Technische Universität München

Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Fachgebiet Biostatistik

Data scientific approaches to contemporary clinical risk tool construction

Johanna Elisabeth Tolksdorf

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

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Aurélien Tellier
Prüfer der Dissertation:	1. Prof. Donna P. Ankerst, Ph.D.
	2. Associate Prof. Jonathan Gelfond, Ph.D.

Die Dissertation wurde am 24.06.2019 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 03.10.2019 angenommen.

Summary

Patients and doctors are interested in making informed decisions whether to perform a prostate biopsy or not. It is therefore important to get individualized risk assessments of prostate cancer before biopsy, which constitutes a main focus of this thesis. The data analyzed comprises 8492 observations and stems from the ongoing US Prostate Biopsy Collaborative Group, which gathers individual-level prostate cancer data from ten international institutes. We investigated improvements of current prediction models for prostate cancer using six established risk factors, including age, prostate-specific antigen, digital rectal examination, African ancestry, first degree family history, and prior negative biopsy, by incorporating cohort heterogeneity, enabling contemporary data and allowing for more flexible model structures.

In order to ensure a valid comparison of the new approaches with existing risk tools, this thesis extensively discusses common validation methods. We particularly investigated their variability structure and derived for the first time analytical confidence intervals for clinical net benefit curves. To compare models, we implemented these validation methods using among others a 252-permutation-based sampling plan of all ways to split the ten available sites into five used for training a model and five for testing.

One of the main advantages of multiple cohorts is the possibility of a resultant more generally applicable risk calculator, which can reach a broader range of patients. Furthermore, the combination of multiple cohorts reduces the overall recruitment time and can therefore ensure contemporary data. However, patients from distinct sites may no longer be homogeneous, due to differing clinical assessments or population types. We implemented three contemporary approaches to integrate data from multiple sources for performing logistic regression. The first approach simply pools individual-level data from all sites, the second performs a cohort specific random effects model, and the third a traditional random effects meta-analysis that builds models separately to each cohort. We could not identify a single method to outperform the other approaches. This result supports the common practice of simply pooling data across diverse cohorts. Furthermore, meta-analyses performed equivalently with the additional advantage of scalability: Models can be built locally at individual cohort sites with summaries transported for centralization, making an overall data storage redundant and reducing organizational workload.

The risk prediction tool of the Prostate Cancer Prevention Trial (PCPT) is available online and widely used. However it utilized data from the 1990s and the considered observations were thus based on six-core biopsies and outdated grading systems. We used multinomial logistic regression, adjusted for diverse missingness structures, to investigate the benefits of updating the PCPT risk calculator by using contemporary data. In addition to internal cross validation within North American sites we performed external validation on selected European cohorts. The new models showed superior performance, in particular improved calibration and clinical net benefit. These results stress the importance of contemporary data as these become available.

Comparison of standard logistic regression with more flexible machine learning methods, like k-nearest neighbor, random forests and artificial neural network approaches, showed only small differences in the setting of merely six covariates. However, the investigation identified random forests and artificial neural networks as suitable approaches, disregarding a shortfall in interpretability compared to logistic regression. Whereas random forests were easier to implement, artificial neural networks showed better calibration.

The resulting model of this thesis is a multinomial logistic regression based on 8492 observations and the covariates age, prostate-specific antigen, digital rectal examination, African ancestry, first degree family history, and prior negative biopsy. It is used by doctors and patients with the online risk tool available on the Memorial Sloan Kettering Cancer Center website and the Cleveland Clinic Risk Calculator Library.

Zusammenfassung

Fundierte Entscheidungen über eine mögliche Prostatabiopsie haben für Patienten und Ärzte einen besonderen Stellenwert. Der Schwerpunkt dieser Arbeit liegt daher auf der individuellen Risikobewertung für Prostatakrebs. Die analysierten Daten umfassen 8492 Beobachtungen aus der fortlaufenden US Prostate Biopsy Collaborative Group, welche von zehn internationalen Instituten Prostatakrebsdaten auf Patientenebene sammelt. Es wurden Erweiterungen aktueller Vorhersagemodelle für Prostatakrebs untersucht, unter Verwendung der sechs etablierten Risikofaktoren Alter, prostataspezifisches Antigen, digitale rektale Untersuchung, afrikanische Abstammung, Familienanamnese ersten Grades und vorherige negative Biopsie. Hierfür wurde Heterogenität zwischen Kohorten einbezogen, aktuelle Daten verwendet und flexiblere Modellstrukturen ermöglicht.

Um einen validen Vergleich der neuen Ansätze mit bestehenden Risikovorhersagen zu gewährleisten, wurden in dieser Arbeit konventionelle Validierungsmethoden ausführlich diskutiert. Es wurden insbesondere deren Variabilitätsstrukturen untersucht und erstmalig analytische Konfidenzintervalle für klinische Nettonutzen-Kurven hergeleitet. Diese Validierungsmethoden wurden zum Vergleich der Vorhersagemodelle implementiert, indem unter anderem ein auf 252 Permutationen basierender Stichprobenplan verwendet wurde. Dieser teilt wiederholt die zehn verfügbaren Standorte in je fünf Standorte zum Trainieren und fünf zum Testen eines Modells auf.

Ein Hauptvorteil bei der Verwendung mehrerer Kohorten besteht in der Möglichkeit eines allgemeineren Risikorechners, mit welchem ein breiteres Spektrum von Patienten erreicht werden kann. Darüber hinaus wird durch die Kombination mehrerer Standorte die Rekrutierungszeit insgesamt verkürzt, wodurch aktuelle Daten sichergestellt werden. Möglicherweise dürfen jedoch Patienten verschiedener Kohorte aufgrund unterschiedlicher klinischer Beurteilungen oder Bevölkerungsgruppen nicht mehr als homogen betrachtet werden. Es wurden drei Ansätze implementiert, um Daten aus mehreren Quellen für die Durchführung der logistischen Regression zu integrieren. Der erste Ansatz fasst die Daten von allen Kohorten auf Patientenebene unmittelbar zusammen, der zweite führt ein standortspezifisches Modell mit Zufallseffekten durch und der dritte eine herkömmliche Metaanalyse, bei der für jede Kohorte ein separates Modell erstellt wird. Es konnte keine Methode identifiziert werden, welche die anderen Ansätze signifikant übertrifft. Dieses Ergebnis unterstützt die gängige Praxis, Daten unmittelbar über verschiedene Kohorten hinweg zusammenzufassen. Darüber hinaus erzielten Metaanalysen gleichwertige Ergebnisse, mit dem zusätzlichen Vorteil der Skalierbarkeit: Modelle können lokal an den jeweiligen Standorten erstellt werden, sodass lediglich Zusammenfassungen zur Zentralisierung übertragen werden müssen. Dadurch wird eine zentrale Speicherung der Daten auf Patientenebene überflüssig und der Arbeitsaufwand verringert.

Der Risikorechner des Prostate Cancer Prevention Trials (PCPT) ist online verfügbar und weit verbreitet. Hierfür wurden Daten aus den neunziger Jahren verwendet, und somit basieren die betrachteten Beobachtungen auf Sextantenbiopsien und veralteten Bewertungssystemen. Um die Vorteile einer Aktualisierung des PCPT-Risikorechners mithilfe aktueller Daten zu untersuchen, wurde in dieser Arbeit eine multinomiale logistische Regression verwendet, welche verschiedene Strukturen an fehlenden Daten berücksichtigt. Zusätzlich zu einer internen Kreuzvalidierung innerhalb nordamerikanischer Standorte, wurde an ausgewählten europäischen Kohorten eine externe Validierung durchgeführt. Die neuen Modelle zeigten überlegene Ergebnisse, insbesondere eine bessere Kalibrierung und einen erhöhten klinischen Nettonutzen. Diese Resultate unterstreichen den Mehrwert, aktuelle Daten zu verwenden, sobald diese verfügbar sind.

Der Vergleich der logistischen Standardregression mit flexibleren Methoden des maschinellen Lernens, wie k-Nearest Neighbor Methoden, Random Forests und künstliche neuronale Netzwerke, ergab nur geringe Unterschiede bei der Verwendung von lediglich sechs Risikofaktoren. Bei der Untersuchung wurden Random Forests und künstliche neuronale Netzwerke als geeignete Ansätze ermittelt, wobei ein Mangel an Interpretierbarkeit im Vergleich zur logistischen Regression außer Acht gelassen wurde. Während Random Forests einfacher zu implementieren waren, zeigten künstliche neuronale Netze eine bessere Kalibrierung.

Das aus dieser Arbeit resultierende Modell ist eine multinomiale logistische Regression, die auf 8492 Beobachtungen und den Risikofaktoren Alter, prostataspezifisches Antigen, digitale rektale Untersuchung, afrikanische Abstammung, Familienanamnese ersten Grades und vorherige negative Biopsie basiert. Dieses Modell wird von Ärzten und Patienten mit dem Online-Risikorechner verwendet, welcher auf der Website des Memorial Sloan Kettering Cancer Centers und in der Cleveland Clinic Risk Calculator Library verfügbar ist.

Table of Contents

List of Figures

List of Tables

Acronyms

AIC	Akaike information criterion
ANN	artificial neural network
AUC	area under the receiver operating characteristic curve
BIC	Bayesian information criterion
CI	confidence interval
DRE	digital rectal examination
ERSPC	European Randomized Study of Screening for Prostate Cancer
GEE	generalized estimation equation
HLS	Hosmer-Lemeshow statistic
IF	impurity function
i.i.d.	independent and identically distributed
IPD	individual patient data
KNN	k-nearest neighbor
LASSO	least absolute shrinkage and selection operator
lpsa2	log base two transformed prostate-specific antigen
MRI	magnetic resonance imaging
MSE	mean squared error
MSKCC	Memorial Sloan Kettering Cancer Center
PBCG	Prostate Biopsy Collaborative Group
PCPT	Prostate Cancer Prevention Trial
PCPTRC	Prostate Cancer Prevention Trial risk calculator

PSA	prostate-specific antigen
RF	random forest
ROC-curve	receiver operating characteristic curve
SEER	Surveillance, Epidemiology, and End Results Program
SNP	single nucleotide polymorphism
UCSF	University of California San Francisco
USPSTF	U.S. Preventive Service Task Force
VA	Veterans Affairs

Contributing manuscripts

Several parts of this thesis have been published by or submitted to peer reviewed journals. The articles were written in collaboration with co-authors and they are listed in the following, along with the candidate's contribution.

Chapter 3 "Optimal Integration of Heterogeneous Cohorts for Global Prostate Cancer Risk Assessment"

Johanna Tolksdorf, Michael W. Kattan, Stephen A. Boorjian, Stephen J. Freedland, Karim Saba, Cedric Poyet, Lourdes Guerrios, Amanda De Hoedt, Michael A. Liss, Robin J. Leach, Javier Hernandez, Emily Vertosick, Andrew J. Vickers, Donna P. Ankerst. Visualization and internal validation strategies for building multi-cohort prediction models with an application to prostate cancer. Submitted, 2018.

Abstract

Background: Compared to models derived from single cohorts, those based on data from multiple cohorts should result in increased accuracy and applicability. Integrating data from multiple diverse clinical practices into an optimized model requires specialized techniques above that typically considered for single cohort models. The aim of this report was to develop a strategy as recently applied to create an online risk prediction tool for prostate cancer.

Methods: We created models for high-grade prostate cancer risk using six established risk factors. The data comprised 8492 prostate biopsies collected from ten institutions, 2 in Europe and 8 across North America. We calculated area underneath the receiver operating characteristic curve (AUC) for discrimination and the Hosmer-Lemeshow test statistic (HLS) for calibration. We implemented a 252-permutation-based sampling plan of all ways to split the ten cohorts into five used for training a model and five for testing to compare models, and calculated AUC and HLS statistics for omission of cohorts in training sets on individual cohorts as test sets.

Results: High-grade disease prevalence ranged from 18% in Zurich (1863 biopsies) to 39% in UT Health San Antonio (899 biopsies). Visualization revealed outliers in terms of risk factors, including San Juan VA (51% abnormal digital rectal exam), Durham VA (63% African American), and Zurich (2.8% family history). Exclusion of any cohort did not significantly affect the AUC or HLS in the 252-permutations, nor did the choice of prediction model (pooled, random-effects, meta-analysis). Excluding the lowest-prevalence Zurich cohort from training sets did not statistically significantly change the AUC or HLS for any of individual cohorts, except for Sunnybrook, where the effect on the AUC was minimal. Therefore the final multivariable logistic model was built by pooling the data from all cohorts using logistic regression. Higher prostate-specific antigen and age, abnormal digital rectal exam, African ancestry and a family history of prostate cancer increased risk of high-grade prostate cancer, while a history of a prior negative prostate biopsy decreased risk (all p-values < 0.004).

Conclusions: We have outlined a multi-cohort model-building internal validation strategy that is easily implemented and applicable for developing generally applicable risk tools.

Candidate's contribution

Statistical analysis of the data, discussion of results, and creation of figures and tables.

Chapter 4 "Development of a contemporary prostate cancer risk prediction model and comparison to the current standard"

Donna P. Ankerst, Johanna Straubinger, Katharina Selig, Lourdes Guerrios, Amanda De Hoedt, Javier Hernandez, Michael A. Liss, Robin J. Leach, Stephen J. Freedland, Michael W. Kattan, Robert Namg, Alexander Haese, Francesco Montorsi, Stephen A. Boorjian, Matthew R. Cooperberg, Cedric Poyet, Emily Vertosick, Andrew J. Vickers. A Contemporary Prostate Biopsy Risk Calculator Based on Multiple Heterogeneous Cohorts. European Urology 74(2): 197-203, 2018 (Ankerst, Straubinger, et al. 2018).

Abstract

Background: Prostate cancer prediction tools provide quantitative guidance for doctor-patient decision-making regarding biopsy. The widely used online Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) utilized data from the 1990s based on six-core biopsies and outdated grading systems.

Objective: We prospectively gathered data from men undergoing prostate biopsy in multiple diverse North American and European institutions participating in the Prostate Biopsy Collaborative Group (PBCG) in order to build a state-of-the-art risk prediction tool.

Design, setting, and participants: We obtained data from 15 611 men undergoing 16 369 prostate biopsies during 2006–2017 at eight North American institutions for model-building and three European institutions for validation.

Outcome measurements and statistical analysis: We used multinomial logistic regression to estimate the risks of high-grade prostate cancer (Gleason score ≥ 7) on biopsy based on clinical characteristics, including age, prostate-specific antigen, digital rectal exam, African ancestry, firstdegree family history, and prior negative biopsy.We compared the PBCG model to the PCPTRC using internal cross-validation and external validation on the European cohorts.

Results and limitations: Cross-validation on the North American cohorts (5992 biopsies) yielded the PBCG model area under the receiver operating characteristic curve (AUC) as 75.5% (95% confidence interval: 74.2–76.8), a small improvement over the AUC of 72.3% (70.9–73.7) for the PCPTRC ($p<0.0001$). However, calibration and clinical net benefit were far superior for the PBCG model. Using a risk threshold of 10%, clinical use of the PBCG model would lead to the equivalent of 25 fewer biopsies per 1000 patients without missing any high-grade cancers. Results were similar on external validation on 10377 European biopsies.

Conclusions: The PBCG model should be used in place of the PCPTRC for prediction of prostate biopsy outcome.

Patient summary: A contemporary risk tool for outcomes on prostate biopsy based on the routine clinical risk factors is now available for informed decision-making.

Candidate's contribution

Statistical analysis of the data, and creation of figures, tables and R-algorithm. Some parts of Ankerst, Straubinger, et al. 2018 are also included in the description of the data in Chapter 1.2.

1 Introduction

Based on U.S. data from the Surveillance, Epidemiology, and End Results Program (SEER) between 2013 and 2015, 39.3% of men and 37.7% of women had a lifetime risk of being diagnosed with cancer (Noone et al. 2018). In 2016, cancer was the second leading cause of death in the U.S. after heart diseases, counting for approximately one-fifth of all male and female deaths (Heron 2018). Stratified by age, cancer constituted at least the second most common cause for death for all groups, except for ages between 10 and 24 years. For ages between 45 and 64 years, it was the leading cause of death. In Germany the proportion of cancer-related deaths has remained nearly constant since the end of 1990, with approximately 28% for men and 22% for women (Barnes et al. 2016).

Decades of cancer research has helped to gain knowledge of the disease, leading to improvements in diagnosis as well as treatment (Kibbe et al. 2017). The increasing amount of cancer-related data and big data has resulted in the need for improvements in data management, analysis and interpretation. This thesis addresses in particular the use of patient data to perform individualized risk assessments of prostate cancer based on prostate biopsy outcomes.

1.1 Prostate Cancer

Figure 1 depicts for every country the most commonly diagnosed cancer for males in 2012 and shows the broad expansion of prostate cancer with 87 countries (Torre et al. 2016). Whereas prostate cancer is less common in Asia, it is the leading cancer type in most countries in North- and South-America and in Australia. Using SEER and the National Program of Cancer registries data, Siegel et al. 2019 and DeSantis et al. 2019 estimate new cancer

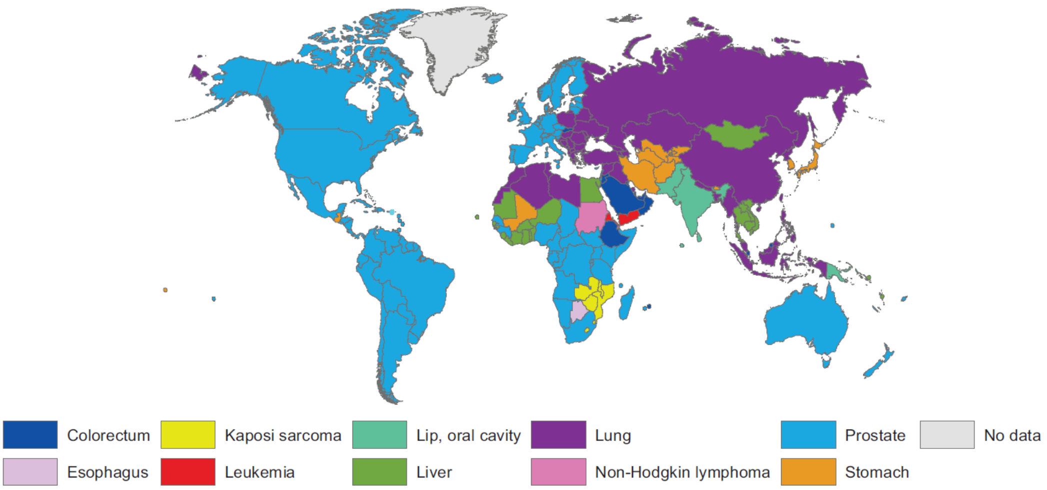
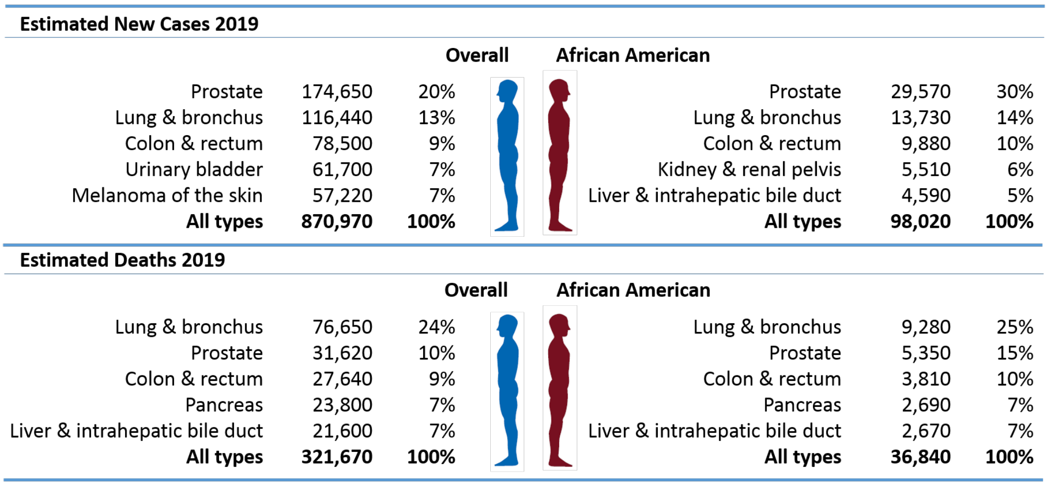


Figure 1 : World map of most commonly diagnosed cancers for males in 2012 (based on Torre et al. 2016).

cases and deaths in 2019 for the U.S. male population and U.S. African Americans, respectively. Figure 2 shows the estimated five most common types of cancer. The types lung and bronchus as well as colon and rectum are the estimated second and third most commonly new diagnosed cancers. Percentages of all cancer types except for prostate are similar for the overall U.S. population and African Americans, but the most common cancer prostate increases from 20% to 30% in African Americans. Prostate cancer is furthermore the second leading cause of mortality and rates also jump 5% in African Americans in comparison to all other cancers, where it remains the same. In fact, Torre et al. 2016 state that the highest prostate cancer incidence rates across registries around the world are held by African Americans, which also have the second highest mortality rates after Trinidad and Tobago.



Estimated New Cases 2019

	Overall		African American		
Prostate	174,650	20%	Prostate	29,570	30%
Lung & bronchus	116,440	13%	Lung & bronchus	13,730	14%
Colon & rectum	78,500	9%	Colon & rectum	9,880	10%
Urinary bladder	61,700	7%	Kidney & renal pelvis	5,510	6%
Melanoma of the skin	57,220	7%	Liver & intrahepatic bile duct	4,590	5%
All types	**870,970**	**100%**	**All types**	**98,020**	**100%**

Estimated Deaths 2019

	Overall		African American		
Lung & bronchus	76,650	24%	Lung & bronchus	9,280	25%
Prostate	31,620	10%	Prostate	5,350	15%
Colon & rectum	27,640	9%	Colon & rectum	3,810	10%
Pancreas	23,800	7%	Pancreas	2,690	7%
Liver & intrahepatic bile duct	21,600	7%	Liver & intrahepatic bile duct	2,670	7%
All types	**321,670**	**100%**	**All types**	**36,840**	**100%**

Figure 2 : Five most common cancer types for estimated new cases and deaths in 2019 for U.S. males (based on Siegel et al. 2019, DeSantis et al. 2019). Estimates are rounded to the nearest 10 and ranking is based on modeled projections. Basal cell and squamous cell skin cancers and in situ carcinoma except in the case of urinary bladder are excluded.

Screening for early signs of prostate cancer has enabled earlier and therefore more effective treatment for prostate cancer. Work by Stamey et al. 1987 on blood markers for adenocarcinoma of the prostate formed the basis for prostate-specific antigen (PSA) screening, introduced in the U.S. in the late 1980s. This practice considerably increased detection of prostate cancer as depicted in Figure 3 (Siegel et al. 2019). Welch and Albertsen 2009 investigated this peak in more detail and showed the evolution of prostate cancer incidence between 1986 and 2005 relative to 1986, stratified by age groups (supplementary Figure A.1). This analysis revealed an increase in detection within younger age groups and a decrease in elderly men after the introduction of PSA screening, which corresponds to the expectation of earlier detection. We might explain the drops in incidence in Figure 3 around 2010 by the U.S. Preventive Service Task Force (USPSTF) recommendation statements concerning screening for prostate cancer. In 2008 they recommended against PSA-based screening in men aged 75 years or older and adjusted their statement in 2012 to hold for all age groups (Moyer 2012). The current recommendation from 2018 further advocates the incorporation of additional factors to assess whether PSA screening is appropriate (Grossman et al. 2018). Reasons for these recommendations include concerns about overdiagnosis and overtreatment. Based

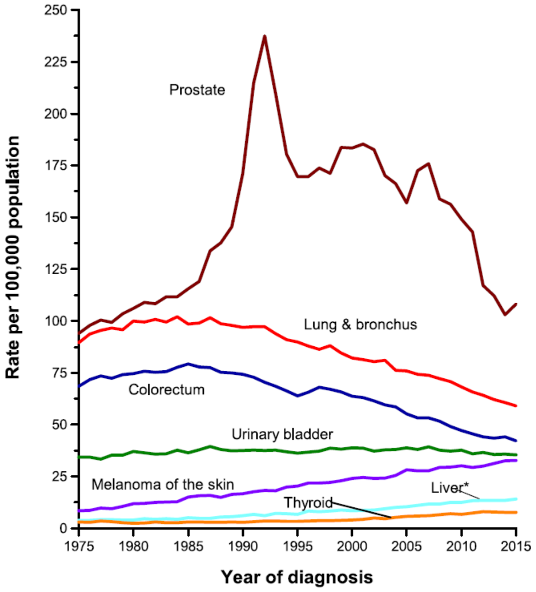


Figure 3 : Rates of selected cancer types in U.S. male population from 1975 to 2015 (based on Siegel et al. 2019). Rates are age adjusted to the 2000 U.S. standard population and adjusted for delays in reporting. * Includes intrahepatic bile duct.

on a literature review, Fenton et al. 2018 point to disadvantages, such as false positive PSA tests, resulting in unnecessary biopsies with their accompanying hassles. The authors furthermore report an increase in prostate cancer specific worries after a negative biopsy result following an increased PSA value. However, they also summarize randomized clinical trials that suggest prevented metastatic prostate cancer and death.

To address the problem of false positive assessments in PSA-based screening, multivariate risk predictions incorporating both, PSA and additional risk factors have been proposed (Chiu et al. 2017, I. M. Thompson, Leach, et al. 2014). Up to now there exists a vast amount of different models to be further discussed in Section 4.1. An evidence report and systematic review to update the USPSTF recommendation statement from 2012 in particular investigated the Prostate Cancer Prevention Trial (PCPT) and European Randomized Study of Screening for Prostate Cancer (ERSPC) risk calculators in 14 studies with a total of 28 cohorts and 48,234 biopsies (Fenton et al. 2018). The report found both risk tools to improve discrimination of PSA-based screening. For calibration however, underestimation and overestimation of both risk calculators occurred in several populations.

A patient might undergo a prostate biopsy if, for instance, a previous PSA test showed increased values, a physician found abnormalities during a digital rectal examination (DRE) or due to a high predicted risk from a model based on a combination of several risk factors. During the biopsy, the doctor uses a needle to gather a number of tissue samples from the prostate. A subsequent analysis identifies and classifies possible cancer. A systematic review by Loeb et al. 2013 of 213 references summarizes complications arising for this procedure, including bleeding, urinary tract infection, acute urinary retention, erectile dysfunction and fever. The most common complication is bleeding, whereby the authors point to broad ranges of reported rates of 10.0-84.0%, 1.1-93.0% and 1.3-45.0% for blood in urine, blood in sperm and rectal bleeding, respectively. Reasons for these differences across studies might include diverse definitions of bleeding, varying methods of data collection and cultural issues. However, Bjurlin et al. 2014 note that most men undergoing biopsy do not assess bleeding as major or even moderate events. Furthermore, Carlsson et al. 2011 and Pinsky et al. 2014 found that prostate biopsy does not significantly increase the risk of death within three centers of the ERSPC study comprising 50,194 men and the Prostate, Lung, Colorectal and Ovarian trial including 37,345 patients, respectively.

Following a positive prostate biopsy, several treatment options are available, whereby the primarily employed ones are surgery and radiation (Scarpato and Albertsen 2016). Even though Scarpato and Albertsen 2016 report an associated mortality risk reduction, these radical methods bear several risks to consider. These include surgical complications, perioperative mortality, urinary incontinence and erectile dysfunction (Fenton et al. 2018). Along with these disadvantages, many men with prostate cancer never experience any treatment-related symptoms (Grossman et al. 2018). The same study reports that men experiencing death from prostate cancer have a median age of 80 years and more than 66% are older than 75 years. Due to these circumstances, immediate radical treatment after positive biopsy for less significant prostate cancers has been questioned and trends toward active surveillance as a third treatment option have emerged (Parker 2004, Cooperberg et al. 2011). Patients electing active surveillance undergo close monitoring with PSA, DRE, magnetic resonance imaging (MRI) and repeated biopsies, followed by radical treatment if signs of more aggressive cancer appear (Bokhorst et al. 2016). Parker 2004 suggests that two-thirds of men might be able to avoid the harms of radical treatment without compromising their survival. More recent prospective clinical long-term studies confirm comparable survival of active surveillance and immediate radical treatment for low-risk patients (Klotz et al. 2015, Hamdy et al. 2016). In contrast to active surveillance, which aims to individualize treatment, watchful waiting intends to avoid treatment in elderly men with limited life expectancy; findings from these two approaches should not be confused (Carter 2012, Cooperberg et al. 2011, Herden and Weissbach 2018).

Overall, several important decisions arise along the path before prostate cancer detection until treatment. To begin, a symptomatic men may decide to undergo PSA-based screening for prostate cancer. Current USPSTF recommendations state that men between 55 and 69 years should discuss the decision to screen, based on their background and preferences, with their clinicians (Grossman et al. 2018). Furthermore, referral to biopsy should include an individual assessment of benefits and harms in light of the patient's characteristics (I. M. Thompson, Leach, et al. 2014). The choice of an appropriate treatment option is necessary in case of a positive biopsy result. Since there exists no consensus regarding an optimal practice, the patient's opinion is of particular importance (Carter 2012). Similar needs for decisions without one correct answer arise in several medical areas and they are the focus of shared decision-making (Elwyn et al. 2017). This practice describes situations where physicians decide along with patients concerning the use of diagnostic tests as well as therapeutic interventions, and can be described with the quote "no decision about me without me" (Ryan and Cunningham 2014). Patients are encouraged to consider diverse options, based on consultations by responsible clinicians, and vice versa, to communicate their preferences. Therefore, it is essential to have adequate and easy-to-understand information on harms and benefits. In particular, respective probabilities help the patients to weight advantages and disadvantages (Stiggelbout et al. 2012). Due to this development, it is not surprising that studies in PubMed on clinical prediction models have notably gained interest between 1970 and 2005 (Steyerberg 2009). In this thesis we investigate in particular risk prediction tools to support the decision concerning referral to prostate cancer biopsy.

1.2 Description of data

In this thesis prospective and retrospective data from the Prostate Biopsy Collaborative Group (PBCG) were used to build a risk prediction model for outcomes of prostate biopsy, and independent retrospective data from Hamburg were used for validation. Data from the PCPT risk calculator (PCPTRC) were assessed to build the existing risk model for comparison. The data are described in the following, with more details provided in Tables A.1 and A.2 of the Appendix.

1.2.1 PBCG data set

The PBCG collects prostate biopsy outcomes and risk factors from several international clinics with the exclusion criterion of a prior biopsy positive for prostate cancer. Patients may contribute multiple biopsies negative for cancer. The cohorts include the tertiary referral centers Cleveland Clinic (ClevelandClinic), Mayo Clinic (MayoClinic), San Raffaele (SanRaffaele), Zurich, Memorial Sloan Kettering Cancer Center (MSKCC), and University of California San Francisco (UCSF); the Veterans Affairs (VA) cohorts Durham (DurhamVA) and San Juan (SanJuanVA); and the sites Sunnybrook and UTHealth, which are consortia including main hospitals, tertiary referral centers and associated community urology providers. SanRaffaele and Zurich are European cohorts, whereas the remaining sites are located in North America. Figure 4 depicts the evolution of the respective number of biopsies collected. Prospective data collection started in 2014 and retrospective data are available between 2006 and 2014 for the cohorts MSKCC, Sunnybrook, UCSF, UTHealth and Zurich. The peak of Sunnybrook in 2015 might be explained by some centers joining the consortium for only one year and then dropping again.

Models analyzed in this thesis aim to support the decision of whether a patient should undergo prostate biopsy. We are therefore interested in prediction of clinically relevant outcomes of the biopsy using the standard risk factors collected in the clinic. As previously discussed, patients with less aggressive cancer might not benefit from treatment and detection of significant cancer might be more important. As a result we are furthermore interested in assess-

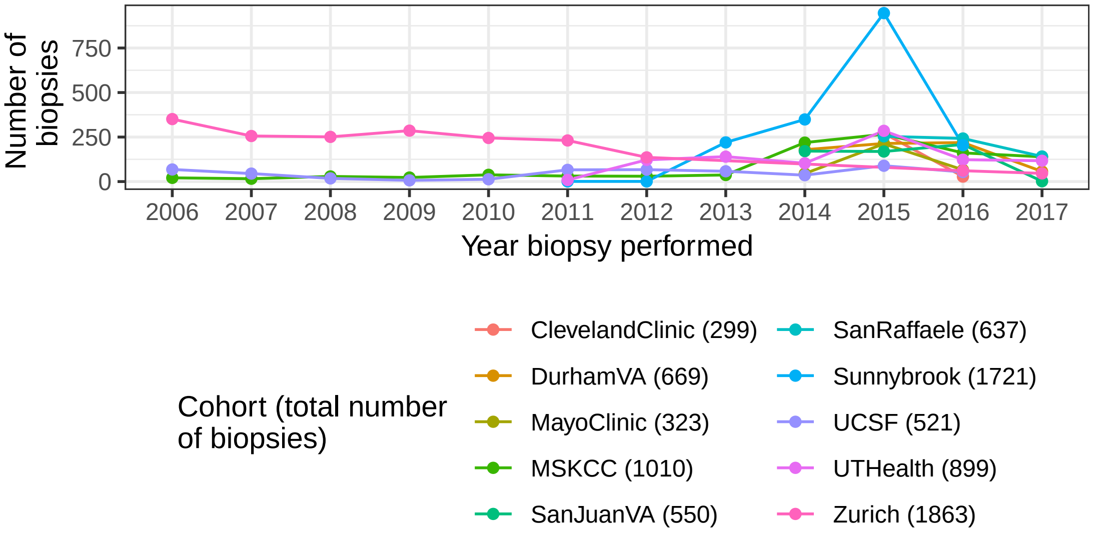


Figure 4 : Evolution of the number of patients from 2006 to 2017 for the PBCG cohorts. Data before 2014 were collected retrospectively and from 2014 onwards, prospectively.

ment of the severity of prostate cancer in case the biopsy is positive. Several classifications for severe prostate cancer exist and we discuss these in Section 4.1. For the analysis performed in this thesis, we decided to focus on a characterization based solely on the Gleason score as this is the most widely used in literature to date (I. M. Thompson, Ankerst, et al. 2006, Roobol et al. 2013, Chan et al. 2016).

The Gleason score is a universal grading scheme introduced by Donald Gleason in 1966 (Samaratunga et al. 2017, Gleason 1966). Several changes have been made to the originally proposed version, with latest improvements through the International Society of Urological Pathology consensus conferences in 2005 and 2014 (Epstein, Allsbrook, et al. 2005, Epstein, Egevad, et al. 2016). The criterion rests entirely upon the tumor architecture classified as Gleason grade one to five, summarized in Figure 5. Recommendations for reporting the Gleason score for a needle core biopsy suggest adding the most common grade observed across the considered cores to the highest occurring grade (Gordetsky and Epstein 2016). Biopsies utilizing six cores, as proposed by Hodge et al. 1989, have been the most popular biopsy method, whereby current practice predominantly utilizes eight to thirteen cores (Ceylan et al. 2014). Ankerst, Till, Boeck, Goodman, Tangen, Feng, et al. 2013 thereby found that higher numbers of cores correlate with increasing risk of positive biopsies. The change in clinical practice from six to more cores, along with its implication on the prevalence of positive biopsies, might pose the need to update existing analyses with contemporary data, which we further discuss in Chapter 4.

For the analyses of this thesis we define high-grade prostate cancer as Gleason score seven or higher, and low-grade cancer otherwise, following the classification of previous publications (I. M. Thompson, Ankerst, et al. 2006, Nam et al. 2007). Figure 6 shows the resulting prevalences of overall cancer, constituting low- and high-grade, and high-grade cancer. Visi-

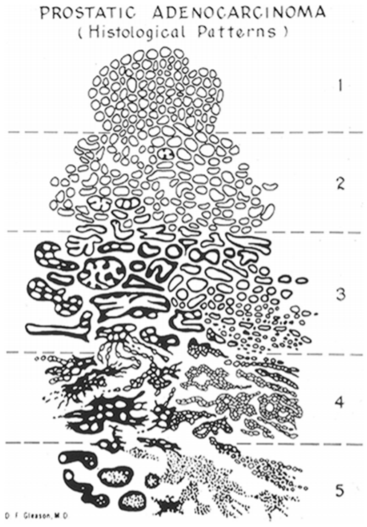


Figure 5 : The original Gleason grading system diagram (Gordetsky and Epstein 2016).

ble discrepancies are present across the diverse sites, with the highest rates for DurhamVA and UTHealth, and the lowest observed prevalence in Zurich.

The PBCG collects several diverse variables, all of which have potential predictive value. We concentrated on the six variables age, DRE, family history, prior biopsy, PSA and race since they are the established risk factors used in the most highly accessed online risk tools (Section 4.1). They incur minimal cost and inconvenience as they are routinely collected in medical practice. More specific variables from the PBCG suffer from a high proportion of missing values as will be shown later in the thesis. Since PBCG data acquisition is still ongoing and the expected increased amount of overall observations might compensate for the missing values, we delay analyses of these variables for future work. Chapter 6 discusses possible approaches. Figures 7 and A.2 in the Appendix provide an overview of the standard risk factors used in this thesis. These figures provide insights into heterogeneities across cohorts with respect to their risk factor distributions and their association to high-grade and overall cancer.

As previously discussed, PSA is a biomarker introduced for prostate cancer detection and a common tool for prostate cancer screening and active surveillance. For the PBCG, the PSA measurement at time of biopsy, or the closest measurement prior to biopsy, is reported and shows an anticipated strong association with high-grade prostate cancer across all PBCG cohorts (Figure 7). Here we have categorized patients into common PSA cutpoints of 4 and 10 ng/ml for referral to biopsy and great concern for cancer, respectively.

DRE is a procedure where the physician inserts his finger through the anus into the rectum to examine the patient's prostate for lumps, a subjective process depending among others on the experience of the doctor. Even though it can not provide an exact assessment on whether prostate cancer is present or not, Figure 7 shows a tendency for higher percentages

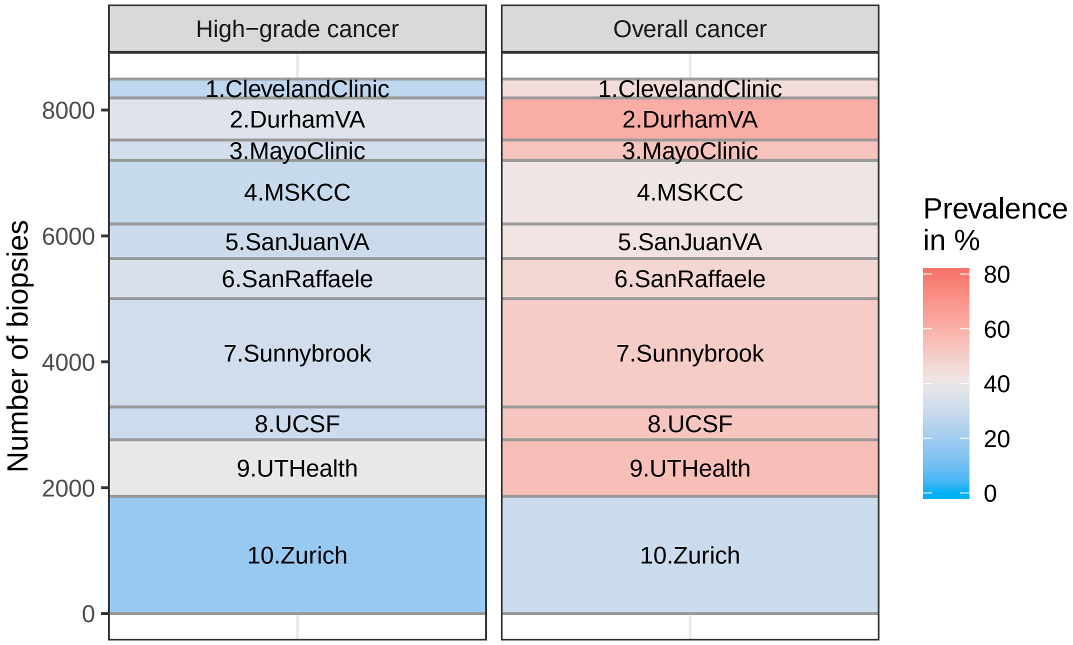


Figure 6 : Prevalence of overall and high-grade cancer for individual cohorts.

of high-grade cancer for positive DRE results. The near indifference for ClevelandClinic and SanJuanVA is anomalous and emphasizes the diversity of the sites. Several cohorts contain considerable amounts of missing DRE values, whereby low prevalences of high-grade prostate cancer for DurhamVA, MSKCC and UCSF, and high prevalences for SanJuanVA and Sunnybrook are found for patients with missing DRE.

