. The Amsterdam and Bethesda criteria that were developed to identify HNPCC families, are used in several questionnaires and serve to identify families at "hereditary" or "increased familial" risk [65]. Another in practice often used criterion for familial CRC risk is having a first-degree relative with CRC. This very simple rule has a positive predictive value of 20 %, as approximately 20 % of all CRC cases show familial clustering [45, 72]. Another study in France reveals an amount of approximately 10 % with a familial history of CRC [4].

NACRC Questionnaire The German society "Netzwerk gegen Darmkrebs e.V." (Network against colorectal cancer (registered society) – here "NACRC" is shortly used) [52] developed a questionnaire specifically to identify familial CRC risk. It consists of only four yes/no questions regarding familial risk of CRC [36]. The four questions are in detail:

1. Do you have a first-degree relative diagnosed with CRC?
2. Do you have a first-degree relative diagnosed with CRC before 50 years of age?
3. Do you have a first-degree relative diagnosed with intestinal polyp before 50 years of age?
4. Do you have at least three first-degree relatives diagnosed with one of the following cancers: colorectal cancer, stomach cancer, endometrial cancer?

Question 1 and 2 and question 1 and 4 overlap at least partly. If no question is answered with "yes", the person respectively the family belongs to normal population. If the first question is answered with "yes", the person is susceptible for familial risk. If one or more of the questions 2 to 4 are answered with "yes", the person or family is at risk for hereditary forms of CRC [36]. The more questions are answered with "yes", the higher is the risk for CRC. Therefore, the number of "yes" answers serves as risk score of the questionnaire in this thesis. The higher the score, the more probable the regarded family is a risk family. Coarse knowledge about the family's history of CRC is sufficient to fill in this questionnaire.

The NACRC questionnaire was validated on a sufficiently large sample with respect to consistent answers (reliability) [58]. A questionnaire was sent to persons of 30 to 54 years and insured in one statutory health insurance (Betriebskrankenkasse BKK) in Essen, Germany. Persons then identified to be at increased risk were followed up by a second questionnaire that was validated by a telephone interview.

Compared with the first questionnaire, question 1 was answered marginally less often with "yes" in the telephone interview. Questions 3 and 4 were answered with "yes"

with quite stable rates. Question 3 was answered much more often with "don't know" in the first questionnaire [58]. This is assumed to be due to a more detailed explanation on the telephone. Question 2 was not analysed in the study.

The authors emphasize that the questionnaire is not meant to replace anamnesis [58], but should only support the medical consulting process.

A small study validated the accordance of the patients' answers to the NACRC questionnaire with the answers of their general practitioner [43]. Again, question 3 was problematic to answer by both the general practitioner and the patients. Sensitivity and specificity of question 1 was desirably high, taking the answers of the general practitioner as gold standard [43].

NCI Questionnaire There is a risk prediction tool for colorectal cancer provided by the US-American National Cancer Institute (www.cancer.gov) [20, 56]. It gives validated risks for CRC for several ethnic groups and for a range of relevant age starting with age 50. Several risk associated and preventive factors are assessed using case-control studies [20]. They are mainly described in section 3.1.3. A single question relates to the history of CRC in first-degree relatives. The risk model was validated using a big study on health and diet (National Institutes of Health – American Association of Retired Persons (AARP) diet and health study) [56] showing slightly weaker associations between the risk factors and CRC. Cancer patients were identified using probability linkage to cancer registries [56]. The number of observed and expected patients in the single risk factor categories was mostly similar to the original data base the risk prediction tool was based on. The AUC was around 0.6 for both sexes [56].

4
The “Family Study”

This chapter contains explanations of the study setting and gives an overview over the data.

4.1 Study Setting

This cross-sectional study was conducted in the catchment area of the Munich Cancer Registry (MCR) from September 2012 until June 2014. The full name was in German “Familien schützen und stärken – Umgang mit familiärem Darmkrebs”, which roughly translates as “To Protect and to Strengthen Families – Handling of Familial Colorectal Cancer”. The shortly named `family study` was approved by the ethical review committee of the Ludwig-Maximilians-Universität München.

The inclusion criterion was newly diagnosed CRC in patients younger than 70 living in the catchment area of MCR. Diagnoses from 1st January 2012 onwards were defined as “newly diagnosed”. Informed consent of the patients was required. Patients should have been recruited through general practitioners or other treating physicians. Participating patients were asked to give a family tree containing parents, siblings and children with their name and address. Relatives were asked to also fill in family trees. This way the family trees were expected to grow and to contain more detailed information. The names and addresses of the relatives were anonymised. The family’s CRC history was then provided by an anonymous record linkage [50] between the MCR data base and the reported characteristics of the family members. Participating patients were also asked to fill out the NACRC questionnaire (see section 3.2.3).

The aim of this study was not to identify specific risk factors or the hereditary mechanism, but an identification of “risk families” and estimation of the prevalence of “risk families” in an incident population of CRC cases.

In a population of incident cases, the probability of uncovering familial structures was expected to be higher than in the general population. The transferability from the incident study population to the general population needed to be analysed. This analysis was done by means of a simulation study (see section 6.1).

Given the known incidences from MCR and a compliance of 70 %, 1820 patients younger than 70 years were expected to get enrolled in two years. Because the recruiting phase was extended to two and a half years, 2275 patients were expected.

Given a proportion of 20 % for all CRC cases to occur in familial clusters [9, 21, 27, 45, 67, 72], 455 familial structures were expected.

4.2 Overview Of the Data Set

Overall, 792 patients were recruited. This was already below the expected number of 2275 patients. Using the inclusion criteria of a diagnosis in 2012 or later and age at diagnosis of 70 at maximum, only 288 patients remained. This corresponded to a participation rate of 13 %. Due to this poor participation rate, it was decided to use all patients with their families for analysis, even if the diagnosis was older than 2012 and even if the patients were older than 70 years at time point of diagnosis.

Those 792 patients reported 4296 family members (including the recruited patients themselves). In 181 cases, the “family” consisted only of the recruited patient. Out of the 181 families with only one member, 150 were living in the catchment area of MCR. After exclusion of those one-person “families”, 611 families remained for an analysis with meaningful information in terms of a more or less extensive family tree (“complete data set”). An anonymous record linkage [50] between the MCR data base and the reported characteristics of the family members (names, sex, address) was only possible for family members living in the catchment area of MCR at the time point of a possible CRC diagnosis. All reported family members were living in this area in 73 families with 461 members (“local family data set”). A total of 538 families with 3654 members had relatives living outside the MCR catchment area. For these persons no cancer history could have been derived from a record linkage with the MCR data. A flow chart of the included families and their members can be found in figure 4.1.

Non-obligatory reporting of cancer diagnoses to the MCR as well as delayed reporting were limitations of the record linkage. The presented data were thus an intermediate version. Both reasons may have led to false-negatives in the data. False-positives were also possible, as the anonymous record linkage used an error tolerant method [50].

If a person had more than one CRC diagnosis, the earliest was used. During data collection, missing values were filled in by a subsequent telephone interview if possible. After the record linkage, values still missing for CRC cases were filled in from the MCR data base if possible. If the age at diagnosis was even then still missing, age at death or age at enrolment was used. The age of persons with missing values in e. g. birth year or year of death, was fixed at 100. Persons were set to “dead”, if the living status was unknown and the age was 100 and more. Persons were set to “dead” without respect to

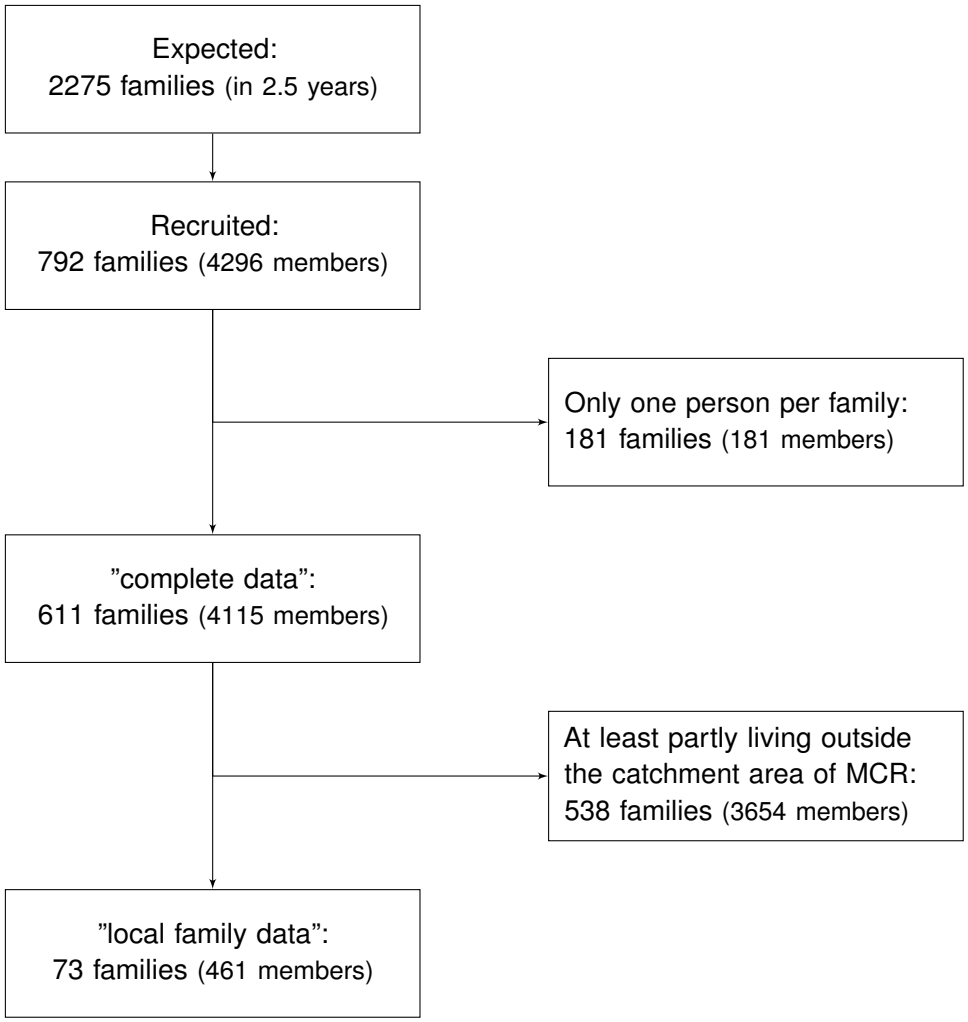

Figure 4.1: Flowchart of the `family study` depicting the numbers of families and their members (including recruited patients) in each data set.

living status, if the age was 110 and more.

The proportion of men among the recruited patients was 59 %, according to values known from Germany (54 % [35]) and Upper Bavaria (53 % [71]). The `family study` was not representative with respect to age of the recruited patients, as they were younger than CRC patients in the general German population: For women, the mean age of the recruited patients was 64.74 years (median 66.5 years). The mean age for women with newly diagnosed CRC in Germany was 75 in 2010 [35]. The mean age for recruited male patients in the `family study` was 66.45 years (median 68 years), whereas the mean age at diagnosis was 71 years in Germany 2010 [35].

The mean number of family members in the original 792 families of the `family study` was 5.42 (standard deviation 3.55, median 5) including the recruited patients. The mean number of family members for families from the MCR catchment area was 2.74 (standard deviation 3.00, median 1).

4.3 Descriptive Analysis of the Complete Data Set

Sex and Age of the CRC Patients After exclusion of the "families" with only one member, there were more men than women among the remaining 611 recruited patients (59.08 % versus 40.92 %). Mean age of the male patients was 66.16 years (median 67 years, standard deviation 9.11), mean age of the female patients was 63.98 years (median 66 years, standard deviation 13.62). However, there were 77 respectively 63 missing values (21.33 % respectively 25.20 %).

The participating patients reported partly also CRC diagnoses among their relatives. In total, there were 669 CRCs in 611 families reported. Exactly 500 CRC cases were found in the MCR data base. By means of the anonymous record linkage, 473 CRC diagnoses could have been verified. That means, additional 27 cases were found in the MCR data base, which were not reported by the recruited patients. On the other hand, 169 reported CRCs were not found in the MCR data base.

Regarding only the verified CRC cases, 60.68 % were men. Age structure was not changing in men nor in women. There were 149 families without any verified CRC diagnosis. Nine families had two verified CRC cases and one family had three diagnoses verified. Mean age at CRC diagnosis was 72.19 years among the persons in those ten families with multiple CRC cases (median 71 years, standard deviation 8.66).

Family Structure The numbers of family members ranged from 2 up to 22. The mean number of family members was 6.74 including the recruited patient (median 6 members, standard deviation 2.98). The relatives were equally distributed regarding their sex.

There were several complex family trees in the study. However, some family parts were sometimes unfortunately living outside the catchment area of the MCR. So, no verification of CRC diagnoses could have been done. There were also some simple family trees in the study. The family trees of four randomly chosen families of the `family study` are shown in figure 4.2 to visualise the structure of the data set. The

respective Lexis diagrams are in figure 4.3. A Lexis diagram represents the life course of members of a cohort (here: of a family) and out-standing events like birth or death. The age is usually plotted against calendar time. A Lexis diagram can only be filled, if there are no missing values in age and year of diagnosis or death.

The recruited patient always had ID 1001 within each family. His or her parents and the persons of that generation had numbers in the 900s. His or her siblings and other persons in this generation had other numbers in the 1000s. The children generation had numbers in the 1100s and so on. First-degree relationships are marked by connecting lines in figure 4.2.

For example, family 129 in figure 4.2 consisted of five persons: the recruited patient (ID 1001) and his daughter (ID 1104), his brother (ID 1003) and his parents (father ID 905 and mother ID 902). Addresses outside the catchment area of the MCR were stated for his parents and his brother. His mother was reported to have CRC as well, but due to her residence outside the catchment area of the MCR, it was impossible to verify this diagnosis. The mother received her CRC diagnosis at the age of 75, the originally recruited patient, i. e. her son, was diagnosed at the age of 67. The daughter turned 40 in 2014. Unfortunately, the birth year of the father was missing. So, in the Lexis diagram (fig. 4.3) no line could be drawn for him. No family member was reported to be dead as e. g. in family 113.

NACRC Questionnaire The NACRC questionnaire (see section 3.2.3) was also given to the recruited patients. They should have filled it in without their own CRC diagnosis. I. e. question 1 should only be "yes", if there was at least one other CRC diagnosis among the first-degree relatives. In some families, there was more than one questionnaire returned. The one with less missing values was kept. The statements were verified using the record linkage with the MCR data base. Question 1 was set to "yes", if there was a verified CRC in the family. Question 2 was handled analogously, if there was a verified CRC and the person was younger than 50 at diagnosis. The questions 3 (adenoma in relatives) and 4 (related cancers in first-degree relatives) were not verified using the record linkage.

Not all families returned the questionnaire and not all filled it in completely. The question answered most often was question 1 with 597 out of 611 possible replies (97.71 %). This was followed by question 2 with 596 out of 611 possible replies (97.55 %) and question 4 with 595 out of 611 possible replies (97.38 %). Question 3 was already in the validation often answered with "don't know" [58] (see section 3.2.3). As the possibility of answering "don't know" was not given in the `family study`, which was only a "yes"/"no" decision, more people skipped this question and produced missing values. So, only 578 out of 611 possible replies (94.60 %) were given for the difficult question 3.

Regarding the proportion of "yes" answers, question 1 was also the one with the highest proportion, namely 17.42 % (104 out of 597 answers). Question 2 was answered very rarely with "yes", i. e. only in 1.68 % (10 out of 596 answers). Question 3 concerning

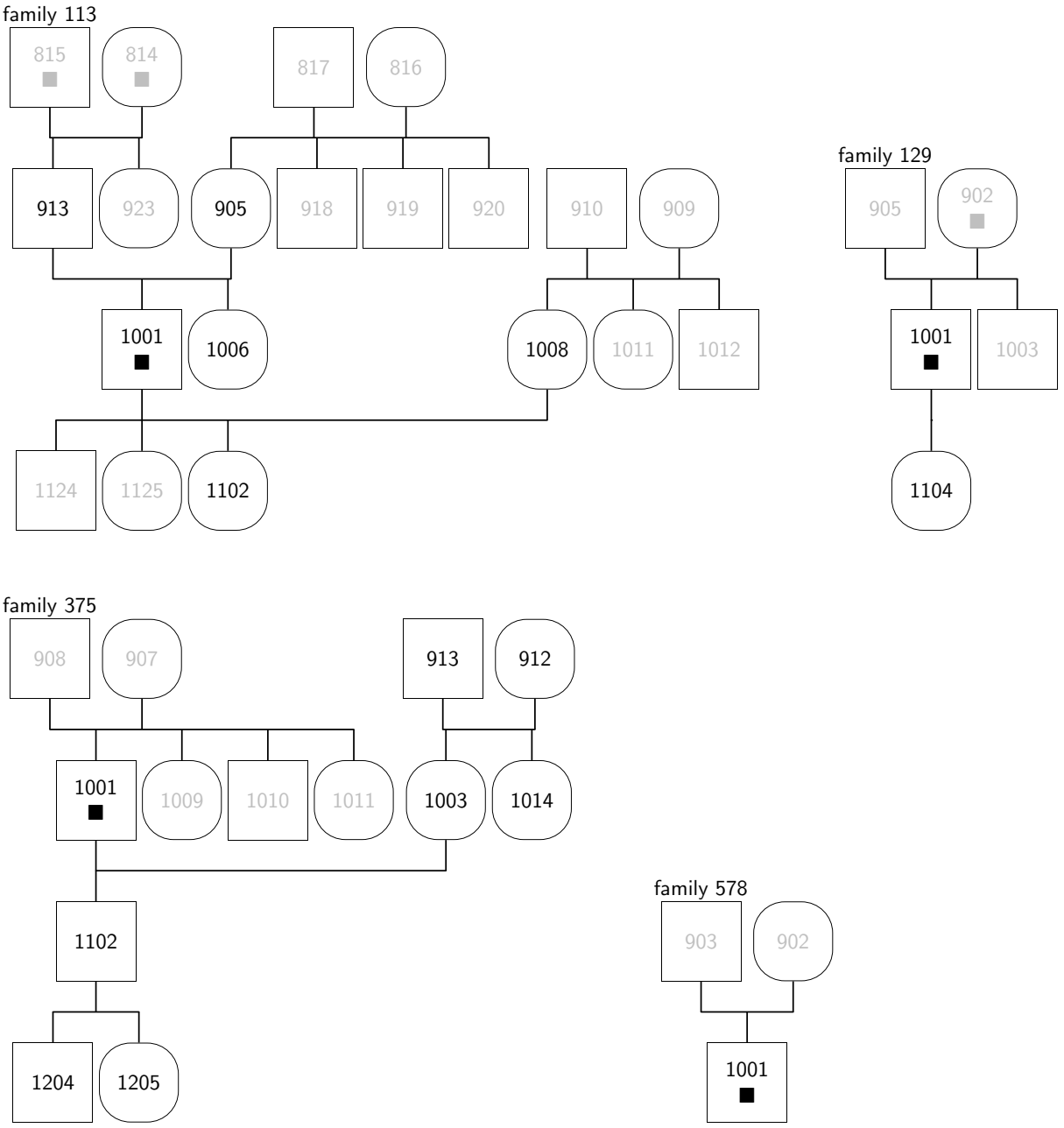

Figure 4.2: Family trees of some random families from the `family study`. Squares denote males, circles denote females. CRC patients are marked by a filled square symbol. Grey colour or grey ID numbers mark residence outside the catchment area of the MCR. First-degree relationships are marked by connecting lines.

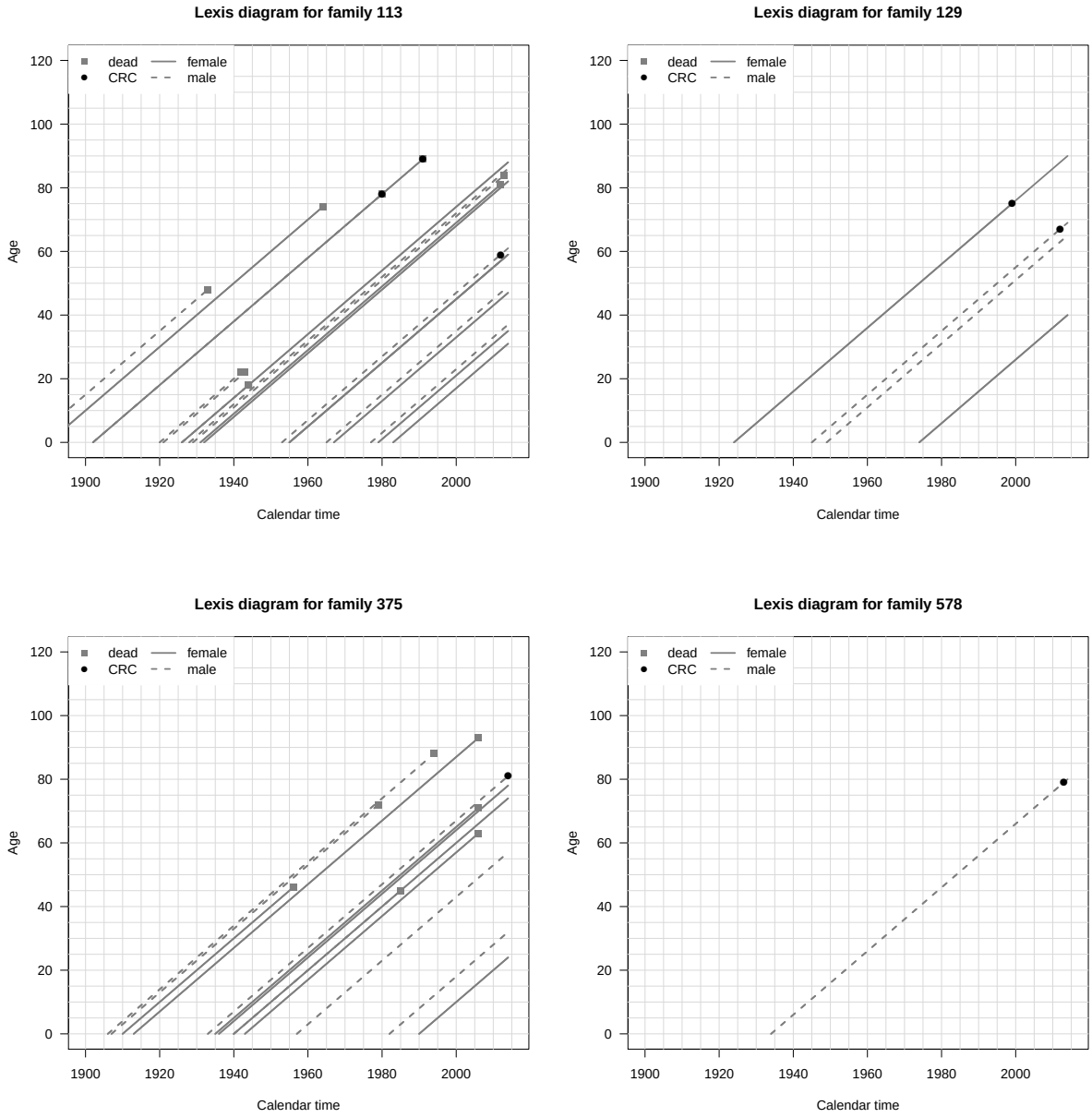

Figure 4.3: Lexis diagrams of some random families from the `family study`. No lines are drawn for persons for whom it was not possible to get any information about age and/or calendar time.

adenoma in relatives younger than 50 was answered in 6.40 % with "yes" (37 out of 578 answers). Question 4 was the question with the second-most "yes" answers with 8.07 % (48 out of 595 answers).

A complete verification of the family members and their CRC diagnoses was not feasible for the complete data set. This may have been the reason why only 7 out of those 104 (6.73 %), who stated "yes" for question 1, could have been verified, i. e. there should have been another CRC diagnosis among the first-degree relatives. The other way round, 491 out of those 493 (99.59 %), who stated "no" for question 1, could have been verified. Additionally, there were 14 answers missing. For 2 families, another CRC case could have been found in the MCR data base. All 586 "no" statements regarding question 2 could have been verified. That means, there was no other patient younger than 50 at CRC diagnosis found in the MCR data base for any family of the local family data set. However, also for the ten families stating "yes", no other patient younger than 50 at CRC diagnosis could have been found. Also for those families, no young CRC case was found in the MCR data base. Fifteen answers were missing.

4.4 Descriptive Analysis of the Local Family Data Set

The local family data set contained only those families, whose members lived all within the catchment area of MCR. With this data set, a verification by means of record linkage was meaningful, as only the inhabitants of this area and their cancer diagnoses were reported to MCR. A total of 73 families with 461 members remained for this analysis.

Sex and Age of the CRC Patients There were less women than men in the local family data set among the recruited patients (35 % versus 65 %). The male patients recruited were slightly younger in the local family data set than in the complete data set with a narrower standard deviation (mean 62.79 years, median 62.5 years, standard deviation 9.11). The mean of the female patients was comparable to the complete data set (62.12 years, median 66 years, standard deviation 13.32). The proportion of missing data was higher in the local family data set: 14 (30 %) respectively 8 (31 %) missing values.

As in the complete data set, there were more CRCs reported than patients recruited due to reporting of CRCs of relatives. Namely, there were 79 CRCs reported (73 recruited patients). There were 65 CRCs found in the data base of the MCR by means of the anonymous record linkage. In total, 56 CRCs could have been verified. I. e., 23 CRCs could not have been found by record linkage and there were nine additional CRCs. However, one needs to be aware of false-negative and -positive hits in the MCR data base (see section 4.2).

Given only the verified CRCs, the proportions of the sexes did not change. Mean age of males with verified CRC diagnoses was 63.47 years (median 63.5, standard deviation 9.34). The mean age of female persons with verified CRC diagnoses was 62.90 years. A median of 67.5 years gave a hint to a slightly skewed age distribution (standard deviation 12.72).

In 21 families, no CRC diagnosis could have been verified. There was one verified CRC diagnosis in 48 families and there were two verified CRCs in four families. As a clustering of CRCs was defined as "risk family", there were only four "risk families" in the local family data set. The age structure of those patients within the four families with more than one CRC was comparable to the complete data set (mean 71.62, median 71, standard deviation 7.33).

Family Structure The number of members per family in the local family data set ranged from 2 to 16. The mean number of members per family was 6.32 (including the recruited patient), with a median of 6 and a standard deviation of 2.92. Those numbers were well comparable to the complete data set.

No example family from figure 4.2 was in the local family data set, as there were at least two members living outside the catchment area of MCR in the families shown (marked by grey colour).

NACRC Questionnaire As mentioned before, the NACRC questionnaire (see section 3.2.3) was also given to the recruited patients. Also for the local family data set, the statements of the families were verified as far as possible using the record linkage with MCR.

Only two families did not return the questionnaire at all. Questions 1, 2 and 4 were all answered by 71 out of 73 families (97 % each). Question 3 regarding adenomas in relatives younger than 50 was answered by 69 of the 73 patients respectively their families (95 %).

A proportion of 14 % of the questionnaires was answered with "yes" in question 1 (10 out of 71). No one did answer to question 2 with "yes". In 12 % of the cases, question 3 got a "yes" answer (8 out of 69). Question 4 got "yes" answers in 10 % of the families (7 out of 71).

For the local family data set, a complete verification of the family members and their CRC diagnoses was theoretically feasible. So, the statements regarding question 1 and 2 could be verified. As mentioned above, the recruited patients should have filled in the questionnaire without counting their own CRC diagnosis. All but 1 of those 61, who stated "no" for question 1, could have been verified. Out of the ten patients respectively families, who stated "yes", i. e. there should have been another CRC diagnosis among the first-degree relatives, only three could have been verified. For seven of those ten families, no CRC diagnosis besides the recruited patient was found within the MCR data base. All 71 statements regarding question 2 could have been verified. That means, there was no other CRC diagnosis younger than 50 found in the MCR data base for any family of the local family data set. The two missing answers were evaluated to have also no other CRC case younger than 50 in the MCR data base.

5
Methods

This chapter contains a short overview of the approach of Bayesian inference in general. Furthermore, it provides information about the calculation of risk scores for being a "risk family" and how to estimate it from data by means of grid search and of expectation-maximisation (EM) algorithm.

5.1 Bayesian Inference in General

Bayesian inference uses observed data to estimate an underlying distribution of a parameter θ. First, a prior distribution of the parameter, $\mathbb{P}(\theta)$, is assumed. The observed data x are used to calculate the likelihood $L(\theta) = \mathbb{P}(x \mid \theta)$ as a function of the parameters to estimate. The prior is kind of "updated" by the likelihood, which is influenced by the observed data, to get the posterior function: The normalized product of those two functions represents the posterior distribution

$$\mathbb{P}(\theta \mid x) = \frac{\mathbb{P}(x \mid \theta) \cdot \mathbb{P}(\theta)}{\sum_{\theta^\star \in \Theta} \mathbb{P}(x \mid \theta^\star) \cdot \mathbb{P}(\theta^\star)}, \tag{5.1}$$

computed with Bayes' theorem [25]. There exist several options to get usable estimates like posterior mean, i. e. the expectation of the posterior, or posterior median [25]. The choice of the prior function can be influenced by prior information about the parameters to estimate. It can also be uninformative and flat. Priors are in a way subjective, but the choice of the prior co-determines the shape of the posterior.

Due to increasing computational power (velocity and memory), modern ways of Bayesian inference become available to users. This includes Gibbs sampling, Monte Carlo methods and many others [22].

5.2 Bayesian Risk Score

In this thesis, a method to calculate a Bayesian risk score of being a "risk family" is proposed. It is focused on the familial cases of CRC clustering since genetic testing already exists for the hereditary cases. The aim is to know if any member of a family is a risk carrier or not. A "risk family" is then defined as a family, where at least one member carries the risk property. This can be e. g. a combination of genes causing this familial risk or shared lifestyle or other properties (see section 3.2). For calculation, the risk carrier property is defined as a binary latent class variable r_i. It represents the property of being a familial risk carrier: $r_i = 1$ if person i is indeed a risk carrier, and $r_i = 0$ otherwise. The occurrence probability, i. e. the prevalence, of r_i is to be estimated. The prevalence is denoted by p_1. The property r_i is not known due to the mentioned causes above (see section 3.2).

The individual CRC risk is calculated using r_i. If a person is no risk carrier, i. e. $r_i = 0$, the CRC risk equals the general population based (age and sex specific) CRC incidence rate. This information can be extracted from a cancer registry or some other statistical data bases. If a person is a risk carrier with $r_i = 1$, the individual CRC risk is modified according to the prevalent genetic model. As mentioned before, r_i is unknown. So it is not possible to calculate the familial CRC risk directly. However, a posterior probability of being a risk family can be calculated by means of Bayesian methods. Furthermore, it is possible to calculate the posterior of a specific risk carrier property constellation. The family's CRC history and the family tree serve as observed data and therefore build up the likelihood together with the assumed penetrance model (see section 5.2.1). The prior is determined by knowledge from literature that forms the prevalence assumption *a priori* and the assumed inheritance mechanism (see section 5.2.2).

Transferring the calculation of the posterior (5.1) to the actual application of estimating r_i, the posterior looks like

$$\mathbb{P}\left(\boldsymbol{R} = \boldsymbol{r} \mid \boldsymbol{X}\right) = \frac{\mathbb{P}\left(\boldsymbol{X} \mid \boldsymbol{R} = \boldsymbol{r}\right)\mathbb{P}\left(\boldsymbol{R} = \boldsymbol{r}\right)}{\sum_{\boldsymbol{r}^\star \in \mathcal{R}} \mathbb{P}\left(\boldsymbol{X} \mid \boldsymbol{R} = \boldsymbol{r}^\star\right)\mathbb{P}\left(\boldsymbol{R} = \boldsymbol{r}^\star\right)}. \tag{5.2}$$

The vector $\boldsymbol{r}$ consists of the single r_i's of the members of a family. An example is $\boldsymbol{r} = (r_1, r_2, r_3, r_4, r_5)$ for a family of five.

The probability $\mathbb{P}(\boldsymbol{R} = \boldsymbol{r} \mid \boldsymbol{X})$ denotes the posterior. The Bayesian risk score is then

$$1 - \mathbb{P}(\boldsymbol{R} = \boldsymbol{0} \mid \boldsymbol{X}), \tag{5.3}$$

i. e. the posterior probability that at least one family member is risk carrier. This is how "risk family" was defined at the beginning of this section.

The data, i. e. the covariates, are denoted by $\boldsymbol{X}$. It consists of the family's CRC history and the family tree. The CRC history is represented by a status variable c_i, age t_i and sex s_i of the single family members. Age t_i takes the value of age at CRC diagnosis, age at death or actual cancer-free age, respectively. The status variable c_i indicates whether the person i has CRC at age t_i ($c_i = 1$) or not ($c_i = 0$). This status variable denotes censoring in normal survival models ($c_i = 1$ if the event is observed, $c_i = 0$ if censored). The family tree may be represented by variables like "generation", "position", "position of father", and "position of mother". The positions of father and mother should obviously be in the generation above that of the respective person.

The term $\mathbb{P}(\boldsymbol{X} \mid \boldsymbol{R})$ is the likelihood. As mentioned before, the penetrance model (see section 5.2.1) determines the shape of the likelihood.