Increasing age is known to be associated with higher rates of prostate cancer detection (Gupta et al. 2017, Welch and Albertsen 2009), and is confirmed in Figure 7.

Highest prostate cancer rates across the world's population are observed for U.S. African Americans (Torre et al. 2016, DeSantis et al. 2019). It is therefore not surprising that patients in cohorts with higher proportions of African ancestry have higher rates of high-grade cancer. It is, however, important to note that several sites have only very small amounts or even no patients with African ancestry. The high rate of missing values for Sunnybrook is concerning, in particular as its percentage of high-grade cancer is more similar to patients with African ancestry than without, suggesting a potential biased lack of reporting.

Family history describes a prior prostate cancer diagnosis in a first-degree relative. Kiciński et al. 2011 suggest that genetic factors, shared environment and/or similar food habits increases the risk of prostate cancer in the index patient. Figure 7 supports an association between family history and high-grade cancer for most cohorts. The prevalence of high-grade cancer for DurhamVA, MSKCC and SanRaffaele is, however, similar for patients with and without family history, and UCSF does not provide any information on family history.

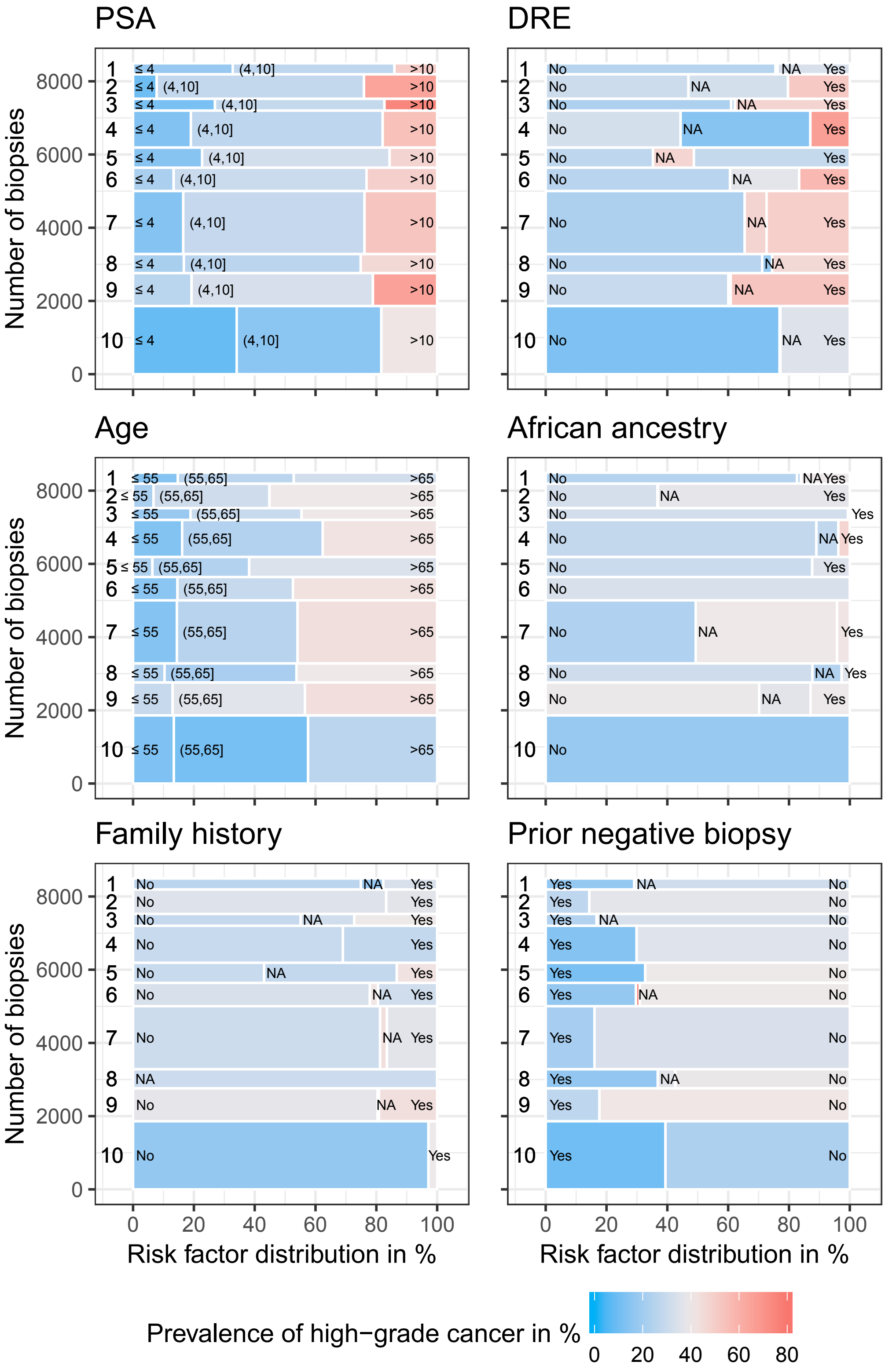


Figure 7 : Percentage of high-grade cancer by risk factor (x) and number of biopsies (y); 1.ClevelandClinic, 2.DurhamVA, 3.MayoClinic, 4.MSKCC, 5.SanJuanVA, 6.SanRaffaele, 7.Sunnybrook, 8.UCSF, 9.UTHealth, 10.Zurich. NA denotes missing values.

Finally, the PCPT has shown a reduced risk of a positive biopsy result for patients with a prior negative biopsy, coinciding with the trends observed across the ten cohorts in Figure 7 (I. M. Thompson, Ankerst, et al. 2006).

1.2.2 Additional data sources

Two additional data sources, from the Martini Klinik in Hamburg, Germany, and the North American PCPT, are available for Chapter 4, with a detailed description in the Tables A.1 and A.2 of the Appendix.

Data from Hamburg comprise retrospective electronic health records and thus suffer from many missing values and potential biases. In particular, for most biopsies it was not possible to determine whether the patient already had a prior positive biopsy or not. This was, however, an exclusion criterion of the PBCG. Furthermore, Hamburg is, with $n = 7877$ recorded biopsies, almost as large as the ten other PBCG sites combined, comprising $n = 8492$ observations. We therefore excluded Hamburg from model building due to a potential bias and only used it as a validation set.

To illustrate the importance of basing risk tools on contemporary data, Chapter 4 assessed the PCPTRC. Ankerst, Hoefler, et al. 2014 describe this risk tool, with its underlying data of $n = 6664$ observations, in detail. The PCPT required an age of at least 55 years, a PSA$\leq$3 ng/ml and a normal DRE to enter the study. Annual screening of the patients proceeded with referral to biopsy in case of an elevated PSA value $\geq$4 ng/ml or an abnormal DRE. The trial included an end of study biopsy after seven years, even for patients with normal PSA and DRE values throughout the study (I. M. Thompson, Goodman, Tangen, Parnes, et al. 2013).

The distribution of biopsy outcomes considerably differs between the PBCG collective, the clinical site Hamburg and the heavily screened PCPT study, as demonstrated in Table 1. In particular, PBCG and Hamburg have considerably lower percentages of patients without cancer and more high-grade cancer cases than the PCPT. Figures 8 and A.3 of the Appendix additionally provide a comparison of the standard variables with their associated biopsy outcomes for the three data sources. The PCPT consists of more patients with small PSA values, normal DRE results, no African ancestry and no family history, therefore showing characteris-

Study	No cancer	Low-grade cancer	High-grade cancer
PBCG	49.8	17.7	32.5
Hamburg	43.3	19.8	36.9
PCPT	82.1	14.1	3.8

Table 1 : Comparison of biopsy outcome prevalences in percent.

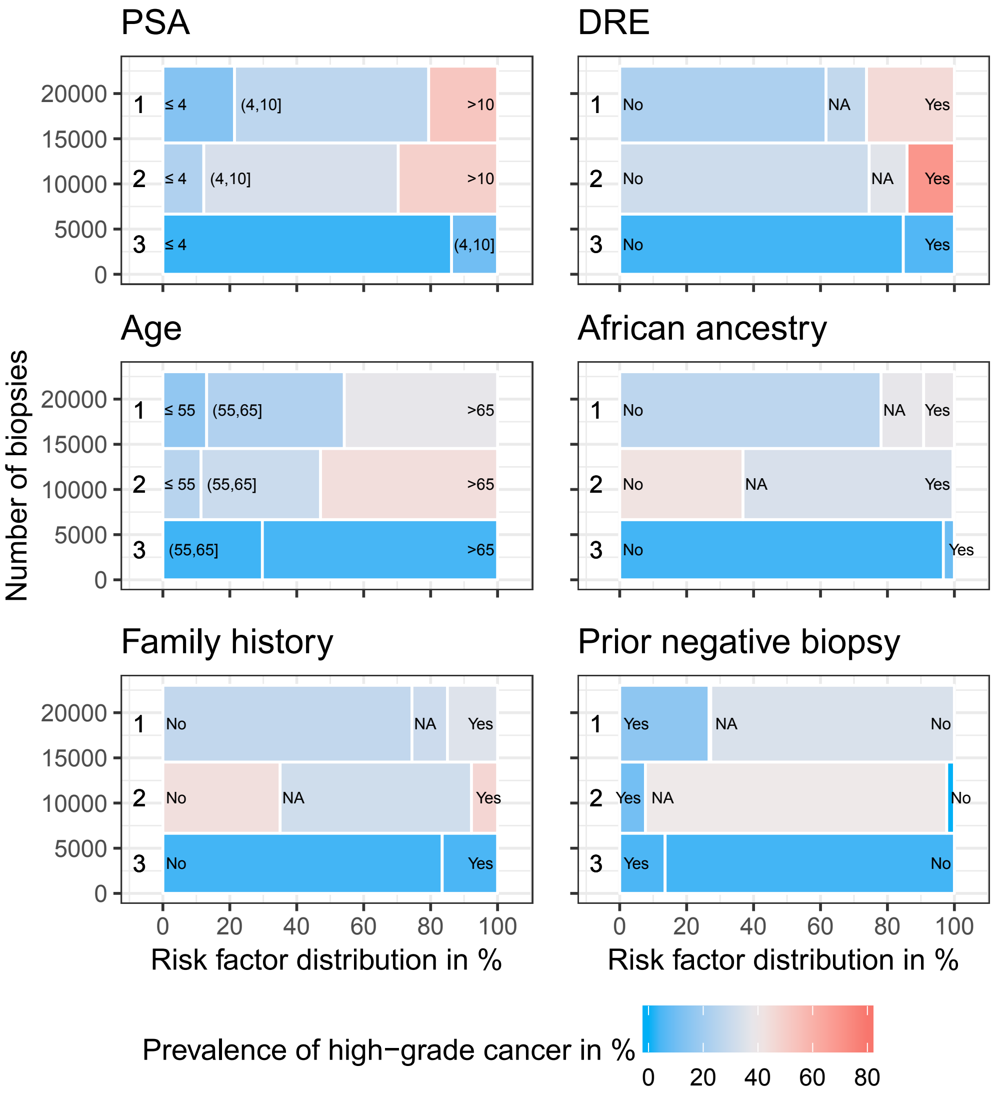


Figure 8 : Percentage of high-grade cancer by risk factor (x) and number of biopsies (y); 1.PBCG, 2.Hamburg, 3.PCPT. NA denotes missing values.

tics which are considered to be associated with lower prevalences of high-grade cancer. The clinical site Hamburg is missing 11.3%, 57.2%, 62.7% and 90.0% of the variables DRE result, family history, race and prior negative biopsy, making it hard to make population comparisons. Being a high-throughput referral clinic makes Hamburg more similar to the PBCG, where its percentages are more in alignment.

1.2.3 Exclusions and missing data imputation

For the PBCG data we exclude four observations due to missing Gleason scores and one patient with missing age. The Hamburg data exclusions were performed locally on site. For the development of the PCPTRC by Ankerst, Hoefler, et al. 2014, observations with a PSA value >10 ng/ml were excluded. All numbers within this thesis refer to the cleaned data sets with the stated exclusions.

Figure 7 displays missingness for the individual standard risk factors of the various PBCG cohorts, and shows almost no missing values of PSA and prior negative biopsies. DRE results are missing throughout most cohorts, with DurhamVA and MSKCC showing particularly high amounts. Race is predominantly missing for Sunnybrook. Finally family history is not reported for UCSF at all, shows missing values for approximately half of the observations from SanJuanVA, but is provided for most biopsies of the remaining sites.

For the models in Chapters 3 and 5 we performed median imputations, imputing a negative DRE result, no family history, no prior biopsy, and non African-American for race. In Chapter 4, the PCPTRC and the analogously developed PBCG model handle missing values for DRE, family history and prior negative biopsy with diverse partial models. We furthermore impute the variable race of the PBCG and Hamburg data by its median non-African American value.

1.3 Objectives of the thesis

In light of the current developments in the field of prostate cancer and the available data, the following aims will be addressed in this thesis.

1. Will incorporation of cohort heterogeneity improve validation of a global risk prediction model?
2. Should existing risk models be updated as soon as contemporary data are available?
3. Can more flexible machine learning methods improve traditional regression approaches for small sets of established risk factors?

The aims 1-3 are considered in Chapters 3-5, respectively, along with a discussion of the existing literature and the obtained results. The methods used for evaluation of the developed concepts are introduced in Chapter 2. At last, in Chapter 6 further approaches considered useful for future work based on the previous findings are outlined.

2 Methods for model evaluation

In this chapter we describe common evaluation measures for discrimination, calibration and net benefit, as proposed by Steyerberg 2009. We comprehensively derive confidence intervals (CIs) for the considered methods to enable a meaningful comparison between the models developed in the following Chapters 3-5.

2.1 Research in context

Evaluation of risk prediction models requires performance measures. These are particularly important for comparing models and assessing the usefulness of new markers incorporated into existing models (Pencina, D'Agostino, and Vasan 2008, D'Agostino 2006, Greenland and O'Malley 2005, Cook 2010). We furthermore advocate the use of CIs for the considered methods to show the reliability of resulting model assessments.

In terms of discrimination, the area under the receiver operating characteristic (ROC) curve (AUC) and the concordance-statistic (c-statistic), which coincide for binary outcomes, have a long history and are commonly used for evaluation of risk models (Bamber 1975). However, criticism is present, in particular if it is used as the only measure of performance (Cook 2007, Pepe and Janes 2008, Vickers and Elkin 2006, Hilden 1991, Kowall et al. 2013, Pennello et al. 2016, Wu et al. 2015, Pencina and D'Agostino 2015). The AUC equals the probability that for a randomly chosen patient with and without the outcome, named case and control, respectively, the considered model predicts a higher risk for the former. It is a summary measure across all possible risk thresholds between 0 and 100%. In practice, however, only a very restricted range of thresholds represents an area where accuracy is of concern as the decision for further diagnostic testing based on those thresholds could go either way.

For risk prediction of prostate cancer biopsy results, as a basis to decide whether a patient should be referred to biopsy or not, we assume the range from 5 to 25% to be relevant. Patients with a risk prediction beneath or above these thresholds would be advised to either not or definitely undergo further testing, respectively. The AUC might therefore be primarily influenced by thresholds never used in practice.

A further concern is that information on consequences are not included. In practice a false negative result might be more harmful than a false positive one. In this case, a model with higher specificity (lower false positive rate) and slightly worse sensitivity (higher false negative rate) across thresholds compared to another model might result in higher AUC, but can be considered less useful. Moreover, the AUC is only based on the ordering of predicted risks, not their actual values. As a consequence, multiplication of all probabilities with some scalar does not change the AUC.

Another concern with the AUC is its insensitivity with respect to additional strong predictors. The AUC hardly improves if new markers are added to an existing model, even though the predictive ability of the model might change. Consider for instance a change in risk prediction from 10 to 15% and from 9 to 5% for a case and non-case patient, respectively. The ordering of risks among these two patients remains identical and the AUC is unlikely to change, however, the new model might be preferred since there is a larger distance between the risks for this case/control pair of patients. CIs for the AUC, for instance based on DeLong et al. 1988, are routinely reported for differences in AUCs of diverse models or after addition of new markers to existing models (Gengsheng and Hotilovac 2008, Robin et al. 2011).

The ROC-curve represents the more transparent graphic behind the AUC, as the latter equals the integrated area beneath the ROC-curve, resulting in a single summary value. Like the AUC itself, it is commonly used and diverse papers also discuss estimates of corresponding CIs, for instance based on bootstrap or empirical likelihood methods (Su et al. 2009, Zhou and Qin 2005, Demidenko 2012, Hall et al. 2004, Martínez-Camblor et al. 2018, Lahiri and Yang 2017). The ROC-curve plots sensitivity versus 1-specificity of all possible thresholds between zero and one. It is typically non-transparent in regards to thresholds of the risk model corresponding to the values of sensitivity and specificity as these are not usually displayed on the graph. As discussed for the AUC, most thresholds are irrelevant for use in prostate cancer clinical prediction. We therefore analyze sensitivity and specificity individually, by plots for thresholds between 5 and 25% and corresponding CIs based on variance estimates for proportions. In contrast to other approaches, we do not assume the number of cases and controls to be fixed, but rather adjust for their variation across cohort.

It is reasonable to augment analysis of discrimination by other evaluation methods, like assessment of model calibration. Calibration describes how well observed and predicted values agree. The Hosmer-Lemeshow statistic (HLS) compares these within groups of patients, often defined by deciles of predicted risk (Hosmer and Lemeshow 1980). For more detailed visualization of observed outcomes versus predicted risks smoothing techniques enable calibration curves over the full range of predicted risks (Austin and Steyerberg 2014). Cleveland 1979 and Cleveland and Devlin 1988 discuss variation and resulting t-based approximate CIs for the smoothing technique of locally weighted regression, which we use in this thesis.

Net benefit curves, introduced by Vickers and Elkin 2006, are a rather new method to evaluate clinical usefulness of a risk prediction model. Several researchers already make use of them for model evaluation and comparison (Augustin et al. 2012, Pulleyblank et al. 2013, Allyn, Ferdynus, et al. 2016, Allyn, Allou, et al. 2017, Zastrow et al. 2015, Kondo et al. 2018). Requiring only the test data set and no additional information, they incorporate clinical consequences, and therefore address one of the criticisms of the AUC. They serve as a decision tool for whether to use a prediction model at all or which of several models to choose. Following the decision-theoretic justifications of Claxton 1999, Vickers, Cronin, et al. 2008 argue that CIs attached to the net benefit curves have only limited use for choosing between models. They recommend that the model with the best expected outcome should be selected, independently of statistical significance of difference in net benefit compared to other models. However, the authors acknowledge exceptions, such as when comparing a new method with current clinical practice or evaluating whether further research is of value. Simple bootstrap solutions are used to calculate CIs (Vickers, Cronin, et al. 2008, Talluri and Shete 2016, Kerr, Brown, et al. 2016). However, we develop analytic pointwise asymptotic CIs using that net benefit curves are a function of the prevalence of disease, and sensitivity and specificity of the risk tool.

A single evaluation method cannot address the multiple ways a model can fail to perform adequately in a new population. It is therefore useful to employ several measures to address different aspects of model validation. We use sensitivity and specificity, along with the summary measure AUC, for discrimination, calibration curves and the HLS summary for calibration, and net benefit curves for clinical utility. However, many further established measures exist, new methods constantly emerge and previous techniques evolve over time (Hilden 1991, Shapiro 1977, Tjur 2009, Nagelkerke 1991, Billheimer et al. 2014, Baker et al. 2009, Pennello et al. 2016, Steyerberg et al. 2010, Paul et al. 2013). They are introduced and adjusted to meet specific validation aspects of interest and to address disadvantages of former measures. Broadly known statistics include the Brier score, calculated as the mean of squared differences between the true outcomes and their respective predicted values, the discrimination slope, defined as the absolute difference in average risk prediction between cases and non-cases, and calibration-in-the-large and calibration slopes as quantitative enhancements to the graphical calibration curve (Brier 1950, Yates 1982, Cox 1958, Steyerberg et al. 2010). Recent developments are, for instance, predictiveness curves discussed by Pepe, Feng, Y. Huang, et al. 2008 and Y. Huang et al. 2007 to simultaneously display predictiveness and classification performance of risk markers or models. An extension to the net benefit curve is the weighted area under the net benefit curve, introduced by Talluri and Shete 2016 as a summary measure using an estimated distribution of threshold probabilities. Furthermore, Pencina, D'Agostino, and Vasan 2008 propose two new methods to evaluate the usefulness of an additional marker, the net reclassification improvement and the integrated discrimination improvement. These are broadly discussed and further developed in diverse literature sources (Pepe, Feng, and W. Gu 2008, Greenland 2008, Ware and Cai 2008, Cook 2008, Kerr, McClelland, et al. 2011, Z. Huang et al. 2016, Pencina, D'Agostino, and Steyerberg 2011, Pencina, D'Agostino, and Demler 2012 and Li et al. 2013).

2.2 Discrimination

Discrimination describes how well a risk prediction model differentiates between individuals with and without the outcome. In this thesis we primarily consider the outcome of high-grade cancer versus no or low-grade cancer. Discrimination is usually measured by sensitivity, specificity and the resulting ROC-curve along with its summary statistic, the AUC, and will be done so here.

For sensitivity, alternatively called the true positive rate (TPR), we consider all patients with the outcome high-grade cancer and report the proportion of patients that are correctly identified to have the outcome. However, a risk prediction model does not group patients into the categories with and without the outcome, but assigns them risks between zero and one. Sensitivity therefore depends on a threshold c that classifies all patients with predicted risk $> c$ as cases and $\leq c$ as controls:

$$\begin{aligned} p_{sens}(c) = TPR(c) &= P(predicted\, risk > c | high - grade\ cancer) \\ &= \frac{P(predicted\, risk > c\ and\ high - grade\ cancer)}{P(high - grade\ cancer)}. \end{aligned} \tag{2.1}$$

Due to its definition as a probability, $p_{sens}(c)$ takes values between zero and one, whereby high values are desirable. Let n_{total}, n_{cases} and $n_{cases,predicted\ risk>c}$ be the total number of patients, the number of patients with high-grade cancer and the number of patients with high-grade cancer and predicted risk $> c$, respectively. With these we estimate $p_{sens}(c)$ by

$$Sens(c) = \frac{\frac{n_{cases,predicted\ risk>c}}{n_{total}}}{\frac{n_{cases}}{n_{total}}} = \frac{n_{cases,predicted\ risk>c}}{n_{cases}}. \tag{2.2}$$

In order to report CIs for $p_{sens}(c)$, centered at the observed sensitivity $Sens(c)$, we have to derive the corresponding variance $Var(Sens(c))$. Let p_{prev} denote the true prevalence of the outcome. We assume that the total sample size n_{total} is fix and the number of patients with the outcome, n_{cases}, is random and follows a binomial distribution:

$$\begin{aligned} & n_{cases} \sim Bin(n_{total}, P(high - grade\ cancer)) \\ \Leftrightarrow \quad & n_{cases} \sim Bin(n_{total}, p_{prev}), \qquad (2.3) \\ & E(n_{cases}) = n_{total} p_{prev}, \qquad (2.4) \\ & Var(n_{cases}) = n_{total} p_{prev}(1 - p_{prev}). \qquad (2.5) \end{aligned}$$

With this the observed prevalence, $Prev = \frac{n_{cases}}{n_{total}}$, has mean

$$E(Prev) = E\left(\frac{n_{cases}}{n_{total}}\right) = \frac{1}{n_{total}} E(n_{cases}) \overset{(2.4)}{=} \frac{1}{n_{total}} n_{total} p_{prev} = p_{prev} \tag{2.6}$$

and variance

$$\begin{aligned} Var(Prev) &= Var\left(\frac{n_{cases}}{n_{total}}\right) = \left(\frac{1}{n_{total}}\right)^2 Var(n_{cases}) \\ &\overset{(2.5)}{=} \frac{1}{n_{total}^2} n_{total} p_{prev}(1 - p_{prev}) = \frac{p_{prev}(1 - p_{prev})}{n_{total}}. \end{aligned} \tag{2.7}$$

We use the binomial distribution conditioned on n_{cases} to describe the number of patients with high-grade cancer and a risk prediction greater than the considered threshold:

$$\begin{align} & n_{cases,predicted\ risk>c}|n_{cases} \sim Bin(n_{cases}, P(predicted\ risk > c|high - grade\ cancer)) \nonumber \\ \Leftrightarrow \quad & n_{cases,predicted\ risk>c}|n_{cases} \sim Bin(n_{cases}, p_{sens}(c)), \tag{2.8} \\ E(n_{cases,predicted\ risk>c}|n_{cases}) &= n_{cases}p_{sens}(c), \tag{2.9} \\ Var(n_{cases,predicted\ risk>c}|n_{cases}) &= n_{cases}p_{sens}(c)\left(1 - p_{sens}(c)\right). \tag{2.10} \end{align}$$

This results in the following conditional mean and variance for the observed sensitivity:

$$\begin{align} E\left(Sens(c)|n_{cases}\right) &= E\left(\left.\frac{n_{cases,predicted\ risk>c}}{n_{cases}}\right| n_{cases}\right) \nonumber \\ &= \frac{1}{n_{cases}}E\left(n_{cases,predicted\ risk>c}|n_{cases}\right) \nonumber \\ &\overset{(2.9)}{=} \frac{1}{n_{cases}}n_{cases}p_{sens}(c) = p_{sens}(c), \tag{2.11} \end{align}$$

$$\begin{align} Var(Sens(c)|n_{cases}) &= Var\left(\left.\frac{n_{cases,predicted\ risk>c}}{n_{cases}}\right| n_{cases}\right) \nonumber \\ &= \frac{1}{n_{cases}^2}Var(n_{cases,predicted\ risk>c}|n_{cases}) \nonumber \\ &\overset{(2.10)}{=} \frac{1}{n_{cases}^2}n_{cases}p_{sens}(c)\left(1 - p_{sens}(c)\right) \nonumber \\ &= \frac{p_{sens}(c)(1 - psens(c))}{n_{cases}}. \tag{2.12} \end{align}$$

For calculating the unconditional variance of the observed sensitivity, we use the law of total variance to get

$$\begin{align} Var(Sens(c)) &= Var\left(E(Sens(c)|n_{cases})\right) + E\left(Var(Sens(c)|n_{cases})\right) \nonumber \\ &\overset{(2.11),(2.12)}{=} Var(p_{sens}(c)) + E\left(\frac{p_{sens}(c)(1 - p_{sens}(c))}{n_{cases}}\right) \nonumber \\ &= 0 + \frac{p_{sens}(c)(1 - p_{sens}(c))}{n_{total}}E\left(\frac{n_{total}}{n_{cases}}\right) \nonumber \\ &= \frac{p_{sens}(c)(1 - p_{sens}(c))}{n_{total}}E\left(\frac{1}{Prev}\right) \nonumber \\ &= \frac{p_{sens}(c)(1 - p_{sens}(c))}{n_{total}}\frac{1}{p_{prev}}, \tag{2.13} \end{align}$$

which we estimate by $\widehat{Var}(Sens(c)) = \frac{Sens(c)(1-Sens(c))}{n_{cases}}$. In the last step we approximate $E(\frac{1}{Prev})$ by $\frac{1}{p_{prev}}$ based on the following considerations. The binomial distributed random variable n_{cases} can be rewritten as the sum of n_{total} independent bernoulli trials:

$$n_{cases} = \sum_{i=1}^{n_{total}} X_i, \tag{2.14}$$

$$X_i \sim Ber(p_{prev}) \quad i.i.d, \tag{2.15}$$

$$E(X_i) = p_{prev}, \tag{2.16}$$

$$Var(X_i) = p_{prev}(1 - p_{prev}). \tag{2.17}$$

By the central limit theorem the mean of X_i, $i = 1, ..., n$ converges in distribution to a normal distribution for $n \to \infty$:

$$\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - E(X_i)\right) \xrightarrow{d} \mathcal{N}(0, Var(X_i)), \tag{2.18}$$

$$\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - p_{prev}\right) \xrightarrow{d} \mathcal{N}(0, p_{prev}(1 - p_{prev})), \tag{2.19}$$

where $\xrightarrow{d}$ denotes convergence in distribution. By assuming n_{total} large enough and p_{prev} and $(1 - p_{prev})$ bounded away from 0 and 1, $Prev = \frac{1}{n_{total}} \sum_{i=1}^{n_{total}} X_i$ has approximately a normal distribution with mean p_{prev} and variance $\frac{p_{prev}(1-p_{prev})}{n}$. Let $f(x) = \frac{1}{x}$ with existing continuous derivative $\frac{\partial f(x)}{\partial x} = -\frac{1}{x^2} \neq 0$ for $x > 0$. Since p_{prev} and $p_{prev}(1 - p_{prev})$ are finite and assumed non-zero, we can apply the delta rule to approximate $f(Prev) = \frac{1}{Prev}$ by a normal distribution with mean

$$E\left(\frac{1}{Prev}\right) = f(p_{prev}) = \frac{1}{p_{prev}} \tag{2.20}$$

(Boos and Stefanski 2013).

A further basic discrimination metric is given by the specificity for the threshold c, which is defined similarly to the sensitivity to be

$$\begin{aligned} p_{spec}(c) &= P(predicted\ risk \leq c | no\ high - grade\ cancer) \\ &= \frac{P(predicted\ risk \leq c\ and\ no\ high - grade\ cancer)}{P(no\ high - grade\ cancer)}. \end{aligned} \tag{2.21}$$

Given a patient without high-grade cancer, specificity describes the probability of categorizing him correctly as not having high-grade cancer. Set $n_{controls}$ to be the number of patients without high-grade cancer, and $n_{controls,predicted\ risk \leq c}$ to be the number of patients without high-grade cancer and with predicted risk $\leq c$. The observed specificity becomes

$$Spec(c) = \frac{\frac{n_{controls,predicted\ risk \leq c}}{n_{total}}}{\frac{n_{controls}}{n_{total}}} = \frac{n_{controls,predicted\ risk \leq c}}{n_{controls}}. \tag{2.22}$$

An alternative to the specificity is the false positive rate (FPR), describing the proportion of patients without high-grade cancer that are incorrectly predicted to have the outcome:

$$FPR(c) = P(predicted\ risk > c | no\ high - grade\ cancer) = 1 - p_{spec}(c). \qquad (2.23)$$

Replacing p_{prev}, $Prev$, $p_{sens}(c)$, $Sens(c)$, $n_{cases,predicted\ risk>c}$ and n_{cases} in (2.3)-(2.13) with $(1 - p_{prev})$, $(1 - Prev)$, $p_{spec}(c)$, $Spec(c)$, $n_{controls,predicted\ risk\leq c}$ and $n_{controls}$ results in the variance estimate for specificity:

$$\widehat{Var}(Spec(c)) = \frac{Spec(c)(1 - Spec(c))}{n_{controls}}. \qquad (2.24)$$

We display discrimination graphically by separate curves for $Sens(c)$ and $Spec(c)$, with the threshold c on the x-axis and approximate corresponding pointwise 95%-CIs by $Sens(c) \pm 2 * \sqrt{\widehat{Var}(Sens(c))}$ and $Spec(c) \pm 2 * \sqrt{\widehat{Var}(Spec(c))}$, respectively. The ROC-curve, plots $TPR(c) = Sens(c)$ on the y-axis versus $FPR(c) = (1 - Spec(c))$ on the x-axis for all possible thresholds $c \in [0, 1]$ and forms the basis for the AUC. However, with this combined display of both measures we lose insights about individual thresholds.

Graphical evaluations are not always feasible, in particular, if several curves have to be summarized or compared at once. Therefore we integrate the area underneath the ROC-curve, resulting in the AUC. This statistic summarizes the previously discussed discrimination measures and takes values between zero and one. A good predictive model has low $FPR(c)$, as well as high $TPR(c)$, for all cutoffs c. We therefore favor values in the upper left corner of the ROC-curve, and a resulting high AUC. For a non-informative model that predicts the outcome no better than by 50:50 chance, the ROC-curve equals the diagonal line from (0,0) to (1,1), which leads to an AUC value of 0.5. For an intuitive interpretation we consider a randomly chosen patient with the outcome and another one without the outcome. The AUC equals the probability that the patient with the outcome has a higher predicted risk than the other patient. As for all evaluation methods, it is important to not merely report the point estimate itself, but also the corresponding CI. Therefore we use the estimated variance of the AUC derived by DeLong et al. 1988, based on theory of U-statistics by Hoeffding 1948. In the following we comprehensively outline the derivation of this estimation approach, which is implemented in the R package pROC using the algorithm proposed by Sun and W. Xu 2014 (Robin et al. 2011).

Let $X_i,\ i = 1, ..., n_{cases}$ and $Y_j,\ j = 1, ..., n_{controls}$ be the predicted risks for patients with and without high-grade cancer, respectively. Assume the X_i and Y_j are all independent with respective distributions F and G. Let $c_k,\ k = 1, ..., K$ and $c_k \neq c_l$ for $k \neq l$ denote the ordered set of all unique values of X_i and Y_j. Define c_0 to be an arbitrary value smaller than c_1. For every c_k the corresponding point on the observed ROC-curve is given by $\left(\frac{Number\ of\ Y_j > c_k}{n_{controls}}, \frac{Number\ of\ X_i > c_k}{n_{cases}}\right)$. The resulting curve ranges from $(1, 1)$ for c_0 to $(0, 0)$ for c_K. Figure 9 shows an example for $K = 5$ diverse cutoffs. In order to calculate the observed AUC with the trapezoidal rule, we first calculate the area of the individual trapezoids A_k for all

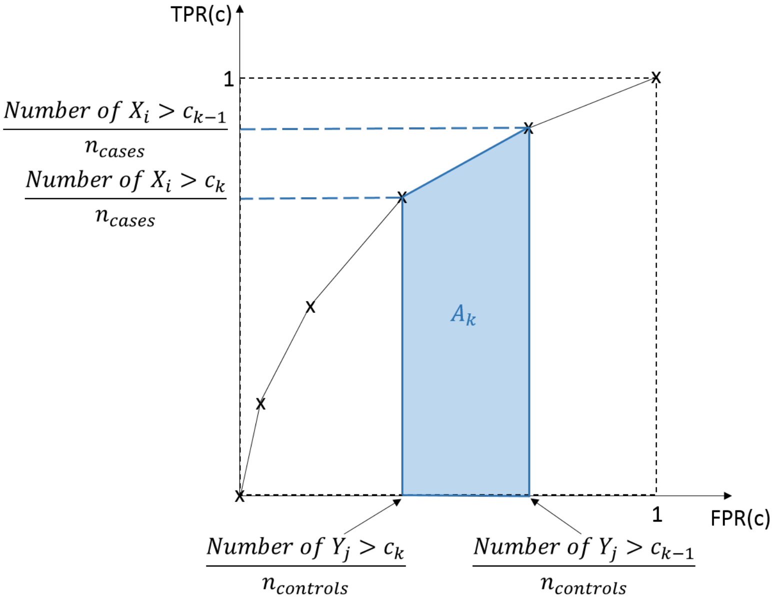


Figure 9 : Schematic representation of ROC-curve for $K = 5$ diverse c_k.

$k = 1, ..., K$:

$$\begin{aligned}
A_k &= \frac{1}{2}\left(\frac{Number\ of\ Y_j > c_{k-1}}{n_{controls}} - \frac{Number\ of\ Y_j > c_k}{n_{controls}}\right) \\
&\quad \cdot \left(\frac{Number\ of\ X_i > c_{k-1}}{n_{cases}} + \frac{Number\ of\ X_i > c_k}{n_{cases}}\right) \\
&= \frac{1}{2} \cdot \frac{Number\ of\ Y_j = c_k}{n_{controls}} \\
&\quad \cdot \left(\frac{Number\ of\ X_i > c_k}{n_{cases}} + \frac{Number\ of\ X_i = c_k}{n_{cases}} + \frac{Number\ of\ X_i > c_k}{n_{cases}}\right) \\
&= \frac{Number\ of\ Y_j = c_k}{n_{controls}} \cdot \frac{Number\ of\ X_i > c_k}{n_{cases}} \\
&\quad + \frac{1}{2} \cdot \frac{Number\ of\ Y_j = c_k}{n_{controls}} \cdot \frac{Number\ of\ X_i = c_k}{n_{cases}} \\
&= \frac{1}{n_{controls} n_{cases}} \sum_{j=1}^{n_{controls}} I_{Y_j = c_k} \sum_{i=1}^{n_{cases}} I_{X_i > c_k} \\
&\quad + \frac{1}{2 n_{controls} n_{cases}} \sum_{j=1}^{n_{controls}} I_{Y_j = c_k} \sum_{i=1}^{n_{cases}} I_{X_i = c_k} \\
&= \frac{1}{n_{controls} n_{cases}} \sum_{j=1}^{n_{controls}} \sum_{i=1}^{n_{cases}} \left(I_{Y_j = c_k} I_{X_i > c_k} + \frac{1}{2} I_{Y_j = c_k} I_{X_i = c_k}\right),
\end{aligned} \tag{2.25}$$

with $I_x = \begin{cases} 1 & if\ x \\ 0 & else \end{cases}$ denoting the indicator function for a logical expression x. The observed

AUC, denoted by $\widehat{AUC}$, is the sum of all trapezoids:

$$\begin{aligned}
\widehat{AUC} &= \sum_{k=1}^{K} A_k \\
&= \sum_{k=1}^{K} \frac{1}{n_{controls} n_{cases}} \sum_{j=1}^{n_{controls}} \sum_{i=1}^{n_{cases}} \left(I_{Y_j = c_k} I_{X_i > c_k} + \frac{1}{2} I_{Y_j = c_k} I_{X_i = c_k} \right) \\
&= \frac{1}{n_{controls} n_{cases}} \sum_{j=1}^{n_{controls}} \sum_{i=1}^{n_{cases}} \left(\sum_{k=1}^{K} I_{Y_j = c_k} I_{X_i > c_k} + \frac{1}{2} \sum_{k=1}^{K} I_{Y_j = c_k} I_{X_i = c_k} \right) \\
&= \frac{1}{n_{controls} n_{cases}} \sum_{j=1}^{n_{controls}} \sum_{i=1}^{n_{cases}} \left(I_{X_i > Y_j} + \frac{1}{2} I_{Y_j = X_i} \right) \\
&= \frac{1}{n_{cases} n_{controls}} \sum_{i=1}^{n_{cases}} \sum_{j=1}^{n_{controls}} \Psi\left(X_i, Y_j\right), \qquad (2.26)
\end{aligned}$$

where

$$\Psi(X, Y) = \begin{cases} 1 & Y < X \\ 0.5 & Y = X \\ 0 & Y > X \end{cases}. \qquad (2.27)$$

With this we get that $E\left(\widehat{AUC}\right) = P(Y < X) + 0.5P(Y = X) = AUC$, which reduces to $E\left(\widehat{AUC}\right) = P(Y < X)$ for variables with continuous distributions, as is the case here. This corresponds to the intuitive explanation of the AUC given previously.

In order to calculate CIs for the observed AUC, we derive an estimate of $Var\left(\widehat{AUC}\right)$, which is given by:

$$\begin{aligned}
Var\left(\widehat{AUC}\right) &= Var\left(\frac{1}{n_{cases} n_{controls}} \sum_{i=1}^{n_{cases}} \sum_{j=1}^{n_{controls}} \Psi(X_i, Y_j) \right) \\
&= \frac{1}{n_{cases}^2 n_{controls}^2} \sum_{i=1}^{n_{cases}} \sum_{j=1}^{n_{controls}} \sum_{i'=1}^{n_{cases}} \sum_{j'=1}^{n_{controls}} Cov\big(\Psi(X_i, Y_j), \Psi(X_{i'}, Y_{j'})\big).
\end{aligned} \qquad (2.28)$$

Define

$$\begin{aligned}
\Psi_{00} &= E\big(\Psi(X_i, Y_j)\big) = AUC, && (2.29) \\
\Psi_{11}(x_i, y_j) &= E\big(\Psi(X_i, Y_j) \big| X_i = x_i, Y_j = y_j\big) = \Psi(x_i, y_j), && (2.30) \\
\Psi_{10}(x_i) &= E\big(\Psi(X_i, Y_j) \big| X_i = x_i\big), && (2.31) \\
\Psi_{01}(y_j) &= E\big(\Psi(X_i, Y_j) \big| Y_j = y_j\big), && (2.32)
\end{aligned}$$

and note that all these functions have expectation AUC:

$$
\begin{aligned}
E(\Psi_{00}) &= E(AUC) = AUC, && (2.33)\\
E\big(\Psi_{11}(X_i, Y_j)\big) &= E\Big(E\big(\Psi(X_i, Y_j)\big|X_i, Y_j\big)\Big) = E\big(\Psi(X_i, Y_j)\big) = AUC, && (2.34)\\
E\big(\Psi_{10}(X_i)\big) &= E\Big(E\big(\Psi(X_i, Y_j)\big|X_i\big)\Big) = E\big(\Psi(X_i, Y_j)\big) = AUC, && (2.35)\\
E\big(\Psi_{01}(Y_j)\big) &= E\Big(E\big(\Psi(X_i, Y_j)\big|Y_j\big)\Big) = E\big(\Psi(X_i, Y_j)\big) = AUC. && (2.36)
\end{aligned}
$$

Furthermore let

$$
\begin{aligned}
\sigma^2_{00} &= Var(\Psi_{00}) = Var(AUC) = 0, && (2.37)\\
\sigma^2_{11} &= Var\big(\Psi_{11}(X_i, Y_j)\big) = Var\Big(E\big(\Psi(X_i, Y_j)\big|X_i, Y_j\big)\Big) = Var\big(\Psi(X_i, Y_j)\big), && (2.38)\\
\sigma^2_{10} &= Var\big(\Psi_{10}(X_i)\big) = Var\Big(E\big(\Psi(X_i, Y_j)\big|X_i\big)\Big), && (2.39)\\
\sigma^2_{01} &= Var\big(\Psi_{01}(Y_j)\big) = Var\Big(E\big(\Psi(X_i, Y_j)\big|Y_j\big)\Big). && (2.40)
\end{aligned}
$$

With this we get for i, i' in $1, ..., n_{cases}$ and j, j' in $1, ..., n_{controls}$:

$$Cov\big(\Psi(X_i, Y_j), \Psi(X_{i'}, Y_{j'})\big) = \sigma^2_{c_1,c_2}, \tag{2.41}$$

whereby $c_1 = \begin{cases} 1 & i = i' \\ 0 & i \neq i' \end{cases}$, and $c_2 = \begin{cases} 1 & j = j' \\ 0 & j \neq j' \end{cases}$. To show Equation (2.41), we first consider the case $c_1 = c_2 = 1$:

$$Cov\big(\Psi(X_i, Y_j), \Psi(X_{i'}, Y_{j'})\big) = Cov\big(\Psi(X_i, Y_j), \Psi(X_i, Y_j)\big) = Var\big(\Psi(X_i, Y_j)\big) = \sigma^2_{11}. \tag{2.42}$$

For $c_1 = c_2 = 0$, the terms $\Psi(X_i, Y_j)$ and $\Psi(X_{i'}, Y_{j'})$ are independent by definition, we therefore get $Cov\big(\Psi(X_i, Y_j), \Psi(X_{i'}, Y_{j'})\big) = 0 = \sigma^2_{00}$. At last, we have for $c_1 = 1,\ c_2 = 0$, and therefore $i = i'$ as well as $j \neq j'$,

$$
\begin{aligned}
&Cov\big(\Psi(X_i, Y_j), \Psi(X_{i'}, Y_{j'})\big) = Cov\big(\Psi(X_i, Y_j), \Psi(X_i, Y_{j'})\big)\\
&\quad = E\Big(\Big[\Psi(X_i, Y_j) - E\big(\Psi(X_i, Y_j)\big)\Big]\Big[\Psi(X_i, Y_{j'}) - E\big(\Psi(X_i, Y_{j'})\big)\Big]\Big)\\
&\quad = E\Big(\Big[\Psi(X_i, Y_j) - AUC\Big]\Big[\Psi(X_i, Y_{j'}) - AUC\Big]\Big)\\
&\quad = E\Big(E\Big(\big[\Psi(X_i, Y_j) - AUC\big]\big[\Psi(X_i, Y_{j'}) - AUC\big]\Big|X_i\Big)\Big)\\
&\quad = E\Big(\Big[E\big(\Psi(X_i, Y_j)\big|X_i\big) - AUC\Big]\Big[E\big(\Psi(X_i, Y_{j'})\big|X_i\big) - AUC\Big]\Big)\\
&\quad = E\Big(\big[\Psi_{10}(X_i) - AUC\big]\big[\Psi_{10}(X_i) - AUC\big]\Big)\\
&\quad = Var\big(\Psi_{10}(X_i)\big) = \sigma^2_{10}, \qquad (2.43)
\end{aligned}
$$

whereby we used that, conditioned on X_i, the terms $\Psi(X_i, Y_j)$ and $\Psi(X_i, Y_{j'})$ are independent. Analogously we obtain for $c_1 = 0,\ c_2 = 1$

$$Cov\big(\Psi(X_i, Y_j), \Psi(X_{i'}, Y_{j'})\big) = Cov\big(\Psi(X_i, Y_j), \Psi(X_{i'}, Y_j)\big) = \sigma_{01}^2. \tag{2.44}$$

In a next step we consider the number of possibilities $n_{c_1=1,c_2=1}$, $n_{c_1=0,c_2=0}$, $n_{c_1=1,c_2=0}$, and $n_{c_1=0,c_2=1}$ to get the various combinations of c_1 and c_2 for i, i' in $1, ..., n_{cases}$ and j, j' in $1, ..., n_{controls}$. For $c_1 = 1,\ c_2 = 1$ we have n_{cases} possibilities to choose i and $n_{controls}$ to choose j. Since $i' = i$ and $j' = j$, the overall amount is given by $n_{c_1=1,c_2=1} = n_{cases}n_{controls}$. For $c_1 = 0,\ c_2 = 0$ it follows $n_{c_1=0,c_2=0} = n_{cases}n_{controls}(n_{cases} - 1)(n_{controls} - 1)$, since there remain $(n_{cases} - 1)$ choices for i' after having chosen i, for j' analogously. For $n_{c_1=1,c_2=0}$, and analogously $n_{c_1=0,c_2=1}$, we get $n_{cases}n_{controls}(n_{controls} - 1)$ and $n_{cases}n_{controls}(n_{cases} - 1)$ possibilities. With this we get the variance of $\widehat{AUC}$ to be:

$$\begin{aligned}
Var\left(\widehat{AUC}\right) &= \frac{1}{n_{cases}^2 n_{controls}^2} \sum_{i=1}^{n_{cases}} \sum_{j=1}^{n_{controls}} \sum_{i'=1}^{n_{cases}} \sum_{j'=1}^{n_{controls}} Cov\big(\Psi(X_i, Y_j), \Psi(X_{i'}, Y_{j'})\big) \\
&= \frac{1}{n_{cases}^2 n_{controls}^2} \left(n_{c_1=1,c_2=1}\sigma_{11}^2 + n_{c_1=0,c_2=0}\sigma_{00}^2 + n_{c_1=1,c_2=0}\sigma_{10}^2 + n_{c_1=0,c_2=1}\sigma_{01}^2\right) \\
&= \frac{1}{n_{cases}^2 n_{controls}^2} \big(n_{cases}n_{controls}\sigma_{11}^2 + n_{cases}n_{controls}(n_{cases} - 1)(n_{controls} - 1) \cdot 0 \\
&\quad + n_{cases}n_{controls}(n_{controls} - 1)\sigma_{10}^2 + n_{cases}n_{controls}(n_{cases} - 1)\sigma_{01}^2\big) \\
&= \frac{1}{n_{cases}n_{controls}}\sigma_{11}^2 + \frac{n_{controls} - 1}{n_{cases}n_{controls}}\sigma_{10}^2 + \frac{n_{cases} - 1}{n_{cases}n_{controls}}\sigma_{01}^2.
\end{aligned} \tag{2.45}$$

Assume $\sigma_{11}^2 < \infty$ and $\frac{n_{cases}}{n_{total}} \xrightarrow{n_{total}\to\infty} prev \in (0, 1)$. As n_{total} becomes large, we get

$$Var\left(\widehat{AUC}\right) \xrightarrow{n_{total}\to\infty} 0 \cdot \sigma_{11}^2 + \frac{1}{n_{cases}}\sigma_{10}^2 + \frac{1}{n_{controls}}\sigma_{01}^2 \tag{2.46}$$

and

$$\sqrt{n_{total}}\left(\widehat{AUC} - AUC\right) \xrightarrow{d} \mathcal{N}\left(0, n_{total}\left(\frac{1}{n_{cases}}\sigma_{10}^2 + \frac{1}{n_{controls}}\sigma_{01}^2\right)\right). \tag{2.47}$$

In order to show this asymptotic normality, define

$$\hat{\theta}_1^* = \frac{1}{n_{cases}} \sum_{i=1}^{n_{cases}} \big(\Psi_{10}(X_i) - AUC\big), \tag{2.48}$$

$$\hat{\theta}_2^* = \frac{1}{n_{controls}} \sum_{j=1}^{n_{controls}} \big(\Psi_{01}(Y_j) - AUC\big). \tag{2.49}$$

Since $\Psi_{10}(X_i) - AUC$ are independent and independent and identically distributed (i.i.d.) with mean 0 and variance σ_{10}^2, the central limit theorem implies $\sqrt{n_{cases}}\hat{\theta}_1^* \xrightarrow{d} \mathcal{N}\left(0, \sigma_{10}^2\right)$, analogously for $\hat{\theta}_2^*$. The sum of these two approximately normal distributed random variables $\hat{\theta}^* = \hat{\theta}_1^* + \hat{\theta}_2^*$ is again approximately normal with mean 0 and variance $\frac{1}{n_{cases}}\sigma_{10}^2 + \frac{1}{n_{controls}}\sigma_{01}^2$.

It remains to show that $\sqrt{n_{total}}\hat{\theta}^*$ and $\sqrt{n_{total}}\left(\widehat{AUC}-AUC\right)$ are asymptotically equivalent and therefore have the same limiting distribution. For this we show

$$E\left(\left(\sqrt{n_{total}}\hat{\theta}^* - \sqrt{n_{total}}\left(\widehat{AUC}-AUC\right)\right)^2\right) \xrightarrow{n_{total}\to\infty} 0: \tag{2.50}$$

$$\begin{aligned}
&E\left(\left(\sqrt{n_{total}}\hat{\theta}^* - \sqrt{n_{total}}\left(\widehat{AUC}-AUC\right)\right)^2\right)\\
&= n_{total}E\left(\left(\hat{\theta}^*\right)^2\right) + n_{total}E\left(\left(\widehat{AUC}-AUC\right)^2\right) - 2n_{total}E\left(\hat{\theta}^*\left(\widehat{AUC}-AUC\right)\right),
\end{aligned} \tag{2.51}$$

with

$$E\left(\left(\hat{\theta}^*\right)^2\right) = Var\left(\hat{\theta}^*\right) + E\left(\hat{\theta}^*\right)^2 = \frac{1}{n_{cases}}\sigma_{10}^2 + \frac{1}{n_{controls}}\sigma_{01}^2, \tag{2.52}$$

$$\begin{aligned}
E\left(\left(\widehat{AUC}-AUC\right)^2\right) &= E\left(\widehat{AUC}^2\right) - 2AUC\cdot E\left(\widehat{AUC}\right) + AUC^2\\
&= E\left(\widehat{AUC}^2\right) - E\left(\widehat{AUC}\right)^2\\
&= Var\left(\widehat{AUC}\right) \xrightarrow{n_{total}\to\infty} \frac{1}{n_{cases}}\sigma_{10}^2 + \frac{1}{n_{controls}}\sigma_{01}^2,
\end{aligned} \tag{2.53}$$

and

$$\begin{aligned}
&E\left(\hat{\theta}^*\left(\widehat{AUC}-AUC\right)\right) = E\left(\hat{\theta}^*\right)E\left(\widehat{AUC}-AUC\right) + Cov\left(\hat{\theta}^*, \widehat{AUC}\right)\\
&= 0\cdot(AUC-AUC) + Cov\left(\hat{\theta}_1^* + \hat{\theta}_2^*, \widehat{AUC}\right)\\
&= Cov\left(\frac{1}{n_{cases}}\sum_{k=1}^{n_{cases}}\big(\Psi_{10}(X_k)-AUC\big), \frac{1}{n_{cases}n_{controls}}\sum_{i=1}^{n_{cases}}\sum_{j=1}^{n_{controls}}\Psi(X_i,Y_j)\right)\\
&\quad + Cov\left(\frac{1}{n_{controls}}\sum_{l=1}^{n_{conrols}}\big(\Psi_{01}(Y_l)-AUC\big), \frac{1}{n_{cases}n_{controls}}\sum_{i=1}^{n_{cases}}\sum_{j=1}^{n_{controls}}\Psi(X_i,Y_j)\right)\\
&= \frac{1}{n_{cases}^2 n_{controls}}\sum_{k=1}^{n_{cases}}\sum_{i=1}^{n_{cases}}\sum_{j=1}^{n_{controls}} Cov\big(\Psi_{10}(X_k), \Psi(X_i,Y_j)\big)\\
&\quad + \frac{1}{n_{cases}n_{controls}^2}\sum_{l=1}^{n_{controls}}\sum_{i=1}^{n_{cases}}\sum_{j=1}^{n_{controls}} Cov\big(\Psi_{01}(Y_l), \Psi(X_i,Y_j)\big)\\
&= \frac{n_{cases}n_{controls}}{n_{cases}^2 n_{controls}}\sigma_{10}^2 + \frac{n_{cases}n_{controls}}{n_{cases}n_{controls}^2}\sigma_{01}^2\\
&= \frac{1}{n_{cases}}\sigma_{10}^2 + \frac{1}{n_{controls}}\sigma_{01}^2.
\end{aligned} \tag{2.54}$$

We have used that the covariance of $\Psi_{10}(X_k)$ and $\Psi(X_i,Y_j)$ is 0 if $X_k \neq X_i$, and σ_{10}^2 otherwise. For a fixed k there remain $n_{controls}$ possibilities for $\Psi(X_i,Y_j)$ so that $X_k = X_i$. Since there are n_{cases} possibilities of choosing k, we have a total of $n_{cases}n_{controls}$ terms unequal to 0. Analogously we get $Cov\big(\Psi_{01}(Y_l), \Psi(X_i, Y_j)\big) = \begin{cases} 0 & Y_l \neq Y_j \\ \sigma_{01}^2 & Y_l = Y_j \end{cases}$ with $n_{cases} n_{controls}$ possibilities to get $Y_l = Y_j$. With this the convergence of (2.50) is shown.

To get an estimator for $Var\left(\widehat{AUC}\right)$, we consider σ_{10}^2 in more detail:

$$\sigma_{10}^2 = Var\big(\Psi_{10}(X_i)\big) = Var\Big(E\big(\Psi(X_i, Y_j)\big|X_i\big)\Big), \tag{2.55}$$

whereby we estimate $E\big(\Psi(X_i, Y_j)\big|X_i\big)$ by $\psi(X_i) = \frac{1}{n_{controls}} \sum_{j=1}^{n_{controls}} \Psi(X_i, Y_j)$. With this we get the estimate

$$\hat{\sigma}_{10}^2 = \frac{1}{n_{cases} - 1} \sum_{i=1}^{n_{cases}} \big(\psi(X_i) - \bar{\psi}(X_i)\big)^2, \tag{2.56}$$

with $\bar{\psi}(X_i)$ being the mean of $\psi(X_i)$, given by

$$\bar{\psi}(X_i) = \frac{1}{n_{cases}} \sum_{i=1}^{n_{cases}} \frac{1}{n_{controls}} \sum_{j=1}^{n_{controls}} \Phi(X_i, Y_j) = \widehat{AUC}. \tag{2.57}$$

Analogously we estimate σ_{01}^2 by

$$\hat{\sigma}_{01}^2 = \frac{1}{n_{controls} - 1} \sum_{j=1}^{n_{controls}} \left(\frac{1}{n_{cases}} \sum_{i=1}^{n_{cases}} \Psi(X_i, Y_j) - \widehat{AUC} \right)^2 \tag{2.58}$$

to get the overall estimate $\widehat{Var}\left(\widehat{AUC}\right) = \frac{1}{n_{cases}} \hat{\sigma}_{10}^2 + \frac{1}{n_{controls}} \hat{\sigma}_{01}^2$. Since we have shown an approximate normal distribution for $\widehat{AUC}$, we get the resulting 95%-CI

$$\widehat{AUC} \pm z_{0.975} \sqrt{\widehat{Var}\left(\widehat{AUC}\right)}, \tag{2.59}$$

with $z_{0.975} = 1.96$ the 0.975 quantile of the standard normal distribution.

2.3 Calibration

Calibration of a model indicates to which extent predicted and observed risks agree. We consider the HLS as a summarizing measure, and the calibration curve for a graphical display.

For the HLS we group patients by decile of their predicted probabilities, leading to 10 groups of patients ranging from low to high risk. Any other grouping is possible, but less common and derived analogously. Denote the number of patients in group i by n_i, and the average predicted risk in i by $\bar{p}_i$. This leads to an expected number of patients with and without high-grade cancer of $e_{y=1,i} = n_i \bar{p}_i$ and $e_{y=0,i} = n_i(1 - \bar{p}_i)$, respectively, for $i = 1, ..., 10$. We then compare the expected with the observed numbers of high-grade and no high-grade cancer patients in each group, denoted by $o_{y=1,i}$ and $o_{y=0,i} = n_i - o_{y=1,i}$. The HLS equals the sum over the squared differences between the expected and observed values, standardized by the expected values:

$$
\begin{aligned}
\text{HLS} &= \sum_{j=0}^{1}\sum_{i=1}^{10}\frac{(e_{y=j,i}-o_{y=j,i})^2}{e_{y=j,i}} = \sum_{i=1}^{10}\left(\frac{(e_{y=1,i}-o_{y=1,i})^2}{e_{y=1,i}}+\frac{(e_{y=0,i}-o_{y=0,i})^2}{e_{y=0,i}}\right) \\
&= \sum_{i=1}^{10}\left(\frac{(n_i\bar{p}_i-o_{y=1,i})^2}{n_i\bar{p}_i}+\frac{\big(n_i(1-\bar{p}_i)-(n_i-o_{y=1,i})\big)^2}{n_i(1-\bar{p}_i)}\right) && (2.60)\\
&= \sum_{i=1}^{10}\frac{(o_{y=1,i}-n_i\bar{p}_i)^2}{n_i\bar{p}_i(1-\bar{p}_i)}. && (2.61)
\end{aligned}
$$

With this measure we can compare the predicted risks of different models for the same test set, whereby smaller HLS values indicate better fit. However, we avoid comparisons across different test sets, as the HLS is sensitive to the underlying sample size (Kramer and Zimmerman 2007).

For a graphical display of the HLS, we plot the mean predicted risks $\bar{p}_i$ on the x-axis versus the corresponding observed average risks $\frac{o_{y=1,i}}{n_i}$ on the y-axis. We use the central limit theorem to approximate the proportion of observed cases by a standard normal distribution and get the pointwise 95%-CIs $\frac{o_{y=1,i}}{n_i} \pm z_{0.975} * \sqrt{\widehat{Var}\left(\frac{o_{y=1,i}}{n_i}\right)}$, with $z_{0.975} = 1.96$ the 0.975 quantile of the standard normal distribution. For the derivation of $\widehat{Var}\left(\frac{o_{y=1,i}}{n_i}\right)$, we assume the group size n_i fixed, and model the observed number of patients with high-grade cancer in group i by

$$o_{y=1,i} \sim Bin(n_i, p_{prev,i}), \tag{2.62}$$

with $p_{prev,i}$ as the true prevalence of high-grade cancer in group i. With this, the observed average risk in group i, $\frac{o_{y=1,i}}{n_i}$, has variance

$$Var\left(\frac{o_{y=1,i}}{n_i}\right) = \frac{1}{n_i^2}Var(o_{y=1,i}) = \frac{1}{n_i^2}n_i p_{prev,i}(1-p_{prev,i}) = \frac{p_{prev,i}(1-p_{prev,i})}{n_i}, \tag{2.63}$$

which can be estimated by $\widehat{Var}\left(\frac{o_{y=1,i}}{n_i}\right) = \frac{\frac{o_{y=1,i}}{n_i}\left(1-\frac{o_{y=1,i}}{n_i}\right)}{n_i}$.

In order to get a more detailed visualization, we use calibration plots, which show all individual risk predictions on the x-axis instead of group averages. However, the corresponding observed outcomes, to be plotted on the y-axis, are either zero or one. We therefore use locally weighted regression to get a smoothed line through the binary results. This approach is based on Cleveland 1979 and implemented in R with the command loess of the R Stats Package. The idea is to build a weighted regression for every risk prediction, based on observations with similar predicted risk.

First, we consider a single observation i with outcome $y_i \in \{0,1\}$ and predicted risk $p_i \in [0,1]$. For a given smoothing parameter $\alpha \in (0,1]$, we define the neighborhood size r to be α times the overall sample size n, rounded to the next integer. The r predicted risks closest to p_i, including p_i itself, are the neighborhood of p_i. Set $h_i = max(|p_i - p_k|)$, with p_k in the neighborhood of p_i. So h_i is the maximal distance of p_i to one of its neighbors. With this we define weights for every observation $k = 1, ..., n$, depending on the considered observation i:

$$w_k(p_i) = \begin{cases} \left(1 - \left|\frac{p_i - p_k}{h_i}\right|^3\right)^3 & \left|\frac{p_i - p_k}{h_i}\right| < 1 \\ 0 & \left|\frac{p_i - p_k}{h_i}\right| \geq 1 \end{cases}. \tag{2.64}$$

This weight function has the properties that

1. weights for observations in the neighborhood of p_i are greater than zero,
2. weights for observations depend only on their distance to p_i, not their exact location,
3. weights get smaller for observations that are further away from p_i and
4. weights are zero for observations outside the neighborhood of p_i.

For observation i, we estimate regression coefficients $\hat{\beta}_j(p_i),\ j = 0, ..., d$, of a polynomial regression of degree d of the outcomes $y_k \in \{0, 1\},\ k = 1, ..., n$ on the predicted risks $p_k \in [0, 1],\ k = 1, ..., n$. These are fitted by weighted least squares with weights as defined in Equation 2.64, by minimizing

$$\sum_{k=1}^{n} w_k(p_i) \left(y_k - \beta_0(p_i) - \beta_1(p_i)p_k - \beta_2(p_i)p_k^2 - ... - \beta_d(p_i)p_k^d\right)^2. \tag{2.65}$$

For the considered pair of predicted risk and observed outcome, (p_i, y_i), we get the smoothed point $(p_i, \hat{y}_i)$, with

$$\hat{y}_i = \sum_{j=0}^{d} \hat{\beta}_j(p_i)p_i^j, \tag{2.66}$$

the fitted value of the regression at p_i. Note that the smoothed outcome $\hat{y}_i$ can be outside the interval $[0, 1]$. This leads to the commonly observed problem of calibration curves for risk models falling outside the range of values reasonable for risks.

We repeat the previous steps of finding the neighborhood, setting weights and fitting a weighted regression, for all observations $i = 1, ..., n$. Finally we plot the smoothed points $(p_i, \hat{y}_i)$ with connecting lines in between.

In order to add pointwise CIs, we use estimated variances of the $\hat{y}_i$ based on Cleveland 1979. Consider the relation

$$y_i = f(p_i) + \epsilon_i, \qquad i = 1, ..., n, \tag{2.67}$$

where we assume $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d. Let $f : \mathbb{R} \to \mathbb{R}$ be a smooth function that is estimated by $\hat{y}_i$ at the points p_i:

$$\hat{f}(p_i) = \hat{y}_i = \sum_{j=0}^{d} \hat{\beta}_j(p_i)p_i^j, \qquad i = 1, ..., n. \tag{2.68}$$

Cleveland 1979 express $\hat{y}_i$ in Equation (2.68) in terms of a linear combination of the $y_k,\ k = 1, ..., n$ and coefficients $r_{i,k}$, which do not depend on $y_k,\ k = 1, ..., n$:

$$\hat{y}_i = \sum_{k=1}^{n} r_{i,k} y_k, \qquad i = 1, ..., n. \tag{2.69}$$

The derivation is not straightforward and omitted here. Let R be the $n \times n$ matrix with (i, k)th entry $r_{i,k} \in \mathbb{R}$, $\hat{y} = (\hat{y}_1, ...\hat{y}_n)' \in [0, 1]^n$ and $y = (y_1, ..., y_n)' \in \{0, 1\}^n$ with n-variate normal distribution, $y \sim \mathcal{N}_n(f(p), \sigma^2 I)$. Furthermore let $p = (p_1, ..., p_n)' \in \mathbb{R}^n$, $f(p) = \big(f(p_1), ..., f(p_n)\big)' \in \mathbb{R}^n$, $I \in \mathbb{R}^{n\times n}$ the identity matrix of size n and $\epsilon = (\epsilon_1, ..., \epsilon_n)' \in \mathbb{R}^n$. With this we obtain

$$\hat{y} = Ry = \begin{bmatrix} r_{1,1} & \cdots & r_{1,n} \\ \vdots & \ddots & \\ r_{n,1} & \cdots & r_{n,n} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \tag{2.70}$$

which follows a multivariate normal distribution with mean

$$E(\hat{y}) = E(Ry) = E\big(R(f(p) + \epsilon)\big) = RE(f(p)) + RE(\epsilon) = Rf(p), \tag{2.71}$$

and covariance matrix

$$Cov(\hat{y}) = Cov(Ry) = RCov(y)R' = R(\sigma^2 I)R' = \sigma^2 RR', \tag{2.72}$$

since $E(\epsilon) = (E(\epsilon_1), ..., E(\epsilon_n))' = (0, ..., 0)'$. Here we used that the expectation of a vector is defined as the vector of expectations of its components. Define furthermore,

$$\hat{\epsilon} = y - \hat{y} = y - Ry = (I - R)y, \tag{2.73}$$

which follows a multivariate normal with covariance matrix

$$\begin{aligned} Cov(\hat{\epsilon}) &= Cov(y - \hat{y}) = Cov(y - Ry) = Cov\big((I - R)y\big) = (I - R)Cov(y)(I - R)' \\ &= (I - R)\sigma^2 I(I - R)' = \sigma^2(I - R)(I - R)'. \end{aligned} \tag{2.74}$$

We estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{tr\big((I - R)(I - R)'\big)} \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \frac{1}{tr\big((I - R)(I - R)'\big)} \hat{\epsilon}'\hat{\epsilon}, \tag{2.75}$$

with $tr(X)$ the trace of a matrix X. To prove that $\hat{\sigma}^2$ is an approximate unbiased estimate of σ^2, we use that the expectation of a matrix is the matrix of expectations of individual elements, the following properties of the trace

$$tr(AB) = tr(BA), \quad A \in \mathbb{R}^{m\times n},\ B \in \mathbb{R}^{n\times m}, \tag{2.76}$$

$$tr(c) = c, \quad c \in \mathbb{R}, \tag{2.77}$$

$$tr(cA + B) = ctr(A) + tr(B), \quad c \in \mathbb{R},\ A, B \in \mathbb{R}^{n\times n}, \tag{2.78}$$

and the resulting equality

$$E(tr(A)) = tr(E(A)), \quad A \in \mathbb{R}^{n\times n}. \tag{2.79}$$

We consider

$$E\left(\hat{\sigma}^2\right) = \frac{1}{tr\big((I-R)(I-R)'\big)} E(\hat{\epsilon}'\hat{\epsilon}), \tag{2.80}$$

with

$$\begin{aligned} E(\hat{\epsilon}'\hat{\epsilon}) &= E\Big(\big((I-R)y\big)'(I-R)y\Big) = E\left(y'(I-R)'(I-R)y\right) \\ &\overset{(2.77)}{=} E\Big(tr\big(y'(I-R)'(I-R)y\big)\Big) \overset{(2.76)}{=} E\Big(tr\big((I-R)'(I-R)yy'\big)\Big) \\ &\overset{(2.79)}{=} tr\Big(E\big((I-R)'(I-R)yy'\big)\Big) \overset{(2.78)}{=} tr\left((I-R)'(I-R)E(yy')\right). \end{aligned} \tag{2.81}$$

We obtain

$$\begin{aligned} E(yy') &= E\Big(\big(f(p)+\epsilon\big)\big(f(p)+\epsilon\big)'\Big) = E\left(f(p)f(p)' + f(p)\epsilon' + \epsilon f(p)' + \epsilon\epsilon'\right) \\ &= E\left(f(p)f(p)'\right) + E\left(f(p)\epsilon'\right) + E\left(\epsilon f(p)'\right) + E\left(\epsilon\epsilon'\right) = f(p)f(p)' + \sigma^2 I, \end{aligned} \tag{2.82}$$

since $E\left(f(p)\epsilon'\right) = f(p)E\left(\epsilon'\right) = f(p)(0, ..., 0) = 0_{n,n} \in \mathbb{R}^{n\times n}$, a $n \times n$-matrix with all entries equal 0, similarly for $E\left(\epsilon f(p)'\right)$, and $E(\epsilon\epsilon') = \begin{bmatrix} E(\epsilon_1^2) & \dots & E(\epsilon_1\epsilon_n) \\ \vdots & \ddots & \\ E(\epsilon_n\epsilon_1) & \dots & E(\epsilon_n^2) \end{bmatrix} = \sigma^2 I$, as the ϵ_i are i.i.d. with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$. With this Equation (2.81) becomes

$$\begin{aligned} E(\hat{\epsilon}'\hat{\epsilon}) &= tr\left((I-R)'(I-R)E(yy')\right) = tr\left((I-R)'(I-R)(f(p)f(p)' + \sigma^2 I)\right) \\ &\overset{(2.78)}{=} tr\left((I-R)'(I-R)f(p)f(p)'\right) + \sigma^2 tr\left((I-R)'(I-R)\right) \\ &= tr\left((I-R)'(If(p) - Rf(p))f(p)'\right) + \sigma^2 tr\left((I-R)'(I-R)\right). \end{aligned} \tag{2.83}$$

The authors assume negligible bias in the fitted values, discussed in Cleveland and Devlin 1988, therefore assuming $f(p) = E(\hat{y}) = Rf(p)$. We approximate Equation (2.83) with

$$E(\hat{\epsilon}'\hat{\epsilon}) \approx 0 + \sigma^2 tr\left((I-R)(I-R)'\right), \tag{2.84}$$

and by inserting into Equation (2.80), obtain

$$E(\hat{\sigma}^2) \approx \frac{1}{tr\left((I-R)(I-R)'\right)} E(\hat{\epsilon}'\hat{\epsilon}) = \frac{1}{tr\left((I-R)(I-R)'\right)} \sigma^2 tr\left((I-R)(I-R)'\right) = \sigma^2, \tag{2.85}$$

so that $\hat{\sigma}^2$ is approximately unbiased for σ^2.

We obtain the estimated variance

$$\widehat{Var}(\hat{y}_i) = \hat{\sigma}^2 (RR')_{i,i} = \hat{\sigma}^2 \big(r_{i,1}, ..., r_{i,n}\big)\big(r_{i,1}, ..., r_{i,n}\big)' = \hat{\sigma}^2 \sum_{k=1}^{n} r_{i,k}^2. \tag{2.86}$$

Using a t-based approximation as argued in Cleveland and Devlin 1988 and implemented in the employed R package, we obtain the approximate 95%-CI for $\hat{y}_i$:

$$\hat{y}_i \pm t_{0.975,df} \sqrt{\widehat{Var}(\hat{y}_i)} = \hat{y}_i \pm t_{0.975,df} \sqrt{\hat{\sigma}^2 \sum_{k=1}^{n} r_{i,k}^2}, \tag{2.87}$$

with $t_{0.975,df}$ the 0.975 quantile of a t distribution with df degrees of freedom, given by

$$df = \frac{\Big(tr\big((I-R)(I-R)'\big)\Big)^2}{tr\Big(\big((I-R)(I-R)'\big)^2\Big)}. \tag{2.88}$$

Calibration curves should be close to the diagonal line from (0,0) to (1,1), as this indicates similar predicted and observed risks. If the calibration curve is, for instance, beneath this reference line, the predicted risks are higher than the actual observed ones.

2.4 Net benefit

Decision curve analysis takes into account weighting of true positive and false positive classifications and was introduced by Vickers and Elkin 2006. In this section we again consider predictions of the risk of high-grade cancer diagnosis on prostate biopsy. Based on this risk, a patient, or the respective doctor, will decide whether the patient is referred to biopsy. Decision analysis is now based on the benefit of this referral. The decision tree in Figure 10 shows the four possible outcomes for a patient with respective values $e,\ f,\ g,$ and h. The authors refer to $h - f$ as the harm of being referred unnecessarily, so the consequence of a false positive. It consists of the missed benefits $h > 0$ and the harm $f < 0$. They further define $e - g$ as the consequence of avoiding referral if high-grade cancer is present, so for a false negative. The expected net benefit of the decision policy to refer a patient with a predicted risk greater than

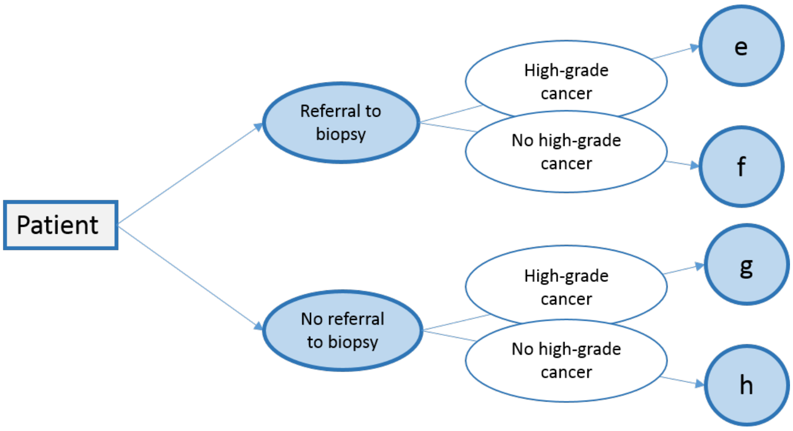


Figure 10 : Binary decision tree for referral to biopsy and diagnosis with high-grade cancer.

some threshold c is defined as

$$\begin{aligned}
&(e-g)P(predicted\ risk > c\ and\ high-grade\ cancer)\\
&-(h-f)P(predicted\ risk > c\ and\ no\ high-grade\ cancer)\\
=&(e-g)p_{sens}(c)p_{prev}-(h-f)(1-p_{spec}(c))(1-p_{prev})\\
=&(e-g)\Big(p_{sens}(c)p_{prev}-\frac{(h-f)}{(e-g)}(1-p_{spec}(c))(1-p_{prev})\Big),
\end{aligned} \tag{2.89}$$

where we used the definition of sensitivity and specificity of Section 2.2. If we consider this quantity to be measured in units of $(e-g)$, this expression simplifies to

$$p_{sens}(c)p_{prev}-\frac{(h-f)}{(e-g)}(1-p_{spec}(c))(1-p_{prev}), \tag{2.90}$$

however, this further complicates the interpretation of the final net benefit measure. Decision theory postulates that the optimal risk threshold c would weight the two potential outcomes of high-grade versus no high-grade cancer equally whether referred to biopsy or not:

$$\begin{aligned}
&ce+(1-c)f=cg+(1-c)h\\
\Leftrightarrow\ &c(e-g)=(1-c)(h-f)\\
\Leftrightarrow\ &\frac{(h-f)}{(e-g)}=\frac{c}{1-c}
\end{aligned} \tag{2.91}$$

(Kerr, Brown, et al. 2016). This reformulation leads to the definition of net benefit as given in Vickers and Elkin 2006:

$$net\ benefit(c)=p_{sens}(c)p_{prev}-\frac{c}{1-c}\big(1-p_{spec}(c)\big)(1-p_{prev}). \tag{2.92}$$

Note that Equation 2.92 does not explicitly include the value of referral, but implicitly accounts for it with the threshold c. This necessitates the assumption that c is chosen rationally, and therefore reflects benefit and harm of referral as in Equation (2.91). Kerr, Brown, et al. 2016 note an additional critical assumption that expected consequences do not depend on the predicted risks. We estimate the net benefit by

$$NetBen(c) = Sens(c)Prev - \frac{c}{1-c}\big(1 - Spec(c)\big)(1 - Prev), \tag{2.93}$$

and obtain estimated net benefit values for every possible threshold $c \in (0, 1)$, resulting in a net benefit curve. Since it remains difficult to interpret units of $net\ benefit(c)$ for a specific risk prediction model, we compare net benefit curves to two other decision rules. The first is the strategy of referring all patients to biopsy independently of their predicted risk. This effectively corresponds to a threshold $c = 0$, $p_{sens}(0) = 0$, $p_{spec}(0) = 1$, and hence to a net benefit of

$$\begin{aligned} net\ benefit\ all(c) &= p_{sens}(c)p_{prev} - \frac{c}{1-c}\big(1 - p_{spec}(c)\big)(1 - p_{prev}) \\ &= 1 \cdot p_{prev} - \frac{c}{1-c} \cdot 1 \cdot (1 - p_{prev}), \end{aligned} \tag{2.94}$$

estimated from the validation set as $NetBenAll(c) = Prev - \frac{c}{1-c}(1 - Prev)$. Despite that Vickers and Elkin 2006 effectively assume the threshold $c = 0$ for p_{sens} and p_{spec}, the net benefit of this rule is still a function of c to capture the relative size of the consequences of a false positive and false negative result. Analogously, we consider the decision rule of referring no patients to biopsy, corresponding to a threshold $c = 1$:

$$\begin{aligned} net\ benefit\ no(c) &= p_{sens}(c)p_{prev} - \frac{c}{1-c}\big(1 - p_{spec}(c)\big)(1 - p_{prev}) \\ &= 0 \cdot p_{prev} - \frac{c}{1-c} \cdot 0 \cdot (1 - p_{prev}) \\ &= 0. \end{aligned} \tag{2.95}$$

For a visual analysis, we plot the net benefit for a considered model and the two simple decision rules on the y-axis, versus all possible cutoffs $c \in (0, 1)$ on the x-axis. Vickers and Elkin 2006 model has clinical value if it has higher net benefit than the strategies of biopsying all or no patients in the range of thresholds reasonable for the considered disease. Analogously, we can compare two models in terms of their net benefit curves.