The distribution $\mathbb{P}(\boldsymbol{R})$ is the prior that is determined from literature in form of the *a priori* assumed prevalence and the chosen inheritance mechanism (see section 5.2.2). The prior gives the probability of a certain risk carrier property constellation $\boldsymbol{R} = \boldsymbol{r}$ without any data observed, i. e. without any knowledge about the family members. Only their relationships, i. e. the family tree has to be known.

The product of likelihood and prior needs to be normalized to get a valid posterior and therefore to be divided by the sum over all possible risk carrier property constellations $\boldsymbol{r}^\star$. The set $\mathcal{R}$ consists of all those for the respective family.

The CRC disease risk can be interpreted as cancer-free survival. The likelihood contribution of a family j is in general in survival analysis

$$\mathbb{P}\left(\boldsymbol{X}_j \mid R\right) = L_j = \prod_{i=1}^{n_j} h(t_i)^{c_i} S(t_i). \tag{5.4}$$

The term h denotes the hazard function, c_i is the status variable as mentioned above ($c_i = 1$ is observed, i. e. CRC diagnosed; $c_i = 0$ otherwise), S is the survival function and the number of members in family j is denoted by n_j. If person i has not yet CRC diagnosed at time t_i, the observation is right-censored. Person i contributes with $S(t_i)$ to the likelihood. This equals the probability of getting CRC after t_i. If CRC is diagnosed at time t_i, then the "event" CRC is fully observed and the contribution is $f(t_i) = h(t_i) \cdot S(t_i)$, i. e. the probability of getting CRC at t_i. The penetrance model forms the functions h and S (see section 5.2.1). The complete likelihood for a set of N families is then the product of the likelihood contribution of those families, as the families are assumed to be independent from each other:

$$L = \prod_{j=1}^{N} L_j. \tag{5.5}$$

As the inheritance and the penetrance of genes causing familial CRC, i. e. the genetic model, are unknown so far, different statistical approaches for genetic mechanisms are formulated in the following sections that cover a range of plausible biological settings.

5.2.1 Penetrance Models

The term penetrance of a gene can be taken as the duration until a disease caused by a gene breaks out. From a statistical point of view, methods for survival analysis like the Weibull distribution or Cox proportional hazards model are appropriate. As mentioned above, those models provide the functions h and S. For simplicity, it is focused on the Weibull model with the two parameters k and λ_0 in the following. Let the standard hazard function h be denoted by $h(t_i) = k{\lambda_0}^k {t_i}^{k-1}$ and let the standard survival function S be $S(t_i) = \exp\left(-{\lambda_0}^k {t_i}^k\right)$.

The familial risk is modelled with a relative risk (RR) approach. That means that persons at familial risk, i. e. risk carriers, have an increased CRC risk compared to the general population [4, 5]. Using incidence rates (hazard rates) from a cancer registry, this can be done age and sex specific. There is also some evidence in the literature [10] that risk carriers get CRC earlier than people at normal risk. Therefore, a time shift in the hazard is also possible instead of RR.

Applying the RR setting, the hazard can be written as

$$h(t_i) = k{\lambda_0}^k {t_i}^{k-1} \cdot RR^{r_i}.$$

The risk respectively hazard of getting CRC is dependent on the risk carrier property r_i. The hazard equals that of the general population if $r_i = 0$, because then it reduces to the standard hazard function. The survival function S is in the RR setting then

$$S(t_i) = \exp\left(-{\lambda_0}^k {t_i}^k \cdot RR^{r_i}\right),$$

according to $S(t) = \exp\left(-\int_0^t h(u)\,\mathrm{d}u\right)$.

Sex is known as a strong modifier of CRC risk [35, 42]. The function h and therefore S are extended for sex risk by means of

$$h(t_i) = k{\lambda_0}^k {t_i}^{k-1} \cdot {RR_{\text{fam}}}^{r_i} \cdot {RR_{\text{sex}}}^{s_i}.$$

The variable s_i describes the sex of person i: $s_i = 0$ for women and $s_i = 1$ for men, as men are more likely to get a CRC diagnosis [35, 42]. The survival function is given by

$$S(t_i) = \exp\left(-{\lambda_0}^k {t_i}^k \cdot {RR_{\text{fam}}}^{r_i} \cdot {RR_{\text{sex}}}^{s_i}\right).$$

The parameters k and λ_0 should be chosen in a way, that they fit the observed incidence rates best, e. g. from a cancer registry. If the incidence rates are observed in age intervals, they can be interpolated.

If the shift variant should be modelled, the time shift τ needs to be added to the age t_i: The hazard is given by

$$h(t_i) = k{\lambda_0}^k \left(t_i + \tau r_i\right)^{k-1}$$

leading to the survival function

$$S(t_i) = \exp\left(-{\lambda_0}^k \cdot \left[\left(t_i + \tau r_i\right)^k - \left(\tau r_i\right)^k\right]\right).$$

Both also reduce to the standard functions for $r_i = 0$, i. e. the time shift has only impact on risk carriers with $r_i = 1$. The only constraint here is $k \neq 1$, since with $k = 1$, t_i cancels out in the hazard function. The time shift variant can also be extended for sex risk to adjust for the higher risk of men:

$$h(t_i) = k{\lambda_0}^k \left(t_i + \tau_{\text{fam}} r_i + \tau_{\text{sex}} s_i\right)^{k-1}$$

and

$$S(t_i) = \exp\left(-{\lambda_0}^k \cdot \left[\left(t_i + \tau_{\text{fam}} r_i + \tau_{\text{sex}} s_i\right)^k - \left(\tau_{\text{fam}} r_i + \tau_{\text{sex}} s_i\right)^k\right]\right).$$

Age t_i, sex s_i and CRC status c_i need to be known for each family member i as well as their relationships. For simplicity, it is focused on the RR setting in the following.

5.2.2 Hereditary Mechanisms

Beside penetrance, the hereditary mechanism determines a genetic model. A very simple mechanism is the so-called "complete risk transmission" (CTM). The children of parents at familial risk inherit the risk carrier property in every case. This points more to shared lifestyle or environment and dietary habits. As it is biologically implausible as genetic inheritance, CTM serves more for thinking and concept. Weakening the assumption of 100 % inheritance probability to get more biologically plausible settings, the "random risk transmission" (RTM) is introduced. The risk carrier property is transmitted with a specific probability $p^\star_{\text{inh}} < 100\,\%$. For simplicity, one global probability is used. Weakening the assumption of a global probability, "risks as random effects" (RRE) can be modelled. Two settings can be formulated: (i) one probability per family, but different probabilities for different families or (ii) one probability for each person such that the probabilities within a family are correlated. Since it is focused on the CTM and RTM setting here, the RRE scenario is not worked out in more detail in the following.

The term $p^\star_{\text{inh}}$ applies globally. For individual persons, p_{inh} is introduced. This probability depends on the risk carrier properties of the parents of a person and $p^\star_{\text{inh}}$. In the CTM setting, $p^\star_{\text{inh}} = 1$, and in the RTM setting $p^\star_{\text{inh}} \in [0;1[$. Let r_{father} and r_{mother} be the risk carrier properties of the father and the mother. According to simple probability

calculation rules, the individual probability of inheritance is

$$\begin{aligned} p_{\text{inh}} &= \mathbb{P}(R_i = 1) = \mathbb{P}(R_i = 1 \mid r_{\text{father}}, r_{\text{mother}}) \\ &= \mathbb{P}(R_i = 1 \mid r_{\text{father}}) + \mathbb{P}(R_i = 1 \mid r_{\text{mother}}) - \mathbb{P}(R_i = 1 \mid r_{\text{father}}, r_{\text{mother}}) \\ &= r_{\text{father}} \cdot p^\star_{\text{inh}} + r_{\text{mother}} \cdot p^\star_{\text{inh}} - \mathbb{P}(R_i = 1 \mid r_{\text{father}}) \cdot \mathbb{P}(R_i = 1 \mid r_{\text{mother}}) \\ &= r_{\text{father}} \cdot p^\star_{\text{inh}} + r_{\text{mother}} \cdot p^\star_{\text{inh}} - r_{\text{father}} \cdot p^\star_{\text{inh}} \cdot r_{\text{mother}} \cdot p^\star_{\text{inh}} \\ p_{\text{inh}} &= r_{\text{father}} \cdot p^\star_{\text{inh}} + r_{\text{mother}} \cdot p^\star_{\text{inh}} - r_{\text{father}} \cdot r_{\text{mother}} \cdot {p^\star_{\text{inh}}}^2 \end{aligned} \tag{5.6}$$

This can be generalised to

$$\mathbb{P}(R = r) = {p_{\text{inh}}}^r \cdot (1 - p_{\text{inh}})^{1-r}. \tag{5.7}$$

This formula is valid for offspring. For the "founders" of a family, i. e. the persons with no known parents in the family tree, the estimated prevalence of risk carriers p_1 is used instead of p_{inh}:

$$\mathbb{P}(R = r) = {p_1}^r \cdot (1 - p_1)^{1-r} \tag{5.8}$$

The prior of a family is then the product of the individual $\mathbb{P}(R = r)$ over all members. An example is given for a family of five:

$$\begin{aligned} \mathbb{P}(\boldsymbol{R} = (r_{\text{father}}, r_{\text{mother}}, r_{\text{child}_1}, r_{\text{child}_2}, r_{\text{child}_3})) &= \mathbb{P}(R_{\text{father}} = r_{\text{father}}) \cdot \mathbb{P}(R_{\text{mother}} = r_{\text{mother}}) \cdot \\ &\quad \mathbb{P}(R_{\text{child}_1} = r_{\text{child}_1}) \cdot \mathbb{P}(R_{\text{child}_2} = r_{\text{child}_2}) \cdot \\ &\quad \mathbb{P}(R_{\text{child}_3} = r_{\text{child}_3}) \\ &= {p_1}^{r_{\text{father}}} \cdot (1 - p_1)^{1-r_{\text{father}}} \cdot \\ &\quad {p_1}^{r_{\text{mother}}} \cdot (1 - p_1)^{1-r_{\text{mother}}} \cdot \\ &\quad {p_{\text{inh}}}^{r_{\text{child}_1}} \cdot (1 - p_{\text{inh}})^{1-r_{\text{child}_1}} \cdot \\ &\quad {p_{\text{inh}}}^{r_{\text{child}_2}} \cdot (1 - p_{\text{inh}})^{1-r_{\text{child}_2}} \cdot \\ &\quad {p_{\text{inh}}}^{r_{\text{child}_3}} \cdot (1 - p_{\text{inh}})^{1-r_{\text{child}_3}} \end{aligned}$$

Father and mother are founders here, as they have no parents within the family considered. The children are offspring of father and mother, so their p_{inh} depends on the risk carrier properties r_{father} and r_{mother}.

The possibility of spontaneous mutations to a risk carrier is excluded for simplicity.

5.2.3 Specifications Used Here

The parameters of the Weibull distribution need to be specified (see section 5.2.1). To do this, age and sex specific incidence rates for ICD-10-Codes C18–C20, i. e. colorectal cancer without anal cancer, are downloaded from MCR [71]. In table 5 of [71], age specific incidences are given for men and women in five-year age classes. To get one-year classes, the incidences are interpolated by fitting a generalized additive model

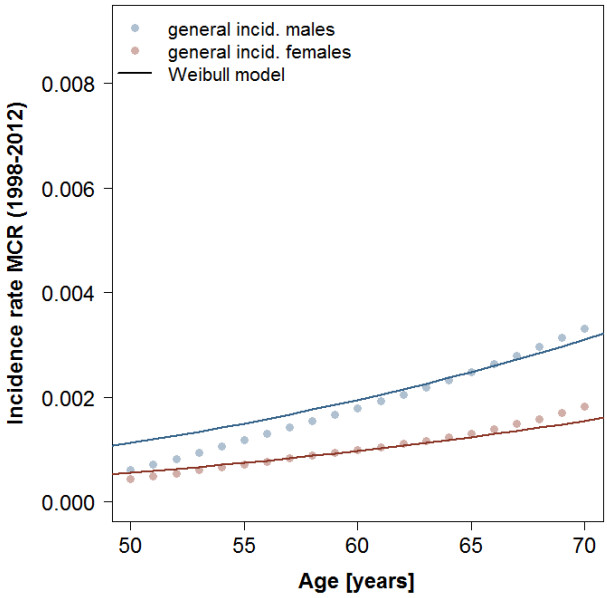

Figure 5.1: Observed incidence rates of CRC (ICD-10 C18–C20) at MCR and Weibull model with parameters $k = 4$ and $\lambda_0 = 0.0058$ and $RR_{\text{sex}} = 2$ in the relevant age interval for the "`family study`" (50 to 70 years).

(GAM) and predicted for every age year. The given incidence of the age classes is set to the youngest age of each interval. Some predictions are below 0 or above 1; they are truncated at the respective bound. Relative sex risk is set to $RR_{\text{sex}} = 2$, as there is some evidence in the literature for this value ([4, 5, 42] and the incidence curves from MCR (see fig. 5.1)). Then Weibull distributions are plotted and visually compared to those predicted age and sex specific rates arising from the application of a GAM to the MCR incidence rates. The best fit in the relevant age range of the `family study` (50 to 70 years) is achieved with Weibull parameters $k = 4$ and $\lambda_0 = 0.0058$ (see fig. 5.1).

Furthermore, $p^{\star}_{\text{inh}}$ needs to be specified or estimated (see section 5.2.2). It determines the hereditary mechanism that contributes to the prior. As it is not the primary aim to estimate the specific hereditary mechanism, the analysis is done at some fixed values. For the CTM setting, $p^{\star}_{\text{inh}}$ is set to 1.00 per definition. Children are supposed to inherit the risk carrier property of their parents in every case. Additionally, three RTM specifications are analysed here. This setting reflects genetic transmission instead of lifestyle factors like CTM. The term $p^{\star}_{\text{inh}}$ can be set to every value between 0 and 1. The probabilities 25 %, 50 % and 75 % are used as $p^{\star}_{\text{inh}}$ here, denoted by RTMo.75, RTMo.50 and RTMo.25 respectively.

5.3 Estimation of the Bayesian Risk Score

As the posterior function can be written down in a closed form, a maximum likelihood (ML) estimation is possible to apply here. A simple, but not the fastest way, is doing it by grid search (see section 5.3.1). Since the posterior is proportional to the likelihood (multiplied with the prior), this approach is valid. If a closed form cannot be derived, the estimation could be done e. g. by means of an expectation-maximisation (EM) algorithm (see section 5.3.2).

5.3.1 Grid Search

Grid search is easy to apply: So, for example, for a two-dimensional problem, one takes for both parameters a set of possible values in the desired coarseness and forms all possible respectively meaningful pairs. The likelihood is then calculated for every pair of parameters of interest. The pair with the highest resulting likelihood is chosen as estimate. Grid search is a straightforward technique to avoid deriving the score function. Unfortunately it is time-consuming. Therefore, a trade-off between coarseness respectively fineness of the grid and computer time has to be found.

If the parameters of interest RR_{fam} and p_1 would be known, calculation of the posterior risk would be possible. Thus, the parameters RR_{fam} and p_1 will be estimated, while the risk carrier properties $\boldsymbol{R}$ are treated as latent variables.

The parameter RR_{sex} is also unknown for Germany, but fixed here at $RR_{\text{sex}} = 2$ (see section 5.2.3). This risk modifier is of secondary interest. Therefore, RR is restricted to RR_{fam} in the following sections. The method to estimate RR described below can easily be extended to the combination of familial and sex risk.

If the risk carrier properties would be known (see section 5.2), the likelihood $\mathbb{P}(\boldsymbol{X} \mid \boldsymbol{R}) = L(RR \mid \boldsymbol{R})$ could be calculated. The prior would be possible to calculate as well, as the prevalence of risk carriers p_1 would then also be known.

If data of N families are given, the likelihood is the product of the single families' contributions, as the families are assumed to be independent from each other: $L(RR \mid \boldsymbol{R}) = \prod_{j=1}^{N} L_j(RR \mid \boldsymbol{R}_j)$ (see (5.5)). The matrix $\boldsymbol{R}$ contains risk carrier constellations $\boldsymbol{R}_j$ of all families. The vector $\boldsymbol{R}_j$ contains the risk carrier properties of each family member: $\boldsymbol{R}_j = \left(R_1, \ldots, R_{n_j}\right)$, e. g. $\boldsymbol{R}_j = (R_1, \ldots, R_5) = (R_{\text{father}}, R_{\text{mother}}, R_{\text{child}_1}, R_{\text{child}_2}, R_{\text{child}_3})$ for a family of five.

The contribution of family j is $L_j(RR \mid \boldsymbol{R}_j)$. Using formula (5.4) and the functions stated in section 5.2.1, the likelihood contribution of family j is

$$L_j(RR \mid \boldsymbol{R}_j) = \prod_{i=1}^{n_j} \left[k{\lambda_0}^k {t_i}^{k-1} \cdot RR^{r_i} \cdot {RR_{\text{sex}}}^{s_i}\right]^{c_i} \left[\exp\left(-{\lambda_0}^k {t_i}^k \cdot RR^{r_i} \cdot {RR_{\text{sex}}}^{s_i}\right)\right].$$

The data used in the likelihood are denoted by t_i (age), r_i (risk carrier property) and s_i (sex) for person i (see section 5.2). The number of members in the family j is labelled with n_j.

However, the risk carrier properties $\boldsymbol{R}$ respectively $\boldsymbol{R}_j$ are unknown in reality. The law of total probability leads to the likelihood of RR:

$$\begin{aligned} P(RR) &= \sum_{\rho \in \boldsymbol{\mathcal{M}}} L(RR \mid \boldsymbol{R} = \boldsymbol{\rho}) \cdot \mathbb{P}\left(\boldsymbol{R} = \boldsymbol{\rho}\right) \\ &= \sum_{\rho \in \boldsymbol{\mathcal{M}}} \left(\prod_{j=1}^{N} L_j(RR \mid \boldsymbol{R}_j) \right) \mathbb{P}\left(\boldsymbol{R} = \boldsymbol{\rho}\right), \end{aligned}$$

where $\boldsymbol{\rho} = (\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N)$ is a specific risk carrier property constellation of the N families and the set $\boldsymbol{\mathcal{M}} = (0,1)^{\sum_j n_j}$ contains all possible risk constellations of the N families.

The function $P(RR)$ needs to be maximised with respect to RR and p_1, the two parameters of interest. As mentioned above, the prevalence p_1 is known, if the risk carrier properties $\boldsymbol{R}_j$ are given. Therefore, with given $\boldsymbol{R}$ only RR needs to be estimated. Standard maximum likelihood (ML) estimation routines lead to a kind of log-likelihood:

$$\begin{aligned} p(RR) &= \log(P(RR)) \\ &= \log\left(\sum_{\rho \in \boldsymbol{\mathcal{M}}} \left(\prod_{j=1}^{N} L_j(RR \mid \boldsymbol{R}_j) \right) \mathbb{P}\left(\boldsymbol{R} = \boldsymbol{\rho}\right) \right). \end{aligned}$$

One can decompose the prior to the product of the single families, as the families are assumed to be independent from each other: $\mathbb{P}\left(\boldsymbol{R} = \boldsymbol{\rho}\right) = \prod_{j=1}^{N} \mathbb{P}\left(\boldsymbol{R}_j\right)$. This leads to

$$\begin{aligned} p(RR) &= \log\left(\sum_{\rho \in \boldsymbol{\mathcal{M}}} \left(\prod_{j=1}^{N} L_j(RR \mid \boldsymbol{R}_j) \right) \prod_{j=1}^{N} \mathbb{P}\left(\boldsymbol{R}_j\right) \right) \\ &= \log\left(\sum_{\rho \in \boldsymbol{\mathcal{M}}} \left(\prod_{j=1}^{N} L_j(RR \mid \boldsymbol{R}_j) \cdot \mathbb{P}\left(\boldsymbol{R}_j\right) \right) \right). \end{aligned}$$

The product $\prod_{j=1}^{N} L_j(RR \mid \boldsymbol{R}_j) \cdot \mathbb{P}\left(\boldsymbol{R}_j\right)$ gets quite small due to multiplication of probabilities. The resulting numbers are numerically not distinguishable from 0. This may lead to computational problems. The solution lies in first logarithmising and then

summing up. Fubini's theorem and simple logarithm calculation rules result in

$$\begin{aligned} p(RR) &= \log\left(\sum_{\boldsymbol{R}_1\in\mathcal{M}}\sum_{\boldsymbol{R}_2\in\mathcal{M}}\cdots\sum_{\boldsymbol{R}_N\in\mathcal{M}}\left(\prod_{j=1}^{N-1}L_j(RR\mid\boldsymbol{R}_j)\cdot\mathbb{P}\left(\boldsymbol{R}_j\right)\right)\left(L_N(RR\mid\boldsymbol{R}_N)\cdot\mathbb{P}\left(\boldsymbol{R}_N\right)\right)\right) \\ &= \log\left(\left[\sum_{\boldsymbol{R}_1\in\mathcal{M}}\cdots\sum_{\boldsymbol{R}_{N-1}\in\mathcal{M}}\left(\prod_{j=1}^{N-1}L_j(RR\mid\boldsymbol{R}_j)\cdot\mathbb{P}\left(\boldsymbol{R}_j\right)\right)\right]\left[\sum_{\boldsymbol{R}_N\in\mathcal{M}}L_N(RR\mid\boldsymbol{R}_N)\cdot\mathbb{P}\left(\boldsymbol{R}_N\right)\right]\right) \\ &= \log\left(\prod_{j=1}^{N}\left[\sum_{\boldsymbol{R}_j\in\mathcal{M}}L_j(RR\mid\boldsymbol{R}_j)\cdot\mathbb{P}\left(\boldsymbol{R}_j\right)\right]\right) \\ &= \sum_{j=1}^{N}\log\left(\sum_{\boldsymbol{R}_j\in\mathcal{M}}L_j(RR\mid\boldsymbol{R}_j)\cdot\mathbb{P}\left(\boldsymbol{R}_j\right)\right) \end{aligned} \tag{5.9}$$

With this artifice it is possible to calculate $p(RR)$ for a grid of values for the two parameters of interest RR and p_1 without numerical struggle. The second parameter p_1 (prevalence of risk carriers) is hidden in $\mathbb{P}\left(\boldsymbol{R}_j\right)$, the prior.

A confidence interval or confidence region around the ML estimate with significance level α can be calculated via

$$\begin{aligned} \widetilde{l}(\boldsymbol{\theta}) &\geq c \\ l(\boldsymbol{\theta}) - l(\widehat{\boldsymbol{\theta}}) &\geq c \\ l(\boldsymbol{\theta}) &\geq c + l(\widehat{\boldsymbol{\theta}}) \\ \text{with } c &= \tfrac{1}{2}\chi^2{}_{1-\alpha}(df) \end{aligned} \tag{5.10}$$

according to standard ML theory. The so-called "normed" likelihood is denoted by $\widetilde{l}$, and $\widehat{\boldsymbol{\theta}}$ stands for the ML estimate. The term df represents the number of parameters, i. e. the length of the vector of parameters of interest $\boldsymbol{\theta}$. Here, $\boldsymbol{\theta} = (RR, p_1)$, i. e. $df = 2$.

This results in a two-dimensional grid search in order to estimate the increased relative risk RR and the prevalence of risk carriers p_1. To implement a grid search, one has to calculate $p(RR)$ for a range of possible values of RR and p_1. This has to be done for every genetic model separately (RTMo.25, RTMo.50, RTMo.75, CTM). By doing this, a "likelihood surface" over the grid of the two parameters RR and p_1 can be plotted, e. g. with a contour plot. The "peak" of it, i. e. the point where the likelihood is maximised, provides the estimated values of the grid searched parameters. Actually, this kind of analysis can be seen as a profile likelihood, since RR_{sex} is kept fixed, too.

5.3.2 Expectation-Maximisation Algorithm

The expectation-maximisation (EM) algorithm [17] or Monte Carlo EM (MCEM) [78] algorithm are alternatives to obtain estimates of the parameters of interest.

The expectation-maximisation (EM) algorithm was introduced by [17] and represents a very flexible tool for ML estimation. It is a computational method to iteratively estimate parameters with incomplete data [17]. Two steps, the expectation step (E-step) and the maximisation step (M-step) alternate until convergence is reached. The incomplete, observed data y are augmented with latent variables to get a complete data set x [78]. The EM algorithm then finds an ML estimate for y using the associated x [17].

The complete data x arise from a distribution $f(x \mid \phi)$. The parameter of interest to estimate via ML is (a transformation of) ϕ. In a very general formulation [17], the pth E-step is to compute

$$Q(\phi \mid \phi^{(p)}) = \mathbb{E}\left(\log(f(x \mid \phi)) \mid y, \phi^{(p)}\right) = \int_u \log\left(f(\phi \mid x)\right) \cdot f(u \mid \phi^{(p)}, y)\, \mathrm{d}u$$

with u as the additional latent variable, i. e. $u = x \setminus y$ [17, 78].

Therefore, the Q function is the expectation of the log-likelihood of the complete data x at ϕ [51]. In the pth M-step, $Q(\phi \mid \phi^{(p)})$ needs to be maximised with respect to ϕ [17]:

$$\phi^{(p+1)} = \arg\max_{\phi} Q(\phi \mid \phi^{(p)}).$$

The start of the algorithm is an initialisation of ϕ as $\phi^{(0)}$. The initialisation is important, as the EM algorithm converges to a local maximum [51]. If the likelihood has more than one maximum, several starting values should be used.

The EM is useful in several settings with data missing at random. It can be used to find ML estimates in data with "classic" missing observations or to discriminate a mixture of distributions or grouping and many other fields [17], since the Q function can be very flexible.

An extension of the general EM is the Monte Carlo EM (MCEM) algorithm, where the Q function is approximated by a Monte Carlo integration [51]: The expectation of the log-likelihood $\log(f)$ that forms the Q function, is approximated by the mean of the log-likelihood of a sample $x^{(l)}$, $(l = 1, \ldots, m)$, from the current target distribution [51]

$$Q(\phi \mid \phi^{(p)}) = \frac{1}{m} \sum_{l=1}^{m} \log\left(f(\phi \mid x^{(l)})\right).$$

Starting with a small number m and increasing it with the number of iterations increases efficiency [78]. This approximation helps, if the expectation, i. e. the Q function, cannot be derived in a closed form [51].

For convergence of an EM algorithm, it is necessary to initialise it with a value lower than the ML estimator (MLE), because an EM update will always be equal or higher than the last step [17, 51]. This ascent property is not present in the MCEM, but there are several rules to have a high probability of ascending [51]. The EM algorithm also converges to a local maximum or even saddle point, which may not be the global maximum [51]. Additionally for the MCEM, the Monte Carlo error adds to the convergence difficulty. Nevertheless, with a carefully chosen initialisation and iteration

number, the (MC)EM algorithm can help to estimate the likelihood, where no classical MLE is available due to unsolvable expectations or functions without closed form.

In the case described here, the missing variable u that adds to the observed data y to form the complete data x, is the latent, i. e. unknown, risk carrier property R_i. The parameter of interest ϕ is a vector containing RR and p_1. The Q function in this context here is $p(RR)$ (cf. equation (5.9)). As the Q function has a closed form, the programming is straightforward. Nevertheless, some computational issues arise. Those specific issues related to the EM algorithm are stated elsewhere [18].

5.4 Simulation Study ("*in silico*")

The previously introduced method of grid search (see section 5.3.1) was tested within a simulation, i. e. *in silico* study. Families of nine persons each were simulated.

Generation of Families One simulated family consisted of the mother, father, their parents and three children with random sex. The sex of the three children was determined by sampling from a binomial distribution with probability 0.5. The grandparents were therefore the founders of the family.

Determination of the Risk Carrier Property Their risk carrier properties were randomly sampled from a binomial distribution with probability p_1. This was possible because the grandparents are assumed to be independent from each other. The risk carrier properties of the parents and the three children were determined according to the chosen hereditary mechanism (CTM or an RTM, see section 5.2.2). The individual probability of inheritance p_{inh} was calculated using (5.6) and the risk carrier properties of the grandparents or parents, respectively. Then, again a binomial distribution was used to sample the individual risk carrier property. Each generation had to be determined after the respective parental generation, because the risk carrier property of the respective parents need to be known for sampling. The individuals of one generation were treated as independent from each other, according to the specification of the hereditary mechanisms (see section 5.2.2).

Determination of Age at CRC Diagnosis The age at CRC diagnosis was determined by means of the chosen penetrance mechanism and the individual risk carrier property, age and sex of the persons. A standard uniform distributed variable u_i was sampled for each person i. Then, u_i was plugged in the inverse survival function of the Weibull distribution with the respective relative risk extensions. This inverse function determines then the age at CRC diagnosis, i. e. the end of cancer free survival:

$$S^{-1}(u_i) = t_i = \frac{\left(\frac{-\log(u_i)}{({RR_{\text{fam}}}^{r_i})\cdot({RR_{\text{sex}}}^{s_i})}\right)^{1/k}}{\lambda_0}$$

The figure t_i was rounded to an integer to match the information of the real data from the `family study`. By the determination of the Weibull distribution (see section 5.2.3), the data were adjusted indirectly to reality given by the incidence rates of MCR. The relative risk RR_{sex} was fixed at 2 (see section 5.2.3).

Determination of Age at Adenoma Onset A question about adenomas in relatives is included in the NACRC questionnaire (see section 3.2.3). Therefore, age of adenoma onset was simulated as well, to be able to “fill in” the questionnaire for the simulated families. The age of adenoma onset was fixed at ten years before age at CRC diagnosis, which was simulated before. This lag of ten years was chosen according to literature [10, 41]. Indirectly, the risk carrier property influenced the development of adenomas, too.

Determination of Age at Death A data set containing mean life expectancy, number of survivors and probability of death of every age for men and women for the year 2009 was downloaded from the database GENESIS-Online (Bayern) [53]. Mean life expectancy and probability of death were conditioned on completed age, i. e. the probability to die within one year given a completed age of x years was provided as well as the mean further life expectancy given a completed age of x years. The data set was augmented by a mean life expectancy of 0 years at age 107 for both sexes. With these data, an additive model (GAM) of age on mean life expectancy was estimated. The predictions arising from that GAM gave smooth mean life expectancies for the ages above 94 years, where no data were available. The probability of death was estimated analogously. Both GAMs were calculated for the two sexes separately. With the probability of death, numbers of survivors and subsequently numbers of deaths were calculated. As the provided probability of death was conditioned on completed age, the unconditioned probability of death needed to be calculated for the simulation. This was done for each sex separately using the number of deaths for each age that was estimated in parts via GAM.

The age of death was sampled for the children of the simulated families using the estimated unconditioned probability of death. For the parents’ and grandparents’ generation, the probability of death needed to be conditioned on being alive at birth of the respective youngest child. So, new conditioned probabilities of death were calculated. The father was required to be at least 34 years old, the mother at least 31 years. Age of death was then sampled with those new calculated conditioned probabilities. The same was done for the grandparents: The grandfather needed to be at least 29 years, the grandmother at least 26 years. Those minimum ages arose from literature research: Husbands are on average three years older than their wives [40], which did not change over decades [40]. Therefore, this age difference between spouses could have been used for both the parents’ and the grandparents’ generation. Mothers born around 1970 gave birth at a mean age of 26, 29 respectively 31 to their first, second, and third child [62]. The values were almost the same for mothers born in the middle 1930s [62]. Thus,

the same numbers were used for both the parents' and the grandparents' generation. However, the grandparents had a lower minimum age, as they only needed to be alive for one child (i. e. the parents), whereas the parents needed to have conceived three children.

Grid Search Binary status variables were created to indicate to pre- or absence of CRC or adenoma: CRC was present, if the simulated age at CRC diagnosis was lower than the simulated age of death. Presence of adenoma was calculated analogously. A variable "age" was created, which was the minimum of age at CRC diagnosis or age of death.

With these simulated data, a grid search according to the method described above (see section 5.3.1) was performed for a grid around the true parameters RR and p_1. The grid for p_1 was set up in steps of 0.1, until deviations of $|0.5|$ from the true parameter were reached. The grid was pruned at 1 and at 0. Those grid steps were set to 0.99 and 0.01, respectively. The grid for RR_{fam} contained deviations up to $|5|$ from the true parameter in steps of 1. The grid was pruned at 0, the minimum was 1.

Calculation of the Posterior The parameters leading to the peak of the "likelihood surface" (see section 5.3.1) were taken as estimates. Those estimates were plugged in formula (5.2) to get the posterior of being a risk family for each family in the simulated data set: The respective data of the family members were plugged in as $\boldsymbol{X}$. The posterior for being *no* risk family was calculated with $\boldsymbol{r} = \mathbf{0}$. The posterior of interest was actually the opposite, i. e. $1 - \mathbb{P}(\boldsymbol{R} = \mathbf{0} \mid \boldsymbol{X})$ (see formula (5.3)), because a family was said to be a "risk family", if at least one member was risk carrier. Thus, one posterior probability per family for being a family with familial CRC risk could have been gathered.

Completion of the Questionnaire The NACRC questionnaire was filled in for every simulated family. To answer the first question ("Do you have a first-degree relative diagnosed with CRC?"), the overall sum of CRC diagnoses in a family was determined. Question 1 was set to "yes", if the sum was greater than 0. Question 2 ("Do you have a first-degree relative diagnosed with CRC before 50 years of age?") was answered analogously. Question 3 ("Do you have a first-degree relative diagnosed with intestinal polyp before 50 years of age?") was answered analogously using the sum of persons with adenoma at age of 49 or younger. As it was assumed for the simulation, that every adenoma was known, but not removed during colonoscopy, this question added to the score of the NACRC questionnaire in a risk-overestimating manner. Data for question 4 ("Do you have at least three first-degree relatives diagnosed with one of the following cancers: colorectal cancer, stomach cancer, endometrial cancer?") were not simulated. Therefore, all families got a "no" recorded for question 4. This partly counteracted the overestimation of question 3.

The number of "yes" answers served as risk score of the NACRC questionnaire. The missing question 4 implicated that the maximum score was 3 instead of a range from 0 to 4 in real life.

Computational Details As specified for the whole analysis done here (see section 5.2.3), the used parameters for the Weibull distribution were $k = 4$ and $\lambda_0 = 0.0058$ according to best overlapping of the Weibull distribution with the age and sex specific incidence rates from MCR (see fig. 5.1). The hereditary mechanisms CTM, RTMo.75 and RTMo.50 as well as RTMo.25, i. e. a probability of inheritance $p^{\star}_{\text{inh}} \in \{100\,\%, 75\,\%, 50\,\%, 25\,\%\}$, were covered. The relative risk of sex was fixed at $RR_{\text{sex}} = 2$ for males, as they have an approximately doubled risk compared to women during the whole lifetime (see fig. 5.1 and section 5.2.3).

The simulation contained $N = 500$ families. Prevalence p_1 was set to 0.2, as a familial background was assumed for approximately 20 % of the CRC cases in Germany [45, 72]. The prevalence was aimed to be estimated in the application using data from the `family study`. The second parameter to be estimated in the application was the familial relative risk RR_{fam}. In the *in silico* study, the relative risk of risk carriers was varied from 2 through 3, 4, 6, 8, 10 and 12, to 15.

All computations were done in R version 3.3.3 [64] on platform `x86_64-pc-linux-gnu` (64-bit) (see section 1).

5.5 Comparison of Bayesian Risk Score with NACRC Questionnaire

The NACRC questionnaire is an ordinal variable, whereas the Bayesian risk is a continous variable. It is possible for both methods to plot a receiver operating characteristic (ROC) curve. The ROC curve is a method to optimise and visualise the classification performance of a continuous variable, where the threshold is varied. It needs a "gold standard" to which the classification of the continuous variable is compared. To construct a ROC curve, the true positive rate is plotted against the false positive rate for every threshold of the continuous variable. The threshold that is classifying best has a threshold with a true positive rate of 1 and a false positive rate of 0, showing an area under the curve (AUC) of 1. The worst classifier has an AUC of 0.5, as it shows equal true and false positive rate and lies therefore on the bisectrix. The classification of the associated continuous variable is then as good as throwing a coin. Furthermore, different test procedures can easily be compared with ROC curves. The better classifier lies above the worse one (i. e. nearer on the perfect point $(0, 1)$) and has therefore a higher AUC.

Here, the Bayesian posterior score (see section 5.2) was compared with the simple questionnaire of the NACRC (see section 3.2.3). The "gold standard" used here was classification as a "risk family". A family was defined as "risk family", if at least one member was risk carrier for the simulation study. For the application to the `family study`, having at least one other CRC case in the family beside the recruited patient was used as criterion. Only verified CRC cases, i. e. those proved by MCR, were used in the application to the `family study` (see section 4). The continuous variable here

was the posterior score respectively the NACRC questionnaire, although the latter was only an ordinal variable. Nevertheless, it was possible to calculate the true and false positive rate of every threshold and to interpolate them to get a curve and an AUC. Confidence intervals for the AUC were also calculated [16]. Those AUCs and their confidence intervals are shown in the results section 6.

6
Results

This chapter shows the results of the grid search in two kinds of simulated data sets, representing the general population and a selected population similar to the `family study`. The method introduced in section 5.2 was also applied to the `family study`; results are shown here. Furthermore, parameters taken from literature were used to estimate the Bayesian risk score in the `family study`. Those results are also presented in this section.

6.1 Simulation Study

The procedure of simulating the data was described in section 5.4. Families of nine persons were generated and risk carrier property, age at CRC diagnosis and age at death were determined for each single member of the 500 families. Afterwards, grid search as described in section 5.3.1 was performed. In the end, the Bayesian risk score (introduced in section 5.2) was calculated and the NACRC questionnaire was filled in using the simulated informations.

6.1.1 Grid Search in the Simulated General Population

According to the setup, the simulated data of this *in silico* study represented an application in the general population. The average of mean age of male patients was 64.10, the average of mean age of female patients across the several simulated data sets was 68.15. There were 60.20 % males among the CRC patients. The 4500 simulated persons

included 656.7 CRC patients in average. This corresponded to a proportion of around 14.59 % that was higher than in the real general population. However, this mean value included also scenarios with a high relative risk RR_{fam} and therefore this high proportion of CRC cases seems eligible. The mean number of CRCs per family was 1.3. Of course, there were also families without any CRC diagnosis at all.

As described in section 5.4, $RR_{\text{sex}} = 2$, $\lambda_0 = 0.0058$, $k = 4$, and $p_1 = 0.2$ were set as default and RR_{fam}varied from 2 to 15.

Some randomly chosen likelihood surfaces (see section 5.3.1) are shown in figure 6.1. The remaining surfaces are given in the appendix (see section B.1.1). The figure illustrates, that grid search was able to find the true parameters. The true parameters lay in nearly all settings at least within the confidence region around the estimate from grid search. This confidence region was calculated according to the formula (5.10) described in section 5.3.1. Despite a relatively coarse grid used, the parameters were found with sufficient accuracy. The likelihood surfaces got more and more pointed, the bigger the value of RR_{fam} got. The surface seemed to have only one mode. That is clearly an advantage with respect to a possible application of the EM algorithm. The shape of the likelihood surfaces did not change much with the increase of $p^{\star}_{\text{inh}}$. This is an advantage, as the accuracy of estimation of the parameters did not change. The estimation of the true parameter worked better for high $p^{\star}_{\text{inh}}$, but the true parameters lay at least within the confidence region for low $p^{\star}_{\text{inh}}$.

As described in section 5.5, the Bayesian risk score was compared to the NACRC questionnaire using ROC curves and AUCs. For each setting of the *in silico* study (i. e. for every combination of those four $p^{\star}_{\text{inh}}$ and the eight RR_{fam} used), the estimated parameters were taken and the posterior score was computed according to the method proposed in section 5.2. The NACRC questionnaire was filled in for each family as described above (see section 5.4). Then, ROC curves were plotted and the area under the curve (AUC) was calculated. The ROC curves are provided in section B.1.2.

The resulting AUCs for the simulated general population consisting of 500 families are shown in table 6.1 with their corresponding confidence intervals. The method of the Bayesian posterior score showed in all settings a higher value of the AUC than the score of the NACRC questionnaire. The AUC for the NACRC questionnaire was in all settings around 0.8, which provided good results with respect to AUC and therefore with respect to differentiation between "risk families" and families without risk carrier properties. The AUC for the Bayesian risk score was often around 0.95, showing very good discrimination between "risk families" and "normal" families using a family anamnesis with subsequent calculation of the posterior score. Both risk scores were therefore able to differentiate those two family types ("risk" and "normal") very well. Increasing RR_{fam} did not change AUCs very much. Increasing the inheritance probability $p^{\star}_{\text{inh}}$ had only negligible effect on the AUC, too. Therefore, results remained stable despite the chosen hereditary mechanism (e. g. CTM with $p^{\star}_{\text{inh}} = 1.00$).

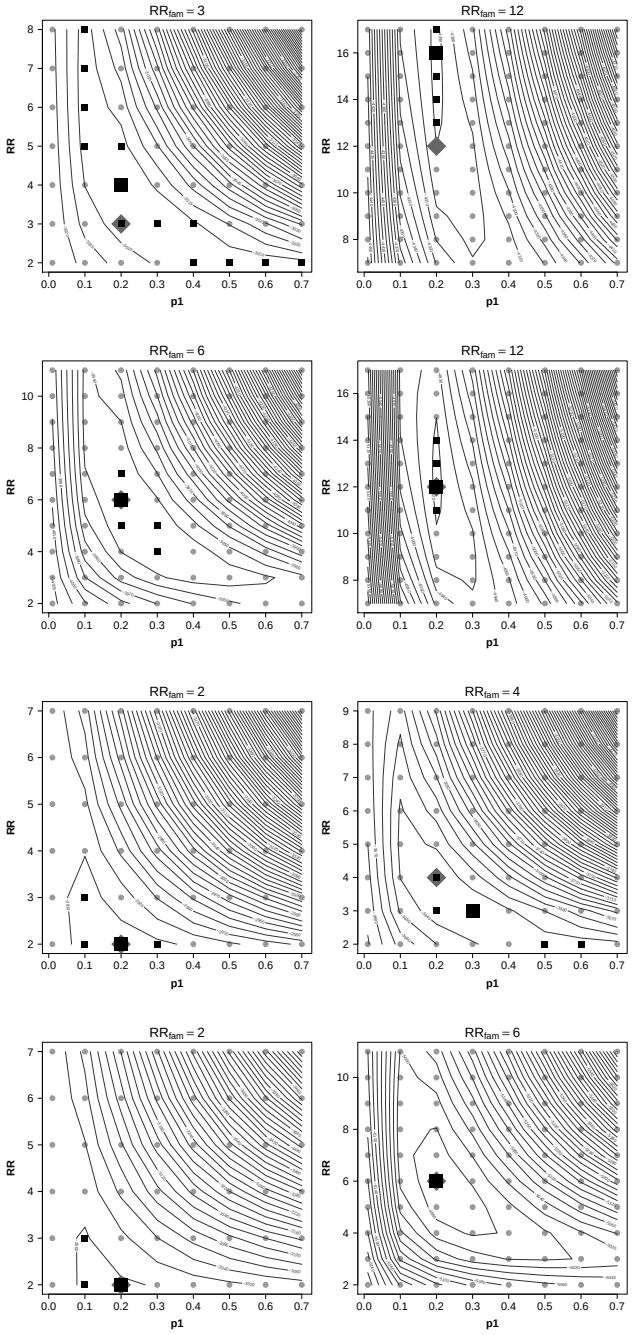

Figure 6.1: Contour plot of randomly chosen likelihood surfaces for the simulated general population. At the top, the probability of inheritance was set to $p^{\star}_{\text{inh}} = 0.25$, the rows beneath contains plots for $p^{\star}_{\text{inh}} = 0.50$, $p^{\star}_{\text{inh}} = 0.75$, and $p^{\star}_{\text{inh}} = 1.00$ at the bottom. The data sets consist of $N = 500$ families. The grey round points show the grid for the search. The big squared black dot shows the maximum of the likelihood surface. The small squared black dots show the likelihood confidence region around the maximum. The grey diamond shows the true parameters, with which the data were simulated. The prevalence was set to $p_1 = 0.2$. Please be aware of different y-axes.

Table 6.1: AUC of ROC curves (see section B.1.2) for several $p^{\star}_{\text{inh}}$ for the simulated general population. The data sets consist of $N = 500$ families. AUC for the NACRC questionnaire is shown, too. The posterior score for the ROC curves was calculated with estimated parameters by means of grid search. 95 % confidence intervals (95 % CI) were calculated using the method proposed by [16].

	$p^{\star}_{\text{inh}} = 0.25$		$p^{\star}_{\text{inh}} = 0.50$		$p^{\star}_{\text{inh}} = 0.75$		$p^{\star}_{\text{inh}} = 1.00$	
RR_{fam}	AUC	95 % CI	AUC	95 % CI	AUC	95 % CI	AUC	95 % CI
2	0.9509	0.9314; 0.9705	1	1; 1	0.9962	0.9930; 0.9994	0.9760	0.9631; 0.9889
quest.	0.8280	0.7989; 0.8571	0.8257	0.7937; 0.8578	0.8158	0.7853; 0.8463	0.7850	0.7517; 0.8183
3	0.9598	0.9424; 0.9773	1	1; 1	0.9971	0.9948; 0.9994	0.9745	0.9627; 0.9863
quest.	0.7782	0.7452; 0.8113	0.7898	0.7586; 0.8210	0.8176	0.7886; 0.8465	0.7965	0.7652; 0.8279
4	0.9391	0.9183; 0.9598	0.9998	0.9993; 1	0.9991	0.9980; 1	0.9809	0.9716; 0.9903
quest.	0.8009	0.7698; 0.8321	0.8083	0.7769; 0.8397	0.7765	0.7433; 0.8097	0.7881	0.7566; 0.8197
6	0.9557	0.9376; 0.9738	0.9980	0.9958; 1	0.9953	0.9910; 0.9995	0.9842	0.9763; 0.9921
quest.	0.8008	0.7690; 0.8325	0.7844	0.7517; 0.8172	0.7775	0.7437; 0.8114	0.7911	0.7585; 0.8237
8	0.9428	0.9233; 0.9623	0.9862	0.9785; 0.9939	0.9973	0.9947; 0.9998	0.9878	0.9813; 0.9943
quest.	0.8042	0.7726; 0.8358	0.8155	0.7842; 0.8468	0.7956	0.7620; 0.8293	0.8448	0.8158; 0.8738
10	0.9601	0.9449; 0.9753	0.9965	0.9939; 0.9991	0.9962	0.9931; 0.9993	0.9914	0.9863; 0.9965
quest.	0.7732	0.7390; 0.8075	0.8160	0.7855; 0.8465	0.8110	0.7795; 0.8425	0.8364	0.8048; 0.8681
12	0.9284	0.9065; 0.9504	0.9851	0.9777; 0.9925	0.9941	0.9897; 0.9985	0.9944	0.9904; 0.9984
quest.	0.7730	0.7388; 0.8073	0.8078	0.7763; 0.8393	0.8112	0.7786; 0.8437	0.8711	0.8420; 0.9003
15	0.9479	0.9299; 0.9658	0.9817	0.9734; 0.9900	0.9928	0.9879; 0.9977	0.9940	0.9899; 0.9980
quest.	0.8001	0.7678; 0.8324	0.8249	0.7944; 0.8554	0.8505	0.8220; 0.8790	0.8661	0.8367; 0.8955

6.1.2 Grid Search in the Simulated Selected Population

The simulated study population re-used the data from the simulated general population (see section 6.1.1). From those 500 families, only the families with at least one CRC case were taken. The families without CRC cases were excluded from the simulated data set above. So, this data sets consisted of less than $N = 500$ families. The simulated selected population aimed to mimic a situation similar to the `family study`. The `family study` consisted only of families with incident CRC cases. The analysis of the simulated selected population illuminated the applicability of the scores in an incident population like the `family study`.

As the same data sets as for the simulated general population were used, mean age and proportion of males among the CRC patients did not change. The mean number of CRCs per family raised to 1.9, as the families without any CRC diagnosis at all dropped out of the data set. For the same reason, the proportion of CRCs among the persons regarded increased to 21.05 %. There were in average 339.2 families in the data sets regarded.

For every combination of $p^{\star}_{\text{inh}}$ and RR_{fam}, likelihood surfaces were plotted after grid search (see section 5.3.1). Randomly chosen likelihood surfaces of the pruned data set of the study population are shown in figure 6.2. The chosen settings are the same as for the simulated general population (see fig. 6.1). The remaining surfaces are given in the appendix (see section B.2.1). In all settings of figure 6.2, the true parameters were missed by grid search. This was also true for those settings in section B.2.1). Especially the prevalence p_1 in the general population seemed to be problematic to estimate properly in a selected population. It was overestimated in every setting. Using RTMo.25 with $p^{\star}_{\text{inh}} = 0.25$, i. e. the weakest inheritance regarded, RR_{fam} was overestimated for small true RR_{fam} values and underestimated for high true RR_{fam} values. As mentioned before, the prevalence p_1 was overestimated in every setting regarding estimation for a general population. It was estimated in most settings to the highest value considered within grid search. This gave at least some evidence, that it maybe would have been estimated even higher, if the grid would have been more expanded. Using RTMo.50 with $p^{\star}_{\text{inh}} = 0.50$ and low true RR_{fam}, RR_{fam} was estimated with only small deviations from the true parameters. This may have been influenced by the relatively coarse grid used. Increasing RR_{fam} worsened the estimation with bigger deviations from the true parameter. This means, the maximum of the likelihood surface, i. e. the estimate from grid search, was further away from the true RR_{fam}. Prevalence p_1 was again overestimated for every setting regarded. The estimate got nearer to the true parameter for increasing RR_{fam}, but never met it. Increasing RR_{fam} worsened the estimation with bigger deviations from the true parameter also with RTMo.75 ($p^{\star}_{\text{inh}} = 0.75$). The prevalence p_1 was overestimated again, but the deviations from the true value 0.2 were smaller for higher RR_{fam}. CTM was the setting with the highest probability of inheritance regarded. It was set to $p^{\star}_{\text{inh}} = 1.00$. The relative risk RR_{fam} was hit relatively good with mostly only minor or no deviation from the true value. The prevalence p_1 was estimated better compared to lower $p^{\star}_{\text{inh}}$, especially for high RR_{fam}. Also the estimation of RR_{fam} worked better for high true values

of RR_{fam}. The grid search estimate, i. e. the maximum of the likelihood surface, was then near to the true parameters. However, the confidence region around the estimate sometimes did not contain the true parameters.

As mentioned before, the risk score using the estimates from grid search and the NACRC questionnaire were compared by means of ROC curves and AUCs (see section 5.5). Also for those settings where the true parameters were obviously missed by grid search, the estimates were taken and the Bayesian risk score (see section 5.2) was calculated for all families in the data sets. The NACRC questionnaire was filled in for every family as described in section 5.4. ROC curves were plotted and AUCs were calculated. ROC curves are depicted in section B.2.2 in the appendix.

The resulting AUCs for the simulated selected population imitating the situation of the `family study` are shown in table 6.2 with their corresponding confidence intervals. Despite the poor overlapping with the true parameters, the AUCs of the Bayesian risk score showed good results. The AUC of the NACRC questionnaire showed lower values around 0.60. These results remained stable for the different probabilities of inheritance $p^{\star}_{\text{inh}}$. This means, the hereditary mechanism (e. g. CTM with $p^{\star}_{\text{inh}} = 1.00$) had not much impact on the results. Increasing RR_{fam} had also no big influence on the AUC respectively ROC curve. These results showed that the choice of a beforehand determined hereditary mechanism does not have much impact on the differentiation ability of the Bayesian posterior score. The performance of the NACRC questionnaire risk score is untouched by the choice of $p^{\star}_{\text{inh}}$ by definition.

6.2 Application to the Family Study

An overview of the `family study` can be found in section 4, including descriptive analysis of both the complete data (see section 4.3) and the local family data (see section 4.4). The latter consisted of the data of the families of the `family study`, who lived completely in the catchment area of the MCR, as described before (see section 4.2).

For the comparison of the NACRC questionnaire with the Bayesian risk score, a kind of gold standard was needed for a ROC analysis (see section 5.5). For the application to the `family study`, having at least one other CRC case in the family besides the recruited patient was used as criterion for being a "risk family". Only verified CRC cases were used for this gold standard to build the ROC curves (see section 4).

6.2.1 Grid Search

Grid search was performed as described in section 5.3.1. Of course, for those real data of the `family study`, no true parameters were known as for the *in silico* study in section 6.1. Grid search was done separately with all four hereditary mechanisms regarded, i. e. for CTM using $p^{\star}_{\text{inh}} = 1.00$ and RTM with $p^{\star}_{\text{inh}} \in \{0.75, 0.50, 0.25\}$ (see section 5.2.3). The Bayesian risk score (see section 5.2) was calculated using the estimates arising from grid

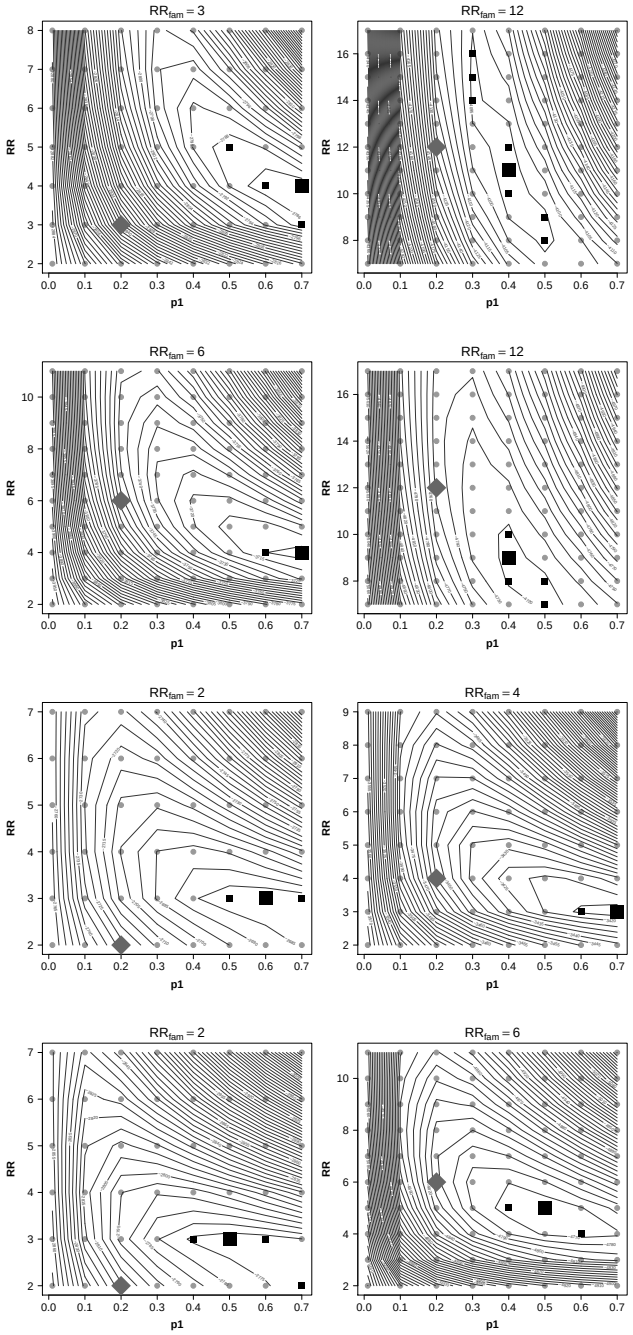

Figure 6.2: Contour plot of randomly chosen likelihood surfaces for the simulated selected population. At the top, the probability of inheritance was set to $p^{\star}_{\text{inh}} = 0.25$, the rows beneath contains plots for $p^{\star}_{\text{inh}} = 0.50$, $p^{\star}_{\text{inh}} = 0.75$, and $p^{\star}_{\text{inh}} = 1.00$ at the bottom. The data sets consist of all families with at least one CRC case out of the $N = 500$ simulated families. The grey round points show the grid. The big squared black dot shows the maximum of the likelihood surface. The small squared black dots show the likelihood confidence region around the maximum. The grey diamond shows the true parameters, with which the data were simulated. The prevalence was set to $p_1 = 0.2$. Please be aware of different y-axes.

Table 6.2: AUC of ROC curves (see section B.2.2) for several $p^{\star}_{\text{inh}}$ for the simulated selected population. The data sets consist of all families with at least one CRC case out of $N = 500$ simulated families. AUC for the NACRC questionnaire is shown, too. The posterior score for the ROC curves was calculated with estimated parameters by means of grid search. 95 % confidence intervals (95 % CI) were calculated using the method proposed by [16].

	$p^{\star}_{\text{inh}} = 0.25$		$p^{\star}_{\text{inh}} = 0.50$		$p^{\star}_{\text{inh}} = 0.75$		$p^{\star}_{\text{inh}} = 1.00$	
RR_{fam}	AUC	95 % CI	AUC	95 % CI	AUC	95 % CI	AUC	95 % CI
2	0.8896	0.8443; 0.9349	0.9998	0.9995; 1	1	1; 1	0.9918	0.9807; 1
quest.	0.5822	0.5304; 0.6340	0.6324	0.5762; 0.6886	0.6161	0.5662; 0.6660	0.5690	0.5174; 0.6206
3	0.9213	0.8879; 0.9546	0.9993	0.9980; 1	1	1; 1	0.9973	0.9945; 1
quest.	0.5450	0.5049; 0.5949	0.5587	0.5113; 0.6061	0.5938	0.5496; 0.6380	0.5931	0.5463; 0.6398
4	0.8884	0.8490; 0.9277	1	0.9998; 1	1	1; 1	0.9997	0.9991; 1
quest.	0.5770	0.5279; 0.6260	0.6215	0.5724; 0.6707	0.5879	0.5411; 0.6347	0.5876	0.5443; 0.6310
6	0.9318	0.9002; 0.9635	0.9980	0.9952; 1	0.9985	0.9960; 1	0.9927	0.9872; 0.9982
quest.	0.5963	0.5468; 0.6457	0.5948	0.5482; 0.6414	0.5840	0.5384; 0.6297	0.6314	0.5887; 0.6741
8	0.9023	0.8677; 0.9369	0.9974	0.9940; 1	0.9982	0.9961; 1	0.9932	0.9873; 0.9991
quest.	0.5822	0.5334; 0.6309	0.6158	0.5686; 0.6631	0.6079	0.5607; 0.6552	0.6586	0.6183; 0.6989
10	0.9531	0.9313; 0.9749	0.9991	0.9976; 1	0.9985	0.9967; 1	0.9915	0.9852; 0.9978
quest.	0.5884	0.5403; 0.6366	0.6443	0.6007; 0.6879	0.6303	0.5864; 0.6741	0.6576	0.6107; 0.7045
12	0.9163	0.8872; 0.9454	0.9913	0.9842; 0.9984	0.9961	0.9923; 0.9998	0.9895	0.9815; 0.9975
quest.	0.6154	0.5690; 0.6619	0.6497	0.6056; 0.6938	0.6268	0.5812; 0.6723	0.6709	0.6211; 0.7207
15	0.9251	0.8972; 0.9530	0.9820	0.9711; 0.9928	0.9909	0.9834; 0.9983	0.9900	0.9830; 0.9969
quest.	0.6232	0.5759; 0.6705	0.6551	0.6115; 0.6987	0.6985	0.6558; 0.7412	0.7109	0.6654; 0.7563

search. The NACRC questionnaire was demanded from recruited patients, or at least their relatives, within the `family study` enrolment process (see section 4.1).

Application to Complete Data Set An overview of the data like age and sex of the CRC patients is given in section 4.3.

The parameters $RR_{sex} = 2$, $\lambda_0 = 0.0058$, and $k = 4$ were used to calculate the Bayesian risk score (see section 5.2.3).

As having at least two verified CRC cases in the family was defined as "risk family", a certain probability of false negatives due to partly impossible record linkage was present here in the complete data set. This was mostly due to family members living outside the catchment area of the MCR and non-obligatory reporting to the MCR (see section 4). The results of the more reliable local family data are given in section 6.2.1.

One family tree was pruned for the execution of grid search due to problems with the working memory. Two parts of this family tree were cut, because they had no connection to the rest of the tree. There was no CRC case in the removed parts.

The likelihood surfaces of the four hereditary mechanisms regarded are shown in figure 6.3. The likelihood surface had only one mode like the surfaces from the *in silico* study. This makes a possible application of e. g. EM algorithm more easier. The shape of the likelihood surfaces did not change much with the increase of $p^{\star}_{inh}$, it only got a little bit more pointed. The grid search in this incident study population of the `family study` showed problems similar to the simulated selected population (see section 6.1.2). The prevalence p_1 was estimated with a quite high value according to the findings of the simulated selected population (see section 6.1.2). The prevalence even reached values of 1 for the lowest probability of inheritance regarded (RTMo.25 with $p^{\star}_{inh} = 0.25$, see fig. 6.3). Additionally, a minimum estimated prevalence of 0.65 was much higher than literature says for the general population (see section 3.2). The estimates of RR_{fam} increased, as $p^{\star}_{inh}$ increased. It reached values of 5.5 for the CTM setting with $p^{\star}_{inh} = 1.00$.

As described in section 5.5 and also done for the *in silico* part (see section 6.1), the Bayesian risk score was compared to the NACRC questionnaire using ROC curves and AUCs. For each of the four $p^{\star}_{inh}$, the estimated parameters were used to calculate the posterior score according to the method described in section 5.2. The NACRC questionnaire was filled in by each family itself (see section 4). Then, ROC curves for both the Bayesian risk score and the NACRC questionnaire risk score were plotted and AUCs were calculated.

The ROC curves can be found in figure 6.4. Table 6.3 shows the AUCs and the respective confidence intervals according to [16] for the Bayesian posterior risk score with the four hereditary probabilities $p^{\star}_{inh}$ as well as for the score of the NACRC questionnaire. In contrast to the simulation study, where even poor results in the grid search led to quite good results in the ROCs respectively AUCs, the method of the Bayesian posterior score showed mostly a lower AUC than the score of the NACRC questionnaire. However, especially for a low false positive rate the posterior score showed as good results as the NACRC score (see fig. 6.4). The four mechanisms

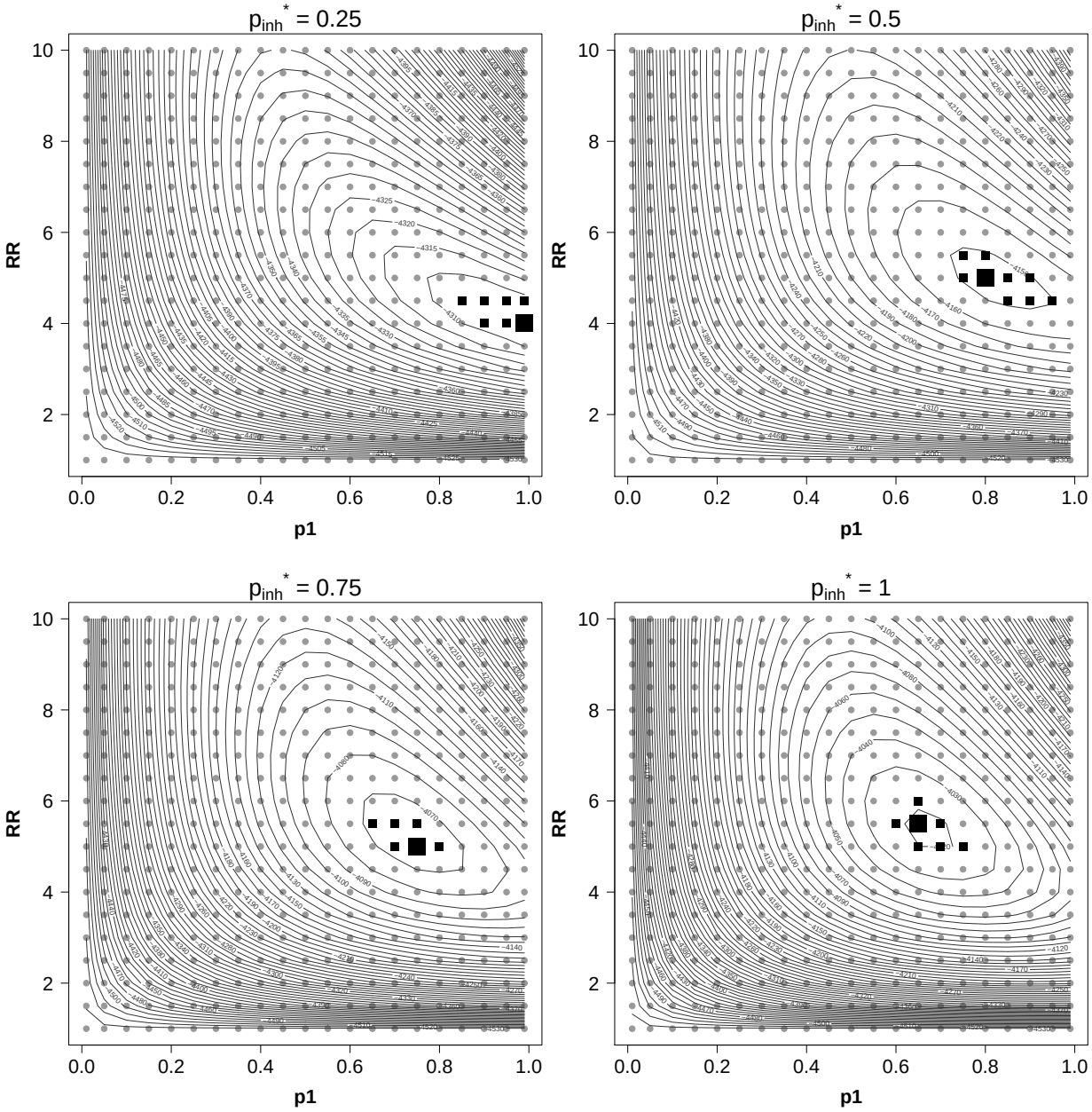

Figure 6.3: Contour plot of the likelihood surface for several $p^{\star}_{\text{inh}}$ for the complete data set of the `family study` ($N = 611$ families, 4115 family members; see also fig. 4.1). The grey round points show the grid. The big squared black dot shows the maximum of the likelihood surface. The small squared black dots show the likelihood confidence region around the maximum.