As for previous validation measures, it is useful to include CIs for net benefit curves. We apply the law of total variance to derive the variance of the observed net benefits, whereby we condition the interior moments on the observed prevalence:

$$Var\big(NetBen(c)\big) = Var\Big(E\big(Netben(c)\big|Prev\big)\Big) + E\Big(Var\big(Netben(c)\big|Prev\big)\Big). \tag{2.96}$$

For the conditional expectation we obtain

$$\begin{aligned} &E\big(NetBen(c)|Prev\big) = E\left(Sens(c)Prev - \frac{c}{1-c}\big(1 - Spec(c)\big)(1 - Prev)\middle| Prev\right) \\ &= E\big(Sens(c)Prev\big|Prev\big) - \frac{c}{1-c}E\Big(\big(1 - Spec(c)\big)(1 - Prev)\Big|Prev\Big) \\ &= Prev \cdot p_{sens}(c) - \frac{c}{1-c}(1 - Prev)(1 - p_{spec}) \end{aligned} \tag{2.97}$$

by linearity and the expectations for $Sens(c)$ and $Spec(c)$ from Section 2.2. Furthermore we use that, if conditioned on prevalence, sensitivity and specificity are independent:

$$\begin{aligned}
&Var\big(NetBen(c)\big|Prev\big) \\
&= Var\left(Sens(c)Prev - \frac{c}{1-c}\big(1 - Spec(c)\big)(1 - Prev)\middle|Prev\right) \\
&= Prev^2 Var\big(Sens(c)\big|Prev\big) + \left(-\frac{c}{1-c}\right)^2 (1 - Prev)^2 Var\big(1 - Spec(c)\big|Prev\big) \\
&= Prev^2 \frac{p_{sens}(c)\big(1 - p_{sens}(c)\big)}{n_{total} Prev} + \left(\frac{c}{1-c}\right)^2 (1 - Prev)^2 \frac{p_{spec}(c)\big(1 - p_{spec}(c)\big)}{n_{total}(1 - Prev)} \\
&= \frac{p_{sens}(c)\big(1 - p_{sens}(c)\big)}{n_{total}} Prev + \left(\frac{c}{1-c}\right)^2 (1 - Prev) \frac{p_{spec}(c)\big(1 - p_{spec}(c)\big)}{n_{total}}. \qquad (2.98)
\end{aligned}$$

With these we compute

$$\begin{aligned}
&Var\Big(E\big(NetBen(c)\big|Prev\big)\Big) \\
&= Var\left(Prev \cdot p_{sens}(c) - \frac{c}{1-c}(1 - Prev)\big(1 - p_{spec}(c)\big)\right) \\
&= Var\left(Prev\left(p_{sens}(c) + \frac{c}{1-c}\big(1 - p_{spec}(c)\big)\right) - \frac{c}{1-c}\big(1 - p_{spec}(c)\big)\right) \\
&= \left(p_{sens}(c) + \frac{c}{1-c}\big(1 - p_{spec}(c)\big)\right)^2 Var(Prev) \\
&= \left(p_{sens}(c) + \frac{c}{1-c}\big(1 - p_{spec}(c)\big)\right)^2 \frac{p_{prev}(1 - p_{prev})}{n_{total}}, \qquad (2.99)
\end{aligned}$$

and

$$\begin{aligned}
&E\Big(Var\big(NetBen(c)\big|Prev\big)\Big) \\
&= E\left(\frac{p_{sens}(c)\big(1 - p_{sens}(c)\big)}{n_{total}} Prev + \left(\frac{c}{1-c}\right)^2 (1 - Prev) \frac{p_{spec}(c)\big(1 - p_{spec}(c)\big)}{n_{total}}\right) \\
&= \frac{p_{sens}(c)\big(1 - p_{sens}(c)\big)}{n_{total}} E(Prev) + \left(\frac{c}{1-c}\right)^2 \big(1 - E(Prev)\big) \frac{p_{spec}(c)\big(1 - p_{spec}(c)\big)}{n_{total}} \\
&= \frac{p_{sens}(c)\big(1 - p_{sens}(c)\big)}{n_{total}} p_{prev} + \left(\frac{c}{1-c}\right)^2 (1 - p_{prev}) \frac{p_{spec}(c)\big(1 - p_{spec}(c)\big)}{n_{total}}. \qquad (2.100)
\end{aligned}$$

Overall, we get the variance

$$\begin{aligned}
&Var\big(NetBen(c)\big) = Var\Big(E\big(NetBen(c)\big|Prev\big)\Big) + E\Big(Var\big(NetBen(c)\big|Prev\big)\Big) \\
&= \left(p_{sens}(c) + \frac{c}{1-c}\big(1 - p_{spec}(c)\big)\right)^2 \frac{p_{prev}(1 - p_{prev})}{n_{total}} \\
&\quad + \frac{p_{sens}(c)\big(1 - p_{sens}(c)\big)}{n_{total}} p_{prev} + \left(\frac{c}{1-c}\right)^2 (1 - p_{prev}) \frac{p_{spec}(c)\big(1 - p_{spec}(c)\big)}{n_{total}} \qquad (2.101)
\end{aligned}$$

with its estimate

$$
\begin{aligned}
&\widehat{Var}\big(NetBen(c)\big) \\
&= \left(Sens(c) + \frac{c}{1-c}\big(1 - Spec(c)\big)\right)^2 \frac{Prev(1-Prev)}{n_{total}} \\
&\quad + \frac{Sens(c)\big(1-Sens(c)\big)}{n_{total}} Prev + \left(\frac{c}{1-c}\right)^2 (1-Prev) \frac{Spec(c)\big(1-Spec(c)\big)}{n_{total}} \\
&= \left(Sens(c) + \frac{c}{1-c}\big(1 - Spec(c)\big)\right)^2 Var(Prev) \\
&\quad + Var\big(Sens(c)\big) Prev^2 + \left(\frac{c}{1-c}\right)^2 Var\big(Spec(c)\big)(1-Prev)^2. \qquad (2.102)
\end{aligned}
$$

It is straightforward to calculate the variance of the observed net benefit for the strategy of referring all patients to biopsy by

$$
\begin{aligned}
Var\big(NetBenAll(c)\big) &= Var\left(Prev - \frac{c}{1-c}(1-Prev)\right) \\
&= Var\left(\frac{1}{1-c} Prev - \frac{c}{1-c}\right) \\
&= \left(\frac{1}{1-c}\right)^2 \frac{p_{prev}(1-p_{prev})}{n_{total}}, \qquad (2.103)
\end{aligned}
$$

along with the estimate $\widehat{Var}\big(NetBenAll(c)\big) = \left(\frac{1}{1-c}\right)^2 \frac{Prev(1-Prev)}{n_{total}}$.

With the derived variances, we can display clinical usefulness in terms of net benefit for all thresholds $c \in (0,1)$, with the corresponding approximated pointwise 95%-CIs $NetBen(c) \pm 2 * \sqrt{\widehat{Var}\big(NetBen(c)\big)}$. For the decision rules of referring all or no patients to biopsy, we obtain $NetBenAll(c) \pm 2 * \sqrt{\widehat{Var}\big(NetBenAll(c)\big)}$ and 0, respectively.

3 Optimal Integration of Heterogeneous Cohorts for Global Prostate Cancer Risk Assessment

The PBCG data set consists of information from several different cohorts. Integration of the data from individual patients across multiple cohorts can bring several advantages for the development of a generally applicable risk prediction model. Nevertheless there also arise difficulties by this diversity. This chapter focuses on different methods to predict the risk of high-grade cancer, whereby all available cohorts are incorporated, and also deals with the challenges arising from the clustered nature of the PBCG data. With this the chapter addresses the first aim of this thesis to assess the benefits of incorporating cohort heterogeneity.

3.1 Research in context

One of the goals of the PBCG consortium is to develop a statistical model that predicts the risk of having a positive biopsy for high-grade prostate cancer by evaluating the observed patient characteristics. A reliable assessment of individual patient risk may reduce the amount of unnecessary prostate biopsies, which is desirable as such examinations are accompanied by high distress for patients (Chapter 1). Models that assess the risk of having a specific disease can be characterized as diagnostic, and their development as well as validation play a pivotal role in medical research (Debray, Riley, et al. 2015).

The PBCG data set combines individual patient data (IPD) from several international cohorts. While most IPD were prospectively collected, some participating cohorts provided additional retrospective data (Section 1.2). This multi-cohort study design can be compared to a primarily prospective IPD meta-analysis (Riley, Lambert, et al. 2010). Even though his chapter focuses on diverse cohorts, the results can be applied analogously to various types of clusters. As Debray, Moons, Ahmed, et al. 2013 discuss, prediction research in general, and therefore also international collaborations such as the PBCG, have become more popular. Hence, adequate risk prediction methods that integrate diverse data are increasingly needed.

3.1.1 Advantages and disadvantages of multi-cohort studies

One of the main advantages of a multi-cohort data structure is the enhanced generalizability of the resulting model (Wynants et al. 2016, Sprague et al. 2009, Debray, Moons, Ahmed, et al. 2013). As the PBCG data comprise patients from ten cohorts from several countries, a broader range of risk factors is included in the study, as shown in Figure 7. The proportion of patients with African ancestry, for instance, exceeds 40% for DurhamVA, whereas other cohorts, such as MayoClinic, SanRaffaele, and Zurich, have at most two patients with African ancestry. Furthermore, cancer prevalence varies between the institutions, as illustrated in

Figure 6. Reasons for this diversity include differences in the biopsy procedure, such as the number of cores taken as well as the subjective assessments of resulting specimens by the local physicians. Typically more participating surgeons and clinical sites increase the generalizability to new patient populations (Sprague et al. 2009). However, this does not always hold. In particular many multi-cohort studies comprise data from several cohorts with exclusively Caucasian patients. In this case even a very high number of cohorts does not help for predicting minorities like African American or Hispanic patients.

Similar to traditional meta-analysis, the multi-cohort approach has the additional positive effect of an increased sample size (Sprague et al. 2009). Single studies for clinical trials can be too small, whereby meta-analysis is a possibility to overcome this problem (DerSimonian and Laird 1986). Advantages of combining several cohorts are not restricted to the increase in number of patients, but also often include a resulting decrease in recruitment time (Wynants et al. 2016, Sprague et al. 2009). This ensures contemporary data, a crucial component for valid risk prediction to be discussed in Chapter 4.

One important issue in conventional meta-analysis is the use of aggregate data, which might decrease power (Tierney et al. 2015). This problem can be dealt with by the use of IPD and the resulting IPD meta-analysis. However, a further disadvantage of meta-analyses, IPD or otherwise, is potential diversity in the study design as well as in the applied methods (DerSimonian and Laird 1986, Riley, Lambert, et al. 2010). Tierney et al. 2015 show that variation in quality across cohorts may also negatively effect reliability of results. These issues are negligible here, since the PBCG is a primarily prospective study. This ensures consistency across the different sites, including standardized assessment of outcomes, homogeneous recording of data and near identical collected variables. Overall, prospectively planned IPD meta-analysis maximize the power of meta-analyses (Riley, Lambert, et al. 2010).

Nevertheless the challenge of clustered data remains in multi-cohort studies due to varying clinical assessments or population types across the different sites (Bouwmeester, Twisk, et al. 2013, Pavlou et al. 2015, Wynants et al. 2016). Debray, Moons, Ahmed, et al. 2013 describe how heterogeneity in study populations primarily manifests in varying baseline risks, measured by the outcome prevalences, or in the association of the outcome with the covariates. Even though we have tried to explain the prevalence differences, shown in Figure 6, by the collected variables, it must be assumed that there remains unexplained clustering in the PBCG data set. Also, differing prevalences of patient characteristics across the various cohorts might have an influence on heterogeneity of the outcome prevalences.

3.1.2 Incorporation of cohort heterogeneity

A systematic review by Bouwmeester, Zuithoff, et al. 2012 showed that regression techniques integrating clustering were not commonly used for prediction models. It is therefore of interest, whether heterogeneity, as present in the PBCG data, can be ignored for the development of a general risk calculator. Several papers agree that the clustered nature of data should be taken into account, whereby different methods for an optimal integration of multiple studies or cohorts are considered (Abo-Zaid et al. 2013, Riley, Lambert, et al. 2010, DerSimonian and Laird 1986, Wynants et al. 2016, Debray, Moons, Ahmed, et al. 2013). Models ignoring heterogeneity of the data might provide attenuated effect estimates, which could even result in missing important diagnostic markers (Burke et al. 2017, Abo-Zaid et al. 2013). Inconsistent model performance as well as reduced generalizability are further possible problems (Debray, Moons, Ahmed, et al. 2013). In addition Pavlou et al. 2015 draw attention to the possibility of resulting incorrect standard errors for the coefficient estimates. In order to incorporate the clustering, various meta-analysis methods will be applied in this chapter and compared to a standard logistic regression model that ignores cohort heterogeneity.

3.1.3 Comparison of two- and one-stage meta-analyses

There are two general approaches for performing a meta-analysis of IPD, termed two- and one-stage meta-analyses. The traditional two-stage approach analyzes data from each cohort separately in a first step. Depending on the type of data, appropriate statistical methods produce aggregate data for each cohort individually. For the resulting summary statistics traditional meta-analysis methods are applied in the second stage to synthesize the available information. In general, the two-stage analyses can be also applied to combine several studies, for which only aggregate data are available. It is therefore not only a possible approach to perform an IPD meta-analysis, but it also represents the standard method for a traditional meta-analysis based on literature search. One of the main advantages is the resulting good documentation (Burke et al. 2017, Riley, Lambert, et al. 2010). Moreover, it is less complex for non-statisticians to use and understand these methods, especially since they are already broadly applied and therefore well known (Riley, Lambert, et al. 2010). The attractiveness of this approach is further enhanced as IPD need not leave the local site where it is analyzed, only aggregate summaries are forwarded for a central analysis. Experience shows that the process of centrally collecting data is time consuming and can be complicated even more by comprehensive data nondisclosure agreements. Since contemporary data are crucial, as discussed in Chapter 4, it is desirable to shorten this process if possible. However, this simplification bears the need for good statisticians at the individual cohorts. It also eliminates the possibility of an overall data cleaning and quality check, therefore increases the diversity across cohorts. An other disadvantage rises in case of imbalanced risk factor distributions. As some cohorts might have the same value of a covariate across all patients or do not report a risk factor at all, no estimates can be obtained for this cohort.

In contrast to this traditional approach, the availability of IPD also enables the possibility to analyze the data of all sites at once, whereby the clustered nature of the data has to be accounted for. This is called a one-stage meta-analysis. It uses a more exact statistical approach, which is especially important if only few studies are given, the number of observations within a study is small, or rare events are present (Abo-Zaid et al. 2013, Burke et al. 2017, Debray, Moons, Abo-Zaid, et al. 2013). In these cases, unstable estimates can occur for the two-stage approach, as the normal assumptions might not be appropriate (Abo-Zaid et al. 2013, Burke et al. 2017, Debray, Moons, Abo-Zaid, et al. 2013, Tierney et al. 2015). Nevertheless Burke et al. 2017 also refer to computational intensity and convergence problems of the one-stage meta-analysis.

Overall, it is commonly agreed, that both methods mostly produce comparable results, particularly when a large number of cohorts and observations are included or single treatment effects are estimated (Abo-Zaid et al. 2013, Burke et al. 2017, Debray, Moons, Abo-Zaid, et al. 2013, Riley, Lambert, et al. 2010, Tierney et al. 2015). However, there can be varying results, whereby Burke et al. 2017 emphasize that the origin of these differences lies mostly in varying model assumptions made for the two approaches. Nevertheless it is advisable to either specify the used methods beforehand or to report the results of both approaches (Burke et al. 2017, Tierney et al. 2015).

3.1.4 Predictions with random effects in one-stage meta-analysis

As previously discussed, two-stage meta-analysis methods are already broadly used and well documented. However, one-stage analyses of IPD lack detailed guidance to date. Methods to incorporate the clustered nature of a data set can be marginal, using for example generalized estimation equation (GEE) methods, which are robust to distribution mis-specifications, or conditional, such as mixed effects models, which require assumption of a normal distribution (Bouwmeester, Twisk, et al. 2013). Both approaches are commonly used, whereby the use of GEEs is less suitable for the prediction of individual patient risks (Pavlou et al. 2015, Bouwmeester, Twisk, et al. 2013). Random intercept models, which incorporate heterogeneous baseline risks for the different cohorts, are the simplest version of mixed effects methods, requiring one additional parameter to be estimated compared to a standard regression model. Random intercept models are discussed for modeling in the presence of clusters in several papers and they are also used in this chapter for comparison (Bouwmeester, Twisk, et al. 2013, Wynants et al. 2016, Debray, Moons, Ahmed, et al. 2013).

In order to get predictions from a random intercept model, one has to distinguish whether the predicted value is for a member of an existing, new or unknown cluster. In the first case it is straightforward to use the estimates for the fixed as well as the random effects (Bouwmeester, Twisk, et al. 2013). Even though a more accurate approach is to integrate over the posterior distribution of the random effects, both methods for existing cluster perform very similarly (Skrondal and Rabe-Hesketh 2009, Pavlou et al. 2015). Debray, Moons, Ahmed, et al. 2013 focus on risk prediction in new study populations, whereby information on new cohorts are incorporated. However, as the models in this thesis should be suitable as a global risk calculator, we concentrate on risk estimation for a patient from an unknown cohort. In practice these risks are often obtained by median predictions that set the random effects to zero, so assuming an average random cohort effect (Wynants et al. 2016). Pavlou et al.

2015 emphasizes that this simplification is technically incorrect to get marginal predictions and proper values are calculated by integrating over the distribution of the random effects, resulting in mean prediction for the population of clusters. However, the mis-calibration will be minor for weakly clustered data, as is often the case for different cohorts. In the subsequent analysis median and mean risk prediction methods for individuals of an unknown cohort are applied and compared to each other.

3.2 Methods for covariate selection

In this chapter only the standard variables PSA, first-degree family history, prior negative biopsy, race, age and DRE, as well as corresponding two way interactions are considered, as we focus on incorporating multiple cohorts rather than on choosing appropriate risk factors. For covariate selection a standard multiple logistic regression model is fit to each combination of 5 cohorts pooled together. In addition, it is built on each of the ten cohorts separately as well as to all cohorts pooled together. For each training set a stepwise selection algorithm based on the measure Bayesian information criterion (BIC) is used in order to choose covariates from the given standard variables and all their corresponding two way interactions.

3.2.1 Bayesian information criterion

The BIC compares models based on their fit to the given data, thereby penalizing for their complexity. It is analog to the Schwarz Criterion and the following derivation is based on Bhat and Kumar 2010 (Schwarz 1978). In this chapter we focus on a set of models $\mathcal{M}_m,\ m = 1, ..., M$, which differ in their included covariates. We consider the inclusion of six risk factors and the resulting 15 possible two-way interactions, therefore a total of 21 covariates. Including an intercept for every model, the number of parameters to be estimated, p_m, for model $\mathcal{M}_m$ therefore ranges between 1 and 22. In order to choose a suitable model among all candidates, we consider the posterior probability of a model, given the observed outcomes of high-grade cancer $y = (y_1, ..., y_n)$ for $i = 1, ..., n$ patients:

$$P(\mathcal{M}_m|y) = \frac{P(y|\mathcal{M}_m)P(\mathcal{M}_m)}{P(y)}, \tag{3.1}$$

by Bayes' theorem. We assume all models equally likely, and therefore use a uniform prior over models, so that $P(\mathcal{M}_m)$ is constant. Similarly $P(y)$ is constant with respect to the model choice, so that both terms can be deleted for the purpose of model selection. The remaining term is given by

$$P(y|\mathcal{M}_m) = \int_{\mathbb{R}^{p_m}} f(y|\theta_m)g_m(\theta_m)d\theta_m = \int_{\mathbb{R}^{p_m}} exp(log(f(y|\theta_m)))g_m(\theta_m)d\theta_m, \tag{3.2}$$

where $\theta_m \in \mathbb{R}^{p_m}$ is the vector of parameters in the model $\mathcal{M}_m$, $f(y|\theta_m)$ the density function of the data given parameters θ_m, and $g_m(\theta_m)$ denotes a prior on θ_m given the model $\mathcal{M}_m$. Next we approximate the log-likelihood $log(f(y|\theta_m))$ by its second order Taylor expansion about its maximum $\hat{\theta}_m$:

$$log(f(y|\theta_m)) \approx log(f(y|\hat{\theta}_m)) + (\theta_m - \hat{\theta}_m)' \left. \frac{\partial log(f(y|\theta_m))}{\partial \theta_m} \right|_{\theta_m = \hat{\theta}_m} + \frac{1}{2}(\theta_m - \hat{\theta}_m)' H_m (\theta_m - \hat{\theta}_m), \tag{3.3}$$

with H_m the Hessian matrix, evaluated at the maximum likelihood estimate $\hat{\theta}_m$, with resulting elements $(H_m)_{ij} = \left. \frac{\partial^2 log f(y|\theta_m)}{\partial(\theta_m)_i \partial(\theta_m)_j} \right|_{\theta_m = \hat{\theta}_m}$. Since $\hat{\theta}_m$ maximizes $log(f(y|\theta_m))$, the second term of Equation (3.3) vanishes, and H_m is negative definite. Let $F_m = -H_m$ be the observed Fisher information matrix. With this Taylor approximation we get

$$\begin{aligned} P(y|\mathcal{M}_m) \approx & \int_{\mathbb{R}^{p_m}} exp(log(f(y|\hat{\theta}_m)) - \frac{1}{2}(\theta_m - \hat{\theta}_m)' F_m (\theta_m - \hat{\theta}_m)) g_m(\theta_m) d\theta_m \\ = & f(y|\hat{\theta}_m) \int_{\mathbb{R}^{p_m}} exp\left(-\frac{1}{2}(\theta_m - \hat{\theta}_m)' F_m (\theta_m - \hat{\theta}_m)\right) g_m(\theta_m) d\theta_m. \end{aligned} \tag{3.4}$$

It is defensible to use the non-informative prior $g_m(\theta_m) = 1$ (Neath and Cavanaugh 2012). With this assumption the integral in Equation (3.4) simplifies to

$$\int_{\mathbb{R}^{p_m}} exp\left(-\frac{1}{2}(\theta_m - \hat{\theta}_m)' F_m (\theta_m - \hat{\theta}_m)\right) d\theta_m, \tag{3.5}$$

for which we substitute $X = \theta_m - \hat{\theta}_m$. Since $\frac{\partial(X + \hat{\theta}_m)}{\partial X} = 1$ we get $\int_{\mathbb{R}^{p_m}} exp(-\frac{1}{2} X' F_m X) dX$.

The observed Fisher information matrix F_m is symmetric and can therefore be diagonalized by $F_m = S' \Delta S$, with Δ a diagonal matrix with the eigenvalues λ_j of F_m. Thereby all λ_j are positive as F_m is positive definite and S is a orthogonal matrix satisfying $S' = S^{-1}$ and $|det(S')| = 1$. We use these properties for the substitution $X = S'U$, as the resulting Jacobian matrix is given by S':

$$\begin{aligned} \int_{\mathbb{R}^{p_m}} exp\left(-\frac{1}{2} X' S' \Delta S X\right) dX = & \int_{\mathbb{R}^{p_m}} exp\left(-\frac{1}{2} U' \Delta U\right) |det(S')| dU \\ = & \int_{\mathbb{R}^{p_m}} exp\left(-\frac{1}{2} \sum_{j=1}^{p_m} \lambda_j U_j^2\right) dU \\ = & \prod_{j=1}^{p_m} \int_{\mathbb{R}} exp\left(-\frac{1}{2} \lambda_j U_j^2\right) dU_j. \end{aligned} \tag{3.6}$$

The resulting integrals are Gaussian integrals with the solutions $\sqrt{\frac{2\pi}{\lambda_j}}$. With this the approximated posterior distribution gets

$$\begin{aligned} P(y|\mathcal{M}_m) \approx & f(y|\hat{\theta}_m)(2\pi)^{p_m/2} \frac{1}{\prod_{j=1}^{p_m} \lambda_j} \\ = & f(y|\hat{\theta}_m)(2\pi)^{p_m/2} \frac{1}{det(F_m)^{1/2}}. \end{aligned} \tag{3.7}$$

In a last step we further investigate the single entries of the observed Fisher information matrix F_m:

$$\begin{aligned}(F_m)_{i,j} =& - \left.\frac{\partial^2 logf(y|\theta_m)}{\partial(\theta_m)_i \partial(\theta_m)_j}\right|_{\theta_m=\hat{\theta}_m} \\ =& - \left.\frac{\partial^2 \sum_{k=1}^n logf(y_k|\theta_m)}{\partial(\theta_m)_i \partial(\theta_m)_j}\right|_{\theta_m=\hat{\theta}_m} \\ =& - \left.\frac{\partial^2 \frac{1}{n}\sum_{k=1}^n nlogf(y_k|\theta_m)}{\partial(\theta_m)_i \partial(\theta_m)_j}\right|_{\theta_m=\hat{\theta}_m} \\ =& - \left.\frac{\partial^2 E(nlogf(y_k|\theta_m))}{\partial(\theta_m)_i \partial(\theta_m)_j}\right|_{\theta_m=\hat{\theta}_m} \\ =& - n\left.\frac{\partial^2 E(logf(y_1|\theta_m))}{\partial(\theta_m)_i \partial(\theta_m)_j}\right|_{\theta_m=\hat{\theta}_m} = n(\tilde{F}_m)_{i,j}, \end{aligned} \tag{3.8}$$

with $\tilde{F}_m$ the Fisher information matrix for a single observation y_k. We have applied the weak law of large numbers on the random variable $nlogf(y_k|\theta_m)$, assuming i.i.d. outcomes $y_k,\ k = 1, ..., n$, as well as large n. Taking the logarithm of Equation (3.7) and multiplying it with minus two we now get:

$$\begin{aligned}-2log(P(y|\mathcal{M}_m)) \approx& - 2log(f(y|\hat{\theta}_m)) - \frac{p_m}{2}log(2\pi) + log(det(F_m)) \\ =& - 2log(f(y|\hat{\theta}_m)) - \frac{p_m}{2}log(2\pi) + log(n^{p_m}det(\tilde{F}_m)) \\ =& - 2log(f(y|\hat{\theta}_m)) - \frac{p_m}{2}log(2\pi) + p_m log(n) + log(det(\tilde{F}_m)) \\ \approx& - 2log(f(y|\hat{\theta}_m)) + p_m log(n) = BIC(\mathcal{M}_m). \end{aligned} \tag{3.9}$$

Assuming large n, we disregard all terms without n in the last step. Since the logarithm is a strictly monotone increasing function, choosing a model with minimal BIC is equivalent to choosing a model with largest approximate posterior probability $P(y|\mathcal{M}_m)$.

A similar frequently used criterion for model selection is given by the Akaike information criterion (AIC), defined as

$$AIC(\mathcal{M}_m) = -2log(f(y|\hat{\theta}_m)) + 2p_m. \tag{3.10}$$

The BIC includes a larger penalty of complex models for $n > e^2 \approx 7.4$, a model selection based on the BIC therefore tends to choose simpler models compared to the AIC. As Hastie et al. 2009 point out, the BIC is asymptotically consistent: In case the true model is included in the set of candidate models $\mathcal{M}_m$, p_m of the true model is finite and remains fixed as n increases, the probability of the BIC choosing this model approaches one as $n \to \infty$. In contrast, in this setting the AIC tends to choose models that are too complex. However, if one of the assumptions is violated, it is difficult to assess consistency and it might be preferred to consider efficiency for some loss function like the mean squared error (MSE) (Vrieze 2012). Then the AIC might be preferred as selection criterion as it is asymptotically efficient. Overall the choice of one of these two selection criteria over the other is based on the situation at hand and for the scope of this thesis we will use the BIC (Shao 1997).

3.2.2 Stepwise selection algorithm

As we consider six risk factors along with their two-way interactions for inclusion, we have a total of 21 possible covariates that we can either include or not. This results in $2^{21} = 2,0097,152$ candidate models. This number slightly reduced, as we require the main effects to be included in case a corresponding interaction term is used. However, it stays resource intensive to calculate the BIC of all possible models and then choose the one with the smallest BIC. Therefore we use a forward-backward stepwise selection algorithm to reduce the number of considered models.

The algorithm starts with the model that includes all possible covariates and calculates the corresponding BIC. In every step it first considers all models that include one additional covariate which is not part of the current model. If there exists at least one model with a lower BIC than the current one, the model with the lowest BIC is chosen, so that the corresponding covariate gets included in the current model. Next all currently incorporated covariates are considered for exclusion. The algorithm therefore calculates the BIC for all models with one covariate removed. If there exists a model with improved BIC, the covariate corresponding to the model with lowest BIC is excluded from the current model. As long as the BIC improves more than a given threshold, this step of considering all single covariates for either inclusion or exclusion is repeated. If the BIC does not change remarkably any more, the current model is considered optimal (Algorithm 1).

Algorithm 1 Stepwise selection

```
procedure STEP(full model)
    $M \leftarrow$ full model                    ▷ model including all possible covariates
    $c \leftarrow$ empty vector
    $\epsilon \leftarrow$ small threshold, e.g.: 0.1
    while $|BIC(M^{**}) - BIC(M)| > \epsilon$ do
        calculate $BIC$ for all models containing one covariate in $c$ added to $M$
        if at least one $BIC < BIC(M)$ then
            $M^* \leftarrow$ model with minimum $BIC$
            $c \leftarrow c$ without the covariate added for model $M^*$
        else$M^* \leftarrow M$
        calculate $BIC$ for all models containing one covariate in $M^*$ removed from $M^*$
        if at least one $BIC < BIC(M^*)$ then
            $M^{**} \leftarrow$ model with minimum $BIC$
            $c \leftarrow c$ with the covariate removed for model $M^{**}$
        else$M^{**} \leftarrow M^*$
    return $M^{**}$
```

3.3 Methods for modeling data from multiple cohorts

Using a suitable set of covariates, chosen according to 3.2, the following models are built and evaluated.

3.3.1 Standard multiple logistic regression model

As the simplest model a standard multiple logistic regression is used. This naive approach pools all cohorts together and an overall model is built on the whole data. Therefore the clustering of patients in different cohorts is ignored. Let y_{ji} denote the binary outcome, high-grade cancer versus low-grade cancer or no cancer, and $X_{ji} = (1, x_{1ji}, ..., x_{Kji})'$ is the $(K+1) \times 1$ covariate vector for the i^{th} observation of the j^{th} cohort, $i = 1, .., n_j$ and $j = 1, ..., J$. The logistic regression model can be written as:

$$
\begin{gathered}
y_{ji} \sim Bernoulli(p_{ji}) \\
log\left(\frac{E(y_{ji}|X_{ji})}{1-E(y_{ji}|X_{ji})}\right) = log\left(\frac{p_{ji}}{1-p_{ji}}\right) = \beta_0^{standard} + \sum_{k=1}^{K} \beta_k^{standard} x_{kji} = \left(\beta^{standard}\right)' X_{ji},
\end{gathered} \tag{3.11}
$$

where $E(.)$ is the expectation, p_{ji} the individual risks for high-grade cancer and $\beta^{standard} = (\beta_0^{standard}, ..., \beta_K^{standard})'$ the regression parameters. For the estimates $\hat{\beta}^{standard}$ of the regression coefficients maximum log-likelihood estimates are used. In the following the superscripts of the coefficients are neglected to ensure readability. The log-likelihood for the standard logistic regression is given by:

$$
\begin{aligned}
l(\beta) &= \sum_{j=1}^{J}\sum_{i=1}^{n_j} log\left[\left(\frac{e^{\beta' X_{ji}}}{1+e^{\beta' X_{ji}}}\right)^{y_{ji}} \left(\frac{1}{1+e^{\beta' X_{ji}}}\right)^{1-y_{ji}}\right] \\
&= \sum_{j=1}^{J}\sum_{i=1}^{n_j} log\left[\frac{e^{\beta' X_{ji} y_{ji}}}{1+e^{\beta' X_{ji}}}\right] \\
&= \sum_{j=1}^{J}\sum_{i=1}^{n_j} \left[y_{ji}\beta' X_{ji} - log\left(1+e^{\beta' X_{ji}}\right)\right].
\end{aligned} \tag{3.12}
$$

In order to maximize Equation (3.12), its derivative has to be set to zero:

$$
\frac{\partial l(\beta)}{\partial \beta} = \sum_{j=1}^{J}\sum_{i=1}^{n_j} \left[y_{ji} X_{ji} - \frac{e^{\beta' X_{ji}}}{1+e^{\beta' X_{ji}}} X_{ji}\right] = 0. \tag{3.13}
$$

In this thesis the solutions to these $K+1$ equations are approximated by the Newton-Raphson algorithm. Therefore the coefficient estimates get iteratively updated by the following rule:

$$\begin{aligned}\beta^{new} &= \beta^{old} - \left[\left(\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta'}\right)^{-1}\frac{\partial l(\beta)}{\partial\beta}\right]\Bigg|_{\beta=\beta^{old}}\\ &= \beta^{old} - \left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left[-\frac{e^{\left(\beta^{old}\right)'X_{ji}}}{\left(1+e^{\left(\beta^{old}\right)'X_{ji}}\right)^2}X_{ji}X'_{ji}\right]\right]^{-1}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left[y_{ji}X_{ji} - \frac{e^{\left(\beta^{old}\right)'X_{ji}}}{1+e^{\left(\beta^{old}\right)'X_{ji}}}X_{ji}\right].\end{aligned} \tag{3.14}$$

The algorithm specified by the updating rule 3.14 can be rewritten as an iteratively reweighted least squares algorithm, which is used in this thesis and implemented in the R statistical package (Hastie et al. 2009). As a starting value $\beta = 0$ can be used and due to the concave form of the log-likelihood convergence is very likely.

One obtains the corresponding estimated covariance matrix $\widehat{\sum}$ of the parameter estimates $\hat{\beta}$ by the negative inverse of the second derivative of the corresponding log-likelihood, evaluated at $\hat{\beta}$ (Hosmer, Lemeshow, and Sturdivant 2013):

$$\begin{aligned}\widehat{\sum} &= \begin{bmatrix} \widehat{Var}\left(\hat{\beta}_0\right) & \widehat{Cov}\left(\hat{\beta}_0,\hat{\beta}_1\right) & \dots & \widehat{Cov}\left(\hat{\beta}_0,\hat{\beta}_K\right)\\ \widehat{Cov}\left(\hat{\beta}_0,\hat{\beta}_1\right) & \widehat{Var}\left(\hat{\beta}_1\right) & \dots & \widehat{Cov}\left(\hat{\beta}_1,\hat{\beta}_K\right)\\ \vdots & \vdots & \ddots & \vdots\\ \widehat{Cov}\left(\hat{\beta}_0,\hat{\beta}_K\right) & \widehat{Cov}\left(\hat{\beta}_1,\hat{\beta}_K\right) & \dots & \widehat{Var}\left(\hat{\beta}_K\right)\end{bmatrix} = \left(-\frac{\partial^2 l\left(\beta\right)}{\partial\beta\partial\beta'}\right)^{-1}\Bigg|_{\beta=\hat{\beta}}\\ &= \left[\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left[\frac{e^{\hat{\beta}'X_{ji}}}{\left(1+e^{\hat{\beta}'X_{ji}}\right)^2}X_{ji}X'_{ji}\right]\right]^{-1}.\end{aligned} \tag{3.15}$$

Algorithm 2 Standard multiple logistic regression model

```
procedure STANDARD($IPD$)
    $myData \leftarrow$ pool all cohorts of $IPD$ together          ▷ cohort information is ignored
    fit standard multiple logistic regression on $myData$ via Newton-Raphson approximation
    $\hat{\beta}^{standard} \leftarrow$ resulting coefficient estimates $\hat{\beta}_0^{standard}, ..., \hat{\beta}_K^{standard}$
    $\widehat{Var}(\hat{\beta}^{standard}) \leftarrow$ resulting variances          ▷ diagonal elements of $\widehat{\sum}$
    return $\hat{\beta}^{standard}$, $\widehat{Var}(\hat{\beta}^{standard})$
```

It is now straightforward to obtain the risk predictions $\hat{p}_{ji}^{standard}$ by

$$\begin{aligned}\hat{p}_{ji}^{standard} &= \frac{exp\left(\left(\hat{\beta}^{standard}\right)' X_{ji}\right)}{1 + exp\left(\left(\hat{\beta}^{standard}\right)' X_{ji}\right)} \\ &= \frac{1}{1 + exp\left(-\left(\hat{\beta}^{standard}\right)' X_{ji}\right)}.\end{aligned} \tag{3.16}$$

Algorithm 3 Prediction for standard multiple logistic regression model

```
procedure PREDICT_STANDARD($\hat{\beta}^{standard}$, $X$)
                                        ▷ $\hat{\beta}^{standard}$: output of Algorithm 2
                                        ▷ $X$: covariate vector of individual patient
    $\hat{p}^{standard} \leftarrow \frac{1}{1+exp(-(\hat{\beta}^{standard})'X)}$
    return $\hat{p}^{standard}$
```

3.3.2 Random intercept model

Whereas the previous method ignored the clustered nature of the data, it is now accounted for by introducing a random intercept. This one-stage meta-analysis is the simplest version of a mixed effects logistic regression model and is suitable in case the heterogeneity between different cohorts is only due to varying outcome frequencies (Debray, Moons, Ahmed, et al. 2013). In this setting only one additional parameter has to be estimated: the random intercept variance $\rho^2_{\beta_0^{random}}$. Thereby the distribution of the random intercept is assumed to be normal with mean 0. The resulting model for the binary outcome y_{ji} is given by

$$\begin{aligned} y_{ji} &\sim Bernoulli(p_{ji}), \\ log\left(\frac{E(y_{ji}|X_{ji})}{1 - E(y_{ji}|X_{ji})}\right) = log\left(\frac{p_{ji}}{1 - p_{ji}}\right) &= \beta_0^{random} + \beta_{0j}^{random} + \sum_{k=1}^{K} \beta_k^{random} x_{kji}, \\ \beta_{0j}^{random} &\sim \mathcal{N}\left(0, \rho^2_{\beta_0^{random}}\right). \end{aligned} \tag{3.17}$$

Similar to the standard logistic regression model, estimation of the fixed regression coefficients $\beta^{random} = \left(\beta_0^{random}, ..., \beta_K^{random}\right)$ of this random intercept model, as well as the between-cohort variance in intercept, $\rho^2_{\beta_0^{random}}$, is performed via maximum likelihood. In the following derivation of the log-likelihood for the random intercept model, the superscripts are again neglected for readability. In a first step a single cohort j is considered. For a given random intercept β_{0j}, the individual patients can be assumed independent, yielding the conditional probability of $(y_{j1}, ..., y_{jn_j})$ by

$$\prod_{i=1}^{n_j} \left(\frac{e^{\beta' X_{ji} + \beta_{0j}}}{1 + e^{\beta' X_{ji} + \beta_{0j}}} \right)^{y_{ji}} \left(\frac{1}{1 + e^{\beta' X_{ji} + \beta_{0j}}} \right)^{1 - y_{ji}}$$

$$= exp \left[\sum_{i=1}^{n_j} \left(y_{ji} \left(\beta' X_{ji} + \beta_{0j} \right) - log \left(1 + e^{(\beta' X_{ji} + \beta_{0j})} \right) \right) \right]. \tag{3.18}$$

In the next step the distribution of the random intercept β_{0j} is incorporated, obtaining the marginal probability of $(y_{j1}, ..., y_{jn_j})$:

$$\int_{\mathbb{R}} exp \left[\sum_{i=1}^{n_j} \left(y_{ji} \left(\beta' X_{ji} + RI \right) - log \left(1 + e^{(\beta' X_{ji} + RI)} \right) \right) \right] \frac{1}{\sqrt{2\pi} \rho_{\beta_0}} e^{-\frac{1}{2\rho_{\beta_0}^2} RI^2} dRI. \tag{3.19}$$

Finally the log-likelihood of the random intercept model is obtained by summarizing over the logarithm of the marginal probabilities of all cohorts:

$$l \left(\beta, \rho_{\beta_0} \right) =$$
$$\sum_{j=1}^{J} log \int_{\mathbb{R}} exp \left[\sum_{i=1}^{n_j} \left(y_{ji} \left(\beta' X_{ji} + RI \right) - log \left(1 + e^{(\beta' X_{ji} + RI)} \right) \right) \right] \frac{1}{\sqrt{2\pi} \rho_{\beta_0}} e^{-\frac{1}{2\rho_{\beta_0}^2} RI^2} dRI. \tag{3.20}$$

Since the integral in Equation (3.20) does not have a closed form solution, it is approximated by an order one Gauss-Hermite quadrature, which is equivalent to a Laplace approximation. This approximation is then maximized in order to obtain the maximum likelihood estimates $\hat{\beta}^{random} = \left(\hat{\beta}_0^{random}, ..., \hat{\beta}_K^{random} \right)$ and $\hat{\rho}_{\beta_0^{random}}$. In this thesis the fitting of the random intercept model is performed in the R package lme4 (Bates et al. 2015).

Algorithm 4 Random intercept model

```
procedure RANDOM($IPD$)
    $myData$ ← pool all cohorts of $IPD$ together, keep cohort information
    fit random intercept regression on $myData$ via Laplace approximation
    $\hat{\beta}^{random}$ ← resulting coefficient estimates $\hat{\beta}_0^{random}, ..., \hat{\beta}_K^{random}$
    $\rho^2_{\beta_0^{random}}$ ← resulting estimate of random intercept variance
    return $\hat{\beta}^{standard}$, $\rho^2_{\beta_0^{random}}$
```

As already discussed in Section 3.1, the focus of the introduced models is on risk predictions for individuals from unknown cohorts, where it is not possible to use coefficient estimates for random intercepts, since these are specific to cohorts used in the analysis. In a simple, but nevertheless widely used approach, one sets the random effects to their median value 0 and performs the prediction analogously to standard logistic regression. Since the inverse of the logistic link $h^{-1}(x) = \frac{1}{1+exp(-x)}$ is a monotonic function, inserting the median of the random intercept, $\beta_{0j}^{random} = 0$, results in the predicted median response for the population of clusters:

$$\hat{p}_{ji}^{random\ zero} = Median(y_{ji}|X_{ji}) = \frac{1}{1 + exp\left(-\left(0 + \left(\hat{\beta}^{random}\right)' X_{ji}\right)\right)}. \tag{3.21}$$

Algorithm 5 Median prediction for random intercept model

```
procedure Predict_Random_Zero($\hat{\beta}^{random}$, $X$)
                                  ▷ $\hat{\beta}^{random}$: output of Algorithm 4
                                  ▷ $X$: covariate vector of individual patient
    $\hat{p}^{random\ zero} \leftarrow \frac{1}{1+exp(-(\hat{\beta}^{random})'X)}$
    return $\hat{p}^{random\ zero}$
```

We compare this simple approach to population-average prediction, which is the correct calculation under the principle of squared error loss. We obtain the risks by integrating over the distribution of the random effects (Skrondal and Rabe-Hesketh 2009, Pavlou et al. 2015):

$$\begin{aligned}
\hat{p}_{ji}^{random\ integration} &= E\left(y_{ji}|X_{ji}\right) = E\left(E\left(y_{ji}|X_{ji}, \beta_{0j}^{random}\right)\right) \\
&= E\left(\frac{1}{1 + exp\left(-\left(\left(\hat{\beta}^{random}\right)' X_{ji} + \beta_{0j}^{random}\right)\right)}\right) \\
&= \int_{-\infty}^{\infty} \frac{1}{1 + exp\left(-\left(\left(\hat{\beta}^{random}\right)' X_{ji} + RI\right)\right)} f\left(RI\right) dRI, \qquad (3.22) \\
f\left(RI\right) &= \frac{1}{\sqrt{2\pi}\hat{\rho}_{\beta_0^{random}}} exp\left(-\frac{1}{2}\left(\frac{RI}{\hat{\rho}_{\beta_0^{random}}}\right)^2\right),
\end{aligned}$$

whereby we used the rule for double expectation. We compute the mean predictions $\hat{p}_{ji}^{random\ integration}$ via numerical integration based on Gauss-Kronrod quadrature implemented in the R package stats.

Algorithm 6 Mean prediction for random intercept model

```
procedure Predict_Random_Integration($\hat{\beta}^{random}$, $\rho^2_{\beta_0^{random}}$, $X$)
                                  ▷ $\hat{\beta}^{random}$, $\rho^2_{\beta_0^{random}}$: output of Algorithm 4
                                  ▷ $X$: covariate vector of individual patient
    calculate $\hat{p}^{random\ integration}$ via numerical integration
    return $\hat{p}^{random\ integration}$
```

3.3.3 Two-stage IPD meta-analysis

In a two-stage approach, individual logistic regression models are built for every cohort separately and then traditional meta-analysis methods combine the results for an overall model. In the first step, a standard multiple logistic regression model, as described in Section 3.3.1, is fit to each of the available cohorts separately:

$$y_{ji} \sim Bernoulli(p_{ji})$$
$$log\left(\frac{E(y_{ji}|X_{ji})}{1-E(y_{ji}|X_{ji})}\right) = log\left(\frac{p_{ji}}{1-p_{ji}}\right) = \beta_{0j} + \sum_{k=1}^{K} \beta_{kj} x_{kji}, \tag{3.23}$$

where y_{ji} is 1 or 0 for patient $i = 1, ..., n_j$ from cohort $j = 1, ..., J$ with or without a positive diagnosis for high-grade cancer, respectively, and p_{ji} is the corresponding risk. The unknown regression coefficients $\beta_{kj},\ k = 0, ..., K$ are estimated for each cohort by $\hat{\beta}_j = (\hat{\beta}_{0j}, \hat{\beta}_{1j}, ..., \hat{\beta}_{Kj})$ and the corresponding estimated within-cohort covariance matrix $\widehat{\sum}_j$ gets:

$$\widehat{\sum}_j = \left(-\frac{\partial^2 l\left(\beta_j\right)}{\partial\beta_j \partial\beta_j'}\right)^{-1}\Bigg|_{\beta_j=\hat{\beta}_j} = \left(\sum_{i=1}^{n_j} \frac{e^{\hat{\beta}_j' X_{ji}}}{\left(1+e^{\hat{\beta}_j' X_{ji}}\right)^2} X_{ji} X_{ji}'\right)^{-1}. \tag{3.24}$$

In the second step the estimates are incorporated in a multivariate random-effects model, which combines the regression coefficients from the various cohorts, thereby accounting for their correlation. It is assumed, that the model parameters follow a multivariate normal distribution across the cohorts, which accounts for between- and within-cohort variability (van Houwelingen et al. 2002, Jackson, White, and S. G. Thompson 2010, Debray, Moons, Abo-Zaid, et al. 2013):

$$\begin{bmatrix} \hat{\beta}_{0j} \\ \hat{\beta}_{1j} \\ \vdots \\ \hat{\beta}_{Kj} \end{bmatrix} \sim \mathcal{N}_{K+1}\left(\begin{bmatrix} \beta_0^{two\ stage} \\ \beta_1^{two\ stage} \\ \vdots \\ \beta_K^{two\ stage} \end{bmatrix}, \begin{bmatrix} \tau_{\beta_0}^2 & \tau_{\beta_0\beta_1} & \cdots & \tau_{\beta_0\beta_K} \\ \tau_{\beta_0\beta_1} & \tau_{\beta_1}^2 & \cdots & \tau_{\beta_1\beta_K} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{\beta_0\beta_K} & \tau_{\beta_1\beta_K} & \cdots & \tau_{\beta_K}^2 \end{bmatrix} + \begin{bmatrix} Var\left(\hat{\beta}_{0j}\right) & Cov\left(\hat{\beta}_{0j}, \hat{\beta}_{1j}\right) & \ldots & Cov\left(\hat{\beta}_{0j}, \hat{\beta}_{Kj}\right) \\ Cov\left(\hat{\beta}_{0j}, \hat{\beta}_{1j}\right) & Var\left(\hat{\beta}_{1j}\right) & \ldots & Cov\left(\hat{\beta}_{1j}, \hat{\beta}_{Kj}\right) \\ \vdots & \vdots & \ddots & \vdots \\ Cov\left(\hat{\beta}_{0j}, \hat{\beta}_{Kj}\right) & Cov\left(\hat{\beta}_{1j}, \hat{\beta}_{Kj}\right) & \ldots & Var\left(\hat{\beta}_{Kj}\right) \end{bmatrix}\right) \tag{3.25}$$

$$\Leftrightarrow \hat{\beta}_j \sim \mathcal{N}_{K+1}\left(\beta^{two\ stage}, \mathcal{T} + \sum\nolimits_j\right), \tag{3.26}$$

where the unknown average regression coefficients are denoted as $\beta_0^{two\ stage}$ to $\beta_K^{two\ stage}$; $\tau_{\beta_k}^2$ and $\tau_{\beta_k\beta_l}$, $k = 1, ..., K$ and $l = 1, ..., K$ are also unknown and describe the heterogeneity between cohorts and the between-cohort covariances, respectively. Superscripts have been neglected for readability at places. For our further considerations we substitute the within-cohort covariance matrix with its estimation $\widehat{\sum}_j$ from step one of the meta-analysis (Equation 3.24) and assume it to be the true covariance structure. This simplification is performed, as incorporating uncertainty resulting from the use of $\widehat{\sum}_j$ instead of $\sum_j$ is behind the scope of this thesis. Therefore the following analyses get conditioned on this commonly used simplification (Jackson, White, and S. G. Thompson 2010, Kacker 2004).

This general model requires many parameter estimations. In particular for meta-analysis based on published studies rather than IPD it is typically not feasible to fit this model, since within-cohort covariances are rarely reported. Even when they are given, or can be calculated in case of available IPD, it remains difficult to estimate between-cohort covariances. Riley, Abrams, et al. 2007 discuss that the corresponding between-cohort correlations,

$$Corr\left(\hat{\beta}_{kj}, \hat{\beta}_{lj}\right) = \frac{Cov\left(\hat{\beta}_{kj}, \hat{\beta}_{lj}\right)}{\sqrt{Var\left(\hat{\beta}_{kj}\right)}\sqrt{Var\left(\hat{\beta}_{lj}\right)}}, \tag{3.27}$$

are often estimated as 1 or -1, therefore at the edge of their parameter spaces. Due to these estimation difficulties, within- and between-cohort covariances are often assumed to be zero, whereby the above model simplifies to a univariate model for each parameter $\hat{\beta}_{0j}, ..., \hat{\beta}_{Kj}$:

$$\begin{aligned} \hat{\beta}_{kj} &\sim \mathcal{N}\left(\beta_k^{two\ stage}, \tau^2_{\beta_k} + \widehat{Var}(\hat{\beta}_{kj})\right), \quad k = 0, ..., K && (3.28) \\ \Leftrightarrow \hat{\beta}_{kj} &\sim \mathcal{N}\left(\beta_{kj}, \widehat{Var}(\hat{\beta}_{kj})\right) \\ \beta_{kj} &\sim \mathcal{N}\left(\beta_k^{two\ stage}, \tau^2_{\beta_k}\right), \quad k = 0, ..., K. && (3.29) \end{aligned}$$

In order to obtain overall estimates $\hat{\beta}_k^{two\ stage}$, it is possible to assume the regression coefficients are either fixed or random across the cohorts. This is equivalent to assuming $\tau^2_{\beta_k}$ to be either zero or not.

For the random effects two-stage IPD meta-analysis, Model (3.29) is used without simplifications. Therefore the regression coefficients of the individual cohorts β_{kj} are assumed normally distributed about an average effect $\beta_k^{two\ stage\ random}$ with variance $\tau^2_{\beta_k}$. Consequently the true coefficients are allowed to vary across cohorts and the estimates $\hat{\beta}_k^{two\ stage\ random}$ can be interpreted as the estimated averages of their distributions (Burke et al. 2017).

Most researchers use the inverse variance method to obtain estimates $\hat{\beta}_k^{two\ stage\ random}$, in order to control for different levels of precision across cohorts. With the notation given in Equation (3.28), it is straightforward to get the weights for each cohort and regression coefficient as the inverse of their corresponding variances:

$$w_{kj} = \frac{1}{\hat{\tau}^2_{\beta_k} + \widehat{Var}(\hat{\beta}_{kj})}. \tag{3.30}$$

With these the $\hat{\beta}_k^{two\ stage\ random}$ are calculated as the weighted average:

$$\hat{\beta}_k^{two\ stage\ random} = \frac{\sum_{j=1}^J \hat{\beta}_{kj} w_{kj}}{\sum_{j=1}^J w_{kj}}, \tag{3.31}$$

$$Var\left(\hat{\beta}_k^{two\ stage\ random}\right) = \frac{1}{\sum_{j=1}^J w_{kj}}. \tag{3.32}$$

The estimated variances $\widehat{Var}(\hat{\beta}_{kj})$ and regression coefficients $\hat{\beta}_{kj}$ are known from step one of the meta-analysis for every cohort j and covariate k. In this thesis the between-cohort variances $\tau^2_{\beta_k}$ are estimated by the DerSimonian and Laird method, whereby their calculations are implemented in R with the package metafor (Kacker 2004, DerSimonian and Laird 1986, Viechtbauer 2010). The DerSimonian and Laird approach is based on method-of-moments estimates. We consider for every covariate k the quantity $b_k = \frac{\sum_{j=1}^J a_{kj}\hat{\beta}_{kj}}{\sum_{j=1}^J a_{kj}}$ with $a_{kj} = \frac{1}{\widehat{Var}(\hat{\beta}_{kj})}$. For the expression $\sum_{j=1}^J a_{kj}(\hat{\beta}_{kj} - b_k)^2$ we get the expected value

$$\begin{aligned}
&E\left(\sum_{j=1}^J a_{kj}(\hat{\beta}_{kj} - b_k)^2\right) = \sum_{j=1}^J a_{kj} E\left((\hat{\beta}_{kj} - b_k)^2\right) \\
&= \sum_{j=1}^J a_{kj}\left[Var(\hat{\beta}_{kj} - b_k) - \left(E(\hat{\beta}_{kj} - b_k)\right)^2\right] \\
&= \sum_{j=1}^J a_{kj}\left[Var(\hat{\beta}_{kj}) + Var(b_k) - 2Cov(\hat{\beta}_{kj}, b_k) - \left(E(\hat{\beta}_{kj}) - E(b_k)\right)^2\right] \\
&= \sum_{j=1}^J a_{kj}\left[Var(\hat{\beta}_{kj}) + Var\left(\frac{\sum_{j=1}^J a_{kj}\hat{\beta}_{kj}}{\sum_{j=1}^J a_{kj}}\right) - 2Cov\left(\hat{\beta}_{kj}, \frac{\sum_{j=1}^J a_{kj}\hat{\beta}_{kj}}{\sum_{j=1}^J a_{kj}}\right)\right. \\
&\quad \left. - \left(E(\hat{\beta}_{kj}) - E(\frac{\sum_{j=1}^J a_{kj}\hat{\beta}_{kj}}{\sum_{j=1}^J a_{kj}})\right)^2\right] \\
&= \sum_{j=1}^J a_{kj}\left[Var(\hat{\beta}_{kj}) + \frac{\sum_{j=1}^J a_{kj}^2 Var(\hat{\beta}_{kj})}{\left(\sum_{j=1}^J a_{kj}\right)^2} - 2\frac{\sum_{l=1}^J a_{kl} Cov(\hat{\beta}_{kj}, \hat{\beta}_{kl})}{\sum_{j=1}^J a_{kj}}\right. \\
&\quad \left. - \left(E(\hat{\beta}_{kj}) - \frac{\sum_{j=1}^J a_{kj} E(\hat{\beta}_{kj})}{\sum_{j=1}^J a_{kj}}\right)^2\right] \\
&= \sum_{j=1}^J a_{kj}\left[Var(\hat{\beta}_{kj}) + \frac{\sum_{j=1}^J a_{kj}^2 Var(\hat{\beta}_{kj})}{\left(\sum_{j=1}^J a_{kj}\right)^2} - 2\frac{a_{kj} Var(\hat{\beta}_{kj})}{\sum_{j=1}^J a_{kj}} - \left(E(\hat{\beta}_{kj}) - E(\hat{\beta}_{kj})\right)^2\right] \\
&= \sum_{j=1}^J a_{kj} Var(\hat{\beta}_{kj}) + \frac{\sum_{j=1}^J a_{kj}^2 Var(\hat{\beta}_{kj})}{\sum_{j=1}^J a_{kj}} - 2\frac{\sum_{j=1}^J a_{kj}^2 Var(\hat{\beta}_{kj})}{\sum_{j=1}^J a_{kj}} \\
&= \sum_{j=1}^J a_{kj} Var(\hat{\beta}_{kj}) - \frac{\sum_{j=1}^J a_{kj}^2 Var(\hat{\beta}_{kj})}{\sum_{j=1}^J a_{kj}}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{J} a_{kj} \left(\tau^2_{\beta_k} + \widehat{Var}(\hat{\beta}_{kj}) \right) - \frac{\sum_{j=1}^{J} a_{kj}^2 \left(\tau^2_{\beta_k} + \widehat{Var}(\hat{\beta}_{kj}) \right)}{\sum_{j=1}^{J} a_{kj}} \\
&= \tau^2_{\beta_k} \sum_{j=1}^{J} a_{kj} + \sum_{j=1}^{J} a_{kj} \widehat{Var}(\hat{\beta}_{kj}) - \tau^2_{\beta_k} \frac{\sum_{j=1}^{J} a_{kj}^2}{\sum_{j=1}^{J} a_{kj}} - \frac{\sum_{j=1}^{J} a_{kj}^2 \widehat{Var}(\hat{\beta}_{kj})}{\sum_{j=1}^{J} a_{kj}} \\
&= \tau^2_{\beta_k} \left(\sum_{j=1}^{J} a_{kj} - \frac{\sum_{j=1}^{J} a_{kj}^2}{\sum_{j=1}^{J} a_{kj}} \right) + \left(\sum_{j=1}^{J} a_{kj} \widehat{Var}(\hat{\beta}_{kj}) - \frac{\sum_{j=1}^{J} a_{kj}^2 \widehat{Var}(\hat{\beta}_{kj})}{\sum_{j=1}^{J} a_{kj}} \right). \qquad (3.33)
\end{aligned}
$$

For the method-of-moments step we now equate the term $\sum_{j=1}^{J} a_{kj} (\hat{\beta}_{kj} - b_k)^2$ with its expectation given by Equation 3.33. Solving for $\tau^2_{\beta_k}$ results in the DerSimonian and Laird estimate $\hat{\tau}^2_{\beta_k}$:

$$
\hat{\tau}^2_{\beta_k} = \frac{\left(\sum_{j=1}^{J} a_{kj} (\hat{\beta}_{kj} - b_k)^2 \right) - \left(\sum_{j=1}^{J} a_{kj} \widehat{Var}(\hat{\beta}_{kj}) - \frac{\sum_{j=1}^{J} a_{kj}^2 \widehat{Var}(\hat{\beta}_{kj})}{\sum_{j=1}^{J} a_{kj}} \right)}{\left(\sum_{j=1}^{J} a_{kj} - \frac{\sum_{j=1}^{J} a_{kj}^2}{\sum_{j=1}^{J} a_{kj}} \right)}. \qquad (3.34)
$$

Since $\tau^2_{\beta_k}$ is a variance and therefore non-negative, its estimate is set to zero in case its computation given by Equation 3.34 gets negative:

$$
\hat{\tau}^2_{\beta_k} = max \left\{ 0, \frac{\left(\sum_{j=1}^{J} a_{kj} (\hat{\beta}_{kj} - b_k)^2 \right) - \left(\sum_{j=1}^{J} a_{kj} \widehat{Var}(\hat{\beta}_{kj}) - \frac{\sum_{j=1}^{J} a_{kj}^2 \widehat{Var}(\hat{\beta}_{kj})}{\sum_{j=1}^{J} a_{kj}} \right)}{\left(\sum_{j=1}^{J} a_{kj} - \frac{\sum_{j=1}^{J} a_{kj}^2}{\sum_{j=1}^{J} a_{kj}} \right)} \right\}, \qquad (3.35)
$$

with $a_{kj} = \frac{1}{\widehat{Var}(\hat{\beta}_{kj})}$.

Algorithm 7 Random effects two-stage IPD meta-analysis

```
procedure 2_STAGE_IPD_MA($IPD$)
    fit standard multiple logistic regression model for each cohort separately
                                                            ▷ compare Algorithm 2
    $\hat{\beta}_j \leftarrow$ resulting coefficient estimates of cohort $j$
    $\widehat{Var}(\hat{\beta}_j) \leftarrow$ resulting within-cohort variances of cohort $j$
    calculate estimated between-cohort variances $\hat{\tau}^2_{\beta_k}$ via DerSimonian and Laird method
    $w_{kj} \leftarrow \frac{1}{\widehat{Var}(\hat{\beta}_{kj}) + \hat{\tau}^2_{\beta_k}}$          ▷ weight for cohort $j$ and covariate $k$
    $\hat{\beta}_k^{two\ stage\ random} \leftarrow \frac{\sum_{j=1}^{J} \hat{\beta}_{kj} w_{kj}}{\sum_{j=1}^{J} w_{kj}}$
    return $\hat{\beta}^{two\ stage\ random} = \hat{\beta}_0^{two\ stage\ random}, ..., \hat{\beta}_K^{two\ stage\ random}$
```

With these estimates, we obtain the resulting risk predictions through

$$
p_{ji}^{two\ stage\ random} = \frac{1}{1 + exp\left(- \left(\hat{\beta}^{two\ stage\ random} \right)' X_{ji} \right)}. \qquad (3.36)
$$

For the fixed effects meta-analysis the between-cohort variance components $\tau^2_{\beta_k}$ are set to

Algorithm 8 Predictions for random effects two-stage IPD meta-analysis

```
procedure PREDICT_TWO_STAGE_FIXED($\hat{\beta}^{two\ stage\ random}$, $X$)
                                        ▷ $\hat{\beta}^{two\ stage\ fixed}$: output of Algorithm 7
                                        ▷ $X$: covariate vector of individual patient
    $\hat{p}^{two\ stage\ random} \leftarrow \frac{1}{1+exp(-(\hat{\beta}^{two\ stage\ random})'X)}$
    return $\hat{p}^{two\ stage\ random}$
```

zero, therefore common regression coefficients $\beta_k^{two\ stage\ fixed}$ are assumed across all cohorts. Equation (3.28) reduces to

$$\hat{\beta}_{kj} \sim \mathcal{N}\left(\beta_k^{two\ stage\ fixed}, \widehat{Var}(\hat{\beta}_{kj})\right), \quad k = 0, ..., K \tag{3.37}$$

and the formulas from the inverse variance method are simplified to become

$$w_{kj}^* = \frac{1}{\widehat{Var}(\hat{\beta}_{kj})}, \tag{3.38}$$

$$\hat{\beta}_k^{two\ stage\ fixed} = \frac{\sum_{j=1}^J \hat{\beta}_{kj} w_{kj}^*}{\sum_{j=1}^J w_{kj}^*}, \tag{3.39}$$

$$Var\left(\hat{\beta}_k^{two\ stage\ fixed}\right) = \frac{1}{\sum_{j=1}^J w_{kj}^*}. \tag{3.40}$$

Algorithm 9 Fixed effects two-stage IPD meta-analysis

```
procedure 2_STAGE_IPD_MA($IPD$)
    fit standard multiple logistic regression model for each cohort separately
                                        ▷ compare Algorithm 2
    $\hat{\beta}_j \leftarrow$ resulting coefficient estimates of cohort $j$
    $\widehat{Var}(\hat{\beta}_j) \leftarrow$ resulting within-cohort variances of cohort $j$
    $w_{kj}^* \leftarrow \frac{1}{\widehat{Var}(\hat{\beta}_j)}$          ▷ weight for cohort $j$ and covariate $k$
    $\hat{\beta}_k^{two\ stage\ fixed} \leftarrow \frac{\sum_{j=1}^J \hat{\beta}_{kj} w_{kj}^*}{\sum_{j=1}^J w_{kj}^*}$
    return $\hat{\beta}^{two\ stage\ fixed} = \hat{\beta}_0^{two\ stage\ fixed}, ..., \hat{\beta}_K^{two\ stage\ fixed}$
```

The resulting risk prediction is given by:

$$\hat{p}_{ji}^{two\ stage\ fixed} = \frac{1}{1 + exp\left(-\left(\hat{\beta}^{two\ stage\ fixed}\right)' X_{ji}\right)}. \tag{3.41}$$

Algorithm 10 Predictions for fixed effects two-stage IPD meta-analysis

```
procedure PREDICT_TWO_STAGE_FIXED($\hat{\beta}^{two\ stage\ fixed}$, $X$)
                                        ▷ $\hat{\beta}^{two\ stage\ fixed}$: output of Algorithm 9
                                        ▷ $X$: covariate vector of individual patient
    $\hat{p}^{two\ stage\ fixed} \leftarrow \frac{1}{1+exp(-(\hat{\beta}^{two\ stage\ fixed})'X)}$
    return $\hat{p}^{two\ stage\ fixed}$
```

3.4 Methods for evaluation of model fit

In order to compare the several model-based risks introduced in the previous Section 3.3, objective criteria are necessary to evaluate their operating characteristics for predicting the binary outcome of high-grade cancer versus no or low-grade cancer on biopsy. Therefore the methods discussed in Chapter 2 are used for thresholds ranging from 5-25%, and smoothing of resulting curves is done via locally weighted regression as described in Section 2.3.

We are interested in optimizing the accuracy of predictions for new patients rather than the question of how good the model fits a given dataset. Therefore, it is important not to fit and evaluate the model on the same data, but to split the data into a training and separate test set. As we are interested in the performance of the model when it is used for a new or unknown study population, we perform cross validation across cohorts: All patients of one cohort will be assigned to either the test or the training set. With this approach it is possible to evaluate whether the model derived on the cohorts in the training set leads to good predictions in independent cohorts. However, it is important to maximize the amount of data used for developing as well as for validating the model. This is accomplished by repeatedly splitting the data, fitting a model on the training set and evaluating this model on the test set. In this thesis, we apply two different methods of splitting the data, as both have some specific advantages.

First, we perform leave-one-cohort-out cross validation, for which each cohort is left out once as a test set. The logistic model is fit to the data of the remaining 9 cohorts, and predictions are made for the hold out test set. We perform this repeatedly for every cohort in our data set, and pool predictions for the 10 cohorts as hold out set for graphical displays.

Second, we consider a permutation-based validation, whereby all 252 possibilities of splitting the 10 cohorts into two sets, each containing five cohorts, are used. This has the advantages that the test set does not consist of one single cohort, which might be quite different to the remaining ones and it allows a bigger test set with cohort heterogeneity. However, this validation method prevents the efficient use of visual evaluation methods, such as sensitivity-, specificity-, calibration-, and net benefit-curves. Nevertheless, we utilize the AUC and HLS values of the 252 cross validation in order to support and extend results from the leave-one-cohort-out cross validation.

The considered cross validations are used to evaluate the model fit and to compare it between the different model approaches. Once an approach is chosen and validated, a final model is derived by the use of all available data at once.

3.5 Results

In this section, we compare the results of the models from Section 3.3 using the validation metrics of Chapter 2. We first determine the relevant covariates, which will be used for all models, and then compare the different model types using the same set of variables.

3.5.1 Underlying data

The data used for this analysis are from the ten PBCG cohorts ClevelandClinic, MayoClinic, SanRaffaele, Zurich, MSKCC, UCSF, DurhamVA, SanJuanVA, Sunnybrook and UTHealth, comprising 8492 biopsies from 8247 patients. A detailed description of this data can be found in Section 1.2.1 and Tables A.2 and A.1 of the appendix. Figure 11 illustrates the heterogeneity between cohorts in terms of prevalence of individual risk factors and their corresponding odds ratios for high-grade cancer. Zurich is amongst the cohorts with lowest prevalence of patients with PSA greater than 4, abnormal DRE, age greater than 65, African ancestry and positive family history, respectively, and highest prevalence of patients with prior negative biopsy. Therefore Zurich comprises a group of patients with characteristics in favor of no high-grade cancer, which aligns with its overall low proportion of high-grade cancer (6). However, its unique distribution of risk factors might rise the need for models incorporating cohort heterogeneities.

In general, individual cohorts potentially cause problems in the standard pooling approach, if they are an outlying cohort in terms of prevalence or odds ratio for one of the considered risk factors. For DRE SanJuanVA is an outlier in both dimensions, having a proportion of 59.2% abnormal DRE results compared to 21.5-39.5% , and a corresponding univariate odds ratio of 1.1 compared to 1.5 for ClevelandClinic and 2.4-4.0 for the other cohorts. The odds ratio of all cohorts combined is pulled towards the low values of ClevelandClinic and SanJuanVA, however, since both outlying cohorts have small sample sizes of 299 and 550 patients, their influence might not be problematic.

With a proportion of 63.2%, in contrast to 0-16.3%, DurhamVA is an outlying cohort in terms of prevalence of patients with African ancestry. However, its odds ratio aligns with the estimates of the two other cohorts UTHealth and SanJuanVA (1.0-1.3), whereby none of these were significant. For the remaining three cohorts that had enough patients with African ancestry to calculate odds ratios, the results are significant and range from 2.0-2.3. The univariate odds ratio for all cohorts combined is given by 1.7, therefore between both clusters.

For the covariate family history Zurich is an outlier in terms of prevalence, 2.8% compared to 16.7-33.1%, and univariate odds ratio, 3.2 compared to 0.9-1.8. Due to Zurichs big sample size the overall prevalence and odds ratio is heavily influenced by this individual cohort. This influential outlier potentially causes problems in risk prediction methods ignoring cohort heterogeneities, it might be even reasonable to exclude Zurich for model fitting. This will be

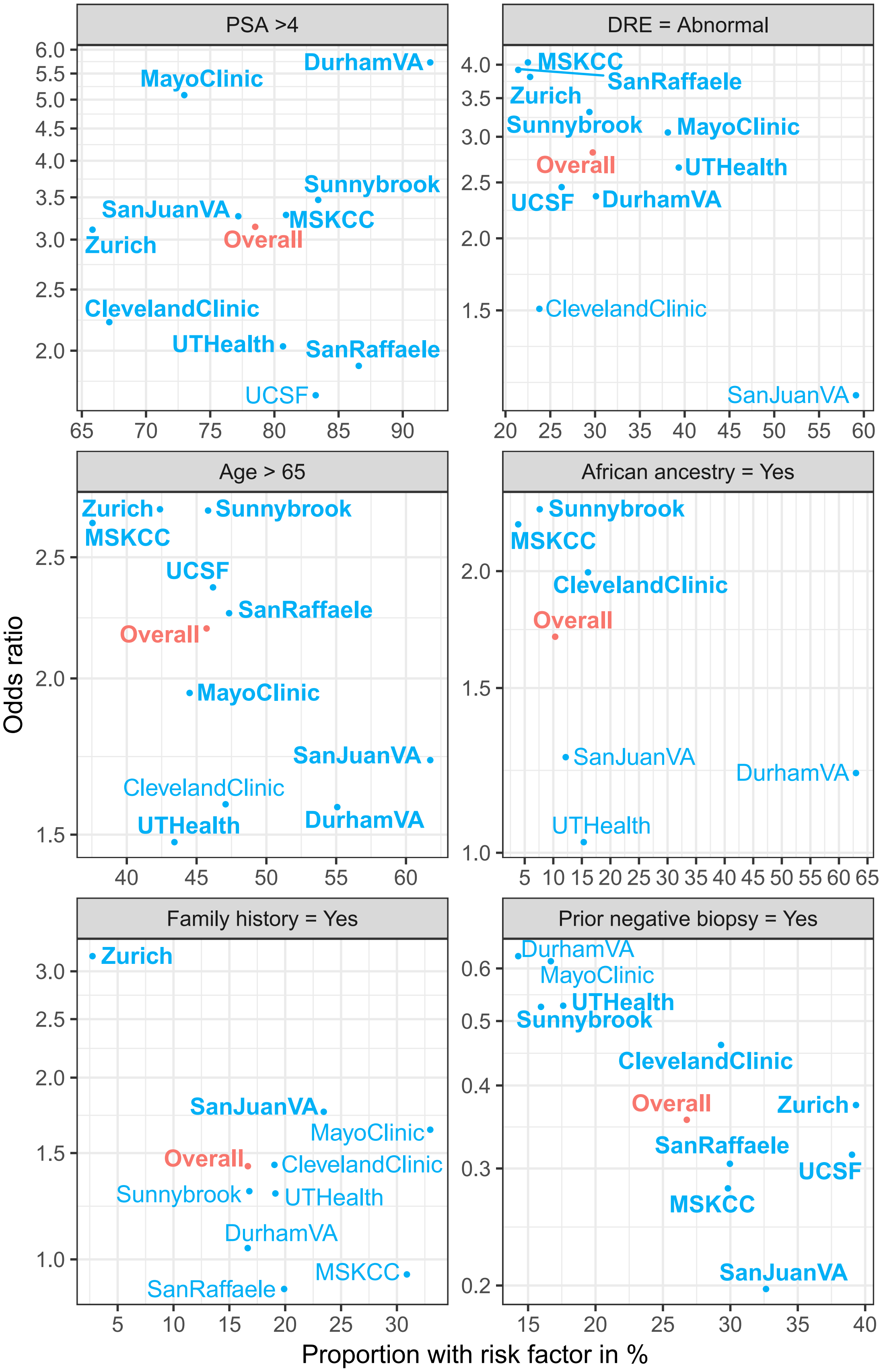


Figure 11 : Univariate odds ratios for individual risk factors plotted against their proportions. Continuous variables PSA and age are dichotomized for comparability, and significance at the 5% level is indicated in bold. Overall displays all cohorts pooled together. Data not shown for African ancestry for Zurich, SanRaffaele, MayoClinic and UCSF as numbers of patients with this risk factor were too low for reliable estimates of the odds ratios. Furthermore family history for UCSF is not shown as this risk factor is not reported for UCSF. Missing values have been excluded for calculation of univariate odds ratios and proportions.

further investigated in the following analysis.

For the other risk factors no cohort can be identified as an outlier and the odds ratio estimates for all cohorts pooled fall mainly in the middle of the individual cohort estimates.

For improving stability of the model-fitting algorithms, some transformations have been performed, which are used throughout this section. The variable PSA is log base two transformed, denoted by lpsa2. The continuous variables lpsa2 and age are then rescaled for the model implementations by subtraction of the corresponding mean values and division by the respective standard deviations. This is done for the whole data set at once, so the same transformation is used for all splits in training and test set.

Missing values are imputed by their corresponding median, resulting in non African ancestry for race, normal for DRE, and no for prior biopsy and family history.

3.5.2 Covariate selection

For covariate selection the approach discussed in Section 3.2 is utilized. Figure 12 displays the selected covariates for the 252 possibilities for choosing 5 out of 10 cohorts, and Figure 13 shows the resulting covariate selection for every single cohort as well as for all cohorts pooled together. The variables age, DRE, family history, lpsa2 and prior biopsy are chosen in most of the 252 possible combinations (Figure 12). Even though the variable race is chosen in less than half of the possible models, it has historically been a significant predictor (Ankerst,

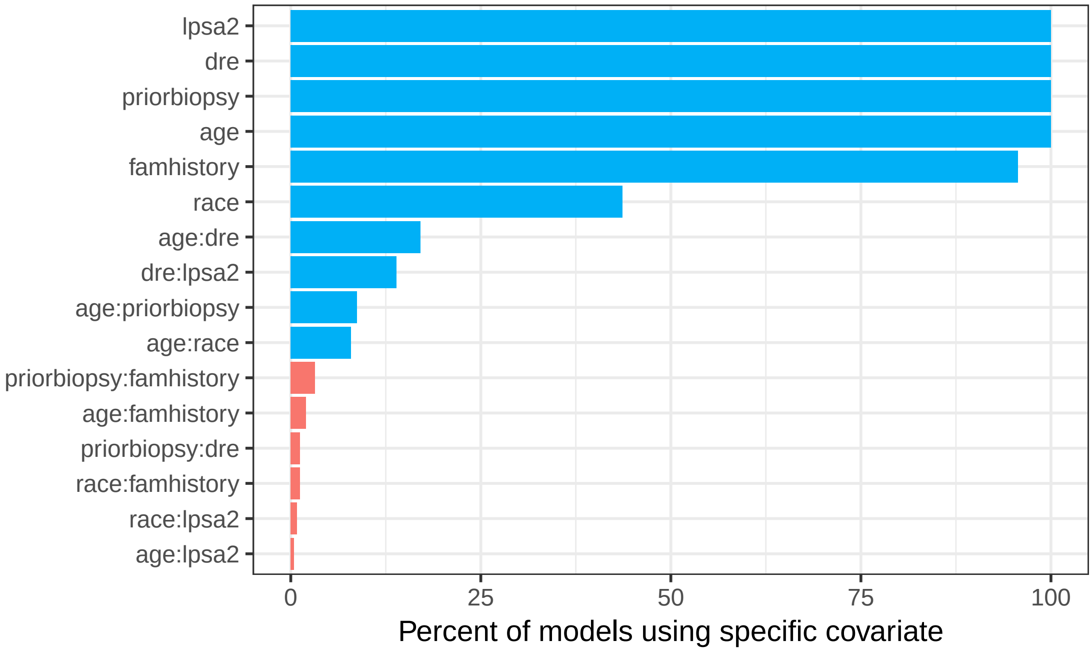


Figure 12 : Variables and interaction terms that have been selected at least once among the 252 training sets are shown, along with their frequency of selection. Frequency of at least 5% is indicated in blue in contrast to less than 5% in red.