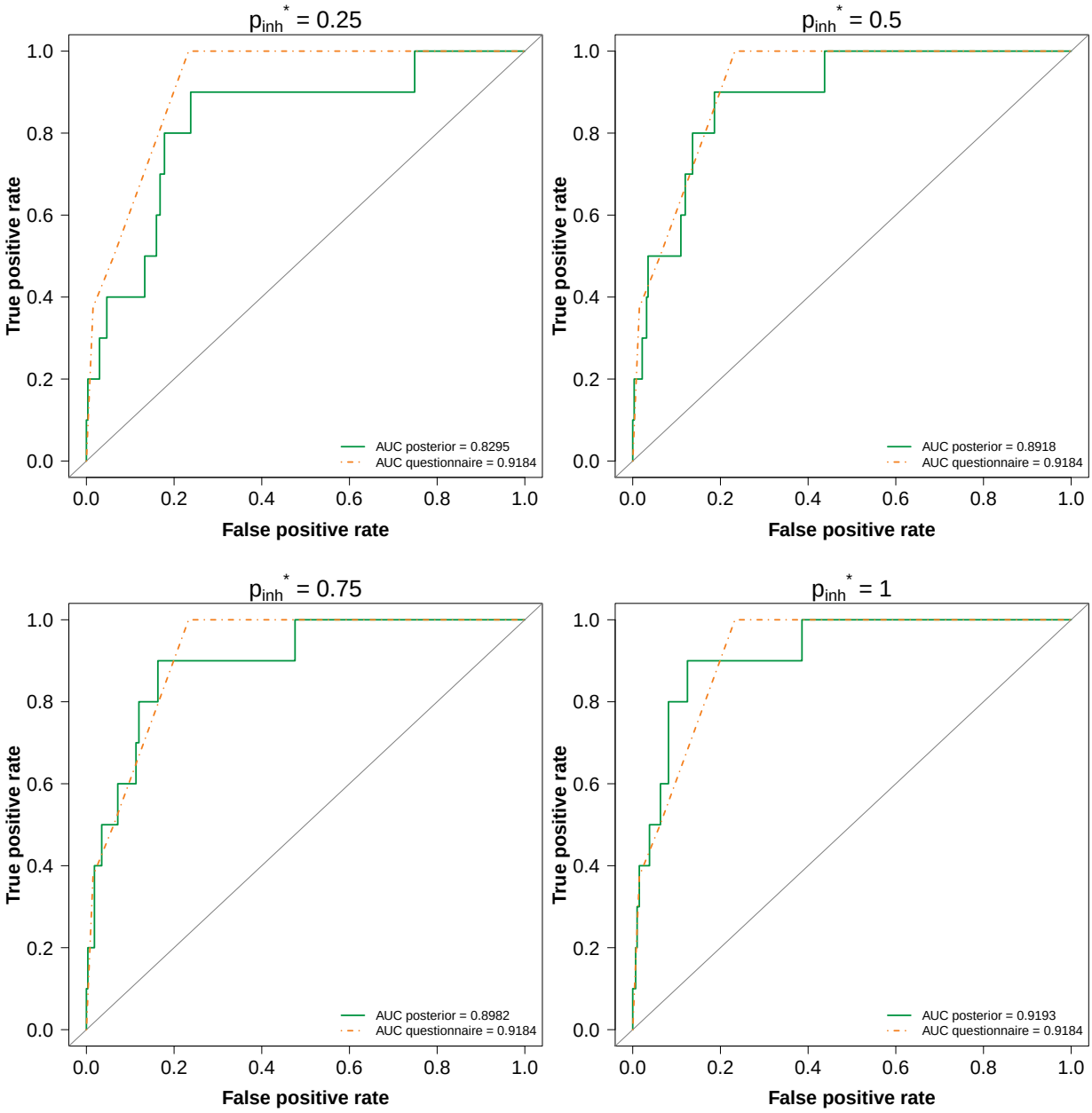

Figure 6.4: ROC curves for several $p^{\star}_{\text{inh}}$ for the complete data set of the `family study` ($N = 611$ families, 4115 family members; see also fig. 4.1). The ROC curve for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with estimated parameters by means of grid search.

RTMo.25, RTMo.50, RTMo.75 and CTM were quite similar in their results. The AUCs were between 0.82 and 0.92, the latter for the CTM setting with $p^{\star}_{\text{inh}} = 1.00$. Varying the inheritance probability $p^{\star}_{\text{inh}}$ had only small effect on the AUC. The NACRC questionnaire had an AUC of 0.92, too, and additionally showed a comparatively narrow confidence interval. Confidence intervals overlapped due to the similar AUCs, but never included 0.5. This provided very good results with respect to differentiation between "risk families" and families without presumed familial burden.

Table 6.3: AUC of ROC curves (see fig. 6.4) for several $p^{\star}_{\text{inh}}$ for the complete data set of the `family study` ($N = 611$ families, 4115 family members; see also fig. 4.1). The AUC for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with estimated parameters by means of grid search. 95 % confidence intervals (95 % CI) were calculated using the method proposed by [16].

$p^{\star}_{\text{inh}}$	AUC	95 % CI
0.25	0.8295	0.6925; 0.9664
0.50	0.8918	0.8086; 0.9751
0.75	0.8982	0.8086; 0.9877
1.00	0.9193	0.8470; 0.9916
quest.	0.9184	0.8670; 0.9698

Application to Local Family Data Only Analogously to the complete data set, an overview of characteristics of the data like age and sex of the CRC patients is given in section 4.4.

As described before (see section 5.2.3), the parameters used for the calculation of the Bayesian risk score were $RR_{\text{sex}} = 2$, $\lambda_0 = 0.0058$, and $k = 4$. A "risk family" was defined as having at least two verified CRC cases in the family. Verification was assumed to work better in the local family data set than in the complete data set by definition. Of course, no one can be sure that the reported family members were all members of this family in the generations and family tree arms regarded. So, no one knows if indeed the complete family lived in the MCR area. Non-obligatory reporting of cancer cases to the MCR remained still a problem. Nevertheless, the results presented in this section here are presumably more reliable than those from section 6.2.1. The probability of false-negative "risk families" was lower than in the complete data set, as there were for sure people included living outside the catchment area without reporting of CRC cases to MCR.

The likelihood surfaces of the four hereditary mechanisms considered in this application are shown in figure 6.5. Again, the shape of the likelihood surface did not change very much with varying $p^{\star}_{\text{inh}}$. The problems first seen in the simulation study (see section 6.1.2) and confirmed in the application to the complete data set (see section 6.2.1) can be seen here, too: The grid search showed very high prevalence p_1 especially for low inheritance $p^{\star}_{\text{inh}}$. The minimum estimated prevalence was 0.70, which was higher

than in the complete data set and far higher than in the literature found for the general population, but one has to note that it was estimated in a selected population (see section 3.2). Again, p_1 was estimated to 1 for the lowest probability of inheritance $p^{\star}_{\text{inh}} = 0.25$ (see fig. 6.5). The relative risk RR_{fam} was estimated to values between 5 and 7. The estimates of RR_{fam} increased with increasing $p^{\star}_{\text{inh}}$.

The Bayesian posterior score was compared to the NACRC questionnaire using ROC curves and AUCs (see section 5.5). The parameters estimated by means of grid search were used to calculate the posterior risk score (see section 5.2). The NACRC score was provided by the questionnaires filled in by the families enrolled (see section 4). The sum of "yes" answers was the resulting risk score of the NACRC questionnaire (see section 3.2.3). With those two risk scores, a ROC curve was drawn for each score and AUCs were calculated for easy comparing the risk scores.

The ROC curves are depicted in figure 6.6. The resulting AUCs and their corresponding confidence intervals for the Bayesian posterior score using four hereditary probabilities $p^{\star}_{\text{inh}}$ and for the NACRC questionnaire are given in table 6.4. The results got a little bit better here than for the complete data set (see tab. 6.3). This may have been due to less false-negative "risk families" and therefore due to better verification possibility in the local family data set by means of record linkage with the MCR data base. However, one has to be aware of only four "risk families" in the local family data set (see section 4.4).

The method of the Bayesian risk score led to AUCs between 0.89 and 0.92. Again, the RTMo.25 setting with $p^{\star}_{\text{inh}} = 0.25$ showed the lowest AUC with 0.89. The hereditary mechanisms RTMo.50, RTMo.75 and CTM were very similar in their results confirming only small effect of varying the inheritance probability $p^{\star}_{\text{inh}}$ on the AUC. Results remained stable despite the chosen hereditary mechanism. No confidence interval included the critical value 0.5. The NACRC score remained quite stable at an AUC of 0.9020 with a slightly broader confidence interval confirming good discrimination ability between "risk" and "normal" families.

Table 6.4: AUC of ROC curves (see fig. 6.6) for several $p^{\star}_{\text{inh}}$ for the local family data set of the `family study` ($N = 73$ families, 461 family members; see also fig. 4.1). The AUC for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with estimated parameters by means of grid search. 95 % confidence intervals (95 % CI) were calculated using the method proposed by [16].

$p^{\star}_{\text{inh}}$	AUC	95 % CI
0.25	0.8877	0.7675; 1
0.50	0.9239	0.8291; 1
0.75	0.9239	0.8272; 1
1.00	0.9239	0.8248; 1
quest.	0.9020	0.7939; 1

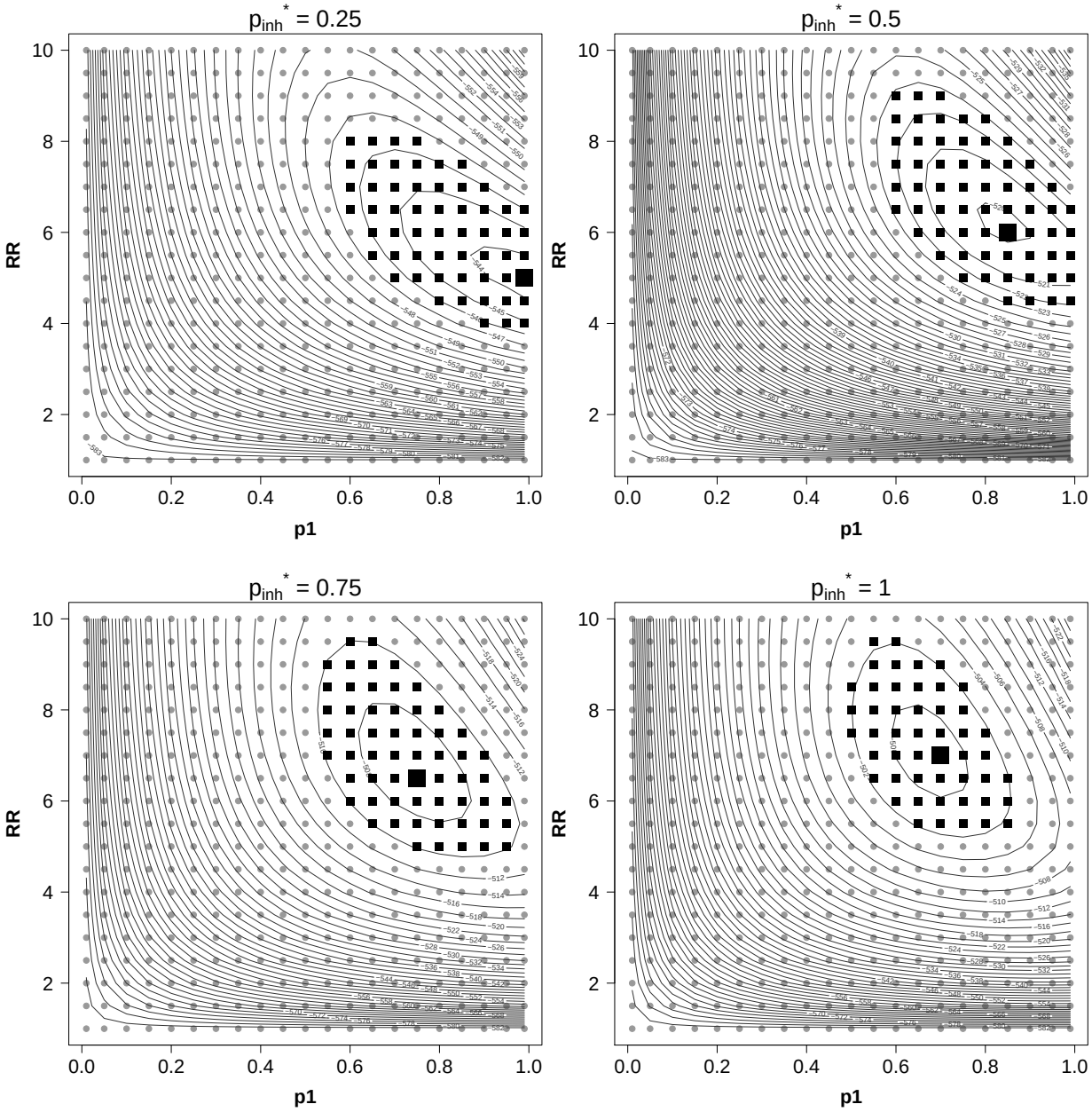

Figure 6.5: Contour plot of the likelihood surface for several $p^{\star}_{\text{inh}}$ for the local family data set of the `family study` ($N = 73$ families, 461 family members; see also fig. 4.1). The grey round points show the grid. The big squared black dot shows the maximum of the likelihood surface. The small squared black dots show the likelihood confidence region around the maximum.

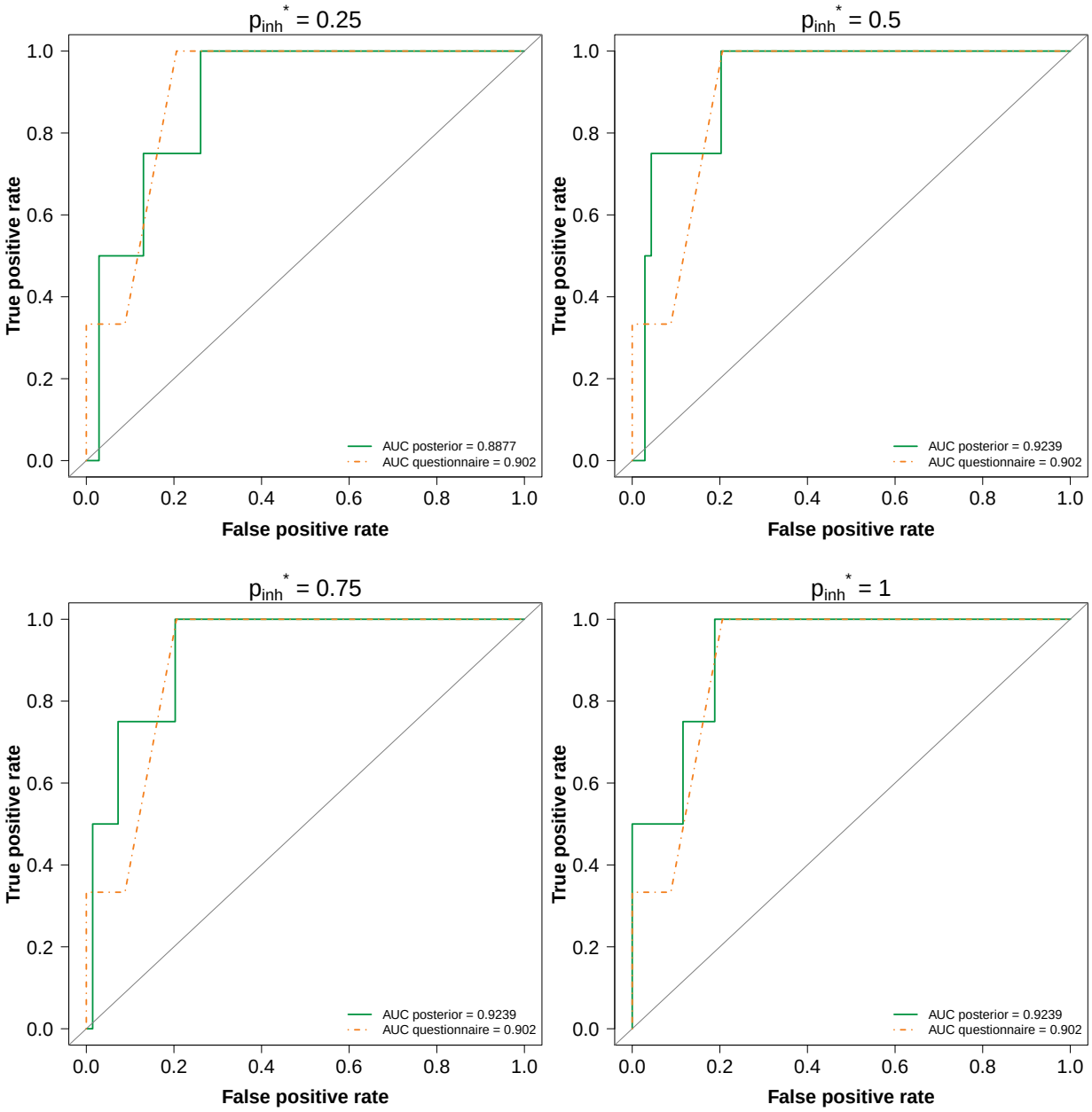

Figure 6.6: ROC curves for several $p^{\star}_{\text{inh}}$ for the local family data set of the `family study` ($N = 73$ families, 461 family members; see also fig. 4.1). The ROC curve for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with estimated parameters by means of grid search.

6.2.2 Plugging in of Parameters

Because the simulation study showed biased results for parameter estimation in a selected population (see section 6.1.2), the application of grid search to the data of the `family study` may not be reliable. In order to get more trustworthy results with respect to the posterior score, parameters from epidemiological research were plugged in the formulas from section 5.2 to calculate a posterior score based on population estimates. For the Weibull distribution, $k = 4$ and $\lambda_0 = 0.0058$ were used as before (see 5.2.3). The prevalence was set to $p_1 = 0.2$, as a familial background was assumed for approximately 20 % of the CRC cases in Germany [9, 21, 27, 45, 67, 72] (see section 3.2). The relative risk for men was set to $RR_{\text{sex}} = 2$ [4, 5, 42] (see section 5.2.3 and fig. 5.1). According to literature, the relative familial risk was also set to $RR_{\text{fam}} = 2$ [1, 5, 21, 26, 37, 69]. As done in the sections before (6.1 and 6.2.1), four probabilities of inheritance were considered: CTM with $p^{\star}_{\text{inh}} = 1.00$, and three RTM settings with $p^{\star}_{\text{inh}} \in \{0.75, 0.50, 0.25\}$.

The parameters explained above were plugged in formula (5.2) to calculate the posteriors of each family both in the complete and in the local family data set. Uncertainty of the parameters plugged in from literature was not accounted for.

Application to Complete Data Set Descriptive analysis of the complete data set of the `family study` is given in section 4.3.

As for grid search (section 6.2.1), having at least two verified CRC cases in the family was used as definition of being a "risk family". Again, the possibility of false-negative "risk families" was present due to family members living outside the catchment area of the MCR and non-obligatory reporting to the MCR (see section 4).

The comparison of the NACRC questionnaire with the Bayesian posterior score was done by means of ROC curves and AUCs (see section 5.5). The ROC curves can be seen in figure 6.7 and the AUCs can be found in table 6.5.

Using the population based parameters from epidemiological research, the posterior score as well as the NACRC questionnaire score showed good results with respect to discrimination between "risk families" and "normal" families without familial CRC burden.

The method of detailed family anamnesis with subsequent calculation of the Bayesian posterior score showed for all four $p^{\star}_{\text{inh}}$ a slightly lower AUC than the score of the NACRC questionnaire. The best, i. e. highest AUC was reached with the hereditary mechanism RTMo.25, i. e. with the lowest $p^{\star}_{\text{inh}}$ regarded (AUC $= 0.90$). However, the results were nearly the same for every scenario, so choice of hereditary mechanism did not affect the efficiency of the posterior score in discriminating between "risk families" and families at no risk. The NACRC questionnaire had also a high AUC of 0.92 and reached therefore slightly better classification quality than the posterior score. The confidence intervals of the AUCs were narrower for the NACRC questionnaire than for the posterior score. No lower bound of a confidence interval reached 0.5.

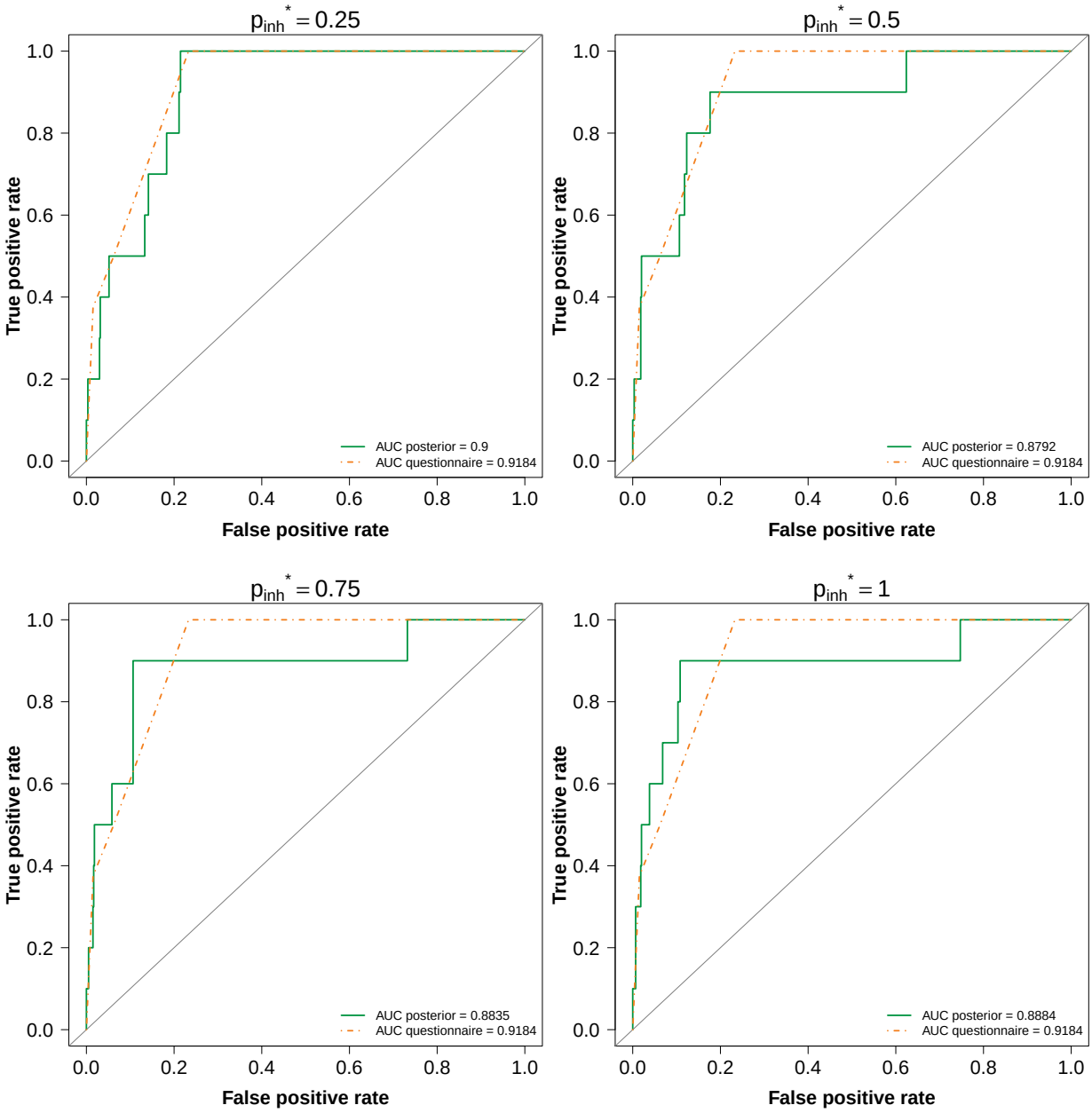

Figure 6.7: ROC curves for several $p^{\star}_{\text{inh}}$ for the complete data set of the `family study` ($N = 611$ families, 4115 family members; see also fig. 4.1). The ROC curve for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with plugged-in parameters.

Table 6.5: AUC of ROC curves (see fig. 6.7) for several $p^{\star}_{\text{inh}}$ for the complete data set of the `family study` ($N = 611$ families, 4115 family members; see also fig. 4.1). The AUC for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with plugged-in parameters. 95 % confidence intervals (95 % CI) were calculated using the method proposed by [16].

$p^{\star}_{\text{inh}}$	AUC	95 % CI
0.25	0.9000	0.8441; 0.9559
0.50	0.8792	0.7623; 0.9961
0.75	0.8835	0.7462; 1
1.00	0.8884	0.7474; 1
quest.	0.9184	0.8670; 0.9698

Application to Local Family Data Only A description of the local family data can be found in section 4.4.

"Risk families" were defined as before by having at least one other proved CRC case within the family besides the recruited patient. This definition led to less false-negative "risk families" in the local data set. Nevertheless, uncertainty about fully reporting of all family members and no obligatory reporting of CRC cases to the MCR may have led to false-negatives even in the local family data set.

The epidemiological parameters (see above) were used to calculate the posterior risk score (see section 5.2). Uncertainty of those parameters was ignored. The NACRC score was provided by the families enrolled by filling in the questionnaire (see section 4). The resulting ROC curves for both risk scores (see section 5.5) can be found in fig. 6.8. The respective AUCs and their corresponding confidence intervals for the Bayesian posterior score using four probabilities of inheritance $p^{\star}_{\text{inh}}$ and for the NACRC questionnaire are given in table 6.6.

The AUC showed a slightly better outcome for plugging in the epidemiological parameters from literature compared to grid search in a selected study population. All AUCs of the Bayesian posterior score were between 0.92 and 0.97. The results were quite similar to each other despite the chosen hereditary mechanism confirming only small effect of varying the inheritance probability $p^{\star}_{\text{inh}}$ on the AUC. However, one has to be aware of the fact that there were only four families defined as "risk family" in the local family data set (see section 4.4). The AUC for the questionnaire lay below the AUCs for the posterior. It was about 0.90 here. That provided a very good ability to discriminate between "risk families" and "normal" families. No confidence interval was containing 0.5 for this application. The confidence intervals of the posterior risk score were narrower than that of the NACRC questionnaire.

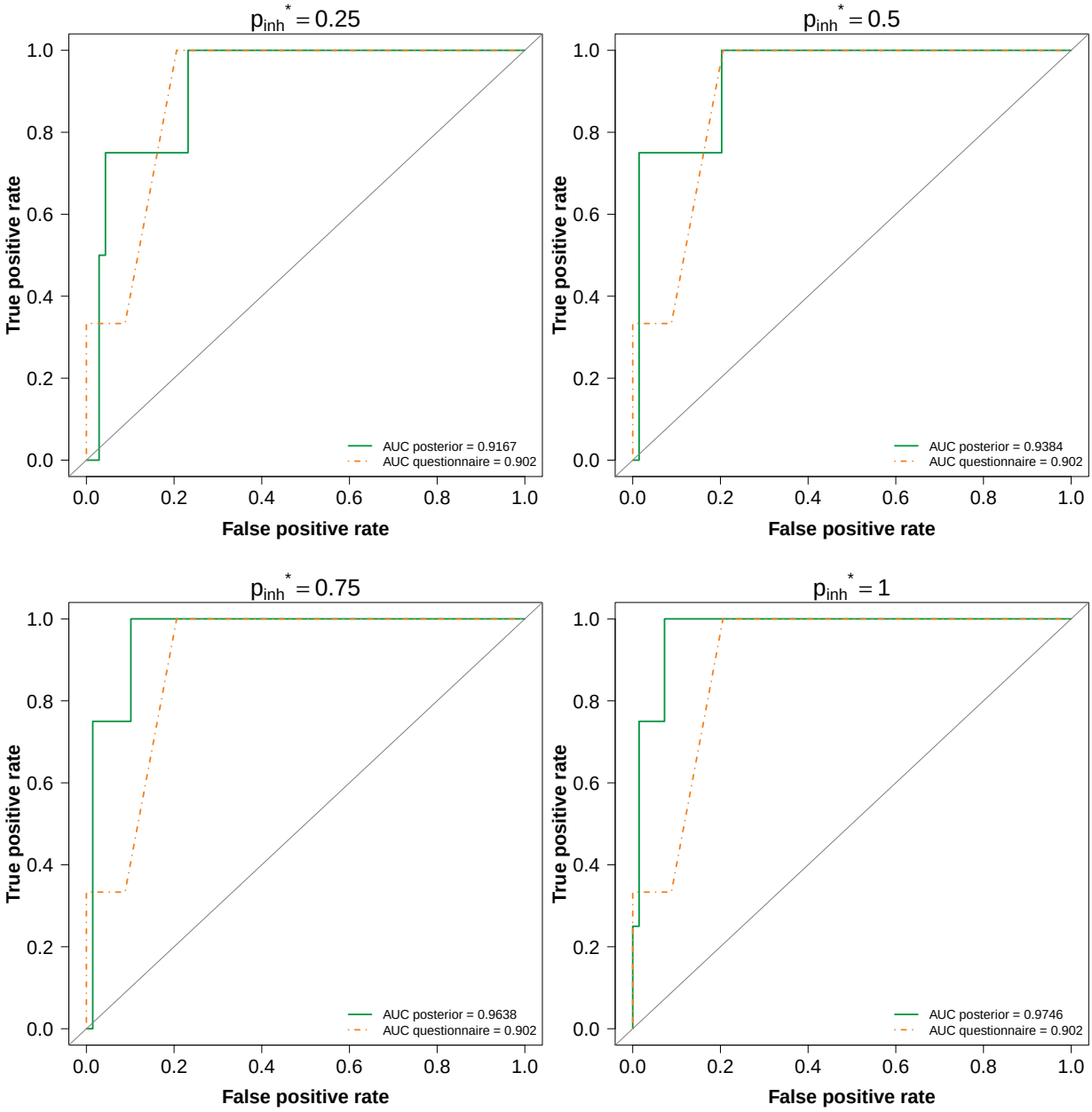

Figure 6.8: ROC curves for several $p^{\star}_{\text{inh}}$ for the local family data set of the `family study` ($N = 73$ families, 461 family members; see also fig. 4.1). The ROC curve for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with plugged-in parameters.

Table 6.6: AUC of ROC curves (see fig. 6.8) for several $p^{\star}_{\text{inh}}$ for the local family data set of the `family study` ($N = 73$ families, 461 family members; see also fig. 4.1). The AUC for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with plugged-in parameters. $95\,\%$ confidence intervals ($95\,\%$ CI) were calculated using the method proposed by [16].

$p^{\star}_{\text{inh}}$	AUC	95 % CI
0.25	0.9167	0.8091; 1
0.50	0.9384	0.8395; 1
0.75	0.9638	0.9102; 1
1.00	0.9746	0.9344; 1
quest.	0.9020	0.7939; 1

7
Discussion

This chapter contains a discussion of the method described in section 5.2 and its estimation (see section 5.3) as well as of the construction of the simulation study described in section 5.4. Furthermore, the `family study` (see section 4) is discussed. The discussion of the results (see section 6) follows at the end of this section.

7.1 Methods

The Bayesian risk score was introduced in section 5.2. Bayes' theorem was used to calculate a posterior risk of being a family with familial CRC burden. For that purpose, the family's cancer history and family tree were needed. In this thesis, several penetrance model (see section 5.2.1) and hereditary mechanisms (see section 5.2.2) were considered. Some parameters like that of the Weibull distribution used as penetrance model were

pre-determined according to known data such as incidence rates from MCR or according to literature (see section 5.2.3).

7.1.1 Bayesian Risk Score

There are tools existing for a similar application with breast cancer (e. g. [34]). Those applications mostly use a frequentist way to analyse (cf. section 5.1). Having an *a priori* probability of being susceptible and an *a posteriori* probability that takes into account the specific CRC history and constellation of the family members within the family tree is one advantage of a Bayesian approach. With a growing data base, the prediction can thus be refined.

The Bayesian risk score provides a probability for being a "risk family", i. e. a family with susceptibility to familial clustering of CRC. Those probabilities are suitable to distinguish between "risk families" and families at no risk for familial CRC as seen in the results (section 6). The risk is melted down into one probability. However, it gives a detailed impression of the extent of the familial burden. This continuous scale Bayesian risk score is therefore more precise than a discrete score arising from e. g. a questionnaire. A cut-off value can be determined for screening decisions. On the other hand, this cut-off value may be determined on an individual basis. Every person can decide about his or her personal risk threshold that can be much more detailed than a discrete score.

The probability of being a "risk family" is modelled depending on the latent class "risk carrier". By using this latent risk variable r_i, it is possible to calculate a posterior risk score. The clue of the suggested approach is, that the persons even within a family are independent from each other, if the risk carrier property r_i is given. This is the case, if r_i is treated as latent variable and regarded as conditional variable like it happens in formula (5.2).

As also filling in a questionnaire needs time and effort to gather the informations needed (e. g. number of relatives with CRC or adenoma, e. g. which degree of relationship have the respective persons), the effort to gather the informations needed to fill a family tree and to set up a CRC history is not much higher. The anamnesis of the family's history of CRC and the family tree is needed to calculate the posterior score. With the posterior score, one is gaining much better information about the CRC risk of the respective family than with a questionnaire.

Another advantage of the suggested Bayesian risk score is the already inherent extension to more complex families with more generations, aunts and uncles, and more. Theoretically, it is easy to extend the method to more generations and broader family trees. Practically, the extension is a challenge to the family anamnesis as it takes time to collect data of high quality regarding the cancer history of a large family. Especially, the topic "colorectal cancer" may be an additional challenge, as CRC is still kind of stigmatised. Records of the family's physician or the counselling of a geneticist may help here. More (high quality) information about a family's CRC history can help to gain more precise estimation of the familial CRC risk. An application to a bigger sample

of families (see also section 7.4) can provide material to supplement e. g. an online prediction tool.