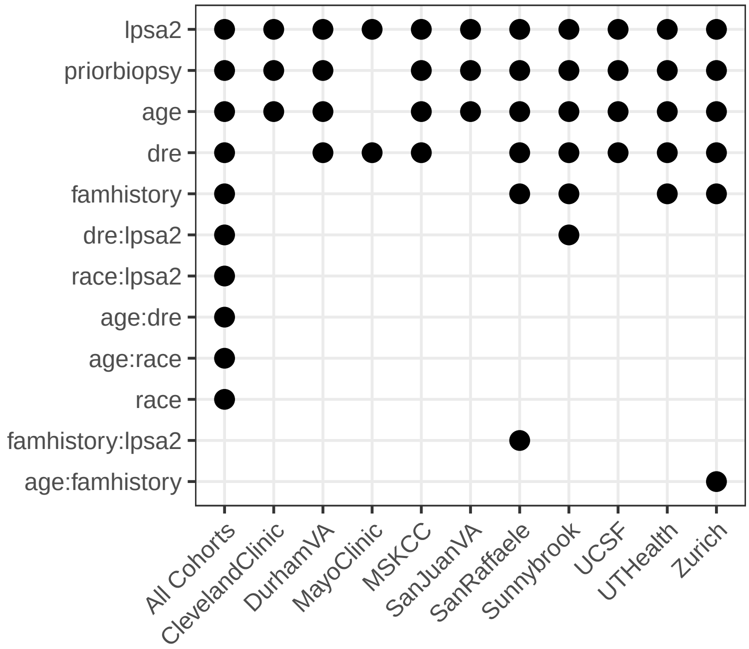


Figure 13 : Covariates selected in every individual cohort and in all cohorts pooled together. Only variables and interaction terms that have been selected at least once are shown.

Hoefler, et al. 2014, Nam et al. 2007). The reason for the low percentage here is that some cohorts have almost no patients with African ancestry.

In order to select among the remaining suggested effects, we further investigated the four interaction terms chosen in at least 5% of the considered models. The interactions of age and DRE, DRE and lpsa2, age and prior biopsy, and age and race are selected by 17.1%, 13.9%, 8.7% and 7.9% of the 252 combinations, respectively. Whereas the interaction between age and prior biopsy is not chosen for any of the single cohorts nor for all combined (Figure 13), the other three interactions at least appear for all cohorts together. Therefore, the interaction term for age and prior biopsy is eliminated.

In order to further justify the selection of covariates, the four interaction effects are displayed in Figure 14. For the interaction term age and prior biopsy no interaction can be seen in Figure 14, therefore we permanently exclude it from the final model. For the interaction of age and DRE, the difference in the risk of high-grade cancer between patients with normal and abnormal DRE results becomes smaller for older patients. For abnormal DRE results, the difference in lpsa2 values for patients with and without high-grade cancer is higher than for normal DRE. Finally the risk of high-grade cancer for patients with no African ancestry strongly depends on age. Whereas the risk for a positive biopsy for high-grade cancer is overall higher for men with African ancestry, it does not depend as much on the age. This relation can also be seen in the corresponding boxplots, whereby the age distribution for African ancestry barely differ between cancer and no cancer patients. In contrast patients without African ancestry have higher risks of high-grade cancer for increasing age. With these considerations the interactions age and DRE, DRE and lpsa2, and age and race are chosen as covariates.

In order to check that the static model based on the previously determined covariates is as good as individualized models, we perform the 252 cross validation of all choices of five cohorts for training versus the remaining five for testing. Thereby, the model fit for a static model including the six main effects and the three interaction terms is compared to an individualized one. For the individualized model, the basic variables and all corresponding two-way interaction terms are considered, whereby the covariates are chosen for every combination of cohorts individually by a stepwise algorithm based on the BIC. The resulting AUC and HLS values are shown in Figure 15. In terms of the AUC the static multivariate regression performs equally well or better in every single test set. In most test sets the HLS is the same for both methods and the small deviations of the individualized model from the static one is in either direction. As the model with the previously selected terms is simpler and performs at least equally well as the individualized ones, this choice of covariates is used for the subsequent sections.

3.5.3 Comparison of models for multiple cohorts

We are interested in the differences between the considered models, in particular in the question whether one model dominates the others and should therefore be used to integrate multiple cohorts. The previous section briefly discussed the choice of covariates, which will be now used to fit the models.

An advantage of the leave-one-cohort-out over the 252 cross validation is the possibility of displaying the overall results in terms of net benefit, sensitivity, specificity and calibration curves. As explained in Chapter 2, these curves display the fit of a model by considering the predicted probabilities. These predictions have been calculated for every cohort separately, based on the model built on the remaining 9 cohorts. Afterwards the probabilities of all cohorts were pooled, so that a single curve can be established for every method. The resulting curves are displayed in Figure 16. All curves are shown for the threshold range of 5 to 25%, as these are reasonable cutoff points for which a patient is usually referred to a biopsy. CIs of the diverse methods are overlapping for all metrics and are neglected for readability.

In terms of net benefit all methods outperform the strategy of biopsying nobody and for all values above the threshold of about 6% they also show a higher net benefit than treating all patients. Corresponding CIs show that this superiority is statistically significant for threshold values higher than 16%. Therefore all methods show a clinical usefulness, however the individual models do not show any detectable differences. Even by considering every cohort separately, only very small differences between the modeling techniques can be seen (Figure A.4(a)). However, the clinical net benefit varies in the differing test sets. Whereas for some cohorts, like MSKCC and Zurich, the net benefit of the prediction models exceeds the one of

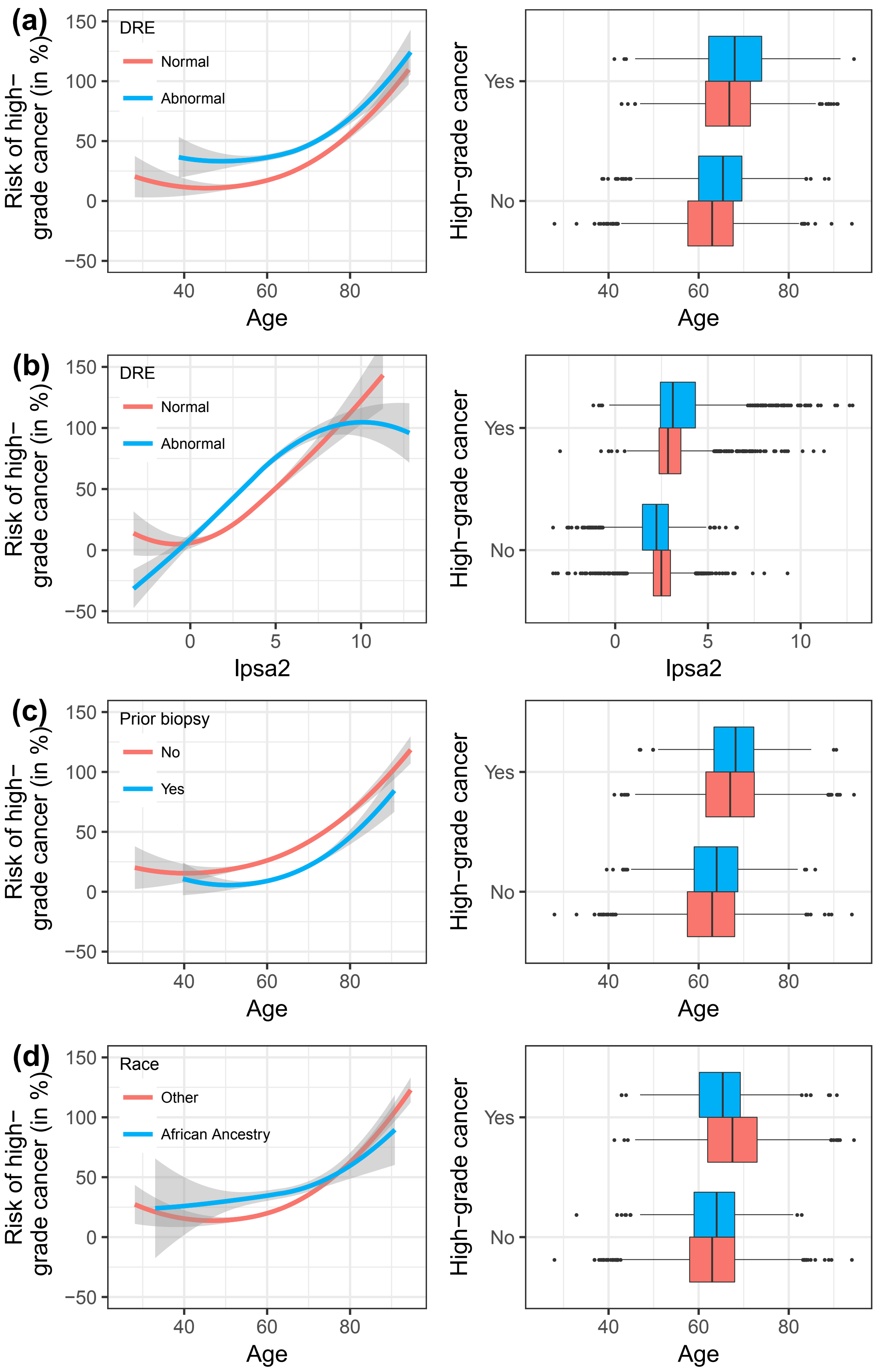


Figure 14 : Interaction effects of age and DRE, lpsa2 and DRE, age and prior biopsy, and age and race. Locally weighted regression curves for the probability of high-grade cancer dependent on the respective covariate are shown along with corresponding boxplots. The graphs are based on the data of all cohorts combined and without rescaling the continuous variables.

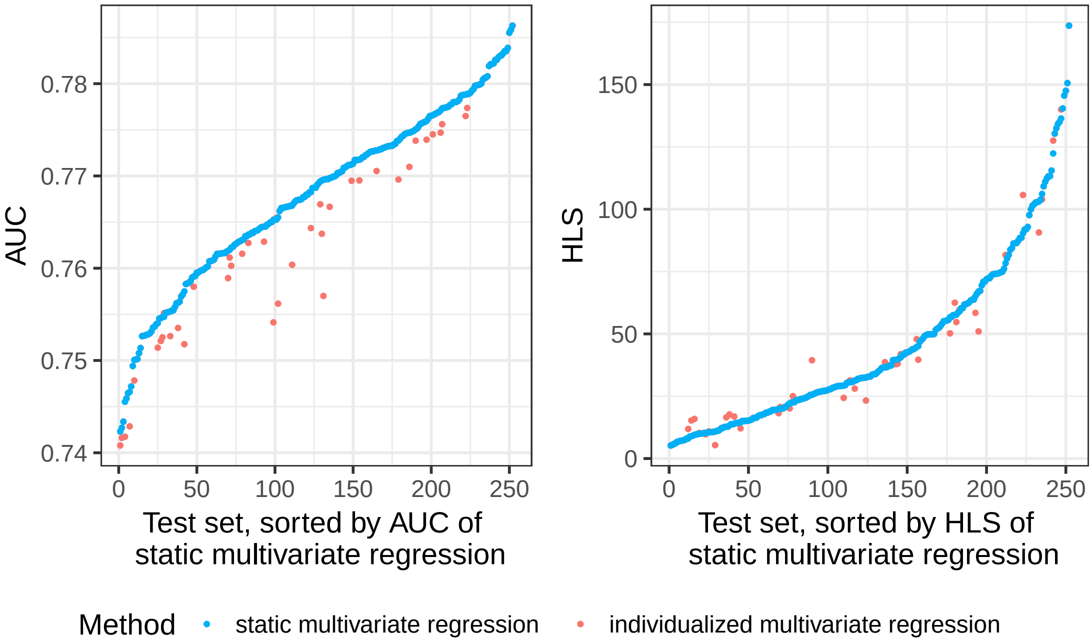


Figure 15 : AUC and HLS values for static and individualized multivariate regressions.

the reference strategy of biopsying all patients for most relevant thresholds, UTHealth does not show superior net benefit up to a cutoff of 20%. These results align with the prevalences of high-grade cancer in the individual cohorts. Zurich has by far the lowest prevalence with 17.7%, followed by ClevelandClinic and MSKCC with 27.1 and 29.2%, respectively. Therefore the strategy of biopsying all patients independent of their individual risks is not optimal for these cohorts. As opposed to this, for UTHealth the highest prevalence of 38.7% is reported.

For the calibration curves, small variations can be detected (Figure 16(b)). The standard model is slightly better calibrated than the other methods, however, corresponding CIs are overlapping. Calibration curves of all methods are close to the diagonal line, indicating good calibration throughout all relevant thresholds. Differences between the methods are also neglectable for the individual cohorts (Figure A.4(b)). Taking CIs into account, deviations from the diagonal line are significant only for UTHealth and Zurich, indicating poor calibration. For the high prevalence cohort UTHealth, predicted risks of all methods are too low for thresholds higher than 10%. Whereas the calibration curves for Zurich lie significantly beneath the diagonal line for predicted probabilities greater than 18%, which indicates that the risk predictions are higher than the actual risks in this Swiss cohort, which is not surprising, as Zurich has a very low prevalence of high-grade cancer compared to all other sites.

Similar results can be observed for the metrics sensitivity and specificity, which indicate overall good discrimination, but it is not possible to determine a significantly superior model, as all CIs are overlapping. Small deviations occur for the standard method, which tends to have lower predicted risks than the other modeling techniques and therefore lower values for

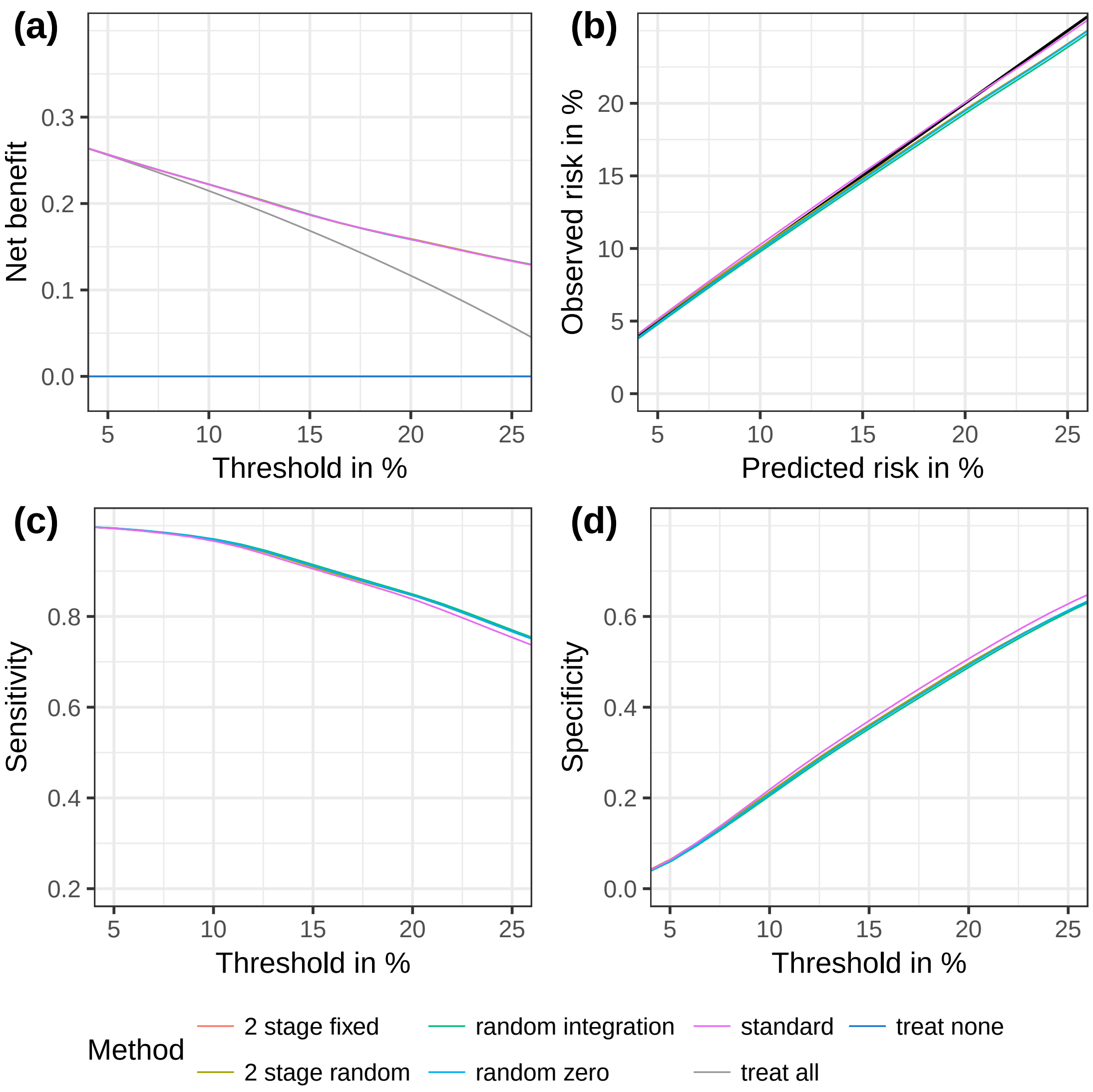


Figure 16 : (a) Net benefit, (b) calibration, (c) sensitivity and (d) specificity curves of the leave-one-cohort-out cross validation.

sensitivity and higher values for specificity for all considered thresholds. Differences in the prediction methods in terms of sensitivity and specificity can not be found in any individual cohort either (Figure A.5).

As the differences between the methods are very small and therefore difficult to detect in the graphical evaluations, we also consider the summarizing statistics AUC and HLS. The AUC for all cohorts pooled together ranges from 77.1% (95%-CI: 76.0%-78.2%) for the standard model to the slightly better 77.2% (95%-CI: 76.1%-78.3%) for the random intercept integration model. The HLS varies between 11.3 for the standard regression model and 23.9 for the two-stage random meta-analysis, whereby low values are preferred for this measure. However, for both measures these differences are marginal.

In Figure 17 AUC and HLS values are displayed for each cohort separately. As described in Section 3.4, for the leave-one-cohort-out cross validation each model is trained on nine

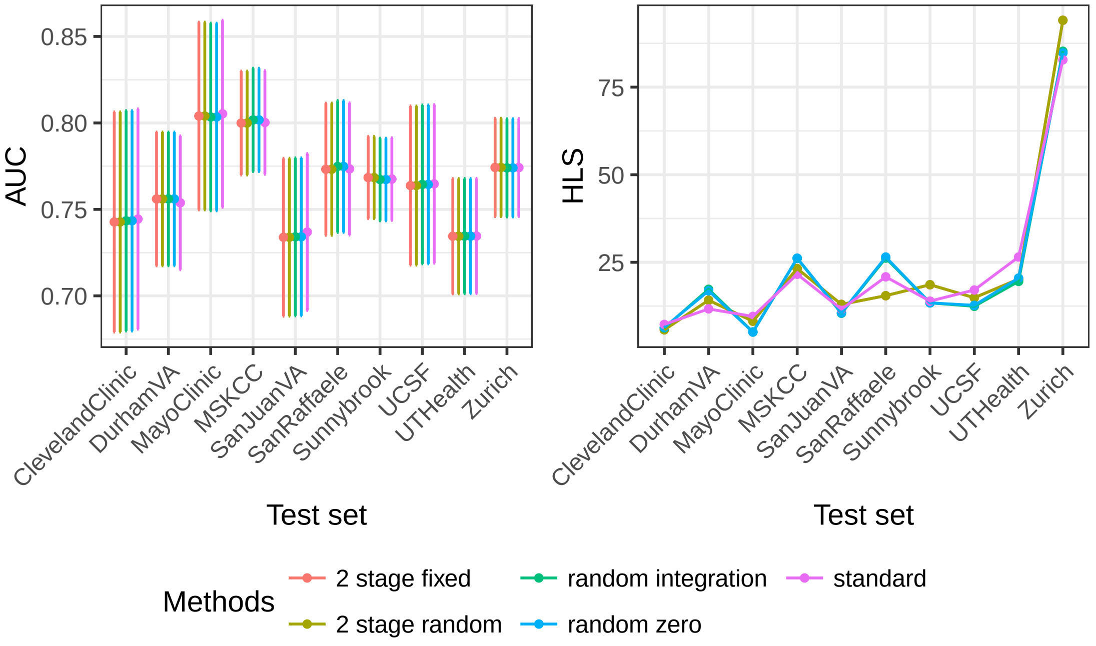


Figure 17 : AUC and HLS values for leave-one-cohort-out cross validation for each of the ten cohorts after prediction based on the remaining nine cohorts. For the AUC values 95%-CIs are included.

cohorts and tested on a single one. Since the cohorts are heterogeneous, it is possible that the test set differs significantly from the remaining training cohorts. For this reason the AUC values vary across the cohorts. Nevertheless, performances of the different models for a single cohort are near identical. Direct comparisons of the HLS across test sets can not be performed, as the HLS depend on the sample sizes of the test set. Therefore it is only possible to compare models within one test set. It is of particular interest that no model under- or outperforms the other models for all test sets, neither in terms of the AUC nor HLS. Furthermore, differences between the models are negligible under consideration of the CIs.

For the 252 cross validation we concentrate on the AUC and HLS values of the 252 different test sets, whereby summarizing boxplots are shown in Figure 18. The AUC values range from 74.1% to 78.8%, they are therefore less extreme than for the leave-one-cohort-out cross validation (Figure 17). This is due to the fact that five cohorts are combined in the test set, heterogeneous cohorts have therefore less influence in the overall evaluation of the corresponding test set. In terms of the AUC only small differences are detectable, in particular the two two-stage methods as well as the two random intercept models have almost identical results, whereby the latter ones show the least variability. The standard approach performs worst. For the HLS the random integration method has the lowest values and it is less variable than all other approaches. Even though all prediction models perform similarly, the random intercept integration model can be therefore determined as the slightly best model in terms of the AUC and HLS, thereby supporting Pavlou et al. 2015 that said it was most correct. Therefore, the random integration model is compared to the remaining methods in more detail in Figure 19. In terms of the AUC, the meta-analysis models two-stage fixed and two-stage

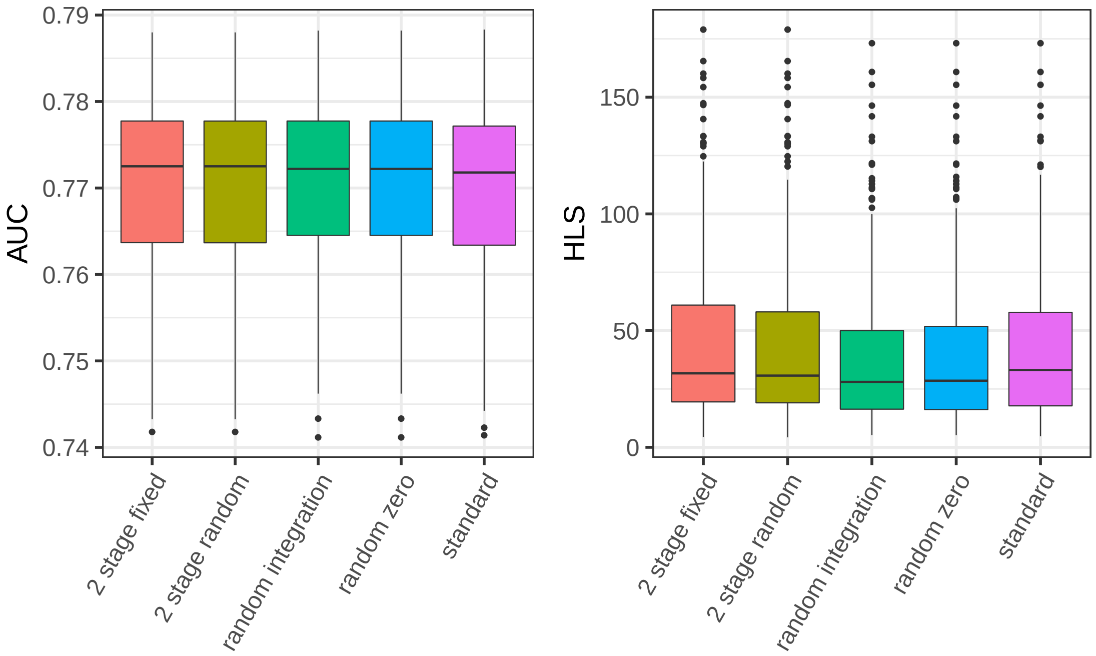


Figure 18 : Boxplots for the AUC (left) and HLS (right) values for the 252 cross validation for all prediction models.

random are slightly better for some test sets and worse for others. Overall no method can be determined to be superior. The random zero method is near identical to the random integration model. The standard multiple logistic regression model performs slightly worse than the random intercept integration model in terms of the AUC and all methods perform worse for the HLS. However, the differences are negligible.

3.5.4 Influence of single cohorts

As for the leave-one-cohort-out cross validation, the 252 cross validation did not reveal any considerable differences between the five different risk prediction methods. However, it is possible to further investigate the modeling results in order to determine the influence of single cohorts. Especially since we are focused on the optimal integration of different cohorts in an overall model, this might be of interest. We investigate the influences of individual sites on the different methods and whether it might be even reasonable to exclude very distinct cohorts. The approach of removing dissimilar cohorts was chosen for instance by the winning team of an online challenge on modeling and predicting prostate cancer results (Pölsterl et al. 2016). In this section, we focus on the evaluation measure AUC, as the HLS depends on the sample size and is therefore inappropriate for the following considerations. Similar to the previous results for the 252 cross validation, the graphical evaluations are not feasible for the following applications.

In order to investigate the influence of large cohorts, Figure 20 displays AUC values sample sizes of the test sets. As expected by the previous analysis, the different risk models show the same behavior. AUC values appear to be overall higher and less spread for larger test

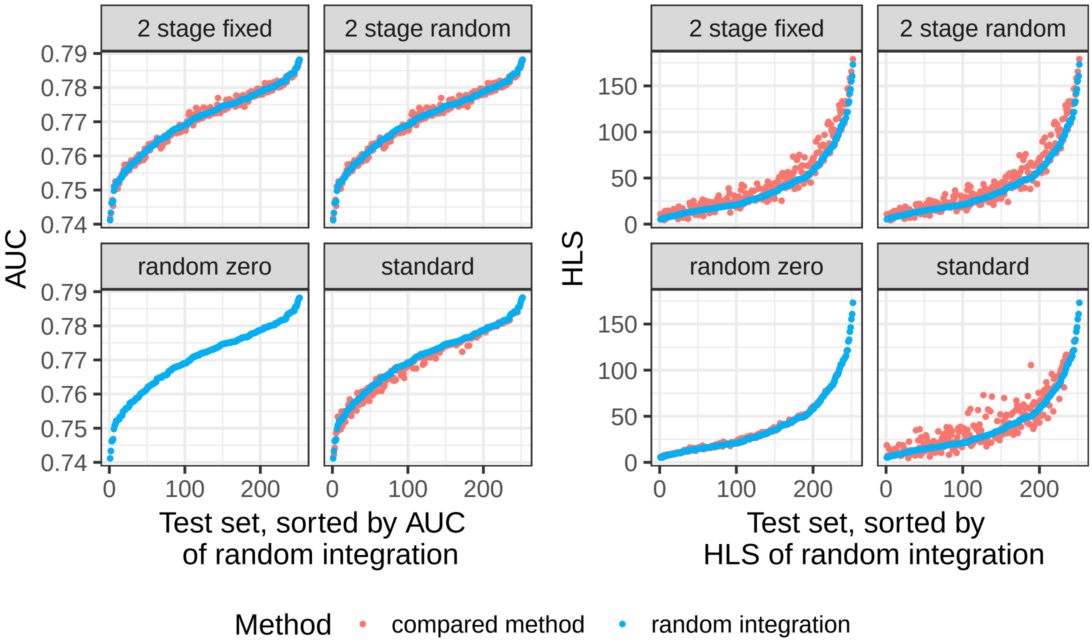


Figure 19 : AUC (left) and HLS (right) values of the methods two-stage fixed, two-stage random, random zero and standard for all 252 test sets are compared to the AUC and HLS values of the method random integration. The test sets are sorted by the AUC and HLS values of the random integration method.

sets. However, as indicated by the color coding, the relationship can be explained by the two largest cohorts, Sunnybrook (n=1721) and Zurich (n=1863). The best results are obtained if Zurich is part of the test set, indicating that the overall size of the test set does not influence the performance as much as the specific cohort composition. Similar conclusions hold for the clusters of test sets where only Sunnybrook or none of the two big cohorts is included. The worse performance for small test sets might be explained by bias due to inclusion of Zurich and Sunnybrook in the training set. This might be especially problematic for Zurich, as in addition to its large size, it also comprises patients with very different characteristics and biopsy results (Table A.2). Later we exclude this cohort in order to further understand its influence on model performance.

The behavior of model performance versus specific cohorts in the test set is further investigated in Figure 21. The results are similar to the leave-one-cohort-out cross validation in Figure 17, with the main difference the smaller range of AUC values for the 252 cross validation, as the influence of a single cohort is smoothed by the other four sites in the test set. The most extreme deviation from the overall median is given by the Zurich cohort. Its inclusion in the test set leads to highest AUC and lowest variability, the latter of which is due to its large size relative to others. Medians of MayoClinic and MSKCC are above the overall median as well, which is not surprising, as these cohorts have a very high AUC for the leave-one-cohort-out cross validation (Figure 17). The worst results are obtained for SanJuanVA and UTHealth. It is interesting that the best results are obtained for Zurich in the test set, which has the lowest prevalence of high-grade cancer with 17.7%. In comparison the worst discrimination is provided by UTHealth, whereby 38.7% of its patients have high-grade can-

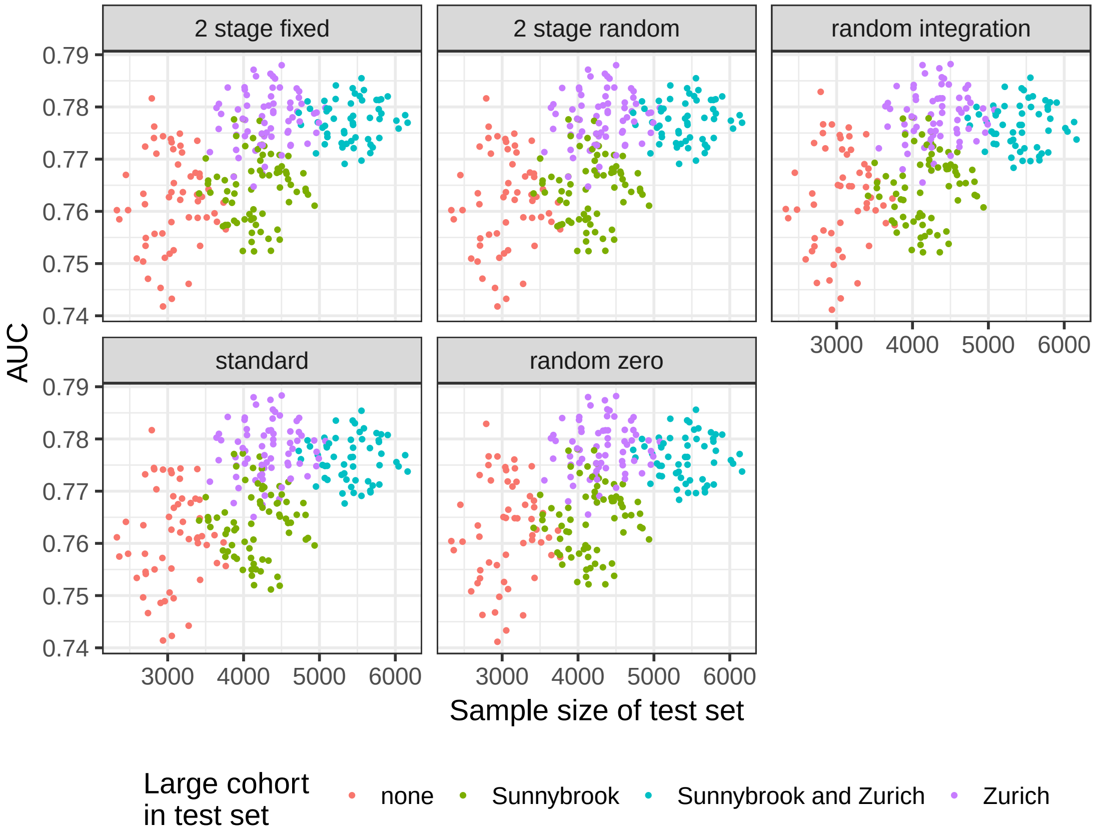


Figure 20 : AUC values for the 252 cross validation, sorted by the sample size of the test sets. Color coding illustrates whether the two largest cohorts, Sunnybrook and Zurich, are included in the test set.

cer, which is the highest percentage of all cohorts. SanJuanVA, however, has a very average prevalence of high-grade cancer and a similar bad performance as UTHealth. The percentages of high-grade cancer patients of the sites MayoClinic and MSKCC are similarly close to the average. An explanation for the different performance relative to cohort risk factors or sample sizes remains lacking, see Figure 6 and 7 in Section 1.2.

As already discussed in Section 1.2, the cohorts differ in terms of their sample sizes, prevalences of high-grade cancer, and patients characteristics. For that reason, the performance of the risk models are influenced by the presence of specific cohorts in the training or test set. It is therefore also of interest whether the model fit improves by leaving individual cohorts out of the training as well as the test set. We performed a permutation analysis comparing differences in AUC values of the 252 cross validation including all cohorts and sequentially excluding each cohort. The resulting differences are given in Figure 22, whereby negative values correspond to an improvement in AUC for excluding the respective cohort. It can be seen that across all cohorts except DurhamVA, the standard method is among the methods with highest values. Exclusion of individual cohorts is therefore less attractive for this method as for the more complicated random intercept and two-stage approaches. However, by taking the percentile intervals into account, these variations are negligible, especially since the total range spans only less than 4% points.

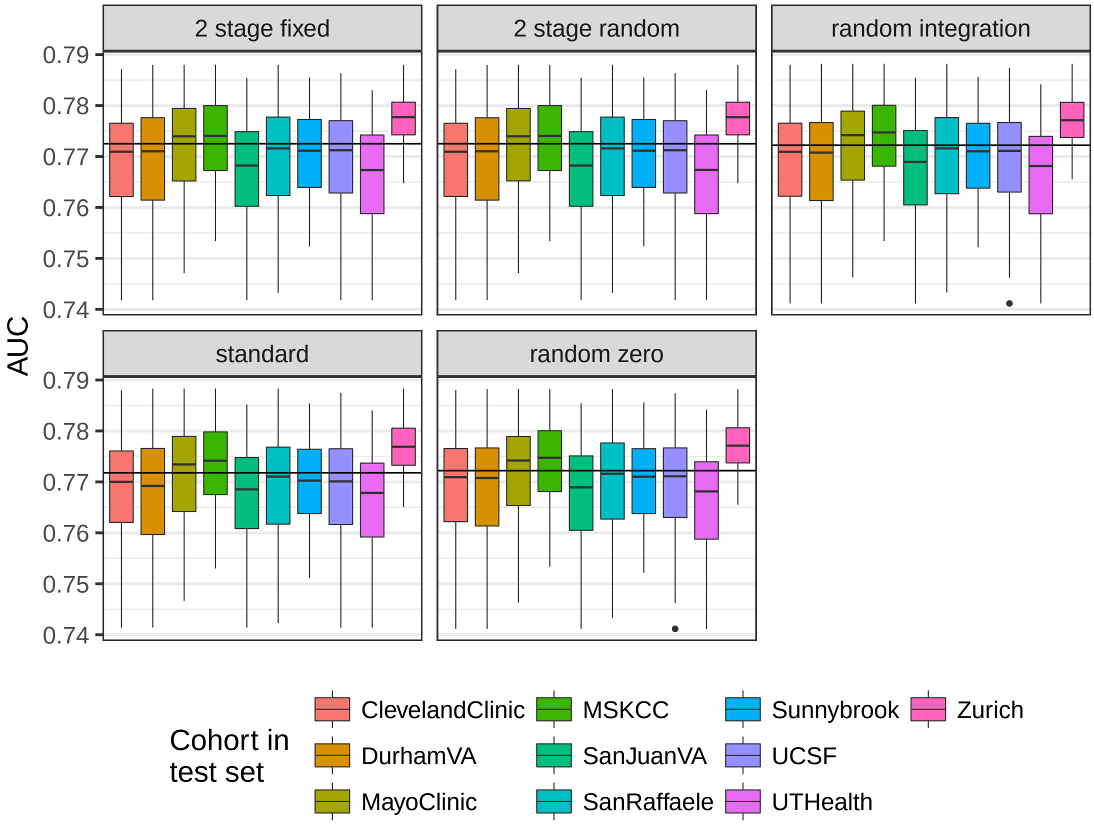


Figure 21 : AUC values for 252 cross validation sorted by cohorts included in the test set. Each boxplot summarizes the AUC values of one method for the 126 test sets in which the considered cohort was in the test set. Horizontal lines correspond to the median of the respective method.

Exclusion of ClevelandClinic, DurhamVA, SanJuanVA as well as UTHealth improved the resulting AUC, whereby SanJuanVA achieved the largest negative difference. The AUC decreased for excluding either MayoClinic, MSKCC or Zurich, whereby exclusion of the latter cohort resulted in highest absolute differences and widest percentile intervals. Interestingly, even though Zurich is different to all other cohorts in terms of its prevalence of risk factors as well as proportion of patients with high-grade cancer, an exclusion leads to a drop in AUC (Figures 6, 11 and 22). Overall, none of the exclusions significantly improved the performance of the different models in terms of the AUC, as the corresponding percentile intervals all cover zero. We further investigated calibration and net benefit curves for the two most extreme cohorts SanJuanVA and Zurich to compare leave-one-cohort-out cross validations with and without the respective cohort. These comparisons showed no significant differences.

As a final measure, we investigate whether a model built on any single cohort, performs consistently better than a model integrating all cohorts. To this purpose a standard multiple logistic regression model is fit to each of the cohorts and evaluated on every remaining cohort. The results are compared to the simplest of the previously discussed methods, the standard model, which pools all sites together without accounting for heterogeneity. For the latter method leave-one-cohort-out cross validation is used in order to get reliable results. The

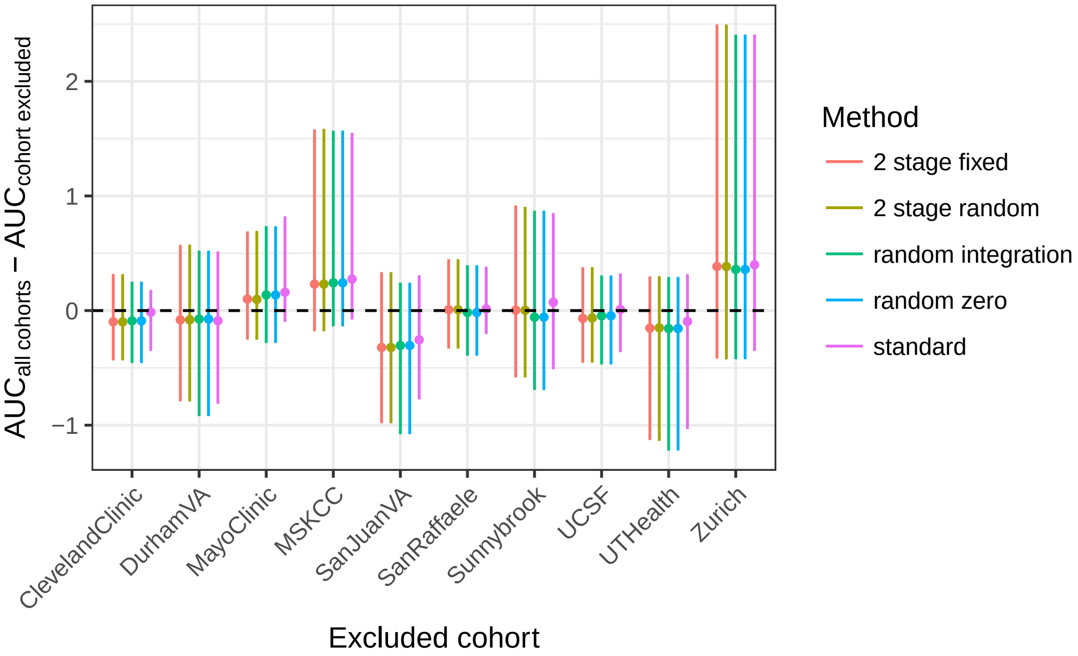


Figure 22 : Median estimates and 95% percentile intervals of permutation analysis for differences in AUC values comparing the 252 cross validation with all cohorts included and the 252 cross validation with the indicated cohort excluded from either the training or test set. Differences are given in percent.

corresponding AUC and HLS values are given in Figure 23. In terms of the AUC the combined model is one of the best methods for every test set, except for SanJuanVA. For this cohort, three individual-cohort models outperform the pooled approach. Nevertheless, none of the models built on single cohorts has a better AUC than the combined method throughout all considered test sets. Similarly, the pooling method has consistently low HLS values, whereas the performance of single models varies across test cohorts. The corresponding net benefit, calibration, sensitivity and specificity curves confirm these observations (Figures A.6, A.7, A.8 and A.9): For single test sets, individual models might outperform the combined approach in terms of calibration and net benefit, but nevertheless, none of them performs consistently better across all test cohorts. The model built on UTHealth shows best sensitivity for all test set, it has, however, lowest specificity. Vice versa, Zurich and MayoClinic perform worst in terms of sensitivity, but achieve high specificity.

3.5.5 Coefficients of the different models

As explained in Section 3.4, in order to derive the final models, all available data are used for estimating the coefficients and their 95%-CIs, whereby the results are shown in Figure 24. The coefficients of standard logistic regressions based on the individual cohorts are colored in red. These are the basis of the two-stage meta-analysis models and it is interesting to compare them to the resulting summarizing models. The two random intercept risk prediction methods have the same coefficients of the fixed effects and are therefore listed only once. Even though the coefficients of the individual cohorts are quite diverse, the summarizing models are very similar, whereby the two-stage models have wider CIs than the one stage

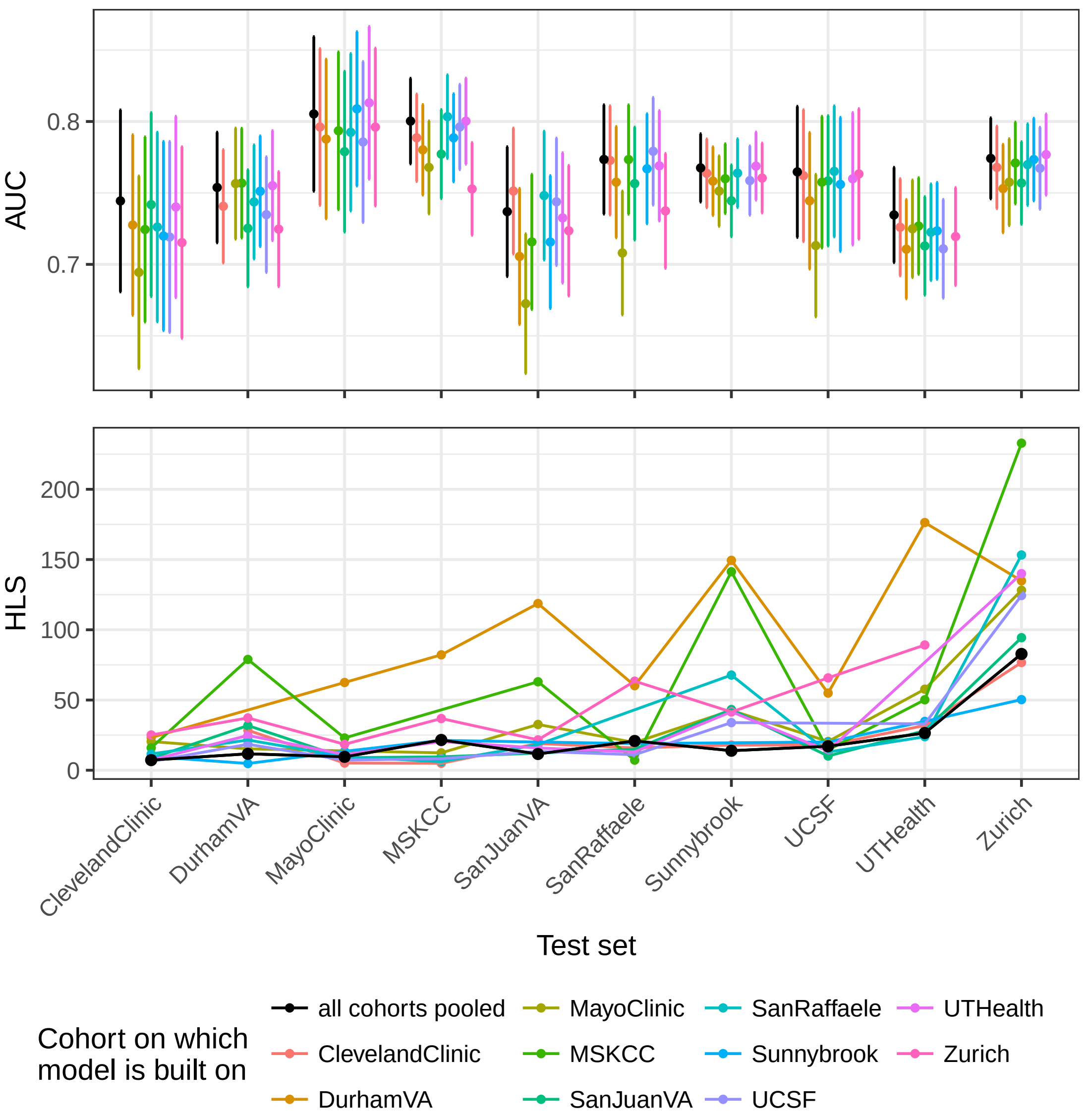


Figure 23 : AUC and HLS values for models built on individual cohorts. Leave-one-cohort-out cross validation results for standard multiple logistic regression are given for comparison in black. For the AUC values; 95%-CIs are included.

methods random intercept and standard.

For the standardized continuous variable age the coefficients range between 0.36 and 0.97, implicating an increased risk of high-grade cancer for older patients. The coefficient most diverse from the other cohorts is given for DurhamVA. However, since this coefficient has a large CI its influence on summary models can be considered marginal. DurhamVA is furthermore the only cohort with a positive coefficient for the interaction of age and DRE. For the other cohorts, age is less influential for patients with an abnormal DRE. For most cohorts the coefficients for age are further reduced for patients with African ancestry. Exceptions are UCSF and Sunnybrook, both cohorts with low prevalences of patients with African ancestry (2.7% and 4.2%, respectively), resulting in wide CIs including 0.

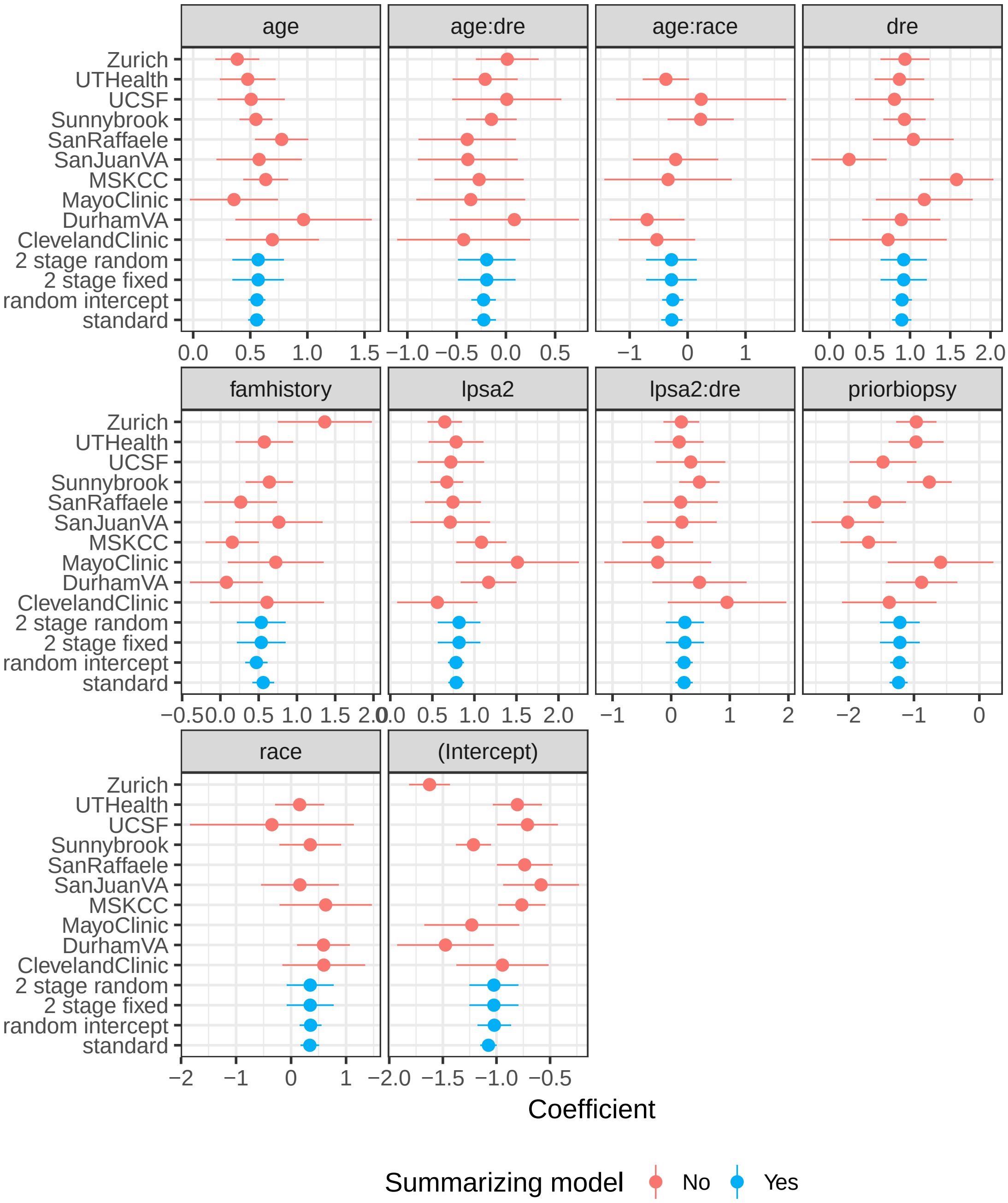


Figure 24 : Coefficients for every individual cohort in red and for the summarizing models in blue. Coefficients for race and the interaction of race and age are not included for Zurich, SanRaffaele and MayoClinic, as there are no or only 2 patients with African Ancestry in these cohorts. Furthermore the coefficient for family history is excluded for UCSF.

A positive DRE is related with an increased risk of high-grade cancer for all cohorts. The smallest coefficient is predicted for SanJuanVA, which is an outlier in terms of proportion of abnormal DRE results (Figure 11). Despite its high prevalence of this risk factor, DRE is no significant covariate for SanJuanVA.

The biggest differences for the summarizing models, even though these are also marginal, can be observed for the covariate family history. For this variable, the coefficient for Zurich is higher than for the other cohorts. Zurich was identified as an outlying cohort for the variable family history in terms of prevalence and univariate odds ratio in Figure 11, and since Zurich is the largest cohort in the data set, this outlying coefficient influences the prediction of coefficients of the summarizing models. It hast the most influence on the standard logistic regression model, followed by the two-stage approaches. The random intercept method is most robust for this outlier.

As for age, the coefficients of lpsa2 are given for the standardized covariate, so for a change in lpsa2 by one standard deviation, given by 1.3. The resulting estimates are in the range of 0.56-1.51, with the lowest value for ClevelandClinic, and the highest for MayoClinic. The interaction between lpsa2 and DRE shows mostly positive coefficients, leading to an increased effect of PSA for patients with an abnormal DRE result.

A prior negative biopsy significantly reduces the risk of high-grade cancer for most cohorts. Only MayoClinic includes 0 in the CI which can be explained by its overall small sample size (n=323).

Except UCSF all cohort, for which a coefficient could be calculated, have positive coefficients for the covariate African ancestry, implying that patients with African ancestry have higher risks for high-grade cancer. However, the only cohort with a significant coefficient is DurhamVA, which has the highest prevalence of patients with African Ancestry with 63.1% (Figure 11). In contrast UCSF has a low prevalence of 2.7%, resulting in a wide CI, undermining the influence of this negative coefficient.

3.6 Discussion

One goal of the PBCG is to derive a global risk calculator to predict the risk of high-grade prostate cancer on biopsy. In order to obtain a general valid tool, several different cohorts were included and the optimal integration of the diverse centers was covered in this chapter. The simplest approach of a standard logistic regression model was compared to one- and two-step meta-analyses, with results validated by internal-external cross validation techniques, and found not to perform less adequately. This is an appealing result, since simple pooling is routinely performed for multi-site clinical trials (Stephan et al. 2002, Yanke et al. 2006): The PCPTRC pooled data from 221 study sites, and has been repeatedly used, validated and updated, ensuring reliability (I. M. Thompson, Goodman, Tangen, Lucia, et al.

2003, I. M. Thompson, Ankerst, et al. 2006, Ankerst, Hoefler, et al. 2014, Ankerst, Boeck, et al. 2012, Grill, Fallah, Leach, I. M. Thompson, Freedland, et al. 2015, Trottier et al. 2011, van den Bergh et al. 2008).

The results of this analysis also support the validity of two-stage methods. This introduces several advantages for big collaborations like the PBCG, as the initial model building as well as consecutive updates of an existing risk prediction tool can be simplified. It becomes redundant to gather the IPD of all cohorts at a central location, as logistic regression models can be built locally at individual sites with only the summary statistics transported for centralization. Analogously it gets easier to include updates of single cohorts, as it is sufficient to use the updated summaries of the respective cohorts and repeat the second step of the meta-analysis. Furthermore it enables easy inclusion of additional published studies for which only data summaries are provided, which would allow further generalization of the given risk prediction tool. With this external enhancement of the data set, the advantages of prospectively designed multi-cohort collaborations become diminished and the resulting challenges of standard meta-analysis have to be addressed, including publication bias as well as poor and selective reporting (Riley, Lambert, et al. 2010, Burke et al. 2017, Debray, Moons, Abo-Zaid, et al. 2013).

An interesting finding of the PBCG analysis was the influence of individual cohorts, in particular the European cohort Zurich. Whereas this site was very different compared to the remaining ones in terms of the large sample size, low baseline risk, and absence of patients with African ancestry, removal of Zurich worsened the overall model fit. Similarly, the exclusion of other cohorts did not improve the model fit noticeably. This leads to the conclusion that in our case it might be useful to include all possible data, which is desirable after the efforts of data collection and in light of the fact that the resulting risk tool should be applied to patients with a broad range of diverse characteristics. In general however, we advise to examine the exclusion of all outlying cohorts. Individual cohorts can be characterized as outliers in terms of unique prevalences of the outcome, which can not be explained by the risk factors, or in terms of the distributions of single risk factors themselves. Pölsterl et al. 2016 argue that model training based on data similar to the test set might be crucial, as for their application a model based on data excluding an outlying cohort outperformed a broad range of other approaches. Also in traditional meta-analysis it is common to include only comparable studies, as otherwise the particular inclusion criteria vary across studies, resulting in different findings (Sigman 2011, Petrosino 2016, Kang et al. 2012, Tabak et al. 1991). Therefore meta-analysis guidelines recommend a detailed documentation (Jain et al. 2012, Moher et al. 2009). We advocate a description of inclusion criteria also for individual cohorts in the field of multi-cohort studies, in particular if cohorts are removed to improve model performance.

Finally, the PBCG analysis supports the conclusion of Pavlou et al. 2015 that median and mean prediction does not vary much in the context of clustering within hospitals. As integration over the random intercept can be omitted for the median prediction, this method is easier to implement and also commonly used (Debray, Moons, Ahmed, et al. 2013, Bouwmeester, Twisk, et al. 2013). It might be interesting to investigate the inclusion of more random effects, as the heterogeneity across cohorts might not be restricted to the baseline risk. On the other hand, more random effects lead to more parameters to be estimated and a more complex structure of the model. The added value to the PBCG model performance might be small, as the clustering in this data set was not extreme and one- and two-stage models had similar results (Section 3.1).

4 Development of a contemporary prostate cancer risk prediction model and comparison to the current standard

The PCPTRC is a widely used online prostate cancer risk assessment tool. However, the underlying model is based on data from the 1990s, which might not be representative for today's population. Thus, a model based on recent data might be a considerable improvement. In this chapter, which is based on Ankerst, Straubinger, et al. 2018, the PCTPRC is validated on patient data from 2006 to 2017 and compared to an updated risk model developed analogously to the PCPTRC for a fair comparison. In the following analyses the second objective of this thesis, whether existing risk tools should get updated as soon as contemporary data are available, is discussed.

4.1 Research in context

Prostate cancer risk prediction models can help clinicians, as well as their patients, to objectively decide whether or not a prostate biopsy should be performed. This enables the possibility of reducing the overall amount of biopsies. An incorrect model, however, can cause severe harm, as patients with a high risk of prostate cancer might decide against a biopsy, due to an unreliable low risk assessment. Alternatively, an incorrectly calculated high risk might lead to unnecessary distress caused by the biopsy. For these reasons it is indispensable to repeatedly validate commonly used risk prediction tools in order to detect and, if possible, correct unreliable models.

The need to update existing risk assessment tools is evident in a broad range of medical areas. DeFilippis et al. 2015 consider the widely used Framingham risk tool for cardiovascular disease risk predictions by Wilson et al. 1998, as well as some of its updates over the years. The authors name possible reasons for a systematic overestimation by these risk scores in recent cohorts. Thereby they discuss changes between older and contemporary study populations, including differences in the significance of individual risk factors and the use of preventive pharmacotherapies. In the area of coronary artery disease, Genders et al. 2011 furthermore reveal limited benefit of the Diamond-Forrester model for current use, as overestimation is evident in contemporary cohorts (Diamond and Forrester 1979). Even though this model is about 40 years old, the authors refer to guidelines that have still included it (Gibbons et al. 2002, Hendel, Berman, et al. 2009, Hendel, Patel, et al. 2006). Another example is the discussion of temporal validation of prediction models in Austin, van Klaveren, et al. 2016 and Austin, van Klaveren, et al. 2017, applied to mortality within 30 days and one year of hospitalization for congestive heart failure. The authors compare data from the time periods 1999 to 2001 and 2004 to 2005, collected at the same hospitals. They detected a lower probability of death in the contemporary data, but mostly comparable influence of the predictor variables in both time periods. In the field of quality of care assessment in intensive care units, Minne et al. 2012a and Minne et al. 2012b emphasize the need to recalibrate prognostic models in time. They furthermore state that corresponding validations should be performed repeatedly over time. The authors identify possible reasons as drifts in population, resulting in differences in patient mix, technologies, and treatment policies. Strobl, Vickers, et al. 2015 and Strobl, I. M. Thompson, et al. 2015 demonstrate the benefit of annual recalibration of prostate cancer risk prediction models for improving accuracy, but less so for discrimination.

Several papers also discuss methods and benefits of updating risk tools with new predictor variables as their influence get discovered or as they become more available (Skates et al. 2001, Raji et al. 2010, Pencina, D'Agostino, and Vasan 2008, K. M. Anderson et al. 1991, T. J. Wang et al. 2006, Ridker et al. 2007, W. Gu and Pepe 2009). Amongst these, Chatterjee et al. 2016, Grill, Ankerst, et al. 2017 and Cheng, Taylor, Vokonas, et al. 2018 illustrate diverse methods to combine information from previous models based on large data sources with information on new risk factors from a single cohort or case-control study. Therefore they enable likelihoood ratio approaches and more general constrained maximum likelihood methods. Applications of model updates using new covariates are for instance in the field of bone lead levels (Cheng, Taylor, Vokonas, et al. 2018). Furthermore, Ankerst, Groskopf, et al. 2008 and Ankerst, Koniarski, et al. 2012 consider integration of the urine marker PCA3, free PSA percentage and [-2]proenzyme PSA into the prostate cancer risk calculator by I. M. Thompson, Ankerst, et al. 2006. Grill, Fallah, Leach, I. M. Thompson, Hemminki, et al. 2015 and Grill, Fallah, Leach, I. M. Thompson, Freedland, et al. 2015 furthermore integrate single nucleotide polymorphisms (SNPs) and detailed family history into the updated prostate cancer model by Ankerst, Hoefler, et al. 2014. For breast cancer, Tyrer et al. 2004 advocate combining the risk factors from the broadly used Gail model by Gail, Brinton, et al. 1989 with genotype information as investigated in Claus et al. 1994 and Parmigiani et al. 1998. However, Gail 2009 and Wacholder et al. 2010 find only modest improvements by incorporating seven and ten SNPs, respectively, into the Gail model. Similarly, inclusion of mammographic density shows modest gain in discriminatory power (Chen et al. 2006).

Like model updates for contemporary data, model adjustments for the population at hand can improve risk predictions (J. Liu et al. 2004, D'Agostino et al. 2001, Hense 2003, Janssen, Moons, et al. 2008, Janssen, Vergouwe, et al. 2009). For instance in predicting invasive breast cancer, Gail, Costantino, et al. 2007, Matsuno et al. 2011, Banegas et al. 2012, Kaur et al. 2004, Marrugat et al. 2003 and Pastor-Barriuso et al. 2013 modify the Gail model and its update in S. J. Anderson et al. 1992 to fit African American, Asian and Pacific Islander American, Hispanic American, American Indian and Alaska Native, and Spanish women, respectively.

Extensive research is performed in the development of prostate cancer risk prediction models, covering model choice, outcome specification, predictors, and underlying population.

Regarding model choice, logistic regression is most commonly employed (Eastham et al. 1999, I. M. Thompson, Ankerst, et al. 2006, Williams et al. 2012, J. Xu et al. 2009, Zheng et al. 2012, Liang et al. 2011, Kranse et al. 2008, Kuo et al. 2013, Lilja et al. 2011, Lindström et al. 2012, Optenberg et al. 1997, Nomura et al. 2012, Gregorio et al. 2007). Extensions to multinomial logistic regression are utilized by Roobol et al. 2013 and Ankerst, Hoefler, et al. 2014, whereas ordinal regression is used by Nam et al. 2007. Both binary and multinomial logistic regression are extended with likelihood ratio analysis for incorporation of specific risk factors by Ankerst, Koniarski, et al. 2012, Ankerst, Hoefler, et al. 2014, Grill, Fallah, Leach, I. M. Thompson, Hemminki, et al. 2015 and Grill, Fallah, Leach, I. M. Thompson, Freedland, et al. 2015. Furthermore, segregation analysis is implemented by Macinnis et al. 2011.

In most prostate cancer risk papers the outcome is focused on the current prostate cancer status, distinguishing between a positive and negative biopsy (Eastham et al. 1999, Kranse et al. 2008, Kuo et al. 2013, Lindström et al. 2012, Nomura et al. 2012, Optenberg et al. 1997, Liang et al. 2011, Zheng et al. 2012, Gregorio et al. 2007). In addition to or instead of this analysis, the outcome of aggressive prostate cancer might be compared to the combined classification of no or non-aggressive cancer. In I. M. Thompson, Ankerst, et al. 2006, aggressive cancer is thereby defined by a Gleason score of seven or higher and in Williams et al. 2012 by a Gleason score of seven or higher, more than three positive cores, or at least 50% tumor involvement in any individual core. Further definitions are given in Lilja et al. 2011 by a clinical stage greater than or equal to T3 or radiographic evidence of bone metastases at diagnosis and in J. Xu et al. 2009, by a Gleason score of eight or higher, a PSA value higher than 50 ng/ml, clinical stage T3/4, N+, or M+. Binary analyses of outcome by Gleason grade are combined in a multinomial outcome of no, non-aggressive and aggressive cancer, whereby the latter two options are defined by Gleason scores smaller than, and higher than or equal to seven, respectively, by Ankerst, Hoefler, et al. 2014 and Nam et al. 2007. In Roobol et al. 2013 the distinction of the two cancer classifications in a multinomial model is defined by Gleason score seven or higher, a PSA of 10ng/ml or higher or a clinical stage greater than T2b. Finally, instead of current cancer status, outcomes are also defined as screening results after 4 or 20-30 years (Roobol et al. 2013, Lilja et al. 2011).

Most variation across prostate cancer risk studies can be found in terms of considered predictors. In most models diverse subsets of the risk factors age, race, PSA value, suspicious DRE, prostate volume, and family history of prostate and other cancers are utilized, whereby most of them are already routinely collected in the clinic (Ankerst, Hoefler, et al. 2014, Eastham et al. 1999, Grill, Fallah, Leach, I. M. Thompson, Hemminki, et al. 2015, Grill, Fallah, Leach, I. M. Thompson, Freedland, et al. 2015, I. M. Thompson, Ankerst, et al. 2006, Kranse et al. 2008, Kuo et al. 2013, Lilja et al. 2011, Lindström et al. 2012, Macinnis et al. 2011, Nam et al. 2007, Nomura et al. 2012, Optenberg et al. 1997, Roobol et al. 2013, Williams et al. 2012, J. Xu et al. 2009, Liang et al. 2011, Gregorio et al. 2007). These common predictors are augmented with information about urinary symptoms by Nam et al. 2007, obesity in terms of body mass index by Williams et al. 2012, geographic regions by J. Xu et al. 2009, detailed family history by Grill, Fallah, Leach, I. M. Thompson, Freedland, et al. 2015 and prior negative prostate biopsies by Roobol et al. 2013, Ankerst, Hoefler, et al. 2014, and I. M. Thompson, Ankerst, et al. 2006. Variations of the risk factor PSA and additional serum markers are given by PSA density, free PSA percentage, [-2]proenzyme PSA, neuroendocrine marker, Dickkopf-1 und human kallikrein 2 (Nam et al. 2007, Gregorio et al. 2007, Williams et al. 2012, Liang et al. 2011, Lilja et al. 2011, Kuo et al. 2013, Ankerst, Koniarski, et al. 2012, Ankerst, Hoefler, et al. 2014). Considerations concerning the prostate transition zone, in terms of transition zone volume and PSA transition zone density, are included by Gregorio et al. 2007. The relation of prostate cancer and SNPs is investigated in other models (Zheng et al. 2012, Grill, Fallah, Leach, I. M. Thompson, Hemminki, et al. 2015, Lindström et al. 2012, Macinnis et al. 2011, J. Xu et al. 2009, Grill, Fallah, Leach, I. M. Thompson, Freedland, et al. 2015).

Analyses are furthermore varied in the underlying study population. They might be limited by specific geographical units, like Sweden, China, Japan, Brazil, Europe or North America, as well as by specific inclusion requirements, such as a given PSA level, age and DRE result (J. Xu et al. 2009, Lilja et al. 2011, Kuo et al. 2013, Zheng et al. 2012, Gregorio et al. 2007, Nomura et al. 2012, Kranse et al. 2008, Williams et al. 2012, Ankerst, Hoefler, et al. 2014). As previously discussed, the time frame in which the biopsies have been performed is a relevant characteristic of the study population. For instance, the analyses of Optenberg et al. 1997 and Eastham et al. 1999 are based on biopsies from 1991-1995 and 1990-1997, respectively, and might therefore be considered outdated.

Among this broad range of models, we identified the PCPTRC and the ERSPC risk calculator as the most commonly used, freely available and user-friendly risk tools (Kranse et al. 2008, Ankerst, Hoefler, et al. 2014). Both are based on large prospective trials, based on a heavily screened population, whereby the ERSPC study includes European centers and the PCPTRC utilizes North American data. The PCPT study required a PSA level of 4 ng\mL or lower, a minimum of 55 years and a normal DRE result for patients to enter the trial. Furthermore, the PCPT required an end-of-study biopsy for all men who had not been diagnosed with prostate cancer within the seven years of the trial.

Since the PBCG data include several North American cohorts with a considerable amount of patients with African ancestry, we chose to validate the PCPTRC on the contemporary PBCG data and compare it with a newer model developed on the PBCG data using cross validation to adjust for overfitting as well as a hold-out PBCG test set. The ERSPC risk tool is less suitable as it comprises a near exclusively white population and previous work has found African ancestry to be a significant risk factor for prostate cancer (Ankerst, Hoefler, et al. 2014, Nam et al. 2007, Yanke et al. 2006).

The PCPTRC is based on data from the 1990's and changes in clinical practice have since then occurred, raising the question as to whether it should be updated or replaced. In the

PCPT, clinicians have primarily utilized six cores as mandated by the protocol, whereas in contemporary practice, such as in the PBCG cohorts, twelve-core biopsies are predominant. A higher number of cores correlates with increased cancer detection on biopsy, thus, this shift in clinical practice may reduce the value of the PCPTRC for contemporary patients (Ankerst, Till, Boeck, Goodman, Tangen, and I. M. Thompson 2013, Ankerst, Till, Boeck, Goodman, Tangen, Feng, et al. 2013, Babayan and Katz 2016, Bjurlin et al. 2014). A further relevant factor is a change in the prostate cancer grading system based on the Gleason score for which the International Society of Urological Pathology first made revisions in 2005, and more recently in 2014 (Kryvenko and Epstein 2016). Several researchers have detected an upwards shift in Gleason grading of prostate cancer (Ghani et al. 2005, Smith et al. 2002, Zareba et al. 2009). Danneman et al. 2015 further suggested that this shift predated the grading system revisions of 2005, and just became more obvious afterwards. This development precipitated an increased assessment of prostate cancer as high-grade disease.

4.2 Methods

This chapter is motivated by the importance of contemporary data and the consequential need for risk model updates. A new risk tool is developed analogously to the PCPTRC to guarantee a fair comparison, with concentration solely on the influence of new data rather than different modeling techniques. The PCPTRC is based on a multinomial logistic regression, in which the risk of high- versus low-grade versus no prostate cancer is predicted using the risk factors PSA (logarithmically transformed), first-degree family history, prior negative biopsy, race, age and DRE. In order to match the PCPTRC, high-grade cancer is defined as Gleason grade greater than or equal to seven, not differentiating between Gleason score 4+3 and 3+4. For model building all cohorts are pooled together, which is current practice for multi-center clinical studies and proves to have no disadvantages for the purpose of risk prediction (Chapter 3).

4.2.1 Multinomial logistic regression

Multinomial logistic regression is an extension of the standard multiple logistic regression model for binary outcomes, discussed in Section 3.3.1. With this modification it is possible to model a nominal outcome with more than two levels. We explain the theoretical concept via the example of a three level outcome. An extension to more categories is straightforward and mainly a matter of notation. In this chapter the outcome corresponds to the biopsy result with the three levels no, low-grade and high-grade prostate cancer, coded as 0, 1 and 2, respectively. Let y_i denote the outcome of patient $i = 1, ..., n$, with $X_i = (1, x_{1i}, ..., x_{Ki})'$ the corresponding covariate vector and n the total number of biopsies within the pooled data set. We assume the y_i are independent for $i = 1, ..., n$.

In order to specify the required logit functions, one of the outcome categories has to be chosen as reference value. An intuitive choice is given by the biopsy result of no cancer, corresponding to $y = 0$. As a result, the two cancer levels low- and high-grade are compared to the baseline of no cancer. A multinomial distribution is assumed for y_i, with probabilities $p_{0i},\ p_{1i},\ p_{2i} \in (0,1)$ for the outcomes 0, 1 and 2, respectively. The multinomial logistic regression model is then expressed by

$$log\left(\frac{p_{1i}}{p_{0i}}\right) = \beta_{10} + \sum_{k=1}^{K} \beta_{1k} x_{ki} = \beta_1' X_i \quad \text{and} \tag{4.1}$$

$$log\left(\frac{p_{2i}}{p_{0i}}\right) = \beta_{20} + \sum_{k=1}^{K} \beta_{2k} x_{ki} = \beta_2' X_i, \tag{4.2}$$

whereby $\beta_j = (\beta_{j0}, \beta_{j1}, ..., \beta_{jK})',\ j = 1, 2$ are the vectors of regression parameters. A comparison of the cancer categories $y = 2$ and $y = 1$ is given by the difference between Equation (4.2) and (4.1):

$$log\left(\frac{p_{2i}}{p_{1i}}\right) = log\left(\frac{p_{2i}}{p_{0i}}\right) - log\left(\frac{p_{1i}}{p_{0i}}\right) = \left(\beta_2' - \beta_1'\right) X_i. \tag{4.3}$$

For an easy representation of the log-likelihood function, the multinomial outcome with three categories is recoded with three binary variables:

$$\begin{aligned} y_i = 0: &\quad y_{0i} = 1, \quad y_{1i} = 0, \quad y_{2i} = 0 \\ y_i = 1: &\quad y_{0i} = 0, \quad y_{1i} = 1, \quad y_{2i} = 0 \\ y_i = 2: &\quad y_{0i} = 0, \quad y_{1i} = 0, \quad y_{2i} = 1. \end{aligned}$$

For the new variables $y_{0i},\ y_{1i}$, and y_{2i}, the equation $\sum_{j=0}^{2} y_{ji} = 1$ holds for all $i = 1, ..., n$. With this it is possible to formulate the log-likelihood as

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} log\left[\left(\frac{1}{1 + e^{\beta_1' X_i} + e^{\beta_2' X_i}}\right)^{y_{0i}} \left(\frac{e^{\beta_1' X_i}}{1 + e^{\beta_1' X_i} + e^{\beta_2' X_i}}\right)^{y_{1i}} \left(\frac{e^{\beta_2' X_i}}{1 + e^{\beta_1' X_i} + e^{\beta_2' X_i}}\right)^{y_{2i}}\right] \\ &= \sum_{i=1}^{n} log\left[\frac{e^{\beta_1' X_i y_{1i}} e^{\beta_2' X_i y_{2i}}}{\left(1 + e^{\beta_1' X_i} + e^{\beta_2' X_i}\right)^{y_{0i} + y_{1i} + y_{2i}}}\right] \\ &= \sum_{i=1}^{n} log\left[\frac{e^{\beta_1' X_i y_{1i}} e^{\beta_2' X_i y_{2i}}}{1 + e^{\beta_1' X_i} + e^{\beta_2' X_i}}\right] \\ &= \sum_{i=1}^{n} \left[y_{1i} \beta_1' X_i + y_{2i} \beta_2' X_i - log\left(1 + e^{\beta_1' X_i} + e^{\beta_2' X_i}\right)\right]. \end{aligned} \tag{4.4}$$

The maximum likelihood estimate $\hat{\beta}$ of the coefficient vector $\beta = (\beta_1', \beta_2')'$ is obtained by setting the partial derivatives of Equation (4.4) to zero:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \left[y_{ji} X_i - \frac{e^{\beta_j' X_i}}{1 + e^{\beta_1' X_i} + e^{\beta_2' X_i}} X_i\right] = 0, \quad j = 1, 2. \tag{4.5}$$

Under regularity conditions, the likelihood function is globally convex, and the resulting maximum likelihood estimate $\hat{\beta}$ unique (Hasan et al. 2016). Analogously to the binary case, the

Newton-Raphson algorithm is used to approximate the solutions of the $2(K+1)$ equalities specified in Equation (4.5). The updating rule of this iterative process becomes

$$\begin{bmatrix} \beta_1^{new} \\ \\ \beta_2^{new} \end{bmatrix} = \begin{bmatrix} \beta_1^{old} \\ \\ \beta_2^{old} \end{bmatrix} - \left(\begin{bmatrix} \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_1'} & \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_2'} \\ \\ \frac{\partial^2 l(\beta)}{\partial \beta_2 \partial \beta_1'} & \frac{\partial^2 l(\beta)}{\partial \beta_2 \partial \beta_2'} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial l(\beta)}{\partial \beta_1} \\ \\ \frac{\partial l(\beta)}{\partial \beta_2} \end{bmatrix} \right) \Bigg|_{\beta_1=\beta_1^{old}, \beta_2=\beta_2^{old}} = \begin{bmatrix} \beta_1^{old} \\ \\ \beta_2^{old} \end{bmatrix} + \Delta, \tag{4.6}$$

with

$$\begin{aligned} \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_1'} &= \sum_{i=1}^{n} -\frac{e^{\beta_1' X_i}(1+e^{\beta_2' X_i})}{1+e^{\beta_1' X_i}+e^{\beta_2' X_i}} X_i X_i', \\ \frac{\partial^2 l(\beta)}{\partial \beta_2 \partial \beta_2'} &= \sum_{i=1}^{n} -\frac{e^{\beta_2' X_i}(1+e^{\beta_1' X_i})}{1+e^{\beta_1' X_i}+e^{\beta_2' X_i}} X_i X_i', \\ \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_2'} &= \frac{\partial^2 l(\beta)}{\partial \beta_2 \partial \beta_1'} = \sum_{i=1}^{n} \frac{e^{\beta_1' X_i} e^{\beta_2' X_i}}{\left(1+e^{\beta_1' X_i}+e^{\beta_2' X_i}\right)^2} X_i X_i', \end{aligned}$$

and Δ set to the negative product of the inverse of the matrix of second derivatives times the vector of first derivatives, evaluated at β_{old}. This algorithm is implemented in R with the package mnlogit (Hasan et al. 2016). In case of $l(\beta^{new}) < l(\beta^{old})$, the previous step is modified to add $\frac{\Delta}{2}$ instead of Δ. This bisecting is repeated until the log-likelihood value of the updated β^{new} is greater than or equal to the value at β^{old}. The original Newton-Raphson algorithm is continued with this modified version of β^{new}. With this procedure the Newton-Raphson iterations converge (Hasan et al. 2016). Using the resulting parameter estimates, the risk predictions $\hat{p}_{0i},\ \hat{p}_{1i}$ and $\hat{p}_{2i}$ are given by

$$\hat{p}_{0i} = \frac{1}{1+e^{\hat{\beta}_1' X_i}+e^{\hat{\beta}_2' X_i}}, \tag{4.7}$$

$$\hat{p}_{ji} = \frac{e^{\hat{\beta}_j' X_i}}{1+e^{\hat{\beta}_1' X_i}+e^{\hat{\beta}_2' X_i}}, \quad j=1,2. \tag{4.8}$$

4.2.2 Missing values

For the PCPTRC missing values are allowed for the variables DRE, family history and prior negative biopsy. This is achieved by fitting marginal models for any combination of missing one, two or all of these covariates. The same procedure for missing values is used to ensure a fair comparison between the existing risk tool PCPTRC and the new model built on the contemporary PBCG data. However, in the PBCG, missing values are also present for the characteristic race, which are imputed by their median non-African ancestry. This imputation is used for building the PBCG model and for validation of both the PBCG model and PCPTRC.

Algorithm 11 Multinomial logistic regression model and its prediction

```
procedure MULTINOMIAL($IPD$)
    $Data \leftarrow$ pool all cohorts of $IPD$ together        ▷ cohort information is ignored
    fit multinomial logistic regression on $Data$ via Newton-Raphson approximation
    $\hat{\beta}_1 \leftarrow$ resulting coefficient estimates for low-grade cancer compared to no cancer
    $\hat{\beta}_2 \leftarrow$ resulting coefficient estimates for high-grade cancer compared to no cancer
    return $\hat{\beta}_1,\ \hat{\beta}_2$
procedure PREDICT_MULTINOMIAL($\hat{\beta}_1,\ \hat{\beta}_2,\ X$)
                                          ▷ $X$: covariate vector of individual patient
    $\hat{p}_0 \leftarrow \frac{1}{1+exp(\hat{\beta}_1' X)+exp(\hat{\beta}_2' X)}$
    $\hat{p}_1 \leftarrow \frac{exp(\hat{\beta}_1' X)}{1+exp(\hat{\beta}_1' X)+exp(\hat{\beta}_2' X)}$
    $\hat{p}_2 \leftarrow 1 - \hat{p}_0 - \hat{p}_1$
    return $\hat{p} = (\hat{p}_0,\ \hat{p}_1,\ \hat{p}_2)$
```

4.2.3 Validation

The models are evaluated by the validation methods described in Chapter 2. The discrimination ability is graphically shown with sensitivity and specificity curves, and summarized by AUC values. Calibration is quantified by the HLS and displayed in calibration plots. To show clinical utility, net benefit curves are used. As these validation metrics are developed exclusively for binary outcomes, solely results of high-grade cancer versus no or low-grade cancer are discussed in detail. Results for overall cancer, with high- and low-grade cancer combined, versus no cancer are shown in the appendix.