The fact that in large families may be only a part of the family tree is at familial risk for CRC, is inherent within the proposed method, too. As it is possible to calculate the probability of a specific constellation of r_i's, this circumstance can be considered. However, it is nearly impossible to catch this circumstance with any questionnaire. Furthermore, using specific constellations of r_i's, it is possible to estimate an individual person's risk by summing up the posterior probabilities of all constellations of r_i's where that person is marked as risk carrier.

Nevertheless, there is no gold standard available for a definition as "risk family", as the causing genes or other causes are unknown until now. Misclassification is another challenge to deal with. However, one does not know so far, if it is indeed misclassification due to the missing gold standard.

The posterior risk score is used to predict if a family is a "risk family", i. e. if there is more than one CRC case in the family. However, this information is also used to calculate the risk score, i. e. there is a kind of circular reasoning in the method.

The proposed method of the Bayesian risk score as well as other methods like a questionnaire can help to identify undiscovered low-penetrance genes [2] by means of analysing the DNA of members of risk families in future studies.

Penetrance Models A genetic model is composed by the penetrance model and the hereditary mechanism. The penetrance model was implemented using a Weibull distribution here (see section 5.2.1). Two variants of the effect of risk carriage on the onset of the disease are assumed in the literature. In the first variant, the familial effect is caused by a raised relative risk (RR), i. e. persons with risk carrier property show an increased rate compared to the general population based age and sex specific CRC incidence rate (hazard rate). In the second variant – in this thesis only briefly touched – risk carriers show an earlier age of onset of the disease, i. e. there is a time shift of the hazard.

A raised incidence or hazard for persons at familial risk compared to the general population is supported by several studies [1, 4, 5, 8, 14, 21, 26, 32, 33, 37, 69]. Alternatives for the penetrance modelling with relative risks RR_{sex} and RR_{fam} would be a time shift modelling or usage of e. g. Cox instead of Weibull models. The time shift approach also finds some evidence in the literature [10], but the "standard" approach is the assumption of a relative risk. Nevertheless, those two variants are not comparable directly, as one cannot transform the one approach into the other. The Weibull model is chosen for computational convenience. It should be analysed in further research, if e. g. a Cox model fits better for this approach, as the most important part of the society, i. e. young men who are most susceptible for (familial) CRC, are met poorly by the Weibull adjustment (see fig. 5.1). Another topic to analyse is the possibly introduced bias from setting the incidence of a age class to the youngest age of this interval (cf. section 5.2.3).

Hereditary Mechanisms The second component of a genetic model is the hereditary mechanism. Four mechanisms were considered here: complete risk transmission (CTM) and random risk transmission (RTM) with three different $p^{\star}_{\text{inh}}$ (25 %, 50 %, 75 %).

As the causing genes or other causes for familial clustering are not known yet, it is difficult to tell if those mechanisms project the reality in a sufficient manner. Those mechanisms considered here are quite simple in the context of genetics and mathematics. Especially the CTM setting is biologically implausible. It points more to causes of familial clustering of CRC cases like common life style, dietary conditions and sports activity. An alternative would be modelling the risk as random effects (RRE) that would even be more biologically plausible than RTM. Two variants could be used: (i) one probability per family, but different probabilities for different families or (ii) one probability for each person such that the probabilities within a family are correlated. Further research should include the topic if RRE can give more precise predictions of being a "risk family".

The chosen probability of inheritance $p^{\star}_{\text{inh}}$ applies to all families and family members regarded instead of an application of individual hereditary probabilities like in RRE setting. However, as results – especially of the simulation study – pointed out, the choice of $p^{\star}_{\text{inh}}$ does not affect the discrimination ability of the Bayesian risk score much (see also section 7.4). Moreover, the different inheritance mechanisms lead to similar results in terms of estimates of relative risk RR_{fam} and prevalence p_1 (see section 7.4).

Specifications Used Here The parameters λ_0 and k of the Weibull distribution used as penetrance model as well as the probabilities of inheritance $p^{\star}_{\text{inh}}$ used and the relative risk for men RR_{sex} were pre-determined.

The parameters of the Weibull distribution were chosen according to the age and sex specific incidence rates observed in the MCR region, where the `family study` was going on. Thus, the parameters were chosen in a way that the respective Weibull distribution fits best the general incidence in the catchment area. The Weibull model is not the most flexible way to recreate those incidence rates. Although this is the best fit that was found, it overestimates the incidence in young males and underestimates it a little bit in the older persons (men and women) (see fig. 5.1). Moreover, recurrence of CRC is included in the MCR incidence rates [71]. They show not only initial diagnoses as considered within the `family study` and the simulation study here. However, recurrence probability is assumed to be low at least for light stages of CRC [12, 24, 63], so these specifications should work in a sufficient way for this application here. Nevertheless, there is some potential to improve the prediction by taking recurrences into account, as recurrence for heavy stages of CRC is much higher [12, 24, 30].

The probabilities of inheritance were chosen in a quite practical way, as the setting of CTM with a probability of inheritance $p^{\star}_{\text{inh}} = 100\,\%$ was used as "thinking model" and the other probabilities were chosen to weaken this assumption in a clearly decreasing manner. As discussed before, the whole concept of those risk transmission assumptions is disputable, but results showed stability regardless of the choice of $p^{\star}_{\text{inh}}$ (see also

section 7.4).

The relative risk for men was set to $RR_{sex} = 2$, as there is some evidence in the literature for this value [4, 5, 42]. Moreover, the fitting to the MCR incidence rates was sufficient with this value (see fig. 5.1).

Estimation of Bayesian Risk Scores The estimation of the parameters needed for the calculation of the Bayesian risk score was done by grid search here. An alternative way to estimate these parameters is an EM algorithm.

The grid search approach is a straight-forward and very flexible method to get estimates of almost any setting one is interested in. The consumption of computational time can be solved by using an accelerated version of the EM algorithm [18]. The EM is also an adequate method to get estimates here, as the likelihood surfaces show only one peak. Thus, the EM cannot be stuck to a local maximum, which is a disadvantage sometimes arising with EM application, but it is almost forced to find the global maximum of the likelihood considered here.

7.1.2 Simulation Study ("*in silico*")

The proposed method of the Bayesian risk score was tested *in silico*, i. e. within a simulation study. Families of nine persons were generated, risk carrier properties were determined as well as age at CRC diagnosis, age at adenoma onset and age at death. The proposed method including grid search and calculation of the risk score was applied.

The number of siblings in the children's generation seems to be appropriate on the one hand, as [3] lists 3.45 members per family, which includes single parents, but does not account for grandparents. The number of children is maybe unrealistic in these days on the other hand, but the children transport least information of all family members, as they are the youngest persons and therefore it is not surprising if they have not developed CRC yet or have not died yet. Nevertheless, further simulations and research in this field are recommended. Different sizes of families like in reality are not regarded yet in this *in silico* study. The "`family study`" consisted of mostly small families with about six members. This is a point to consider in further research with respect to simulation studies. Further investigations are needed to analyse several missing mechanisms, e. g. if parts of the family tree are missing or similar. The determination of the individual risk carrier properties was done according to the chosen hereditary mechanism. As they are not approved yet, this application may fail in comparison to reality in future. The risk situation of the families was known in this *in silico* study, but set aside to simulate lack of knowledge.

The determination of age at CRC diagnosis was done using the Weibull model and the parameters chosen compliant with MCR incidence rates. So the simulated age at CRC diagnosis should fit the reality quite well. Polyps are a challenge in the *in silico* part of this analysis, since it is questionable, if polyps respectively adenomas are already seen in a colonoscopy when they come into existence. So, the simulated adenoma onset ten years before CRC diagnosis (see section 5.4) is maybe too long before and a simulation

of e. g. six years before CRC diagnosis would be closer to reality. The second discussion point with this simulated adenoma onset is the participation rate of 100 % in screening colonoscopy, which is also not close to reality. True participation rates are much lower (about 15 % to 25 % [13, 42]). Additionally, this assumes that screening procedures are used by 100 % of the population and no adenoma are removed during colonoscopy, which is also not true in reality (see section 3.1.1). So, age at adenoma onset is a point with potential to improve, i. e. to get closer to reality.

Another difference to the real data from the `family study` is the fully observed lifetime of each family member. It is not apparent in the `family study` data if living persons without CRC diagnosis so far develop CRC during their further lifetime. Age at death was simulated using data from the general population in Bavaria [53], such that age at death should fit the reality quite well like age at CRC diagnosis. Competing risks for death, i. e. different causes for death are not considered here. Especially, there is no difference between death because of CRC and other death causes. However, the overall probability of death given by [53] does not distinguish between these competing risks.

Grid search and subsequent calculation of the posterior risk score was done according to the proposed procedure in section 5.2 and 5.3.1. The NACRC risk score was compared to the posterior risk score. A family was defined as "risk family", if at least one member was risk carrier. Risk carrier properties were known in this simulation. For the simulated families, the NACRC questionnaire was filled in using the simulated items. The challenge was question 3 (adenoma in first-degree relatives). As the simulation of age at adenoma onset is disputable, also this point is disputable. Additionally, NACRC question 3 concerning adenomas in relatives is answered rarely in practice (see section 4), presumably due to lack of knowledge. In contrast, data for question 4 were not simulated, so the kind of overestimation of question 3 is maybe compensated with this lack of question 4.

Time consumption of grid search is also a disadvantage of this simulation study. Additionally, the impact of the right specified probability of inheritance $p^{\star}_{\text{inh}}$ should be analysed in further research. The actual simulations were done in the ideal situation with known probability of inheritance. Therefore, misclassification of the hereditary mechanism is still a topic to check. Overall, the simulation has some points where it could be improved to obtain more realistic simulated data. The insight in practical relevance of the method is then maybe better. Nevertheless, this simulation study was constructed to get a first insight how the proposed method of the Bayesian risk score works. The results of this *in silico* study already give important hints for the application of the Bayesian risk score (see also section 7.4).

7.2 Screening for Colorectal Cancer

Screening for CRC is in Germany usually done via two colonoscopies starting at the age of 55 [6, 41]. Screening is maybe officially stopped at the age of 75, since then competitive risks of death are numerous and gaining an advantage from cancer

screening is no longer guaranteed [4]. A genetics counsellor is specialised towards screening for hereditary CRC. Screening for familial CRC is usually done by means of a questionnaire. Collecting informations about the family tree and cancer history is also some kind of questionnaire.

7.2.1 Gathering Familial Data with a Questionnaire

Gathering family history data, e. g. with the help of some questionnaire, helps clearly to identify persons at increased risk. A challenge of every questionnaire is the fact, that persons do not remember properly requested topics. Probands remember the occurrence of *no* cancer in their families very good (high specificity), but do not remember properly or even do not know every occurrence of cancer (moderate sensitivity) [47, 83]. Especially the presence of adenomas, i. e. precancerous lesions, is often unknown. Even age at cancer diagnosis is seldom reported [48]. The completion of both the questionnaire as well as the family anamnesis can be supported by the records of the family's general practitioner to obtain especially higher sensitivity. Furthermore, a questionnaire has no perfect sensitivity or specificity, so it may misclassify some persons. This may also be a reason why questionnaires should not replace medical consulting, but only support it [58].

7.2.2 NACRC Questionnaire

If a CRC case occurs in a family, one can e. g. fill in the NACRC questionnaire. If the score is greater than 0, the family is regarded as at least at familial or even at hereditary risk and a screening colonoscopy for all family members may be recommended. A screening colonoscopy for all family members after the Bayesian risk score reaches a certain (personal) threshold can be a valuable alternative, as seen in the results section (see also section 7.4).

The NACRC questionnaire is introduced in detail in section 3.2.3. It contains four questions regarding first-degree relatives diagnosed with CRC, with CRC before 50 years of age, with polyps, and/or with CRC-related cancers. It was developed on behalf of the NACRC [52] to discover persons at familial or hereditary risk. It is assumed to be easy and fast to fill in. Validation with respect to reliability as well as sensitivity and specificity was already done [43, 58]. Details are provided in section 3.2.3.

Question 3 of the NACRC questionnaire (adenomas respectively polyps in first-degree relatives) is problematic to remember for probands. Missing knowledge about those precancerous intestinal lesions in family members seems to be quite common. This was seen in the `family study` here and also in other studies [58].

The NACRC questionnaire was used a little bit differently here than in clinical practice. It is designed for screening for familial and hereditary burden in the general population. Originally, if the first question is answered with "yes", the family is regarded as at familial risk. If at least one of the other questions is answered with "yes", the family is regarded as at hereditary risk. In the application here, the score is used as

continuous or at least ordinal scale, because one has the more evidence for familial risk, the more "yes" answers are given. Hereditary cases are assumed to be known in the affected families. Nevertheless, families with a high NACRC risk score should undergo genetic counselling just to be sure. Excessive care of less affected families can be reduced.

As mentioned before, the NACRC questionnaire is designed for screening in the general population. In this application here, it was also used in the population of families with incident CRC cases – both within the simulation study and within the `family study`. For a discussion of the results, see section 7.4.

The posterior score is working for small as well as very large families. In contrast, the NACRC score leads to disputable results for small families. Question 4 (at least three first-degree relatives diagnosed with a CRC-related cancer) is then maybe impossible to answer with "yes". Therefore, familial risk may be underestimated by the NACRC questionnaire in small families.

If the risk carrier property skips generations, the Bayesian posterior would be able to handle this disadvantage. The size of families and "distance" of the single family members is accounted for in the Bayesian risk score. The NACRC score relates only to direct, i. e. first-degree relatives, and therefore may miss some "risk families".

7.3 The "Family Study"

The `family study` was conducted as cross-sectional study within the catchment area of the MCR running from 2012 until 2014. Newly diagnosed CRC patients were enrolled and asked to give their family tree to the study central. The family's cancer history was gained using an anonymous record linkage with the data base of MCR.

7.3.1 Study Setting

Data especially for Germany are important, since CRC incidence and mortality vary within Europe [11].

The risk for CRC increases considerably reaching the age of 50 [42]. This was the starting point of the inclusion criteria. The originally ending point in age was 70. That is still younger than the mean age of diagnosis in Germany (71 to 75 [35]). However, familial risk seems to be lower for older persons [27]. The inclusion criteria for age were weakened during the course of the `family study` because of the poor participation rate. The interest of participating physicians was low at the beginning despite financial incentive. Therefore, missing or incomplete clarification of the patients and their families may have led to a low participation rate. However, clarification and completeness of data varied across the participating physicians. The small amount of data clearly influences the precision of the parameters in the application of the proposed method (see also section 7.4).

In addition to the poor participation rate, there were mainly small families with a median of only 6 members reported. About a third of the recruited "families" consisted of only one person and were therefore excluded from the analysis, as they are more or less useless to estimate familial risk and do not transport any information about familial CRC risk. The cause for small families may be, that some parts of the family trees were not reported and therefore missing. The study was planned such that family members themselves report their first-degree relatives with a family tree and therefore the family trees of the recruited patient grows. The motivation to participate was varying a lot, so that families with 1 to 22 members were reported. Clearly, the bigger the families are, the better inheritance is estimable. The posterior risk score works better with large families and also the NACRC questionnaire gives more reliable results (see also section 7.2). Moreover, small families result in only a few additional CRC cases despite of the recruited patients. This represents a challenge for the definition of a "risk family" (at least one other CRC case besides the recruited patient). Unfortunately, some recruited patients were not willing to tell their relatives about their CRC diagnosis, so that also those members were not able to participate and help the family trees grow. This is also a point to consider when analysing such data. Additionally, missing knowledge about CRC cases within the family is not that uncommon (see also section 7.2).

Missing contact to other family members and therefore lack of knowledge of their addresses is a problem with respect to the record linkage that used not only the names and birth dates and birth places of the persons (high weight in record linkage), but also the actual address (with a low weight in record linkage). False-negatives are possible due to the study setting, as reporting of CRC cases to the MCR is not obligatory and as the residency of the family members was not restricted to the catchment area of the MCR, but the relatives were maybe living outside this area, where verification of CRC cases is impossible by means of record linkage with the MCR data base. Those few additional CRC cases, i. e. only a few "risk families", can make the estimations unstable. Expansion of the data base is recommended for future studies to get more reliable results. The record linkage used for the `family study` is highly error tolerant [50], but respecting data protection at the same time. Generally, missing or wrong data in the family trees pose a challenge for the record linkage. Thus, the record linkage between the family trees and the data base of the MCR cannot be perfect [50]. Therefore, there may be some false negatives as well as some false positives in the analysed data sets. Another challenge of the data of the `family study` is the originally missing field for age of death in the form of the family tree. In most cases, it could be filled in by calling some relatives by the employed telephone service, but it is partially unknown in the data set. This information is important for "healthy" people, i. e. people who are not diseased with CRC, as this figure marks the CRC-free time. The age of persons, where missing values in e. g. birth year or year of death were not possible to fill in with the help of the telephone service, was fixed at 100. The impact of this fixation is not evaluated in a sensitivity analysis, but is assumed to be small. Maybe some to date undiscovered hereditary cases are included in the real data, too. This may affect the estimation of prevalence p_1 in terms of overestimation.

However, the family study was constructed as a pilot study. The aim of revealing possible strategies to gather data as well as revealing possibilities for improvement was clearly reached. This research should be followed up, as it is relevant for the society: Including the non-participating patients and their families (650 families in two and a half years) and given a lag time of ten years from adenoma to cancer [41], 2600 families with familial burden for CRC may be living within the catchment area of the MCR (containing about 4.5 m people).

7.3.2 Descriptive Analysis of the Complete Data Set

Causes for the poor participation rate were discussed before. The complete data set contained of 611 families with 4115 members and 669 CRC cases in total. There are more men than women affected, according to literature [35, 42]. The proportion of men (59 %) among the recruited patients is slightly higher than reported for Germany (54 % [35]) and Upper Bavaria (53 % [71]). Regarding only verified CRC cases, the proportion is 61 %. It remains unclear, if the prevalence among men is generally higher in Upper Bavaria for unknown reasons, or if men are just more willing to participate in the family study. The mean age of the recruited male patients (66 years) is lower than reported in literature for Germany (71 years [35]). It holds true for the female patients (64 years compared to 75 years in Germany [35]). This fact may be due to the inclusion criteria of the study, as initially younger persons were searched for (to the maximum of 70 years). This was originally constructed in contradiction to a higher mean age at diagnosis in Germany, but weakened during the course of the family study. Additionally, there was a psychological sub-study (not covered here) regarding dealing with a CRC diagnosis of patients in active family phase, i. e. the phase of life when children still living at home. For this sub-study, patients aged 50 years and younger were enrolled. This may also have reduced the mean age of the CRC patients. In total, the family study is assumed to be not representative for the population in Upper Bavaria with respect to age distribution of CRC cases.

There were 500 out of 669 CRC cases verified. This means, about 75 % were found by means of the anonymous record linkage. The remaining quarter not found may be due to address changes since diagnosis including immigration after diagnosis from outside the catchment area of the MCR or other reasons. Typographical errors and different spellings of the same name (both persons and locations) were minimised during record linkage using a so called mapping of different spellings to one. E. g. the family name “Müller” as well as the family name “Mueller” were mapped to “mueller”. With this mapped spelling “mueller”, the record linkage was conducted.

The original thought of the family study was growing family trees with the help of relatives filling out family trees with their first-degree relatives. This worked in some few families very good, leading to a size of 22 members at maximum. However, most families were smaller, i. e. less members were reported leading to a mean size of 6.74 members. It is unclear, if the complete families are reported or if parts of the family trees are still missing. This is assumed to have been reduced to a minimum by means of

the employed telephone service that called the recruited patients and asked for missing data. Nevertheless, sometimes the patients were not willing or not able to fill in the family tree. This circumstance added to a non-complete record linkage makes it difficult to estimate familial risk. The low verification rate for the "yes" answers of question 1 may be due to family members living outside the catchment area of MCR and therefore their diagnoses were not able to be found in the data base. Other reasons may be a false-negative outcome from record linkage. The reason for the high verification rate of the "no" answers may be the same.

7.3.3 Descriptive Analysis of the Local Family Data Set

The local family data set contained only those families, where all members live within the catchment area of the MCR, where a verification by means of record linkage was meaningful. The local family data set consisted of 73 families with 461 members and 79 CRC cases in total. The imbalance between the two sexes raised a little bit compared to the complete data set. There were 65 % men among the recruited patients (54 % in Germany [35] and 53 % in Upper Bavaria [71]). Regarding only verified CRC cases, the proportion did not change. As mentioned before for the complete data set, the reasons for this high proportion of male CRC cases remain unclear. Maybe, the prevalence among men is generally higher in Upper Bavaria or recruiting of patients for the `family study` was unproportionally more successful among men. However, given men are participating in screening procedures less than women [42], this reason seems to be unrealistic. The patients are slightly younger in the local family data set than in the complete data set (mean of 62 years in both sexes compared to 75 years (women) and 71 years (men) in Germany [35]). As for the complete data set, a reason may be the initially search for patients younger than 70 years. Additionally, it remains unclear, if affected persons are generally younger in Upper Bavaria compared to Germany. Maybe there is a slight selection bias towards patients with familial burden, as patients with burden are affected earlier [38] and thus, they bias the mean age towards younger ages. This is in contradiction with the low rate of "risk families" in the data set. Overall, the local family data set of the `family study` is assumed to be not representative for the population in Upper Bavaria with respect to both age and sex distribution of the CRC cases. In total, 56 out of the 79 CRC cases could have been verified (71 %). Since all family members were living within the catchment area of the MCR in the local family data set, this rate is quite disappointing. It was assumed, that the rate is higher than in the complete data set, but this is not true. Address changes may still be a reason for those missing linkages, because if the patient moved since the diagnosis and the recruited patient gave his or her address where he or she is living now, the then actual address at the time of diagnosis is still deposited in the MCR data base. Another reason may be that newer cases are not yet entered in the MCR data base or even not yet reported to MCR. So, the verification rate may be higher in a few months or years using a new record linkage. For this reason, the findings of this thesis are considered as preliminary results. However, the findings are not expected to vary a lot when

using data from a new record linkage. The family structure of the local family data set did not differ much from the complete data set. The mean number of family sizes is comparable (6.32 members vs. 6.74 members in the complete data set). Question 1 of the NACRC questionnaire was answered with "yes" in about 14 %, this could have been verified in only three cases. The proportion is in the range of the numbers found in the literature [9, 21, 27, 45, 67, 72]. The low verification rate is not attributable to a failed record linkage, because addresses may be outside the catchment area of the MCR and therefore they are not possible to be found. Nevertheless, address changes since diagnosis or still incomplete data entering or non-obligatory reporting to the MCR data base may be reasons for it.

7.4 Results

The proposed method of the Bayesian risk score was evaluated in an application to *in silico* data sets as well as to a real data example (`family study`). The risk score may be used in two different settings: screening for familial CRC risk in the general population or in risk populations like the `family study`, i. e. families with an incident CRC case. This was implemented in the simulation study with two data settings: simulated general population and simulated selected population. The selected population mimicked a setting like the `family study`. The application to the `family study` covered grid search in the study population with subsequent calculation of the posterior risk score. Additionally, estimates from literature were taken instead of estimates from grid search in a second application. Both application types were done using the complete and the local family data set of the `family study`.

7.4.1 Simulation Study

The simulating process is described in section 5.4. Families of nine persons were generated and risk carrier property, age at CRC diagnosis and age at death were determined for each member of the 500 families. Afterwards, grid search was done and the Bayesian risk score as well as the NACRC questionnaire score were calculated. Results can be found in section 6.1.

The number of simulated families was $N = 500$ and therefore in the range of the `family study`. However, the single families were slightly bigger in the *in silico* part with nine members compared to the `family study` (median: six family members in both the complete and the local family data set).

Grid Search in the Simulated General Population The average of the mean age of male patients is a little bit too low compared with data from Germany (mean of 64 years compared to 71 years [35], see section 6.1.1 and 6.1.2), but in line with findings from the `family study`. This may be due to the adjustment of the Weibull model that overestimates the risk for men younger than about 60 years (see fig. 5.1). However, these

parameters for the Weibull model fit overall best to the incidence rates from MCR. The higher mean proportion of CRC patients of 15 % compared to literature ([35] reports a lifetime risk of about 6 %) is due to the settings with a high relative risk RR_{fam} that are also included in this mean proportion.

Overall, the grid search showed good results in the simulated general population, although the grid was relatively coarse (see section B.1.1). The true parameters used for simulation of the data were found with sufficient accuracy, i. e. estimates arising from grid search lay in most settings at least within the confidence region around the true parameters. The "easy" estimation was supported by a unimodal likelihood surface.

The AUC for the NACRC questionnaire lay below the AUC for the posterior risk score in the simulated general population (see tab. 6.1). Nevertheless, also the NACRC score reached good levels of about 0.8 in AUC. Maybe it would have reached better AUCs in the simulation, if question 4 of the NACRC questionnaire (related cancers) would have been simulated, too. If RR_{fam} was increasing, the AUCs increased, too. I. e., both scores were able to differentiate "risk families" from "normal" families even better with high RR_{fam}. This seems consistent, as a higher relative risk makes it easier to detect "risk families", as the burden breaks out more often and/or earlier. If $p^{\star}_{\text{inh}}$ was increased, the AUCs remained more or less stable. This suggests, that the chosen hereditary mechanism has only neglectable impact on the result. However, these simulations were done using the known and therefore true probability of inheritance $p^{\star}_{\text{inh}}$. The impact of misclassification of $p^{\star}_{\text{inh}}$ on the results is not analysed yet. This topic is recommended to evaluate in further research.

Grid Search in the Simulated Selected Population The simulated selected population re-used data from the simulated general population. From those 500 families, only the families with at least one CRC case were taken, resulting in about 340 families per data set. Mean age of patients did not change.

The *in silico* study pointed out, that the transfer of results from the general to the study population and vice versa is questionable. The estimation of RR_{fam} and p_1 in the simulated selected population aimed to make a statement for the general population is biased (see section B.2.1). The prevalence p_1 was overestimated in every single scenario. The prevalence of "risk families" is naturally higher in a selected population of families with at least one CRC case. For statements regarding the selected population, the estimation result could be fine. The relative risk RR_{fam} was overestimated for low true RR_{fam} and underestimated for high true RR_{fam}. Maybe the sample size is too small in these selected populations to gain considerable accuracy.

Apart from the bias, the estimation worked best for the CTM setting ($p^{\star}_{\text{inh}} = 1.00$), i. e. the strongest hereditary mechanism regarded that may make estimations easier due to its strength. Generally, estimation of parameters on the level of a selected population instead of on the level of a general population, where the parameters operate, may lead to considerable bias (a kind of selection bias) regardless of the method applied.

The AUC for the NACRC score was lower than the one for the Bayesian risk score for

every setting regarded (see tab. 6.2). The NACRC score reached mainly levels of about 0.65 (up to 0.74). As mentioned before, NACRC questionnaire would maybe discriminate families better if the fourth question may have been simulated, too. The selected families receive obviously nearly the same score, i. e. there is hardly discrimination between "risk families" and "normal" families. Nevertheless, an application to a selected population is not the setting the NACRC questionnaire was designed for. So, failing in discriminating "risk families" from "normal" families may not be overrated in this simulated selected population. Differentiation in "risk families" and families at no risk of the Bayesian posterior risk score was good for all estimates from grid search. Again, AUCs increased with increasing RR_{fam}. As mentioned above, higher relative risk makes it maybe easier to detect "risk families". If $p^{\star}_{\text{inh}}$ was increasing, the AUCs remained more or less stable for the posterior risk score. It seems to have an effect on the NACRC score such that AUCs increased marginally with increasing probability of inheritance. This seems to be justified, as the more family members are indeed risk carriers, the more family members may get CRC and the better "risk families" are detectable. Nevertheless, the chosen hereditary mechanism seems to have only neglectable impact on the result of the Bayesian risk score. This is assumed to be an advantage in practice. As for the simulated general population, the impact of misclassification of $p^{\star}_{\text{inh}}$ on the results is not analysed yet in this *in silico* study.

7.4.2 Application to the Family Study

The data set and the descriptive analysis of the `family study` was already discussed in section 7.3. For both the complete and the local family data set, both grid search and plugging in of parameter estimates found in literature was performed.

Grid Search Grid search and subsequent calculation of the posterior risk score with the estimated parameters was applied to the `family study` as described in section 5.3.1. For the `family study`, no true parameters were known. According to the findings in the simulation study, some bias in the grid search, i. e. overestimation of the parameters RR_{fam} and p_1, was expected.

The results of the grid search in the complete data set for four $p^{\star}_{\text{inh}}$ are shown in fig. 6.3. The prevalence p_1 was presumably overestimated in every setting, as results of the simulated selected population suggested. The lowest prevalence estimated was 0.65, which is much higher than described in literature for the general population (see section 3.2). Prevalence p_1 reached even a value of 1 for the setting with lowest $p^{\star}_{\text{inh}}$ regarded. Estimates of RR_{fam} increased with increasing $p^{\star}_{\text{inh}}$ up to values of 5.5 on the coarse grid used here. This is also higher than expectations arising from literature (see section 3.2). The high value of p_1 may arise from to date undiscovered hereditary cases in the data. However, the high values for both p_1 and RR_{fam} rather seem to be a result of the selected subpopulation the `family study` is representing. Additionally, the small sample size may influence the precision and value of the estimated parameters.

In the simulation study, also poor results in the grid search led to very good results with respect to AUC of the Bayesian posterior score. In the application to the complete data set of the `family study`, the Bayesian risk score showed good but slightly lower AUCs than the score of the NACRC questionnaire for all four $p^{\star}_{\text{inh}}$ (see tab. 6.3). The Bayesian risk score is nevertheless comparable to the NACRC score (see fig. 6.4). AUC of the Bayesian posterior score was highest for the highest probability of inheritance $p^{\star}_{\text{inh}}$ regarded: discrimination showed very good results for the CTM setting (AUC about 0.92). The AUC of the NACRC score showed a very good result of 0.92, too. The confidence intervals of the two risk scores overlapped nearly completely. The AUCs of the Bayesian risk score were slightly lower than in the *in silico* study. Nevertheless, AUCs appear absolutely acceptable for a real data application, especially as there are only a few families with more than one CRC case (669 CRC cases in 611 families), so the number of "risk families" is small. However, some "risk families" are maybe not recognised due to incomplete record linkage with addresses outside the catchment area of the MCR. Despite these thin data the results are quite stable over the several $p^{\star}_{\text{inh}}$. The true genetic model for familial CRC is still unknown. Therefore, it is considered as an advantage, that the choice of the genetic model does not impact the results much. Nevertheless, further research with a bigger sample of families of the general population is recommended to get valid results which can be used in an online prediction tool or similar.

The results of the grid search in the local family data set for the four $p^{\star}_{\text{inh}}$ regarded are shown in fig. 6.5. The problem of overestimation seen in the simulation study and in the application to the complete data set of the `family study` led also in the application to the local family data set to very high prevalence p_1 especially for low inheritance $p^{\star}_{\text{inh}}$. The lowest prevalence estimated was 0.7 (for $p^{\star}_{\text{inh}} = 1.00$), which is even higher than for the complete data set. Prevalence p_1 was estimated higher with decreasing $p^{\star}_{\text{inh}}$. A higher maximum value could also be found for RR_{fam} compared to the application to the complete data set. The estimates increased with increasing $p^{\star}_{\text{inh}}$ up to values of 7. That is also higher than in the application to the complete data. As for the complete data set, the high value of p_1 may be due to unknown hereditary cases. However, this is supposed to arise from overestimation as the simulation study revealed. The reduction to only local families in addition to the quantity of the `family study` may tighten the situation of a selected population.

The AUC of the Bayesian posterior score was again highest for the highest $p^{\star}_{\text{inh}}$ regarded (see tab. 6.4). The AUCs of the Bayesian risk score are marginally higher than in the application to the complete data set and therefore show very good discrimination quality. This seems to be logical, since verification of CRC cases and therefore declaration as "risk family" is biased towards an increased number of false-negatives in the complete data set. The cause may lie in the non-obligatory reporting of cancer cases to the MCR. Furthermore, it is possible that parts or even complete families in the study live outside of the catchment area of the MCR and therefore a verification of CRC cases is not possible. Missing linkage possibility for family members living outside the catchment area of the MCR cannot be excluded due to e. g. address changes, but is reduced to a

minimum (see section 7.3.1). This reduced probability of false-negative "risk families" and thus better verification possibility leads to the assumption, that the results of the application to the local family data set are presumably more reliable than those from the complete data set. However, the problem of non-obligatory reporting of cancer diagnoses to the MCR is still present. Moreover, there are only four "risk families" in the local family data set. The AUC of the questionnaire showed stable results with a high value of 0.90 in comparison to the application in the complete data set. This may show that the NACRC questionnaire is a quite robust tool for identification of risk families. The confidence interval overlapped nearly completely with those of the Bayesian risk score.

Plugging in of Parameters Beside grid search, the calculation of the posterior risk score for the data of the `family study` was also done with epidemiological parameters according to literature. Biased results in the grid search are therefore prevented. Uncertainty of the parameters was not accounted for.

The AUCs of the ROC curves reached high levels of about 0.90 (see tab. 6.5). The NACRC score showed also a good result of about 0.92 in the complete data set. The AUC of the Bayesian risk score was highest for the lowest probability of inheritance ($p^{\star}_{\text{inh}} = 0.25$) regarded. However, the differences between the single $p^{\star}_{\text{inh}}$ settings were small as in the grid search application. I. e. varying the in reality unknown probability of inheritance has not much effect on the discrimination ability of the Bayesian risk score. The confidence intervals overlapped also with the AUC of the NACRC questionnaire. As the values for the parameters used for calculation of the posterior risk score did not arise from presumably biased grid search due to the application in a selected population, the discrimination between "risk families" and families without familial CRC burden worked much better in the plug-in setting than in the grid search setting. Although there were only a few families with more than one CRC case (669 CRC cases in 611 families), AUCs reached high levels and ROC curves are satisfying. These results suggest that the Bayesian risk score leads to adequate identification of "risk families".

The AUCs of the ROC curves for the plugging in of epidemiological parameters for the local family data set were about 0.94 (see tab. 6.6) with very good values for the CTM setting. The NACRC risk score dropped a little bit in discrimination ability to about 0.90 in the local family data set compared to the complete data set. In this setting, the posterior score outreached the NACRC score, but differences were small. The critical value 0.5 was never included in the confidence intervals. The AUCs differed not much at all, since there were even less "risk families" in the local family data set than in the complete data set (73 families with 79 CRC cases). Varying the probability of inheritance $p^{\star}_{\text{inh}}$ had almost no effect on the discrimination ability of the Bayesian risk score. I. e., the specific genetic model does not affect much the estimation of risk for familial CRC. Discrimination ability is high for both Bayesian and NACRC risk score. Even unrealistic respectively biologically implausible settings like CTM with $p^{\star}_{\text{inh}} = 1.00$ showed good results in terms of sensitivity and specificity. With those epidemiological

parameters, the AUCs of the Bayesian risk score are higher than using the estimates from grid search in the selected population of the `family study`. This is assumed to be created by using estimates arising from a general population like those parameters taken from literature, i. e. the parameters are not biased like in the selected population of the `family study`.

7.4.3 Résumé

The Bayesian posterior score proposed here was compared with the score resulting from the questionnaire by NACRC [52], which has only four questions regarding the family's cancer history.

The analysis of the grid search approach by means of the *in silico* study (section 5.4) showed, that this straight forward approach is helpful for the estimation of the interesting parameters RR_{fam} and p_1 in the general population. Even with only 500 families of nine members each and a relatively coarse grid, is it possible to estimate the true parameters with sufficient accuracy. Estimation worked good at least in the general population. However, even with those simple simulations, the *in silico* study showed some bias in the estimation via grid search in the selected population of families with diagnosed CRC cases. The impact of misclassification of $p^{\star}_{\text{inh}}$ was not analysed yet. However, results showed only small impact of varying $p^{\star}_{\text{inh}}$. The AUC i. e. discrimination ability of the posterior risk score was higher than that of the NACRC questionnaire.

The grid search in the data of the `family study` showed presumably some overestimation regarding a statement for the general population. This may be the reason why the NACRC questionnaire score worked marginally better than the Bayesian risk score with respect to AUC. Results of the Bayesian score were better in the application to the local family data set, as there were presumably less false-negative "risk families".

The plugging in of epidemiological parameters resulted in comparable or higher AUCs than the grid search. The performance of the Bayesian risk score was slightly better for the local family data set, but not differing much from the complete data set. The Bayesian risk score reached higher levels of AUC than the NACRC questionnaire for the local family data set and comparable values for the complete data set.

However, the quality and quantity of the data of the `family study` was not very helpful for an application of the method described above to gain knowledge for practice. An application to a bigger sample of families taken from the general population (see also section 7.4) can provide material, with which e. g. an online prediction tool can be filled.

The NACRC questionnaire may lead to misclassification as almost every questionnaire, since sensitivity and specificity are not 100 % [36]. It cannot identify all "risk families" or "risk persons". The other way round, not all persons or families identified as "at risk" are in reality at familial risk. This is also true for the posterior score. Both methods are meant to be used as screening tool and therefore should be applied to the general population. This recommendation is confirmed by the *in silico* study. Calculation of the posterior risk score can be advantageous, even if the hereditary mechanism is just

speculation. As results showed, the choice of $p^{\star}_{\text{inh}}$ does not much affect the discrimination ability of the Bayesian risk score. The questionnaire has also no huge fluctuation in its AUC through the varying settings. It always reaches a good AUC apart from the usage in the simulated selected population. However, this is a scenario the questionnaire was not designed for. Overall, those two methods applied here seem to be at least non-inferior to the NCI questionnaire (see section 3.2.3) that reaches an AUC of about 0.6 [20, 56]. A screening colonoscopy for all family members after the Bayesian risk score reached a certain (personal) threshold can be a valuable alternative to the NACRC questionnaire or the criterion for familial CRC risk of having a first-degree relative with CRC that is often used in practice.

8
Outlook

For further research, the results of the application of the proposed method to data sampled from the general population can be filled in a prediction tool, which may be provided online to a broad public. Then, many people would have the opportunity to get an CRC screening appropriate to their familial risk.

Regarding the methods used here, it is possible to combine several biological plausible models to approximate the still unknown true situation via Bayesian model averaging. This may be better than using one specific setting as done here. Generally, time shift may be an alternative for modelling the increased penetrance for persons at familial risk. The usage of all reported CRC cases instead of only the verified ones via record linkage can provide a sensitivity analysis for another part of the method. Risks as random effects, giving each family member his or her own probability of inheritance, may improve the performance of the Bayesian risk score, although performance is regarded as sufficient here.

Bibliography

[1] N. Andrieu, G. Launoy, R. Guillois, C. Ory-Paoletti and M. Gignoux, *European Journal of Cancer* **39** (2003), 1904.

[2] A. Balmain, *Nature Reviews Cancer* **1** (2001), 77.

[3] BayLfStaD. *Strukturdaten der Bevölkerung und der Haushalte in Bayern 2011 – Teil I der Ergebnisse der* 1 %*-Mikrozensuserhebung 2011 (zusammengefasste Ergebnisse) [Structural Data of the Population and Households in Bavaria 2011 – Part I of the Results of the* 1 % *Microcensus Survey 2011 (Summarized Results)]*. Statistische Berichte, 2012. Available at https://www.statistik.bayern.de/veroeffentlichungen/advanced_search_result.php?keywords=A6201C.

[4] A.M. Benhamiche-Bouvier, C. Lejeune, J.L. Jouve, S. Manfredi, C. Bonithon-Kopp and J. Faivre, *Journal of Medical Screening* **7** (2000), 136.

[5] J.L. Bermejo and K. Hemminki, *Journal of the National Cancer Institute* **97** (2005), 1575.

[6] BMG. *Leistungen der GKV: Früherkennung von Krebs. [Assurance Benefit of the Statutory Health Insurance: Early Detection of Cancer]*. PDF file on website, 2014. Available at http://www.bmg.bund.de/fileadmin/dateien/Downloads/F/Frueherkennung_und_Vorsorgeleistungen_der_GKV/Krebs_Vorsorge-_und_Frueherkennungsleistungen.pdf; December 2014 Version; German.

[7] C.R. Boland, S.N. Thibodeau, S.R. Hamilton, D. Sidransky, J.R. Eshleman, R.W. Burt, S.J. Meltzer, M.A. Rodriguez-Bigas, R. Fodde, G.N. Ranzani and S. Srivastava, *Cancer Research* **58** (1998), 5248.

[8] L. Bonelli, H. Martines, M. Conio, P. Bruzzi and H. Aste, *International Journal of Cancer* **41** (1988), 513.

[9] P. Boyle and B. Levin, editors. *World Cancer Report 2008*. International Agency for Research on Cancer, 2008.

[10] H. Brenner, M. Hoffmeister and U. Haug, *American Journal of Gastroenterology* **103** (2008), 2326.

[11] H. Brenner, M. Hoffmeister and U. Haug, *British Journal of Cancer* **99** (2008), 532.

[12] H. Brenner, M. Kloor and C.P. Pox, *The Lancet* **383** (2014), 1490.

[13] H. Brenner, P. Schrotz-King, B. Holleczek, A. Katalinic and M. Hoffmeister, *Deutsches Ärzteblatt International* **113** (2016), 101.

[14] A.S. Butterworth, J.P.T. Higgins and P. Pharoah, *European Journal of Cancer* **42** (2006), 216.

[15] M. Crabtree, O.M. Sieber, L. Lipton, S.V. Hodgson, H. Lamlum, H.J.W. Thomas, K. Neale, R.K.S. Phillips, K. Heinimann and I.P.M. Tomlinsonn, *Oncogene* **22** (2003), 4257.

[16] E.R. DeLong, D.M. DeLong and D.L. Clarke-Pearson, *Biometrics* **44** (1988), 837.

[17] A.P. Dempster, N.M. Laird and D.B. Rubin, *Journal of the Royal Statistical Society. Series B (Methodological)* **39** (1977), 1.

[18] A. Engelhardt, A. Rieger, A. Tresch and U. Mansmann. *Efficient Maximum Likelihood Estimation for Pedigree Data with the Sum-Product Algorithm* (2017). Accepted at "Human Heredity".

[19] J. Ferlay, H.R. Shin, F. Bray, D. Forman, C. Mathers and D.M. Parkin, *International Journal of Cancer* **127** (2010), 2893.

[20] A.N. Freedman, M.L. Slattery, R. Ballard-Barbash, G. Willis, B.J. Cann, D. Pee, M.H. Gail and R.M. Pfeiffer, *Journal of Clinical Oncology* **27** (2009), 686.

[21] C.S. Fuchs, E.L. Giovannucci, G.A. Colditz, D.J. Hunter, F.E. Speizer and W.C. Willett, *New England Journal of Medicine* **331** (1994), 1669.

[22] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari and D.B. Rubin: *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science, 3. edition. CRC Press, 2013.

[23] German Guideline Program in Oncology. *Evidenced-based Guideline for Colorectal Cancer*. PDF file on website, 2014. Available at http://leitlinienprogramm-onkologie.de/Leitlinien.7.0.html; long version 1.0; AWMF registration number: 021-007OL.

[24] R.J. Heald and R.D.H. Ryall, *The Lancet* **327** (1986), 1479.

[25] L. Held and D.S. Bové: *Applied Statistical Inference – Likelihood and Bayes*. Springer, 2014.

[26] K. Hemminki and X. Li, *International Journal of Cancer* **94** (2001), 743.

[27] K. Hemminki, J. Sundquist and J.L. Bermejo, *Annals of Oncology* **19** (2008), 163.

[28] M. Hirsch-Kaufmann, M. Schweiger and M.R. Schweiger: *Biologie und molekulare Medizin: für Mediziner und Naturwissenschaftler*. 7. edition. Georg Thieme Verlag, 2009.

[29] M. Hoffmeister, S. Schmitz, E. Karmrodt, C. Stegmaier, U. Haug, V. Arndt and H. Brenner, *Clinical Gastroenterology and Hepatology* **8** (2010), 870.

[30] R.H. Huebner, K.C. Park, J.E. Shepherd, J. Schwimmer, J. Czernin, M.E. Phelps and S.S. Gambhir, *The Journal of Nuclear Medicine* **41** (2000), 1177.

[31] T.F. Imperiale and D.F. Ransohoff, *Annals of Internal Medicine* **156** (2012), 703.

[32] D.J.B.S. John, F.T. McDermott, J.L. Hopper, E.A. Debney, W.R. Johnson and E.S.R. Hughes, *Annals of Internal Medicine* **118** (1993), 785.

[33] L.E. Johns and R.S. Houlston, *American Journal of Gastroenterology* **96** (2001), 2992.

[34] M.A. Jonker, C.E. Jacobi, W.E. Hoogendoorn, N.J.D. Nagelkerke, G.H. de Bock and J.C. van Houwelingen, *Cancer Epidemiology, Biomarkers & Prevention* **12** (2003), 1479.

[35] P. Kaatsch, C. Spix, S. Hentschel, A. Katalinic, S. Luttmann, C. Stegmaier, S. Caspritz, J. Cernaj, A. Ernst, J. Folkerts et al., *Beiträge zur Gesundheitsberichterstattung des Bundes* **9** (2013).

[36] A. Katalinic, H. Raspe and A. Waldmann, *Zeitschrift für Gastroenterologie* **47** (2009), 1125.

[37] Z. Kemp, C. Thirlwell, O. Sieber, A. Silver and I. Tomlinson, *Human Molecular Genetics* **13** (2004), 177.

[38] E. Kharazmi, M. Fallah, K. Sundquist and K. Hemminki, *BMJ* **345** (2012), e8076.

[39] P. Klare, S. Ascher, A. Hapfelmeier, P. Wolf, A. Beitz, R.M. Schmid and S. von Delius, *World Journal of Gastroenterology* **21** (2015), 525.

[40] T. Klein, *Zeitschrift für Soziologie* **25** (1996), 346.

[41] F.T. Kolligs, *Deutsche medizinische Wochenschrift* **140** (2015), 1425.

[42] F.T. Kolligs, A. Crispin, A. Munte, A. Wagner, U. Mansmann and B. Göke, *PLoS One* **6** (2011), e20076.

[43] I. Koné, A. Siebenhofer, J. Hartig and J. Plath, *Das Gesundheitswesen* (2016), [Epub ahead of print].

[44] S.J. Laken, G.M. Petersen, S.B. Gruber, C. Oddoux, H. Ostrer, F.M. Giardiello, S.R. Hamilton, H. Hampel, A. Markowitz, D. Klimstra, S. Jhanwar, S. Winawer, K. Offit, M.C. Luce, K.W. Kinzler and B. Vogelstein, *Nature Genetics* **17** (1997), 79.

[45] R. Mendelsohn and A.J. Markowitz, *European Gastroenterology and Hepatology Review* **7** (2011), 251.

[46] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel and F. Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien* (2015). Version 1.6-7.

[47] R.J. Mitchell, D. Brewster, H. Campbell, M.E.M. Porteous, A.H.W. und C. C. Bird and M.G. Dunlop, *Gut* **53** (2004), 291.

[48] H.J. Murff, D. Byrne and S. Syngal, *American Journal of Preventive Medicine* **27** (2004), 239.

[49] J.D. Murken, T. Grimm, E. Holinski-Feder and K. Zerres: *Taschenlehrbuch Humangenetik*. 8. edition. Georg Thieme Verlag, 2011.

[50] D. Nasseh, J. Engel, U. Mansmann, W. Tretter and J. Stausberg, *Studies in Health Technology and Informatics* **205** (2014), 808.

[51] R.C. Neath, *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton* **10** (2013), 43.

[52] N.N. *Netzwerk gegen Darmkrebs e. V. [Network against colorectal cancer – NACRC]*. Website, 2013. Available at www.netzwerk-gegen-darmkrebs.de; 2013 Version; German.

[53] N.N. *GENESIS-Online (Bayern)*. Website, 2014. Available at www.statistikdaten.bayern.de; Version V3.600P4.

[54] N.N. *Bevölkerung auf Grundlage des Zensus 2011 [Population Based on the 2011 Census]*. Table on Website, 2017. Available at www.destatis.de.

[55] N.N. *Mitglieder und mitversicherte Familienangehörige der gesetzlichen Krankenversicherung am 1.7. eines Jahres (Anzahl) [Number of Members and Jointly Insured Family Members of the Statutory Health Insurance on July 1st of the Rspective Year]*. Table on Website, 2017. Available at www.gbe-bund.de.

[56] Y. Park, A.N. Freedman, M.H. Gail, D. Pee, A. Hollenbeck, A. Schatzkin and R.M. Pfeiffer, *Journal of Clinical Oncology* **27** (2009), 694.

[57] G.M. Petersen and J.D. Brensinger, *Oncology* **10** (1996), 89.

[58] C. Pieper, I. Kolankowska and K.H. Jöckel, *European Journal of Cancer Care* **21** (2012), 758.

[59] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models* (2017). Version 3.1-131.

[60] J. Plath, A. Siebenhofer, I. Koné, M. Hechtner, S. Schulz-Rothe, M. Beyer, F.M. Gerlach and C. Güthlin, *Family Practice* **34** (2017), 30.

[61] J. Plath, A. Siebenhofer, S. Schulz-Rothe and C. Güthlin, *Das Gesundheitswesen* (2017), [Epub ahead of print].

[62] O. Pötzsch and D. Emmerling: *Geburten und Kinderlosigkeit in Deutschland – Bericht über die Sondererhebung 2006 "Geburten in Deutschland". [Births and Childlessness in Germany - Report on the Special Ascertainment 2006 "Births in Germany"]*. Statistisches Bundesamt, Wiesbaden, 2008.

[63] J.N. Primrose, R. Perera, A. Gray, P. Rose, A. Fuller, A. Corkhill, S. George and D. Mant, *Journal of the American Statistical Association* **311** (2014), 263.

[64] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2017). R version 3.3.3.

[65] M.A. Rodriguez-Bigas, C.R. Boland, S.R. Hamilton, D.E. Henson, J.R. Jass, P.M. Khan, H. Lynch, M. Perucho, T. Smyrk, L. Sobin and S. Srivastava, *Journal of the National Cancer Institute* **89** (1997), 1758.

[66] C.P. Schaaf and J. Zschocke: *Basiswissen Humangenetik*. 2. edition. Springer-Verlag Berlin Heidelberg, 2013.

[67] R.E. Schoen, A. Razzak, K.J. Yu, S.I. Berndt, K. Firl, T.L. Riley and P.F. Pinsky, *Gastroenterology* **149** (2015), 1438.

[68] T. Sing, O. Sander, N. Beerenwinkel and T. Lengauer, *Bioinformatics* **21** (2005), 3940.

[69] M.L. Slattery and R.A. Kerber, *Journal of the National Cancer Institute* **86** (1994), 1618.

[70] T.M. Therneau. *A Package for Survival Analysis in S* (2015). Version 2.40-1.

[71] TRM. *Tumorstatistik: Basisstatistiken – C18–C21: Darmtumor*. PDF file on website, 2014. Available at http://tumorregister-muenchen.de/facts/base/bC1820G-ICD-10-C18-C20-Darmtumor-Inzidenz-und-Mortalitaet.pdf; March 2014 Version; German.

[72] J. Trojan, *Versicherungsmedizin* **63** (2011), 132.

[73] A. Umar, C.R. Boland, J.P. Terdiman, S. Syngal, A. de la Chapelle, J. Rüschoff, R. Fishel, N.M. Lindor, L.J. Burgart, R. Hamelin, S.R. Hamilton, R.A. Hiatt, J. Jass, A. Lindblom, H.T. Lynch, P. Peltomaki, S.D. Ramsey, M.A. Rodriguez-Bigas, H.F.A. Vasen, E.T. Hawk, J.C. Barrett, A.N. Freedman, and S. Srivastava, *Journal of the National Cancer Institute* **96** (2004), 261.

[74] H.F.A. Vasen, P. Watson, J. Mecklin, H.T. Lynch and the ICG–HNPCC, *Gastroenterology* **116** (1999), 1453.

[75] N. Walach, I. Novikov, I. Milievskaya, G. Goldzand and B. Modan, *Cancer* **82** (1998), 180.

[76] E. Wallace, A. Hinds, H. Campbell, J. Mackay, R. Cetnarskyj and M.E.M. Porteous, *British Journal of Cancer* **91** (2004), 1575.

[77] G.R. Warnes, B. Bolker, L. Bonebakker, R. Gentleman, W.H.A. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz and B. Venables. *gplots: Various R Programming Tools for Plotting Data* (2016). R package version 3.0.1.

[78] G.C.G. Wei and M.A. Tanner, *Journal of the American Statistical Association* **85** (1990), 699.

[79] K. Weigl, L. Jansen, J. Chang-Claude, P. Knebel, M. Hoffmeister and H. Brenner, *International Journal of Cancer* **139** (2016), 2213.

[80] A. Wienecke, B. Barnes, T. Lampert and K. Kraywinkel, *International Journal of Cancer* **134** (2014), 682.

[81] S.J. Winawer, A.G. Zauber, M.N. Ho, M.J. O'Brien, L.S. Gottlieb, S.S. Sternberg, J.D. Waye, M. Schapiro, J.H. Bond, J.F. Panish, F. Ackroyd, M. Shike, R.C. Kurtz, L. Hornsby-Lewis, H. Gerdes, E.T. Stewart and the National Polyp Study Workgroup, *New England Journal of Medicine* **329** (1993), 1977.

[82] S.N. Wood, *Journal of the Royal Statistical Society. Series B (Methodological)* **73** (2011), 3.

[83] A. Ziogas and H. Anton-Culver, *American Journal of Preventive Medicine* **24** (2003), 190.

[84] B. Zöller, X. Li, J. Sundquist and K. Sundquist, *European Journal of Cancer* **50** (2014), 2319.

List of Figures

List of Tables

A
Incidence Rates from MCR

The age and sex specific incidence rates given by MCR were used to determine the Weibull distribution (see section 5.2.3). Those incidence rates are given in the base statistics published online by MCR. They can be downloaded as pdf file from http://tumorregister-muenchen.de/facts/base/bC1820G-ICD-10-C18-C20-Darmtumor-Inzidenz-und-Mortalitaet.pdf. In this thesis, the base statistics for cancer diagnoses with ICD-10 codes C18–C21 were used in the version as of March 2014 that is given on the next site.

The specific table 5 is showing the age specific incidence, DCO rate and proportion of all cancer for the years 1998–2012. The abbreviation "DCO" stands for "death certificate only". It describes a cancer diagnosis where the information about it was given to MCR only by death certificate and not by a report of the treating physician.

The first column gives the age class at diagnosis in years. The second and third column gives the total number of cases for men and women, respectively. The next two columns are the most important columns for this thesis, as they show the age specific incidence for men and women, given in diagnoses per 100 000 persons within the respective class. It describes the risk for disease in the respective age class, as explained in the lines at the bottom. Those columns are followed by the DCO rate and the proportion of all cancers, each in percent for men and women, respectively. At the bottom of the table, sex specific incidences are given in raw numbers and in standardised numbers using world standard (WS), (old) European standard (ES) and German standard (BRD-S).

Table 5

Age-specific incidence, DCO rate and proportion of all cancers
for period 1998-2012

Age at diagnosis Years	Males n	Females n	Males Age-spec. incid.	Females Age-spec. incid.	Males DCO rate n=1078 %	Females DCO rate n=1659 %	Males Prop.all cancers n=146755 %	Females Prop.all cancers n=142297 %
0- 4			0.0	0.0				
5- 9			0.0	0.0				
10-14	1	4	0.1	0.3			0.7	2.5
15-19	4	17	0.3	1.2			1.3	6.4
20-24	8	21	0.5	1.3	12.5		1.4	4.3
25-29	30	42	1.6	2.2			3.4	4.1
30-34	78	65	3.7	3.2			5.5	3.4
35-39	140	116	6.0	5.2		1.7	6.6	3.3
40-44	312	264	12.9	11.5	0.6	0.4	10.4	4.5
45-49	588	488	27.3	23.1	0.9	1.4	11.9	6.1
50-54	1102	813	59.6	43.0	1.8	1.0	13.7	8.0
55-59	1953	1243	114.9	69.8	1.6	1.4	14.4	9.7
60-64	3031	1797	183.9	103.3	1.9	2.0	14.8	11.1
65-69	3665	2089	249.8	130.3	2.6	2.5	14.3	11.8
70-74	3794	2497	327.4	181.1	3.6	4.5	15.5	14.8
75-79	3142	2763	417.0	252.6	5.7	6.6	16.6	17.0
80-84	2236	2925	492.4	338.7	9.0	10.3	17.8	19.8
85+	1594	3414	514.0	416.8	21.7	27.5	17.4	21.3
All ages	21678	18558			5.0	8.9	14.8	13.0
Incidence								
Raw			79.0	64.7				
WS			41.4	25.9				
ES			62.1	39.0				
BRD-S			80.7	50.8				

The age-specific incidence characterizes the disease risk in a particular age group. The age distribution depends on the patient population frequency in each age group and reflects the tangible clinical picture of everyday patients care (see following chart).

B
Results of the Simulation Study

The results of the *in silico* study are given in section 6.1. The detailed procedure of the simulation itself is described in section 5.4. In short, 500 families of nine persons were simulated according to epidemiological research regarding age distribution and cancer incidences. Grid search as described in section 5.3.1 was applied to those simulated data and the Bayesian posterior score as introduced in section 5.2 was calculated. The NACRC score was also calculated using the simulated informations of each family.

B.1 Grid Search in the Simulated General Population

The results for the simulated general population can be found in section 6.1.1. There, only a few randomly chosen likelihood surfaces are shown in figure 6.1. The likelihood surfaces arose from grid search, the method used to estimate the interesting parameters RR_{fam} and p_1. The Bayesian posterior score was used to discriminate between "risk families" and families at no raised risk. The discrimination ability was analysed using ROC curves and AUCs. The AUCs are given in table 6.1 in the main part of this thesis, the ROC curves are given here in section B.1.2.

B.1.1 Likelihood Surfaces

All likelihood surfaces arising from grid search are depicted here, including those already shown in figure 6.1 (section 6.1.1). For each $p^{\star}_{\text{inh}}$, a panel of graphs is shown as there were eight RR_{fam} considered.

$p^{\star}_{\text{inh}} = 0.25$ The resulting likelihood surfaces for $p^{\star}_{\text{inh}} = 0.25$ are given in figure B.1. For low RR_{fam}, the likelihood surface had no clear peak that made it maybe difficult to estimate properly. For increasing RR_{fam}, the peak got more and more pointed. Nevertheless, for $RR_{\text{fam}} = 12$ the true parameters lay not within the confidence region around the estimate.

$p^{\star}_{\text{inh}} = 0.50$ The resulting likelihood surfaces of RTMo.50 with $p^{\star}_{\text{inh}} = 0.50$ are given in figure B.2. Again, the likelihood was quite flat for low RR_{fam} and got more peaked for higher RR_{fam}. Although the peak of the surface was a little bit elongated, estimation by means of grid search worked well.

$p^{\star}_{\text{inh}} = 0.75$ The graphs for $p^{\star}_{\text{inh}} = 0.75$ can be found in figure B.3. The resulting surfaces for $p^{\star}_{\text{inh}} = 0.75$ did not differ much from those of the other $p^{\star}_{\text{inh}}$. Again, the true parameters lay always within the confidence region around the estimated parameters.

$p^{\star}_{\text{inh}} = 1.00$ The graphs of the CTM setting with $p^{\star}_{\text{inh}} = 1.00$ are given in figure B.4. The resulting surfaces had a very similar shape to those of the other $p^{\star}_{\text{inh}}$ regarded. Grid search estimation of the true parameters worked well in the CTM setting. The true parameters were missed in only one setting (with $RR_{\text{fam}} = 3$), but the confidence region included the true parameters.

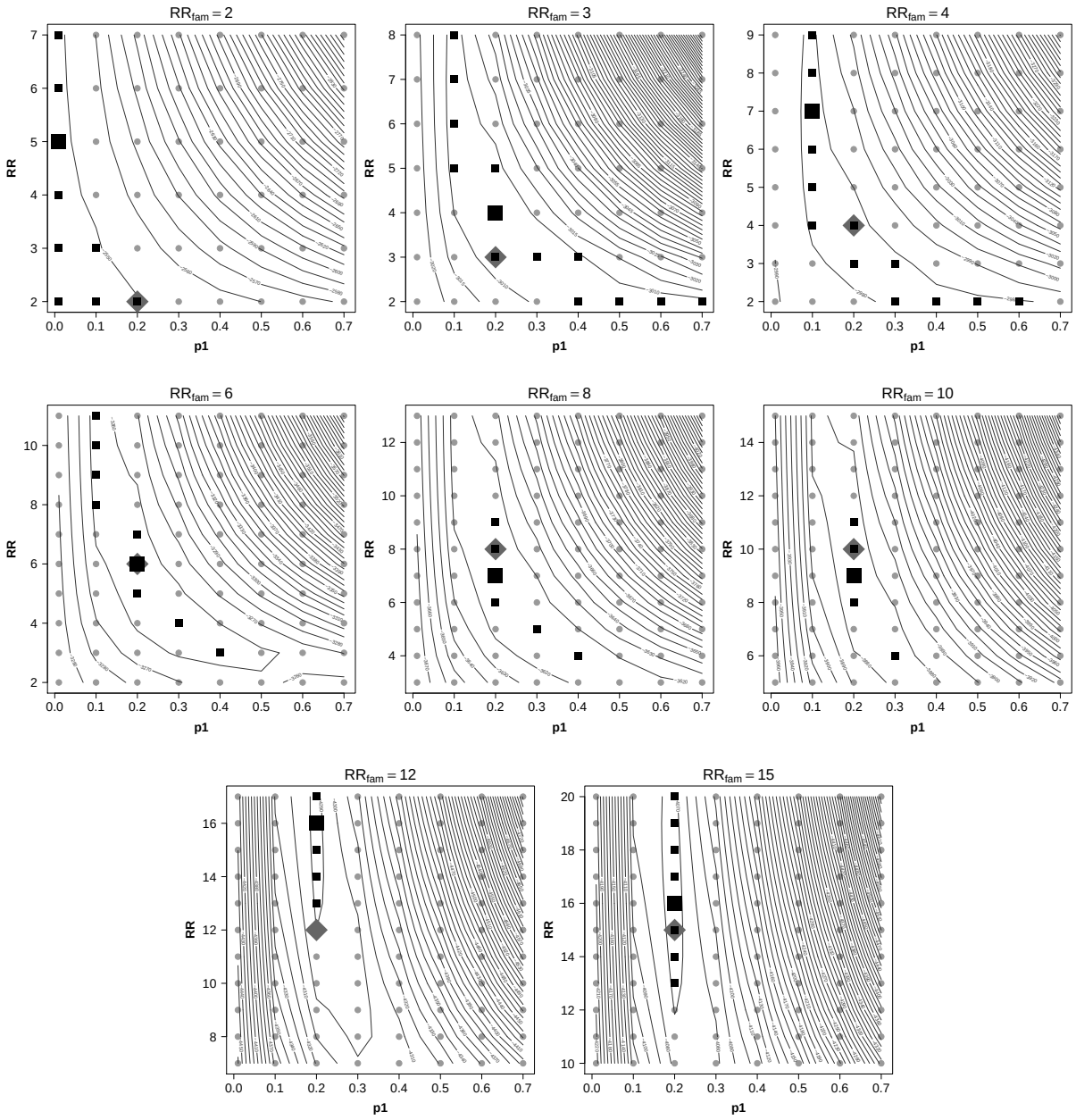

Figure B.1: Contour plot of likelihood surfaces for the simulated general population with probability of inheritance $p^{\star}_{\text{inh}} = 0.25$. The data sets consist of $N = 500$ families. The grey round points show the grid. The big squared black dot shows the maximum of the likelihood surface. The small squared black dots show the likelihood confidence region around the maximum. The grey diamond shows the true parameters, with which the data were simulated. The prevalence was set to $p_1 = 0.2$. Please be aware of different y-axes.

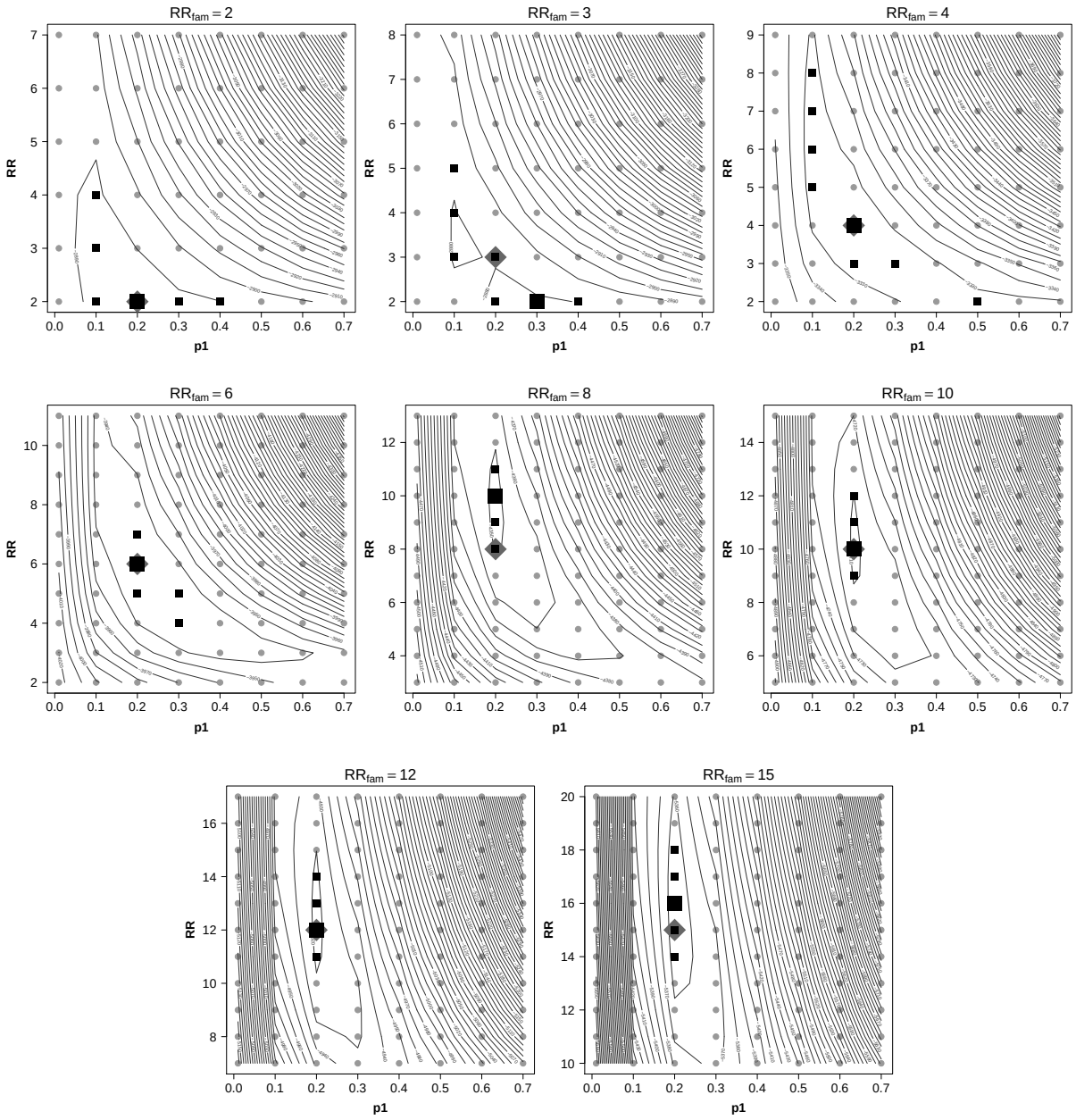

Figure B.2: Contour plot of likelihood surfaces for the simulated general population with probability of inheritance $p^{\star}_{\text{inh}} = 0.50$. The data sets consist of $N = 500$ families. The grey round points show the grid. The big squared black dot shows the maximum of the likelihood surface. The small squared black dots show the likelihood confidence region around the maximum. The grey diamond shows the true parameters, with which the data were simulated. The prevalence was set to $p_1 = 0.2$. Please be aware of different y-axes.

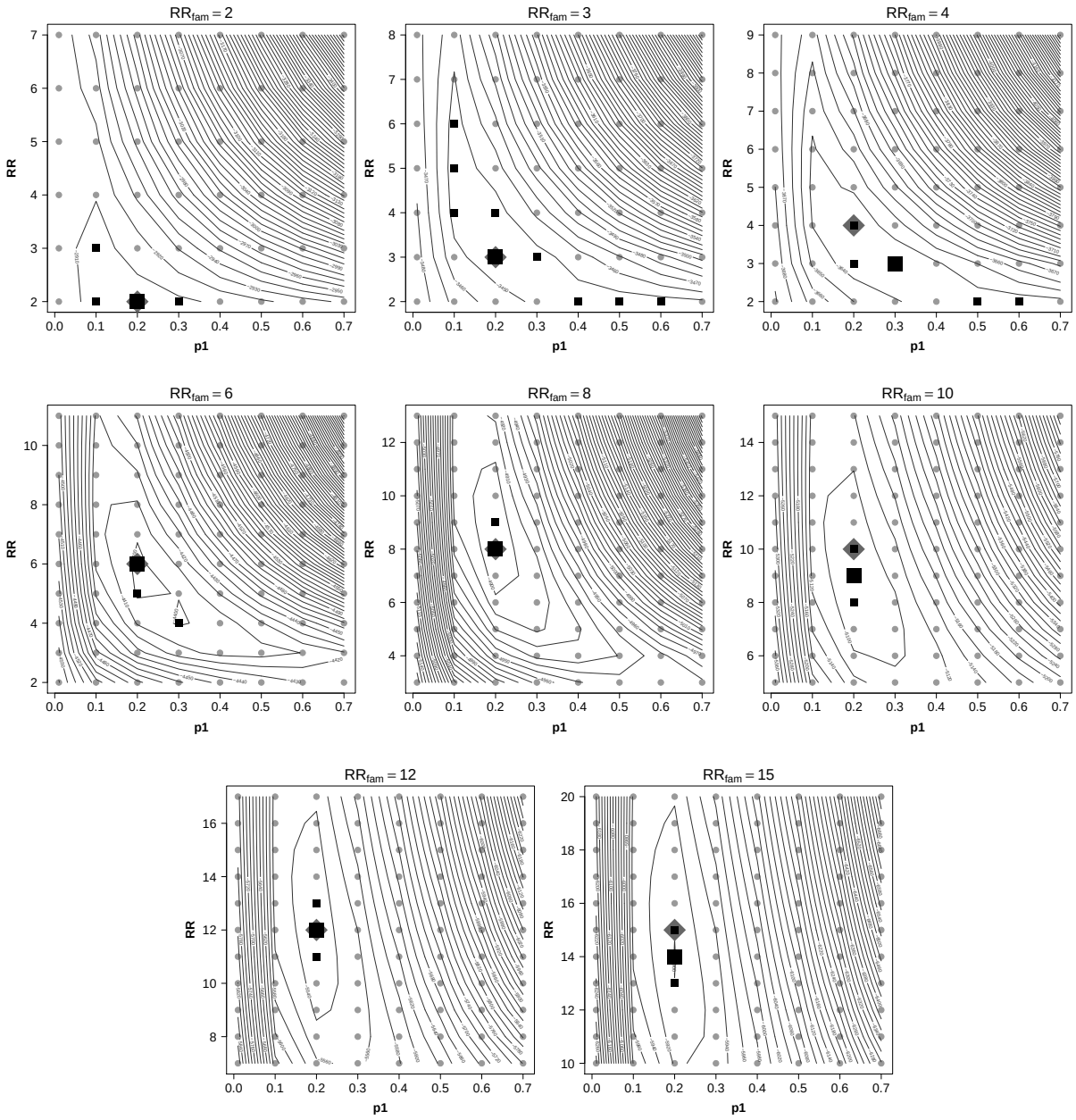

Figure B.3: Contour plot of likelihood surfaces for the simulated general population with probability of inheritance $p^{\star}_{\text{inh}} = 0.75$. The data sets consist of $N = 500$ families. The grey round points show the grid. The big squared black dot shows the maximum of the likelihood surface. The small squared black dots show the likelihood confidence region around the maximum. The grey diamond shows the true parameters, with which the data were simulated. The prevalence was set to $p_1 = 0.2$. Please be aware of different y-axes.

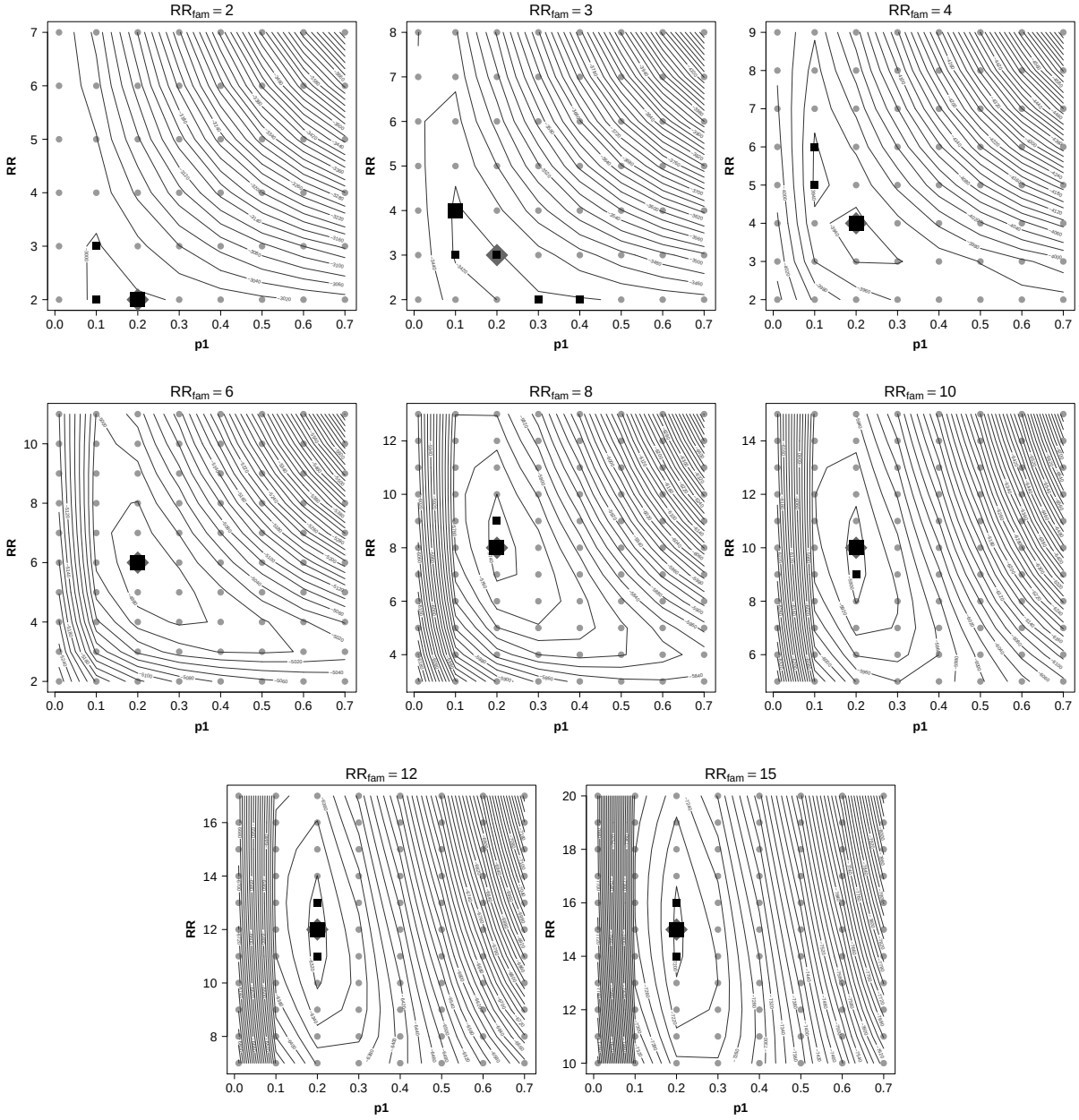

Figure B.4: Contour plot of likelihood surfaces for the simulated general population with probability of inheritance $p^{\star}_{\text{inh}} = 1.00$. The data sets consist of $N = 500$ families. The grey round points show the grid. The big squared black dot shows the maximum of the likelihood surface. The small squared black dots show the likelihood confidence region around the maximum. The grey diamond shows the true parameters, with which the data were simulated. The prevalence was set to $p_1 = 0.2$. Please be aware of different y-axes.

B.1.2 ROC Curves

A Bayesian posterior score was proposed to discriminate between "risk families" and families without predisposition for familial burden. It was compared to the NACRC questionnaire developed on behalf of [52] by means of ROC curves and subsequent calculation of AUCs. AUCs with confidence intervals are given in table 6.1 in the main part of this thesis. ROCs are given here for each $p^{\star}_{\text{inh}}$ separately.

$p^{\star}_{\text{inh}} = 0.25$ The results for RTMo.25 using $p^{\star}_{\text{inh}} = 0.25$ are given in figure B.5. The ROC curves showed a very good result for the posterior score and good results for the NACRC score. Especially for low false positive rates, the Bayesian posterior score showed higher true positive rates than the NACRC score. These findings did not differ much between the single RR_{fam} considered.

$p^{\star}_{\text{inh}} = 0.50$ The results for $p^{\star}_{\text{inh}} = 0.50$ can be found in figure B.6. The AUC was nearly 1 for low RR_{fam} and decreased slightly with increasing RR_{fam}. The NACRC score showed good and stable results with AUCs around 0.8.

$p^{\star}_{\text{inh}} = 0.75$ Figure B.7 shows the ROC curves for RTMo.75 ($p^{\star}_{\text{inh}} = 0.75$). The ROC curves for the Bayesian score reached nearly the perfect line. The ROC curves respectively AUCs of the NACRC score did not differ much from those of the other $p^{\star}_{\text{inh}}$ considered.

$p^{\star}_{\text{inh}} = 1.00$ The results of the CTM setting with $p^{\star}_{\text{inh}} = 1.00$ are given in figure B.8. Again, the ROC curves for both Bayesian posterior score and NACRC score remained stable with almost perfect (AUC near 1) and good discrimination ability (AUC around 0.8).

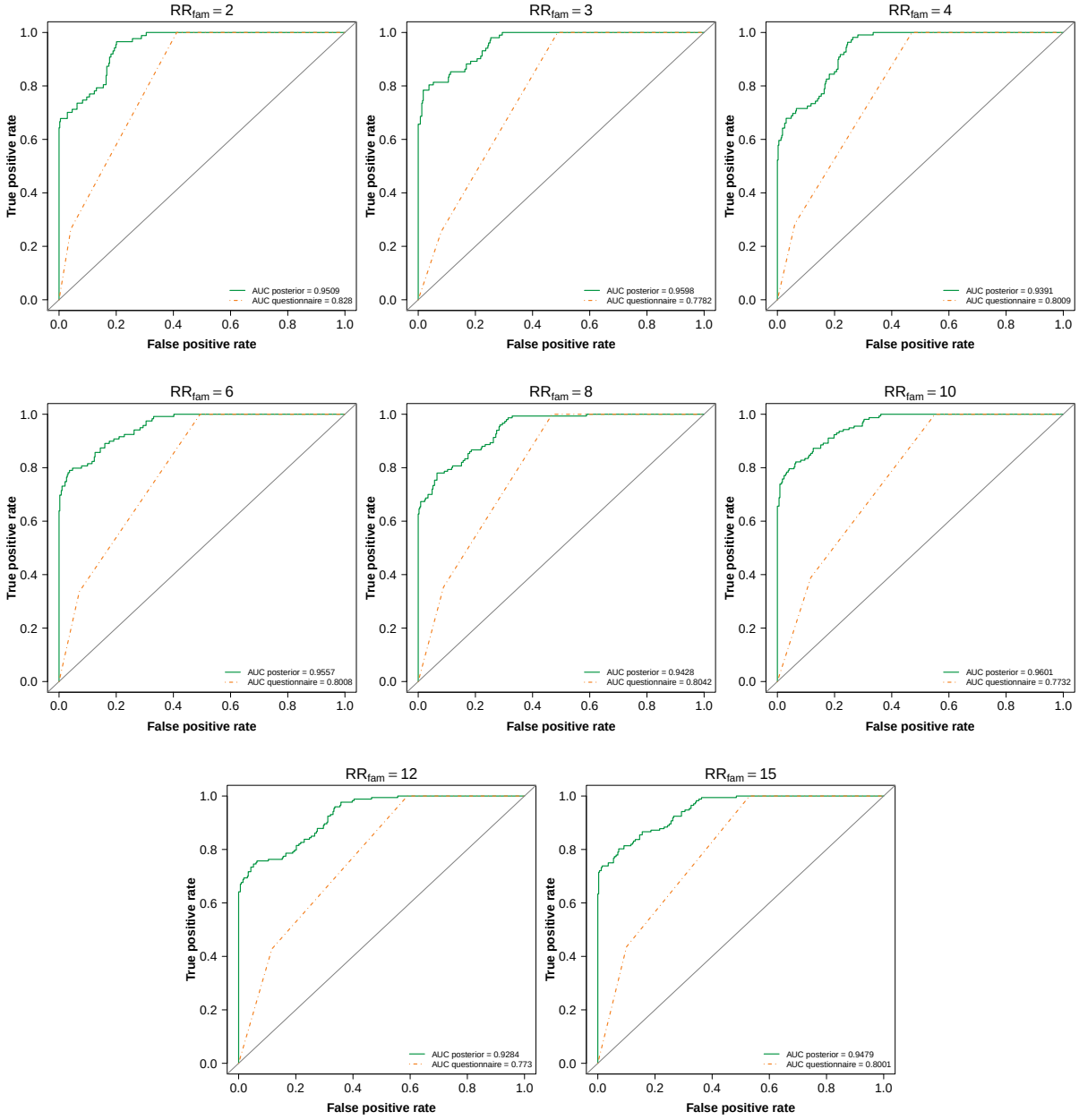

Figure B.5: ROC curves for the simulated general population with probability of inheritance $p^{\star}_{\text{inh}} = 0.25$. The data sets consist of $N = 500$ families. The ROC curve for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with estimated parameters by means of grid search.

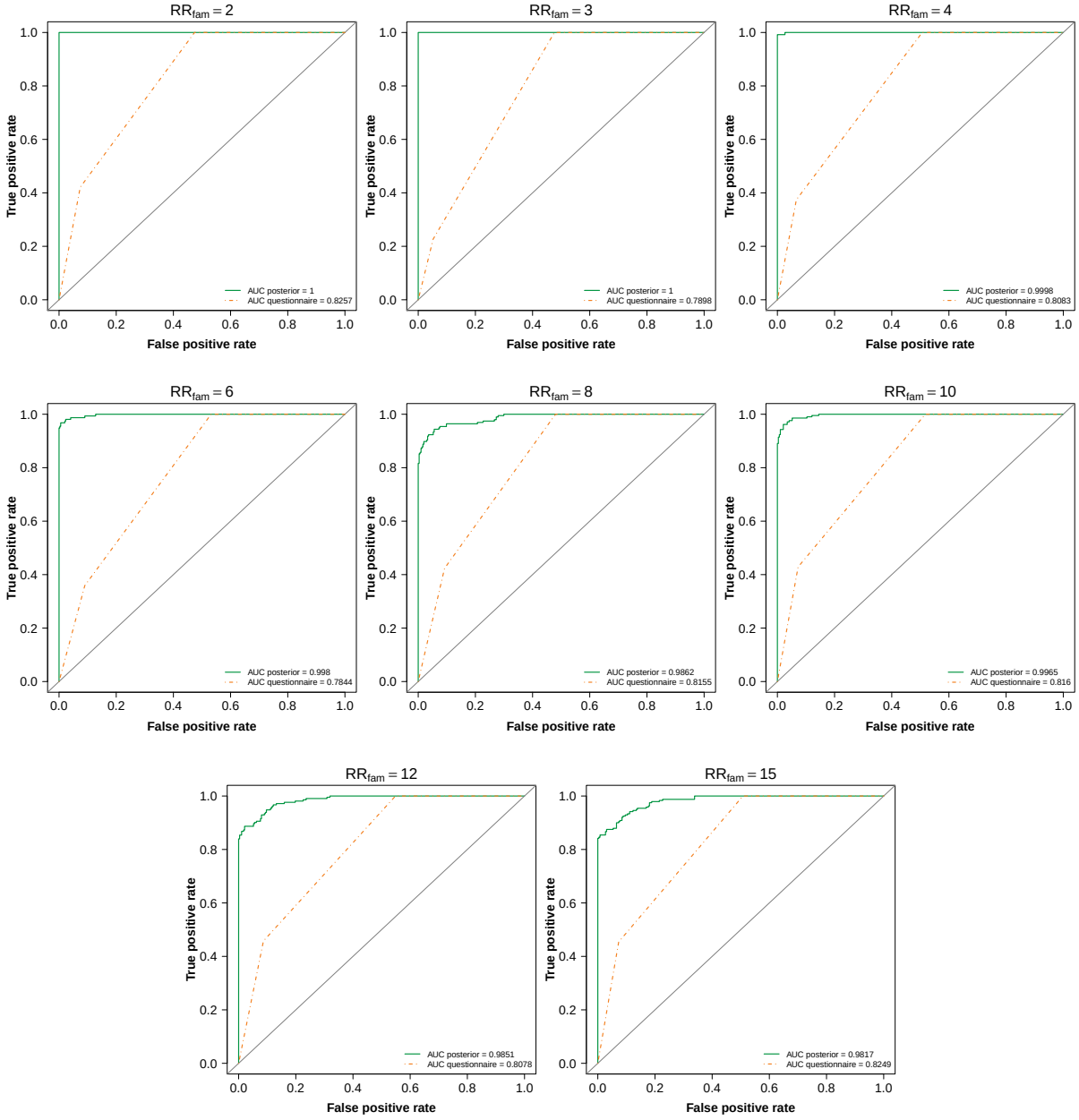

Figure B.6: ROC curves for the simulated general population with probability of inheritance $p^{\star}_{\text{inh}} = 0.50$. The data sets consist of $N = 500$ families. The ROC curve for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with estimated parameters by means of grid search.

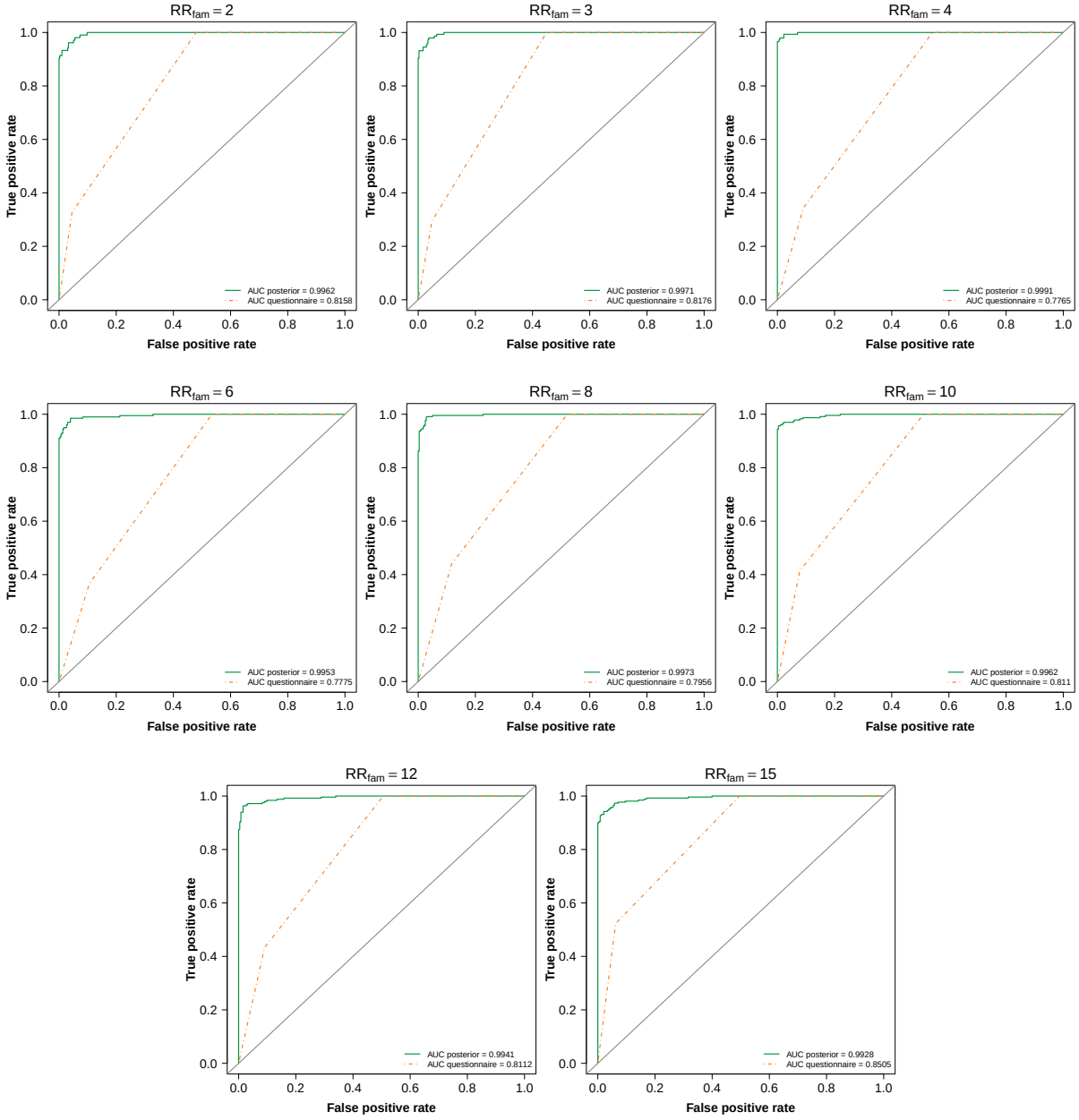

Figure B.7: ROC curves for the simulated general population with probability of inheritance $p^{\star}_{\text{inh}} = 0.75$. The data sets consist of $N = 500$ families. The ROC curve for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with estimated parameters by means of grid search.

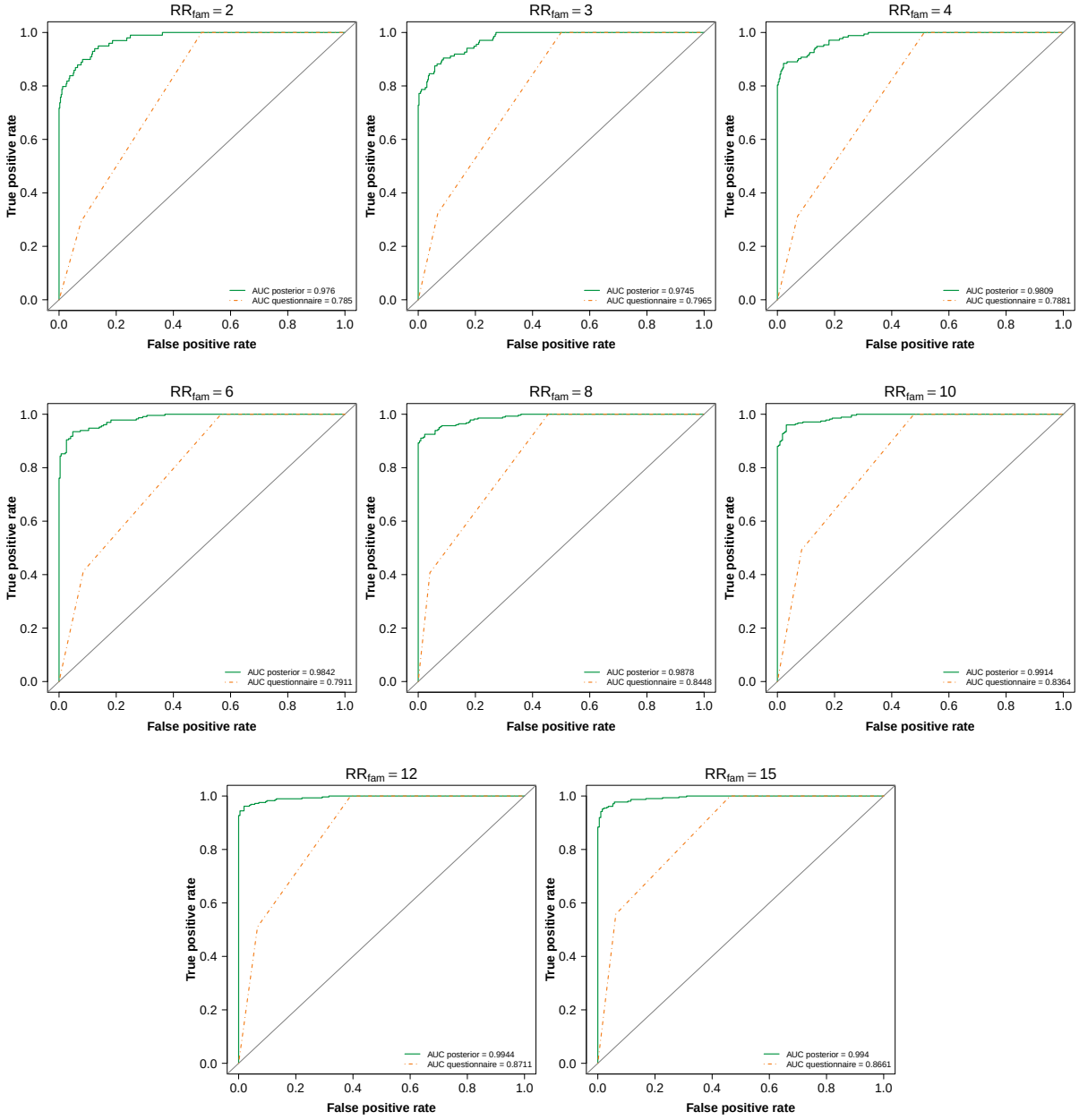

Figure B.8: ROC curves for the simulated general population with probability of inheritance $p^{\star}_{\text{inh}} = 1.00$. The data sets consist of $N = 500$ families. The ROC curve for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with estimated parameters by means of grid search.

B.2 Grid Search in the Simulated Selected Population

Results of the grid search and Bayesian risk score calculation in the simulated selected population can be found in section 6.1.2. During the grid search process, likelihood surfaces can be built. Only some randomly chosen likelihoods are shown in figure 6.2. In this section here, all likelihood surfaces can be found (see section B.2.1). The discrimination abilities of the Bayesian posterior score, introduced in section 5.2, and of the NACRC questionnaire, described in section 3.2.3, were compared by means of ROC cures and AUCS. The AUCs are displayed in section 6.1.2 in table 6.2. The respective ROC curves are shown here (see section B.2.2).

B.2.1 Likelihood Surfaces

As mentioned above, some randomly chosen likelihood surfaces are depicted in figure 6.2 (section 6.1.2). Those are again included in the following figures (one for each $p^{\star}_{\text{inh}}$).

$p^{\star}_{\text{inh}} = 0.25$ The likelihood surfaces for the RTM0.25 setting with $p^{\star}_{\text{inh}} = 0.25$ are given in figure B.9. They had a relatively flat peak, at least for lower RR_{fam}. For high RR_{fam}, the peak was elongated but narrow. The true parameters were not included within the confidence region around the estimate from grid search. The prevalence p_1 seemed to be problematic to estimate properly in a selected population. It was highly overestimated in every setting. However, this seems to be appropriate, as the prevalence of risk carriers is higher in a selected population than in the general population. The parameter RR_{fam} was overestimated for low true RR_{fam} and underestimated for increasing true RR_{fam}.

$p^{\star}_{\text{inh}} = 0.50$ The results of the RTM0.50 setting using $p^{\star}_{\text{inh}} = 0.50$ can be found in figure B.10. The findings of the RTM0.25 setting were true also for the RTM0.50 setting. The parameter p_1 was highly overestimated regarding estimation for a general population. For low true RR_{fam}, the estimation of RR_{fam} worked well with only small deviations from the true value. However, an underestimation was seen for high true RR_{fam}.

$p^{\star}_{\text{inh}} = 0.75$ The graphs for $p^{\star}_{\text{inh}} = 0.75$ are shown in figure B.11. For increasing $p^{\star}_{\text{inh}}$, the deviation from the true parameters became smaller. Nevertheless, the overestimation of p_1 seems to be appropriate, as the prevalence of risk carriers is assumed to be higher in a selected population than in the general population.

$p^{\star}_{\text{inh}} = 1.00$ The likelihood surfaces for the CTM setting with $p^{\star}_{\text{inh}} = 1.00$ are given in figure B.12. The peak of the likelihood surfaces remained relatively flat for high true RR_{fam}. Nevertheless, the results of the other $p^{\star}_{\text{inh}}$ were applicable also for the CTM setting: high overestimation of p_1 regarding the estimation for a general population, rather

good estimation of RR_{fam} for low true $RR_{\text{fam}}$and underestimation of RR_{fam} for higher true RR_{fam}.

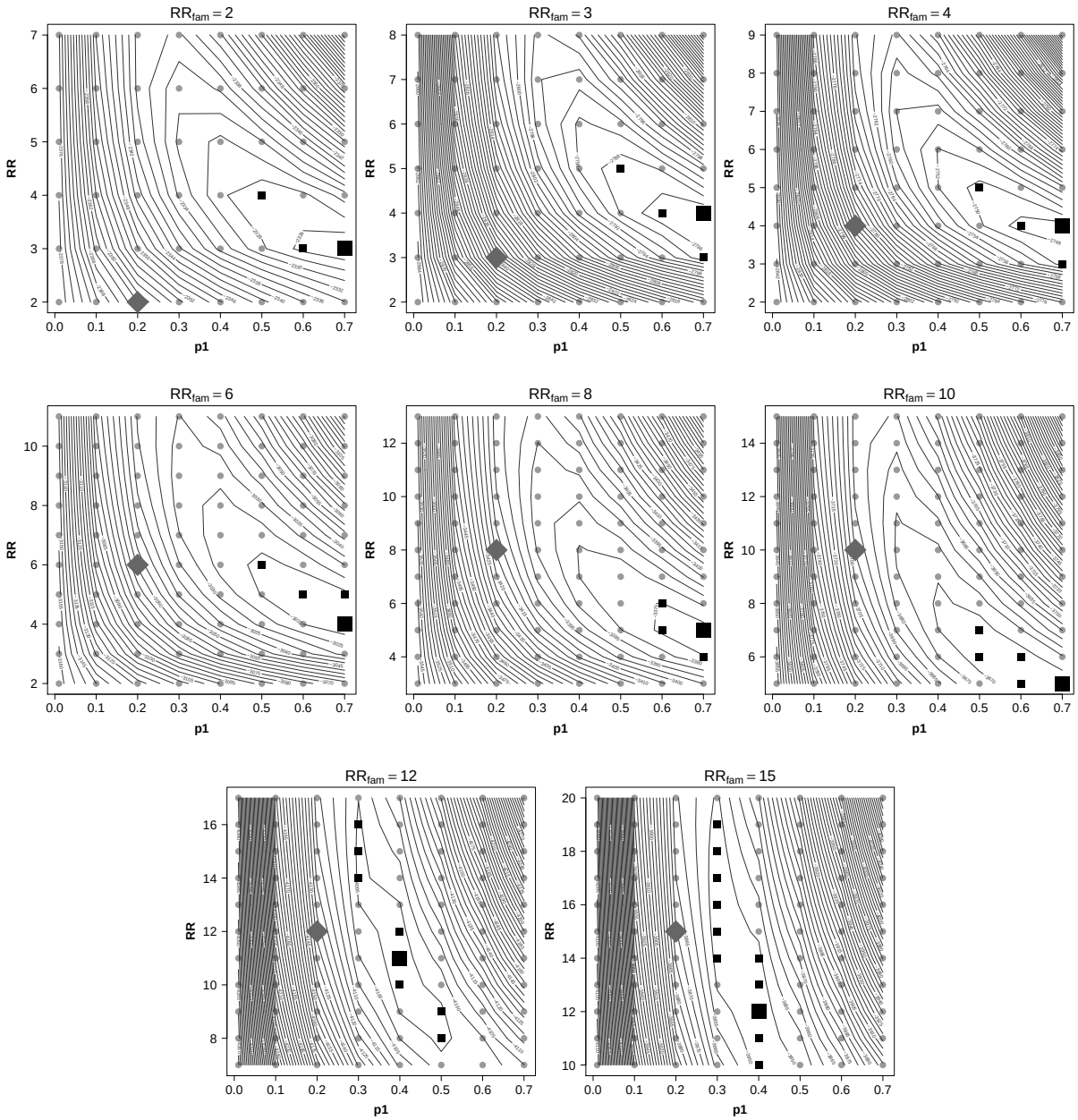

Figure B.9: Contour plot of likelihood surfaces for the simulated selected population with probability of inheritance $p^{\star}_{\text{inh}} = 0.25$. The data sets consist of all families with at least one CRC case out of the $N = 500$ simulated families. The grey round points show the grid. The big squared black dot shows the maximum of the likelihood surface. The small squared black dots show the likelihood confidence region around the maximum. The grey diamond shows the true parameters, with which the data were simulated. The prevalence was set to $p_1 = 0.2$. Please be aware of different y-axes.

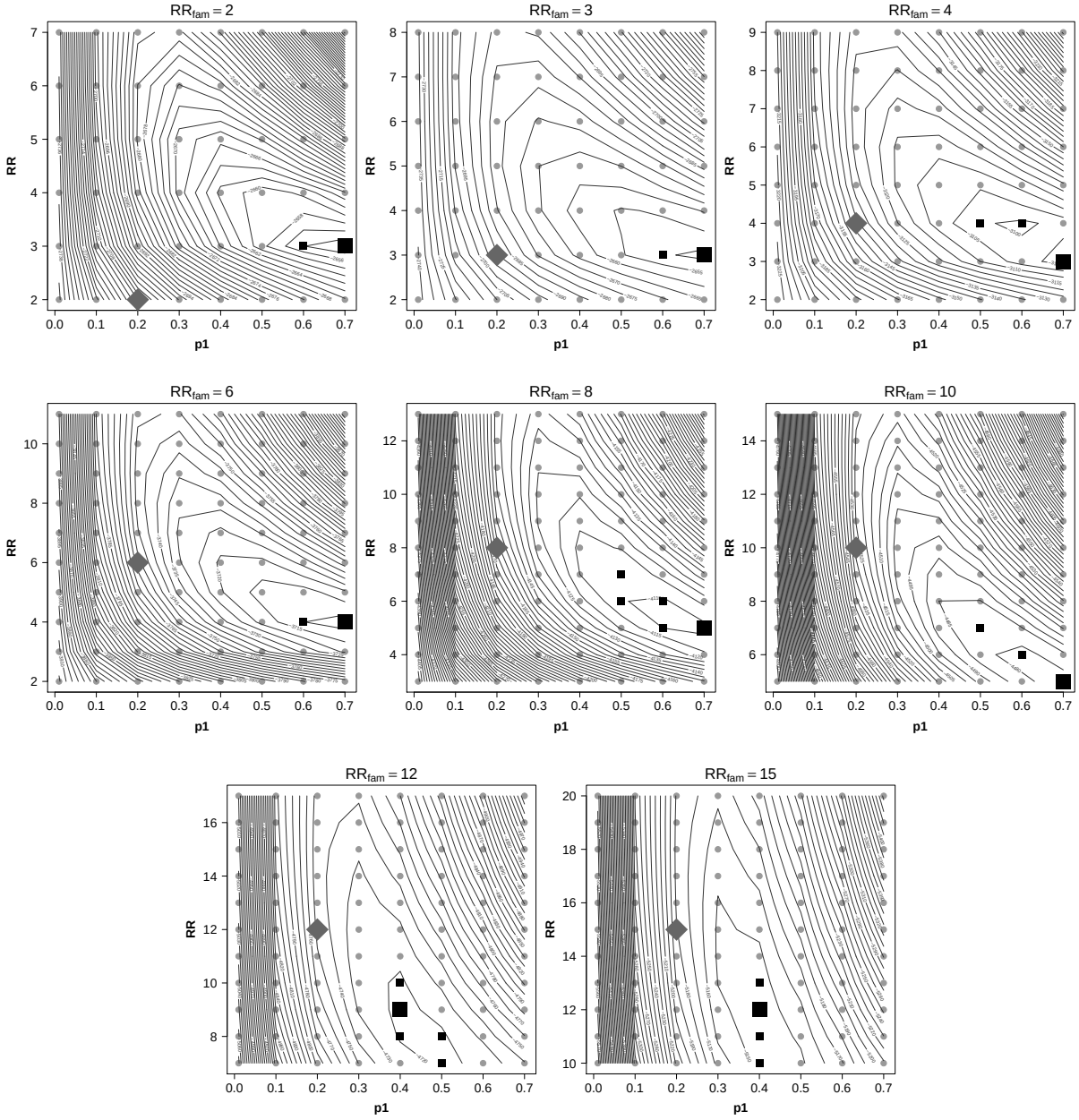

Figure B.10: Contour plot of likelihood surfaces for the simulated selected population with probability of inheritance $p^{\star}_{\text{inh}} = 0.50$. The data sets consist of all families with at least one CRC case out of the $N = 500$ simulated families. The grey round points show the grid. The big squared black dot shows the maximum of the likelihood surface. The small squared black dots show the likelihood confidence region around the maximum. The grey diamond shows the true parameters, with which the data were simulated. The prevalence was set to $p_1 = 0.2$. Please be aware of different y-axes.

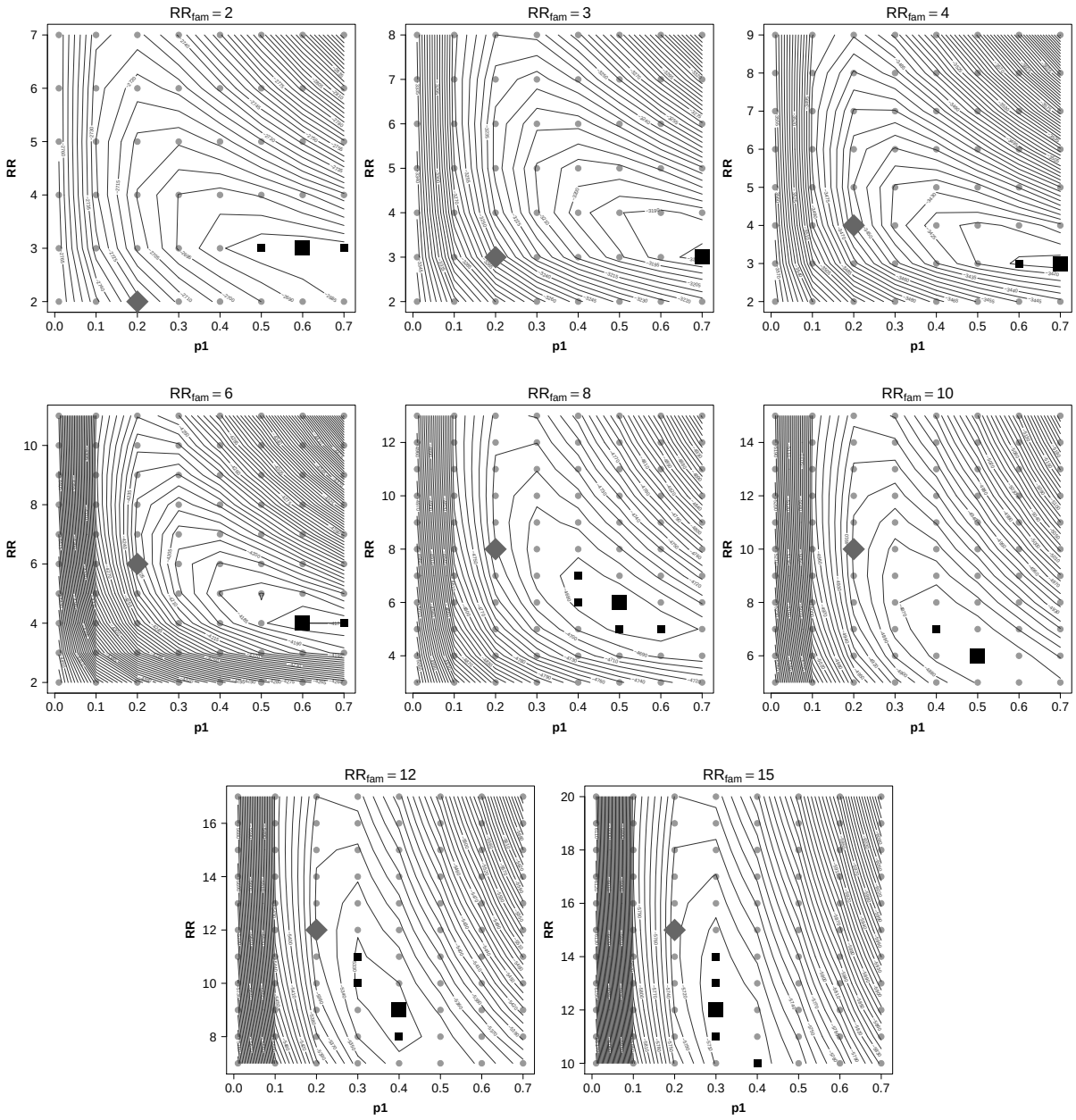

Figure B.11: Contour plot of likelihood surfaces for the simulated selected population with probability of inheritance $p^{\star}_{\text{inh}} = 0.75$. The data sets consist of all families with at least one CRC case out of the $N = 500$ simulated families. The grey round points show the grid. The big squared black dot shows the maximum of the likelihood surface. The small squared black dots show the likelihood confidence region around the maximum. The grey diamond shows the true parameters, with which the data were simulated. The prevalence was set to $p_1 = 0.2$. Please be aware of different y-axes.

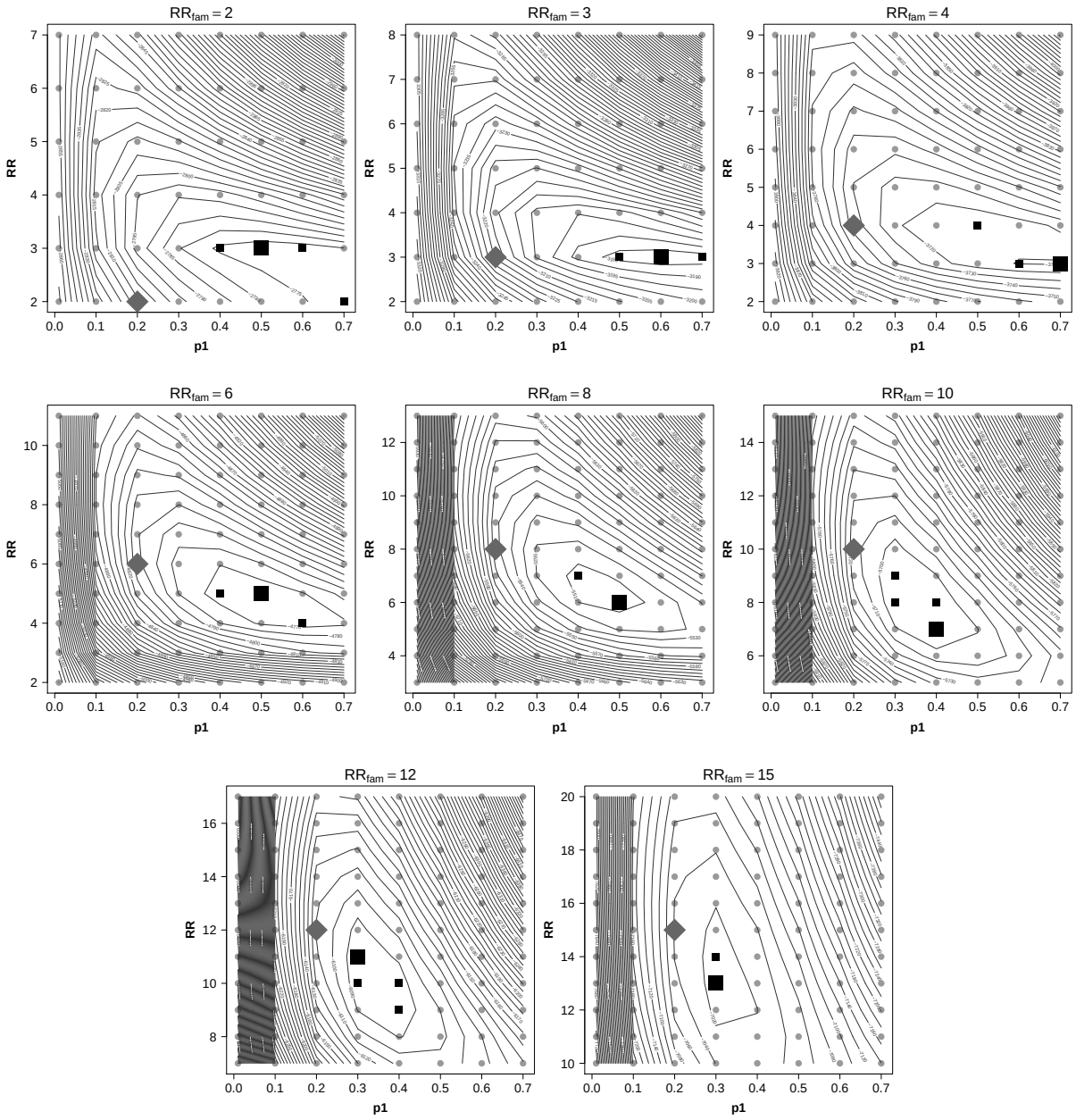

Figure B.12: Contour plot of likelihood surfaces for the simulated selected population with probability of inheritance $p^{\star}_{\text{inh}} = 1.00$. The data sets consist of all families with at least one CRC case out of the $N = 500$ simulated families. The grey round points show the grid. The big squared black dot shows the maximum of the likelihood surface. The small squared black dots show the likelihood confidence region around the maximum. The grey diamond shows the true parameters, with which the data were simulated. The prevalence was set to $p_1 = 0.2$. Please be aware of different y-axes.

B.2.2 ROC Curves

The ability to discriminate between "risk families" and families without predisposition for familial burden of the Bayesian posterior risk score and the score arising from NACRC questionnaire were compared by means of ROC curves and subsequent calculation of AUCs. The AUCs with confidence intervals of the results in the simulated selected population are given in table 6.2 (see section 6.1.2). For each $p^{\star}_{\text{inh}}$, a panel of ROC curves is shown in a separate figure. In each panel, the ROC curve of the NACRC questionnaire is also depicted for comparison.

$p^{\star}_{\text{inh}} = 0.25$ The results for $p^{\star}_{\text{inh}} = 0.25$ are depicted in figure B.13. Because of poor results in the grid search, discrimination quality of the Bayesian risk score dropped a little bit to AUCs around 0.9. The NACRC score decreased to AUCs around 0.6. Remarkably, the Bayesian posterior score showed relatively high true positive rates for low false positive rates.

$p^{\star}_{\text{inh}} = 0.50$ The results for RTMo.50 using $p^{\star}_{\text{inh}} = 0.50$ are given in figure B.14. As the results in the grid search were a little bit better compared to $p^{\star}_{\text{inh}} = 0.25$, the ROC curves respectively AUCs were nearly 1 for $p^{\star}_{\text{inh}} = 0.50$. A good discrimination ability was given even for falsely estimated parameters.

$p^{\star}_{\text{inh}} = 0.75$ The ROC curves for RTMo.75 ($p^{\star}_{\text{inh}} = 0.75$) can be found in figure B.15. The posterior score showed a very good discrimination ability for all RR_{fam} regarded. Results of the NACRC score remained stable with AUCs of around 0.6.

$p^{\star}_{\text{inh}} = 1.00$ Figure B.16 shows the results of the CTM setting with $p^{\star}_{\text{inh}} = 1.00$. Again, missing the true parameters in grid search did not affect the discrimination quality of the Bayesian posterior score. This was true for every RR_{fam} considered. NACRC score AUC was always around 0.6, as its calculation was not affected by varying $p^{\star}_{\text{inh}}$.

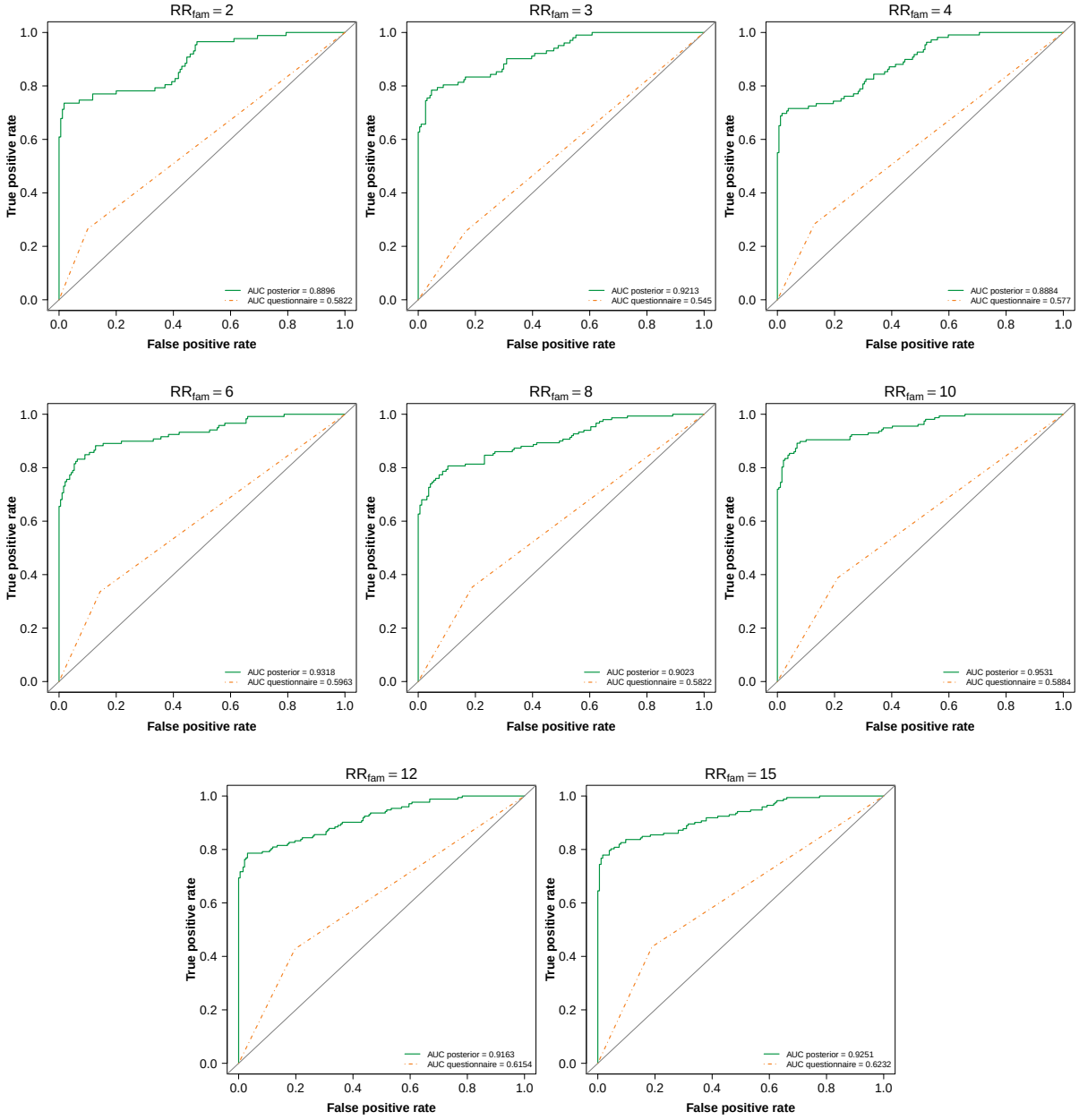

Figure B.13: ROC curves for the simulated selected population with probability of inheritance $p^{\star}_{\text{inh}} = 0.25$. The data sets consist of all families with at least one CRC case out of the $N = 500$ simulated families. The ROC curve for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with estimated parameters by means of grid search.

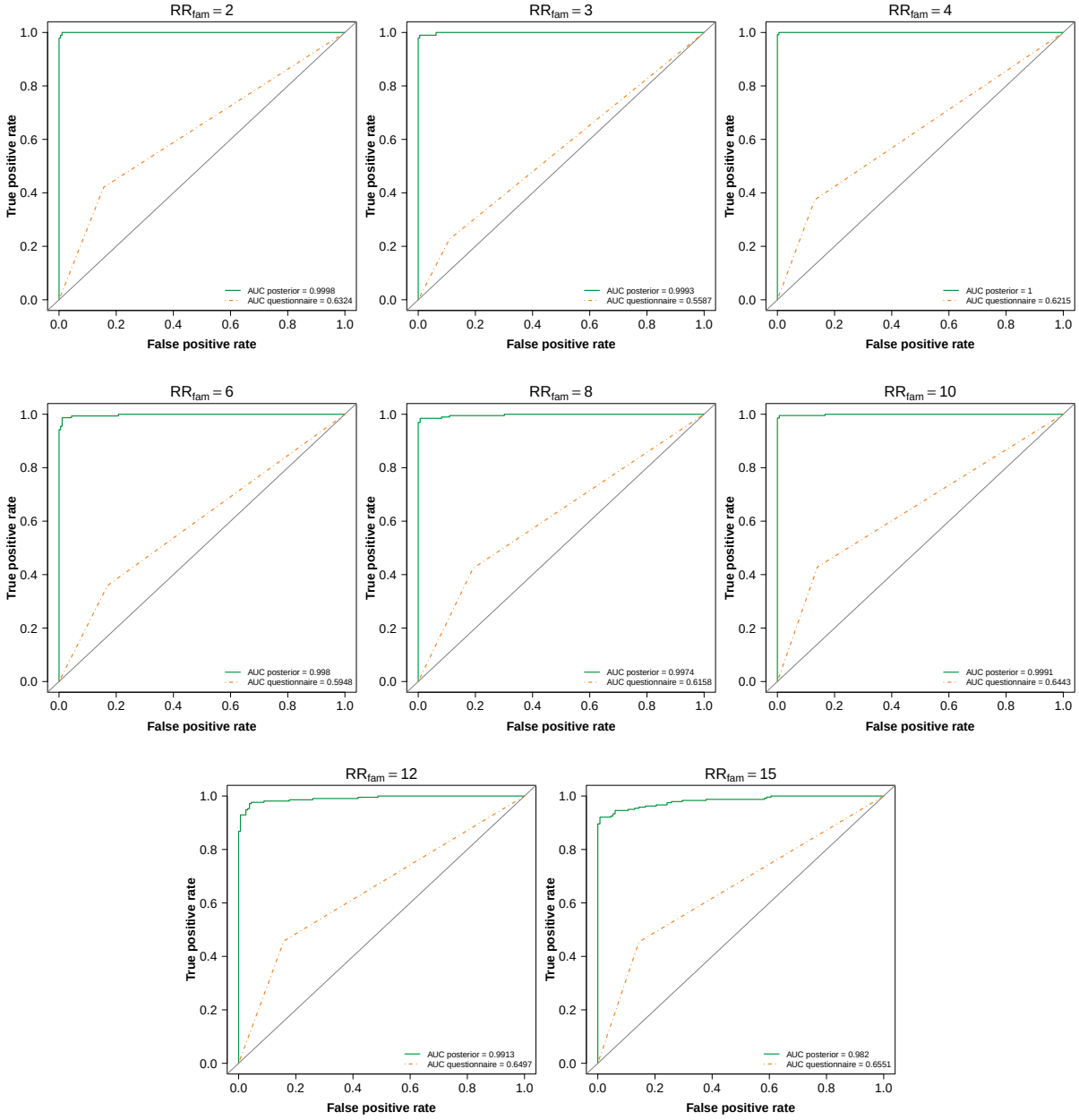

Figure B.14: ROC curves for the simulated selected population with probability of inheritance $p^{\star}_{\text{inh}} = 0.50$. The data sets consist of all families with at least one CRC case out of the $N = 500$ simulated families. The ROC curve for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with estimated parameters by means of grid search.

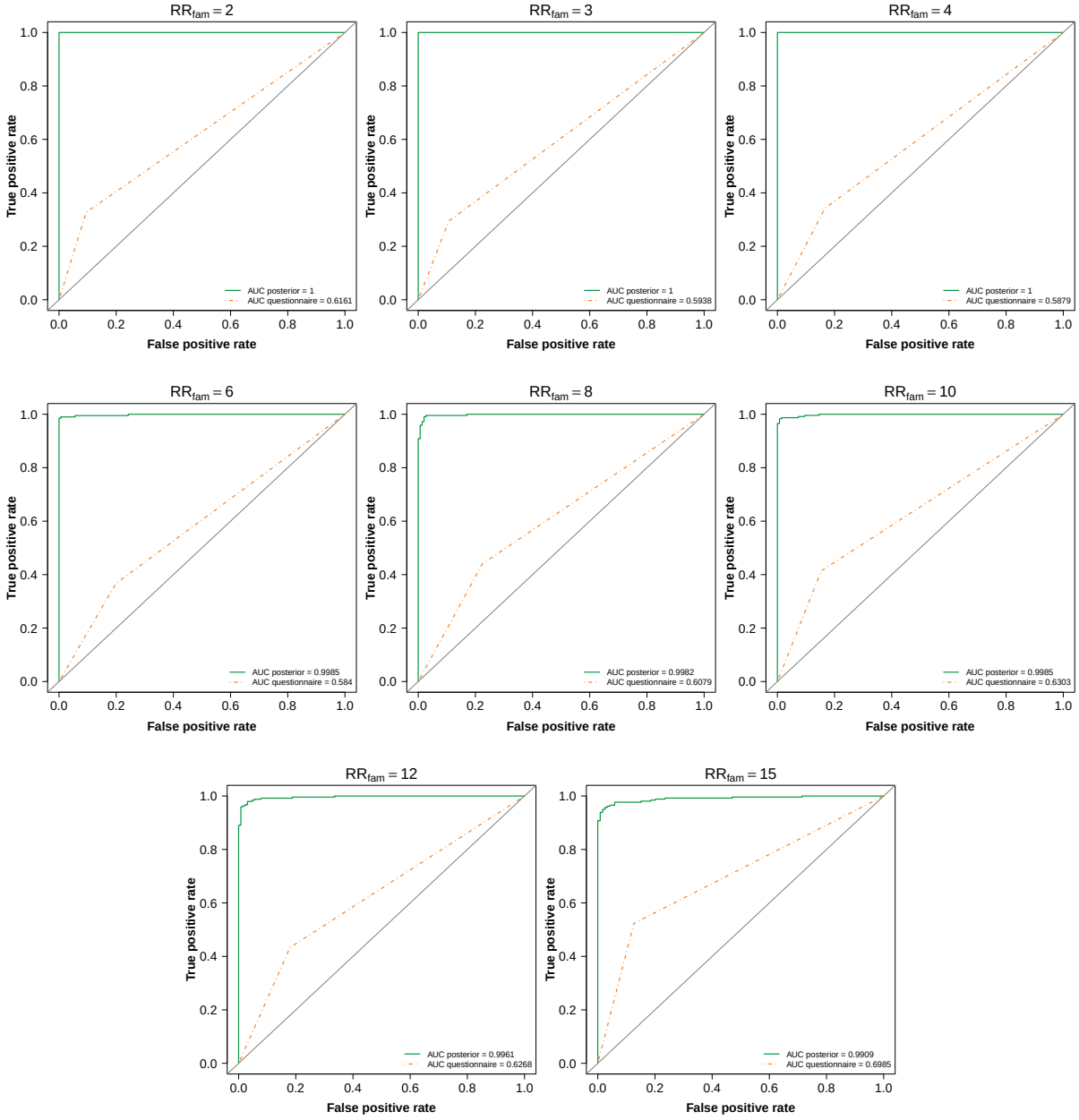

Figure B.15: ROC curves for the simulated selected population with probability of inheritance $p^{\star}_{\text{inh}} = 0.75$. The data sets consist of all families with at least one CRC case out of the $N = 500$ simulated families. The ROC curve for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with estimated parameters by means of grid search.

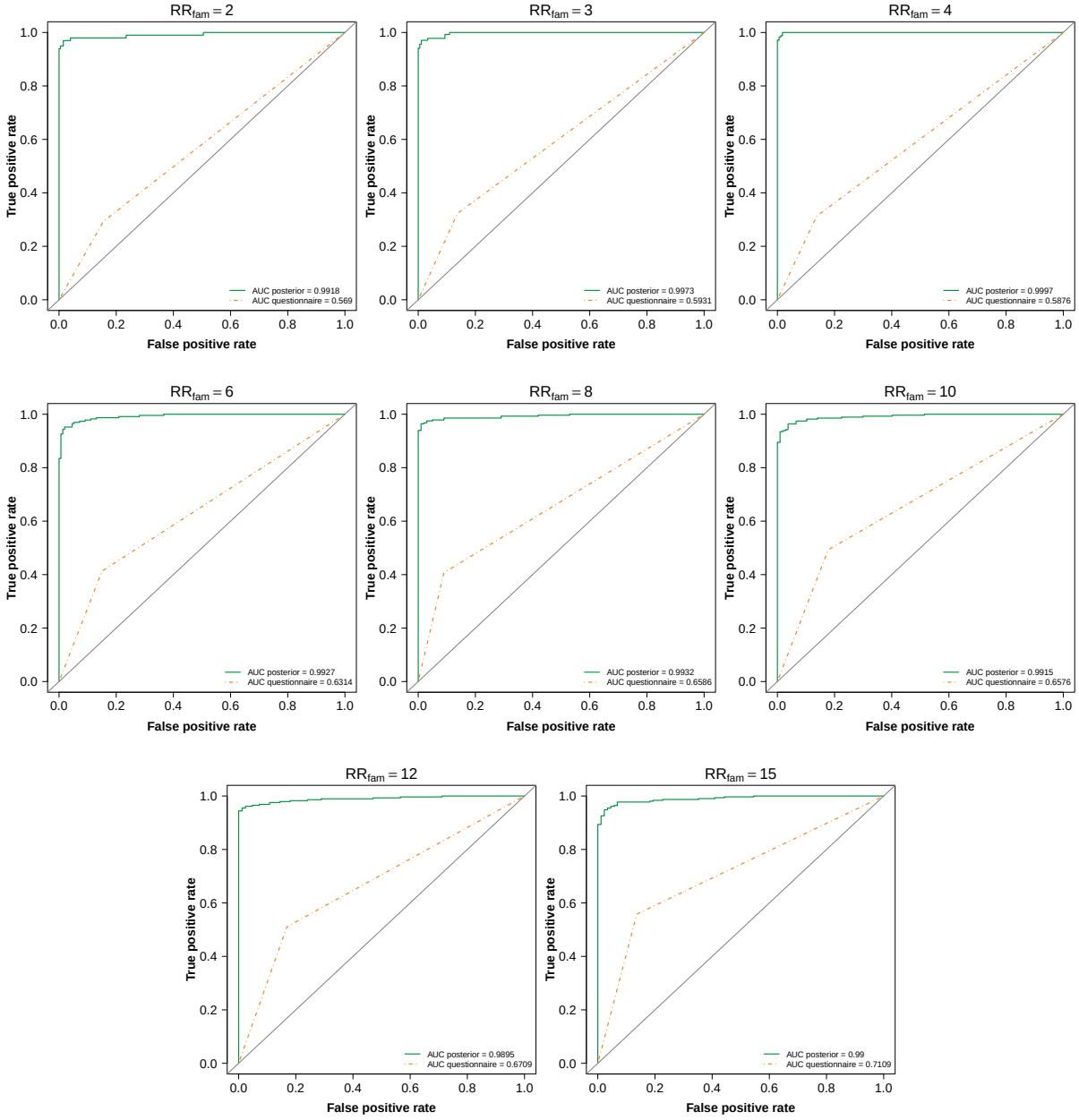

Figure B.16: ROC curves for the simulated selected population with probability of inheritance $p^{\star}_{\text{inh}} = 1.00$. The data sets consist of all families with at least one CRC case out of the $N = 500$ simulated families. The ROC curve for the NACRC questionnaire is shown for comparison. The posterior score for the ROC curves was calculated with estimated parameters by means of grid search.

Eidesstattliche Versicherung

(gemäß § 8 Abs. 2 Pkt. 5 der Promotionsordnung vom 12. Juli 2011)

Hiermit versichere ich an Eides statt, dass die vorgelegte Dissertation selbständig und ohne unerlaubte Beihilfe angefertigt ist.

München, den 12. Juli 2017

Anna Rieger