The PCPTRC is a risk tool based on data from North America, whereas the PBCG combines eight North American and three European cohorts. The European sites include less than 1% of patients with African ancestry, and in Hamburg, African ancestry information is missing altogether on 63% of men. For fair comparisons with the PCPTRC, we decided to build the new model on North American data only. With this partition it is furthermore possible to implement two levels of model validation for the PBCG model: internal-external cross validation within the North American cohorts, and external validation on the remaining European sites. To validate the PBCG model on the individual north American cohorts, we use leave-one-cohort-out cross validation, as described in Section 3.4. For the external validation on the three European cohorts, we use the final PBCG model that is fit on all North American data pooled together. Since the PCPTRC is an external risk tool, it can be directly applied to all cohorts.

4.3 Results

We fit a new multinomial logistic regression model on contemporary North American data and compare its results to the broadly used online available PCPTRC.

4.3.1 Multinomial logistic regression model

The final model is built on all eight cohorts in North America pooled together, comprising 5,992 biopsies. Analogously to the PCPTRC, which was built on 6664 biopsies, marginal models for every combination of missing variables prior negative biopsy, family history and DRE result are fit. Corresponding R-code is given in the appendix Section A.2. Odds ratios for the model with no missing variables, which is based on 4,286 biopsies, are shown in Figure 25 for each of the comparisons high- versus low-grade cancer, high-grade versus no cancer, and low-grade versus no cancer, with further data in Table A.3.

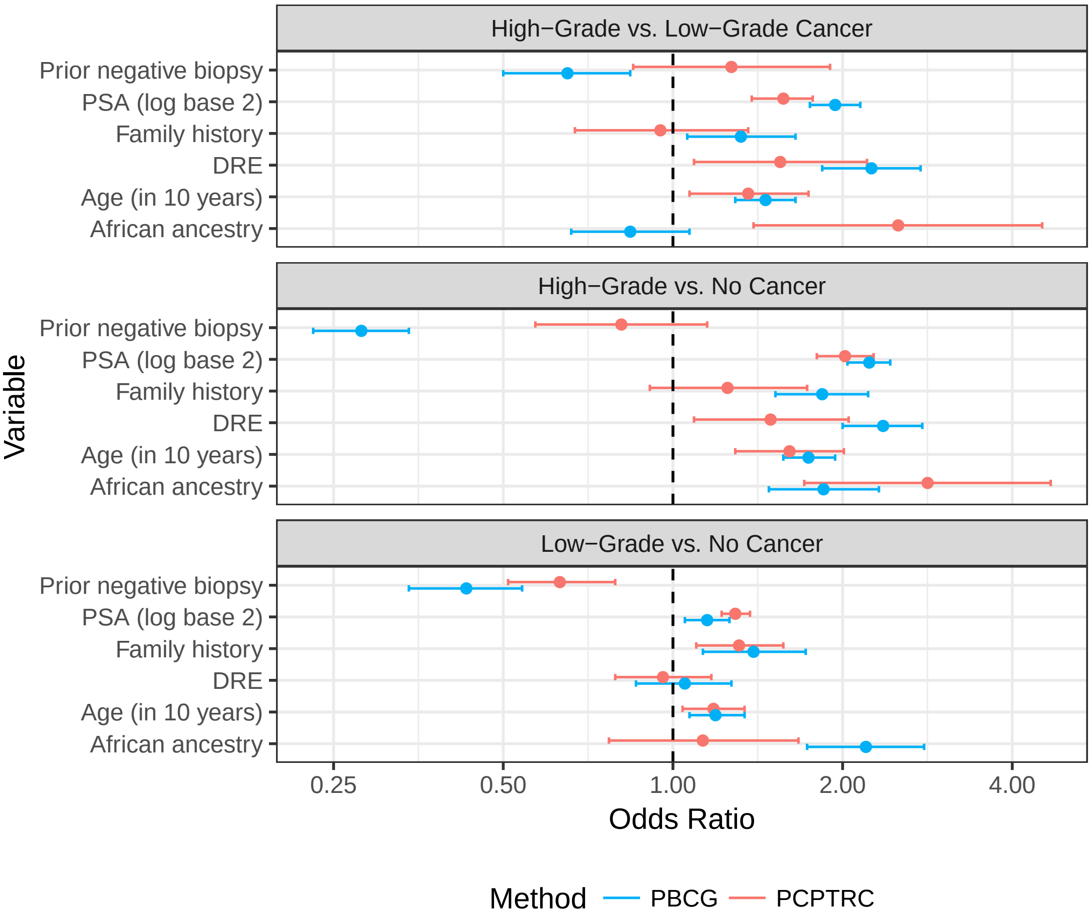


Figure 25 : Comparison between odds ratios of PCPTRC and PBCG for multinomial logistic regression models with no missing data (based on Ankerst, Straubinger, et al. 2018). The PBCG model is built on all eight North American cohorts pooled together. Odds ratios are given for endpoints reported to the left versus reference levels to the right of headers, along with 95%-CIs. Sample size for the PCPT is given by n=6664, with proportions of 82.1% no, 14.1% low-grade and 3.8% high-grade cancer, and for the PBCG by n=4286, with proportions of 47.6%, 18.0% and 34.3%, respectively.

Coefficients for the continuous predictors PSA and age are broadly similar across both studies. Family history and DRE have higher coefficients' values for the PBCG, except for the comparison of low-grade and no cancer. The risk factor prior negative biopsy shows great discrepancies between the studies, thereby having higher values for the PCPT. At last, both studies associate African ancestry with higher cancer risk. Whereas the PCPT finds race to distinguish high- and low-grade cancer, but not low-grade and no cancer, the results of the

PBCG suggest the opposite direction.

Overall, the CIs for the PCPTRC are wider for the endpoint high-grade cancer. This is due to the fact that only 3.8% of the underlying patients have a high-grade cancer diagnosis. In comparison, the PBCG model has smaller CIs, as one third of the biopsy results are high-grade disease.

4.3.2 Model validation

In order to compare performance of the PCPTRC with the newly developed risk tool based on data of the PBCG, two levels of model validation are applied. Separate validations on North American cohorts with 5,992 biopsies, and European cohorts with 10,377 biopsies, are performed (Tables A.2 and A.1). Internal-external cross validation is thereby used to validate the PBCG model on the North American cohorts, and external validation for the European data and the PCPTRC (Section 4.2.3). Results are only shown for high-grade cancer versus the two other outcomes combined. Similar results are obtained for overall versus no cancer, which are shown in the appendix, along with single cohort analyses.

Table 2 summarizes AUC values of both methods, separated by validation set. They show that discrimination is better for North American in comparison to European cohorts and that the PBCG model significantly outperforms the PCPTRC. Also for all individual cohorts, the PBCG achieves higher AUC values compared to the PCPTRC (Figure 26). CIs for all cohorts, except Hamburg, are overlapping due to their smaller sample sizes. The overall inferior performance of the European cohorts is mainly driven by Hamburg, which comprises 75.9% of the European data. With an AUC of 70.3% (CI: 69.1-71.5%) for the PBCG model, it has the worst performance across all cohorts, and with 67.4% (CI: 66.2-68.6%) for the PCPTRC, Hamburg is underperformed only by SanJuanVA with 66.1% (CI: 61.2-71.0%).

	AUC (CI) PCPTRC	AUC (CI) PBCG	P-value
North American cohorts (n=5,992)	72.3% (70.9-73.7%)	75.5% (74.2-76.8%)	<0.0001
European cohorts (n=10,377)	69.7% (68.7-70.8%)	72.9% (71.8-73.9%)	<0.0001
P-value	0.0037	0.0019	

Table 2 : AUC values for high-grade versus low-grade and no cancer with corresponding 95% Delong CI for the PCPTRC and PBCG models. P-values by the Delong test for two correlated ROC-curves are used to compare the PCPTRC and PBCG models (bottom row), and for two uncorrelated ROC-curves to compare the North American and European cohorts (last column). Risk predictions for the North American cohorts by the PBCG model are calculated by leave-one-cohort-out cross validation.

The HLS is also consistently lower for the PBCG in all test cohorts, suggesting better calibration for this model (Figure 26). Calibration curves for the North American and European cohorts, shown in Figure 27, further confirm this result. The PBCG model is well calibrated in the clinical relevant range, only slightly underestimating risk for thresholds greater than 10%.

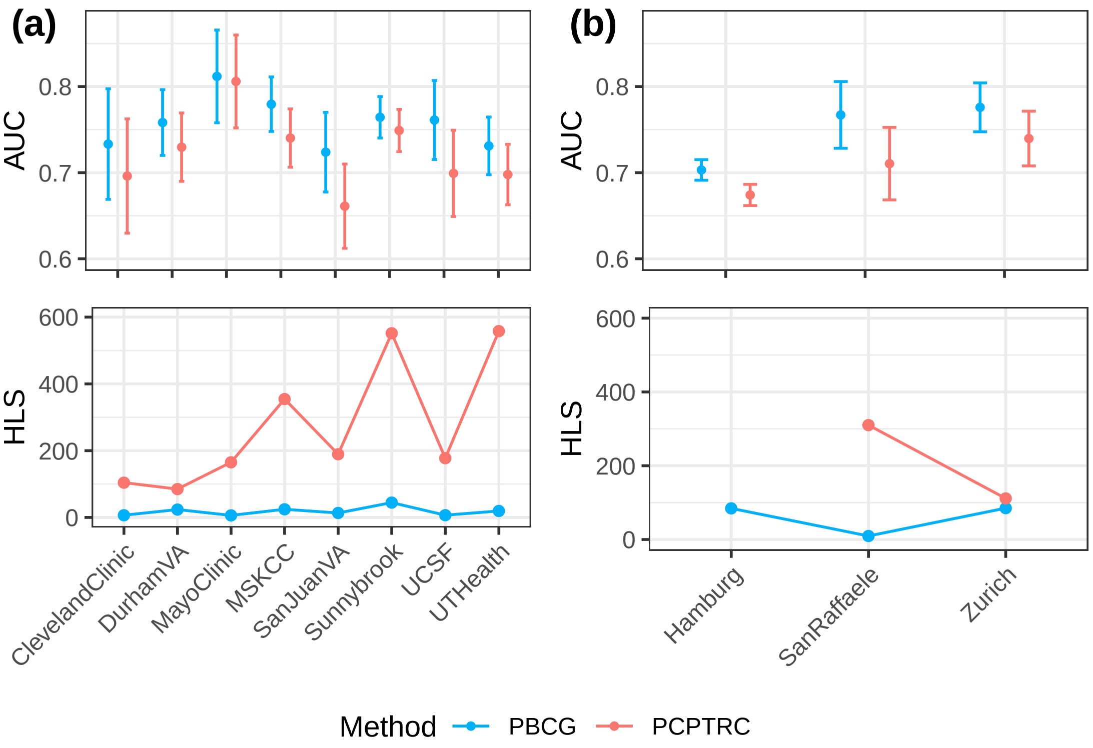


Figure 26 : AUC and HLS values for high-grade cancer of PBCG and PCPT by site. (a) Results of the internal cross validation of the North American cohorts, (b) external validation on the European sites. The HLS for the PCPTRC applied to the Hamburg cohort is neglected as it exceeds 4,000. For AUC higher values are better, while for HLS lower values are preferred. Sample sizes are given by 299 for ClevelandClinic, 669 for DurhamVA, 323 for MayoClinic, 1,010 for MSKCC, 550 for SanJuanVA, 1,721 for Sunnybrook, 521 for UCSF, 899 for UTHealth, 7,877 for Hamburg, 637 for SanRaffaele and 1,863 for Zurich.

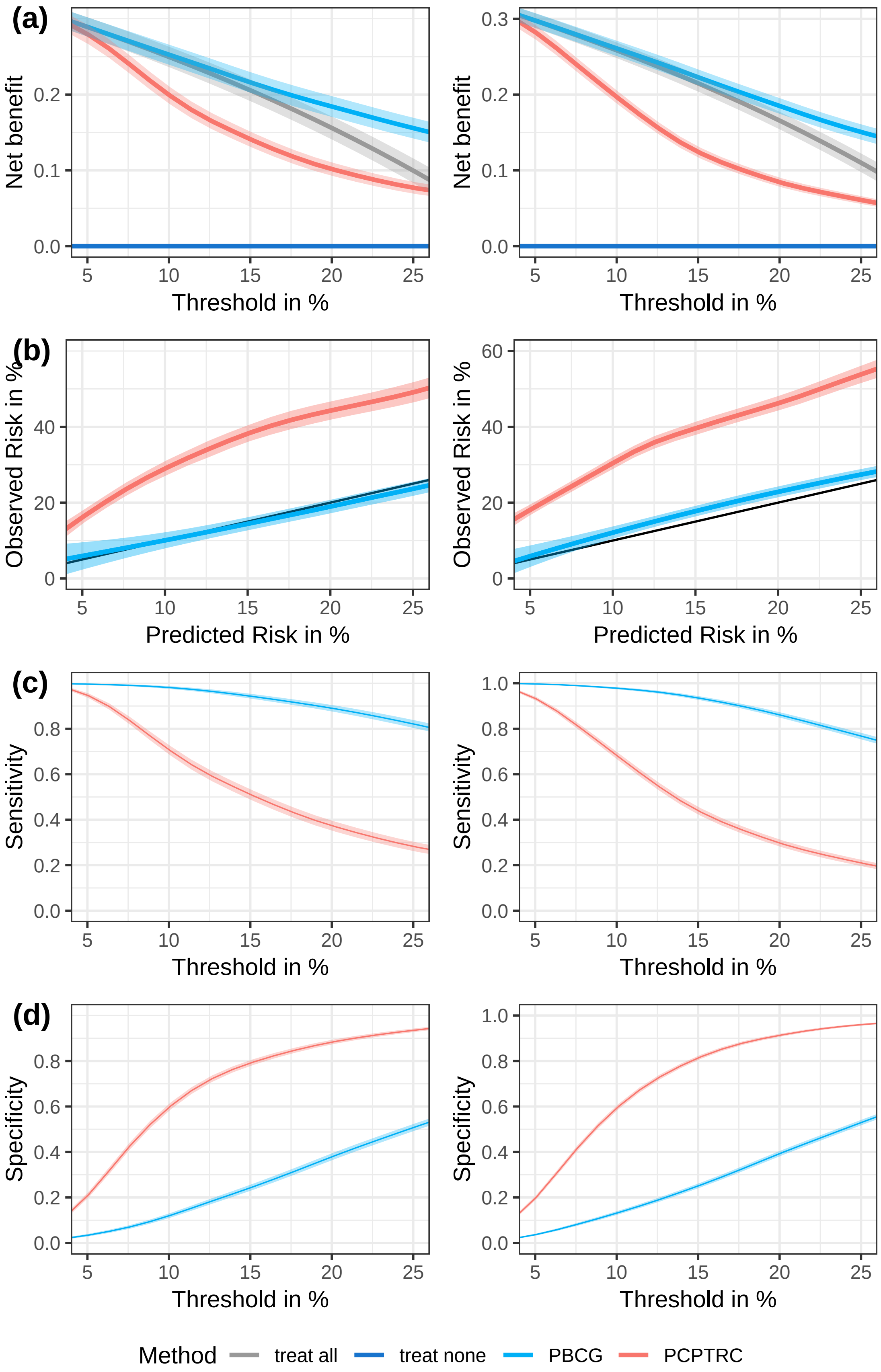


Figure 27 : (a) Net benefit, (b) calibration, (c) sensitivity, and (d) specificity curves for high-grade cancer comparing the PBCG and PCPT models. Results of the internal cross validation of the North American cohorts (left) and external validation with the European sites (right). Strategies of referring all men or none to biopsy are provided in (a) for comparison, pointwise 95%-CIs are shown with shading and black lines in (b) show where predicted risks equal observed risks.

In comparison, the PCPTRC underestimates risk up to about 20% across the whole range of 5-25% in both data sets.

On both data sets, clinical net benefit of the PBCG model exceeds the strategy of biopsying all patients for thresholds of around 10% and higher, with non-overlapping CIs for thresholds higher than about 20% (Figure 27a). Both the PBCG and the strategy of biopsying all patients have superior net benefit to the PCPTRC for all clinically relevant thresholds. Similarly, for all cohorts individually, the PBCG performs at least as well as biopsying all patients in terms of net benefit, even though corresponding CIs often overlap due to smaller sample sizes. Furthermore, the PCPTRC typically has less net benefit than the PBCG and the strategy of biopsying all patients, whereby CIs for the PCPTRC and PBCG methods do not overlap for most cohorts. Zurich is the exception with the highest net benefit for the PCPTRC for thresholds above 20%. Zurich is furthermore the only cohort for which the strategy of biopsying none outperforms another method in terms of net benefit in the range of clinically relevant thresholds. In this case the strategy of biopsying all has a negative net benefit for thresholds of around 17% and higher.

Whereas the PBCG model shows higher sensitivity, the PCPTRC outperforms in terms of specificity for all thresholds between 5 and 25% (Figure 27c, d). This was expected as the model based on the PBCG predicts higher risks than the PCPTRC. CIs of corresponding curves do not overlap and show similar trends for all individual cohorts (Figures A.12 and A.13).

4.4 Discussion

The considered cohorts comprise diverse populations, as described in Section 1.2. Discrepancies in risk tool operating characteristics therefore relate to the difference in cohort designs and inclusion criteria, as well as to changes in clinical practice. Resulting superior performance on the contemporary PBCG data of the newly developed risk tool thereby suggests use of the contemporary risk calculator in modern clinical practice.

The PCPT was a screening study that required a PSA less than or equal to 4 ng/ml and a normal DRE to enter, and then had a mandatory end-of-study biopsy after seven years of annual screening. The resulting risk calculator is therefore based on a predominantly healthy population. In contrast the PBCG considers patients who have undergone biopsy after clinical referral. Consequently, the PBCG model is tailored for men for whom urologists strongly consider a biopsy. In addition, changes in clinical practice have been present as discussed in Section 4.1. The number of cores taken in a biopsy has risen from six to primarily twelve and grading of prostate cancer has shifted towards high-grade assessments. These developments have led to a higher proportion of high-grade cancer in the PBCG study.

The influence of individual risk factors differed between the PCPTRC and PBCG. Odds ratios for prior negative biopsy were lower in the PBCG risk tool, especially for the comparison of high-grade versus no cancer and always less than one, indicating reduction of risk for a patient with a prior negative biopsy. Whereas a prior negative biopsy was always significant for the PBCG model, the PCPTRC found a significant influence only for low-grade versus no cancer, where the odds ratio was also less than one. We assume the number of cores used in the corresponding prior biopsies to be higher for the PBCG cohorts than the PCPT cohorts, which predominantly used 6 cores. An increasing number of biopsy cores relates with a higher prostate cancer detection rate (Ankerst, Till, Boeck, Goodman, Tangen, and I. M. Thompson 2013, Ankerst, Till, Boeck, Goodman, Tangen, Feng, et al. 2013, Babayan and Katz 2016, Bjurlin et al. 2014). As a result, a prior negative assessment could reduce risk more for contemporary data, since it would be more accurate for detecting no disease.

PSA had a significant effect throughout all comparisons and for both methods. The odds ratio for the PBCG model was lower than the PCPTRC for the category low-grade versus no cancer and higher for the two other categories, especially for high- versus low-grade cancer. In 2012 more selective than purely PSA-based biopsy recommendations have been adopted (Moyer 2012). This might have led to the increased association of PSA values with high-grade cancer for the PBCG, which primarily comprises biopsies after this date. Furthermore the PCPT required a PSA value of less than or equal to 4 ng/ml to enter the study, and excluded patients exceeding a PSA value of ten, which impacts the estimates of odds ratios compared to the PBCG, which has PSA values ranging from 0.03 to 7275.09 ng/ml. The PCPT comprises 13.7% of patients with a PSA exceeding four, compared to 83.0% for the PBCG (Tables A.2 and A.1).

Neither method found DRE to be a significant predictor for comparing low-grade and no cancer. In the two other categories the PBCG risk tool estimated a higher association between positive DRE results and high-grade cancer, which might be due to a change in practice standards. Contemporary DREs are more likely to be performed by urologists specialized on prostate cancer instead of general urologists, resulting in more accurate assessments and corresponding higher odds ratios.

The variable family history showed similar significant odds ratios for low-grade versus no cancer between both methods. For the pairs high- versus low-grade and high-grade versus no cancer, family history was not significant in the PCPTRC, but positively associated and significant in the PBCG. Family history is recorded binary, yes if father, brother or son had prostate cancer and no otherwise, and hence does not distinguish aggressiveness of prostate cancer in the first degree relative. The PCPT required documented family history as part of the protocol, but no such routine reporting was implemented across the diverse PBCG sites. PCPT participants thus may have been more likely to report less aggressive prostate cancer in a relative compared to the PBCG, where patients may have only reported aggressive cancers or nothing at all. This might explain the higher odds ratios for the PBCG and advocates the inclusion of cancer aggressiveness in family members to refine the influence of this risk factor.

For instance, Grill, Fallah, Leach, I. M. Thompson, Freedland, et al. 2015 discuss an extension of the binary recording of family history in the PCPTRC by incorporating the number, age and family degree of the relatives in question, and the PBCG already collects these extended family history data in a fraction of sites (Chapter 6).

At last, the coefficients for African ancestry showed great discrepancies between both models. Whereas the odds ratios for comparing high-grade and no cancer were significant for each method, the PCPTRC found race to distinguish between the two cancer groups, but not between low-grade and no cancer. For the PBCG model it was the other way around, implying that race is an important risk factor to predict cancer, but less suitable to distinguish between low- and high-grade disease. The wide CIs for African ancestry of the PCPTRC reflect the small proportion of patients with African ancestry in the underlying data.

5 Comparison of standard logistic regression to more flexible machine learning approaches

Over the last decades the use of machine learning methods has risen in biomedical science and Shariat et al. 2009 summarize various applications in the field of prostate cancer prediction tools (Jensen and Bateman 2011). In this chapter we investigate the performance of three machine learning approaches, namely random forests (RFs), k-nearest neighbor (KNN) methods and artificial neural networks (ANNs). We apply these concepts for risk prediction of prostate cancer biopsy results and compare them with the logistic regression implemented in Chapter 3. For a fair comparison and to address the third aim of this thesis, we merely consider the six standard risk factors used for the logistic regression. Recognizing that a strength of machine learning is in high number of covariates, this small set of predictors furthermore enables a better understanding of the individual steps underlying the machine learning approaches. With this we anticipate to explore whether "off-the-shelf"-methods are generally suitable to predict the risk of high-grade cancer outcome in prostate biopsy. Subsequently we discuss their application for a larger data set as we identify the inclusion of additional predictors as a potential focus of further research in the PBCG (Chapter 6).

5.1 Research in context

For parametric logistic regression it is necessary to completely and correctly specify all important variables, interactions and higher order terms to avoid problems of model misspecification and potentially biased probability estimation (Kruppa, Y. Liu, et al. 2014). Machine learning methods, on the other hand, are often nonparametric and therefore more robust in terms of model specification. Kruppa, Ziegler, et al. 2012 give an overview of popular approaches, along with recommended literature.

The development of most machine learning methods has focused on classification tasks. However, instead of asking "Is the patient more likely to have high-grade prostate cancer than not?", we are more interested in the question "What is the probability the patient has high-grade prostate cancer?", as it provides additional information. In particular, a patient most probably appreciates the difference in risk prediction of 90% compared to 55%, instead of a simple "Yes, you are more likely to have cancer than not". Due to their focus on classification, it is of particular importance to assess considered machine learning approaches with respect to their performance in probability estimation. J. D. Malley, Kruppa, et al. 2012 reformulate probability estimation problems as nonparametric regressions applied to binary outcomes. This connection enables the derivation of properties already known for the latter. The authors consider the targeted estimate of the conditional probability $P(Y = 1|X = x)$, with the binary outcome $Y \in 0, 1$ and predictors $X \in \mathbb{R}^d$, and express it by $P(Y = 1|X = x) = E(Y|X = x) = f(x)$, a regression estimation problem. As a result, statistical learning machines, which perform well for nonparametric regression, are also suitable for probability risk estimation. In particular, consistency has been proven for several statistical learning approaches, and hence can be deduced for probability predictions (Györfi et al. 2002).

An estimate is consistent if it converges to the estimated quantity as the sample size increases. Györfi et al. 2002 thereby distinguishes between weak and strong, as well as between universal and non-universal, consistency. However, Devroye et al. 1996 state that weak and strong consistency are equivalent for most well-behaved rules. Even though machine learning techniques have evolved over the past decades, thereby potentially increasing the advantage of distinction, many researchers refer to weak consistency only or do not differentiate between weak and strong consistency at all (Kruppa, Ziegler, et al. 2012, J. D. Malley, Kruppa, et al. 2012, Biau et al. 2008, Mease and Wyner 2008, Breiman 2004). Similarly, in this thesis we merely focus on weak consistency, in the following referred to as consistency. Let $(X,Y),(X_i,Y_i),\ i = 1,...,n$ i.i.d. random variables, $D_n = \{(X_1,Y_1),...,(X_n,Y_n)\}$ the training set, and $\hat{f}_n(x,D_n)$ an estimate of the regression function $f(x) = E(Y|X = x)$. Consider the random variable $E_n = \int(\hat{f}_n(x,D_n) - f(x))^2\mu(dx)$, with μ the distribution of X and $\hat{f}_n(x,D_n)$ depending on the data D_n. Györfi et al. 2002 calls the sequence $\{\hat{f}_n\}$ consistent for a specified distribution of (X,Y) if $E\left(lim_{n\to\infty}E_n\right) = 0$, with E the expectation with respect to the training set D_n. A sequence is universally consistent if it is consistent for all distributions of (X,Y). This generalization is of particular relevance for the assessment of machine learning methods, as a major advantage of these is that no prior information about the distribution of (X,Y) is required.

Whereas the non-parametric approaches investigated in this chapter largely show universal consistency under general conditions, logistic regressions are only consistent for fully and correctly specified models (Kruppa, Y. Liu, et al. 2014). Stone 1977 proves universal consistency of the KNN approach for $k_n \to \infty$ and $\frac{k_n}{n} \to 0$, with n the total sample size and k_n the number of considered neighbors (Györfi et al. 2002, Devroye et al. 1996). The modification of KNN with bootstrap aggregation also shows consistency under general conditions (J. D. Malley, Kruppa, et al. 2012, Kruppa, Ziegler, et al. 2012). Biau et al. 2008, Breiman 2004 and Meinshausen 2006 prove consistency for certain RFs. Driving forces thereby include a connection to adaptive nearest neighbor methods reported by Lin and Jeon 2006. However, as pointed out by Breiman 2004, RFs are difficult to analyze and are therefore statistically not fully understood. In addition to universal consistency of ANN classifiers, Györfi et al. 2002 also discuss regression ANNs with one hidden layer (Devroye et al. 1996, Farago and Lugosi 1993). They are universally consistent as well, given some constraints on the weights and free parameters of the ANN. For k_n the number of hidden units corresponding to the sample size n, and w_i the weights between the hidden layer and the output, the discussed theorem in Györfi et al. 2002 depends upon $k_n \to \infty$, $\sum_{i=0}^{k_n} |w_i| \leq \beta_n$, $\beta_n \to \infty$, and $\frac{k_n\beta_n^4 log(k_n\beta_n^2)}{n} \to 0$. Even though universal consistency of the general technique was required in the following application, final consistency properties are not further elaborated.

5.2 Methods

Let $(y_1, x_1), ..., (y_n, x_n)$ be a sample of i.i.d. random pairs, with $y_i = 1$ for the biopsy result of high-grade cancer, $y_i = 0$ for a negative or low-grade cancer assessment, and $x_i \in \mathbb{R}^p$ the vector of covariates for patient i. We will not require the random variables to follow a specific distribution.

To compare the different machine learning approaches applied to the PBCG data, we first discuss and analyze each method independently. We thereby select tuning parameters for the individual methods based on the discrimination measure AUC. Subsequently we compare the resulting models with each other and in addition with a standard logistic regression, using the validation metrics of Chapter 2. For all evaluations we implement leave-one-cohort-out cross validations.

For the analysis of this chapter we enable the ten PBCG cohorts ClevelandClinic, DurhamVA, MayoClinic, MSKCC, SanJuanVA, SanRaffaele, Sunnybrook, UCSF, UTHealth and Zurich, resulting in $n = 8492$ observations (Tables A.1 and A.2). As previously discussed, we merely include the six standard risk factors family history, prior biopsy, DRE, African ancestry, age and PSA with median imputations for missing values.

5.2.1 Random forest

Breiman 2001 introduces the definition of RFs as a collection of tree-structured classifiers based on i.i.d. random vectors. Even though this definition comprises some previously developed methods, the common usage of the expression RF refers to the algorithm developed by Breiman 2001. The initial introduction focuses on classification, but modifications for regression, as used in this thesis, are straightforward (Hastie et al. 2009, J. D. Malley, Kruppa, et al. 2012). However we describe both approaches as they might get easily confused for risk predictions of binary outcomes. In particular the implementation of the enabled function in R differs merely in the assigned data type of the response variable.

The regression RF is implemented with the R package randomForest, which is based on Breiman 2001, and summarized in Algorithm 12. Thereby the values of the flexible parameters $mtry$, B and n_{node_min} are chosen to match the following detailed description.

A single regression or classification tree of a RF is grown by successively splitting the data into two subsets. Each split is performed according to one of the covariates and called decision node. A schematic representation of an exemplary regression tree is given in Figure 28. At the starting node a, called root, all observations are considered. The data is now divided into two subsets, named daughter nodes, according to the covariate which minimizes a given impurity function (IF). In the example given in Figure 28, all observations with a log base two transformed PSA value, named lpsa2, smaller than 4.1 are included in the left side of the root

Algorithm 12 Regression and classification random forest (RF)

```
procedure RF(training data, $x$)
    $n \leftarrow$ size of training data
    $p \leftarrow$ number of covariates
    $mtry \leftarrow 2 = max(1, \frac{p}{3})$ rounded to next smallest integer
    $B \leftarrow 500$ number of required bootstrap samples
    $n_{node\_min} \leftarrow 150$
    for i in 1 to B do
        $b_i \leftarrow$ bootstrap sample of size $n$ drawn with replacement from training data
        $T_i \leftarrow$ regression tree grown on $b_i$ by recursively repeating the following steps for each terminal node.
        loop
            Randomly select $mtry$ covariates
            For each selected covariate choose best split-point by minimizing the MSE
            Choose best covariate, with respective split-point, by minimizing the MSE
            Consider to split terminal node into two daughter nodes according to chosen covariate and split-point
            $STOP$ if the node size of a daughter node falls below $n_{node\_min}$
            Set the two daughter nodes to be new terminal nodes
        $p_{i,j_i} \leftarrow$ proportion of high-grade cancer cases in each terminal node $j_i$
        $c_{i,j_i} \leftarrow \begin{cases} 1 & p_{i,j_i} \geq 0.5 \\ 0 & p_{i,j_i} < 0.5 \end{cases}$ dominant class in each terminal node $j_i$
    $risk_{regression} \leftarrow \frac{1}{B}\sum_{i=1}^{B} p_{i,j_i}$ with $j_i$ the resulting terminal node for $x$ in tree $T_i$
    $risk_{classification} \leftarrow \frac{1}{B}\sum_{i=1}^{B} c_{i,j_i}$ with $j_i$ the resulting terminal node for $x$ in tree $T_i$
    return $risk_{regression}$, $risk_{classification}$
```

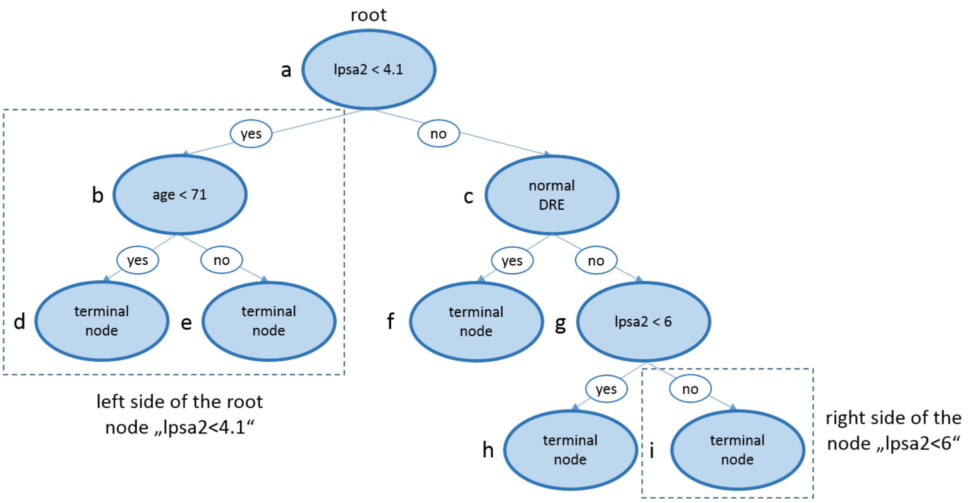


Figure 28 : Schematic representation of a single tree. Split points are chosen arbitrarily and not based on the data.

node and the remaining ones get allocated to the right side. The nodes b and c are called temporary terminal nodes after the first splitting level. For each temporary terminal node a new splitting criteria is chosen, which minimizes the IF based on the data related with this node. For instance, all observations with a lpsa2 smaller than 4.1 are included for node b, resulting in the optimal splitting criteria "age<71". This step is repeated for each temporary terminal node until the chosen split would result into a daughter node with a smaller number of records than an initially determined value n_{node_min}. In this case the node does not get partitioned and is called terminal node. Whereas the partition according to a binary covariate, such as DRE in node c into normal and abnormal, is performed straightforward, an optimal split point has to be selected for continuous variables. This is also performed by minimizing the considered IF. Note that a continuous predictor, such as lpsa2 in the example, can get chosen several times with different splitting thresholds.

In order to obtain a prediction for a new observation, this observation is dropped down the tree and the proportion of high-grade cancer cases in the respective terminal node is considered. For a probability prediction of a regression tree this proportion is reported directly, whereas in a classification tree the class with the highest proportion is returned. In the example of Figure 28, a patient with lpsa2 smaller than 4.1 and an age greater than or equal to 71 will get associated to the terminal node e.

For a specific node, the IF gets minimized to find optimal split points for the continuous covariates and to subsequently choose among the available predictors with their respective split points. For binary outcomes in a classification tree, the IF is given by

$$\begin{aligned} IF &= n_{left}\Psi\left(\frac{1}{n_{left}}\sum_{left} y_i, 1-\frac{1}{n_{left}}\sum_{left} y_i\right) + n_{right}\Psi\left(\frac{1}{n_{right}}\sum_{right} y_i, 1-\frac{1}{n_{right}}\sum_{right} y_i\right) \\ &= n_{left}\Psi\left(\bar{y}_{left}, 1-\bar{y}_{left}\right) + n_{right}\Psi\left(\bar{y}_{right}, 1-\bar{y}_{right}\right), \end{aligned} \tag{5.1}$$

with $\sum_{left}$ the summation over the resulting left side, n_{left} the number of observations and $\bar{y}_{left} = \frac{1}{n_{left}}\sum_{left} y_i$ the mean value of all observations allocated to the left side of a node, and $\sum_{right}$, n_{right} and $\bar{y}_{right}$, defined analogously for the resulting right side of a node (Devroye et al. 1996). Thereby, Ψ has to be defined as a nonnegative function with the properties

1. $\Psi(0.5, 0.5) \geq \psi(\bar{y}, 1-\bar{y})$ for any $\bar{y} \in [0, 1]$,
2. $\Psi(0, 1) = \Psi(1, 0) = 0$ and
3. $\Psi(\bar{y}, 1-\bar{y})$ increasing for $\bar{y} \in [0, 0.5]$ and decreasing for $\bar{y} \in [0.5, 1]$.

A commonly used function for Ψ is given by the Gini index (Devroye et al. 1996, Hastie et al.

2009, Kuhn and Johnson 2013):

$$\Psi(\bar{y}, 1-\bar{y}) = 2\bar{y}(1-\bar{y}). \tag{5.2}$$

For a regression tree the MSE is minimized. Even though regression trees are primarily introduced for continuous outcomes, the MSE can also be used for binary variables. Minimizing the MSE can thereby be considered equivalent to the previously introduced IF combined with the Gini index, termed IF_{Gini}:

$$\begin{aligned}
MSE =& \sum_{left}(y_i - \bar{y}_{left})^2 + \sum_{right}(y_i - \bar{y}_{right})^2 \\
=& \sum_{left}(y_i^2 - 2y_i\bar{y}_{left} + \bar{y}_{left}^2) + \sum_{right}(y_i^2 - 2\bar{y}_{right} + \bar{y}_{right}^2) \\
=& \sum_{left} y_i - \sum_{left} 2y_i\bar{y}_{left} + \sum_{left} \bar{y}_{left}^2 + \sum_{right} y_i - \sum_{right} 2y_i\bar{y}_{right} + \sum_{right} \bar{y}_{right}^2 \\
=& n_{left}\bar{y}_{left} - 2n_{left}\bar{y}_{left}^2 + n_{left}\bar{y}_{left}^2 + n_{right}\bar{y}_{right} - 2n_{right}\bar{y}_{right}^2 + n_{right}\bar{y}_{right}^2 \\
=& n_{left}\bar{y}_{left} - n_{left}\bar{y}_{left}^2 + n_{right}\bar{y}_{right} - n_{right}\bar{y}_{right}^2 \\
=& n_{left}\bar{y}_{left}\,(1-\bar{y}_{left}) + n_{right}\bar{y}_{right}\,(1-\bar{y}_{right}) = \frac{1}{2} IF_{Gini}.
\end{aligned} \tag{5.3}$$

A RF is given by several regression or classification trees aggregated into an overall probability prediction model. To implement the aggregation, B bootstrap samples of the same size as the original training data are drawn with replacement and for each bootstrap replicate an individual tree is grown. However, instead of minimizing the IF across all possible covariates, at each node a different random subset of size $mtry$ of the available covariates is considered. A probability prediction for a new observation is obtained by averaging the output of the resulting terminal nodes in the individual trees. In case of a regression RF these outputs are given by the proportions of high-grade cancer in the respective nodes, therefore values between zero and one. For a classification RF, however, the individual outputs are either zero or one.

For both types of RFs, optimal values of B, $mtry$ and n_{node_min} are treated as tuning parameters, as they depend on the data at hand. It has been shown that RFs can not be overfit by a large value of B, however, their computational intensity increases with large B (Breiman 2001).

The performance of RFs substantially rests upon its random selection of data and predictors. Using an independent bootstrap sample to grow each tree and subsequently aggregating the results, called bagging, leads to improved prediction due to reduced variance (Hastie et al. 2009, Genuer 2012). Furthermore, the random choice of covariates brings advantages, as the usefulness of a single predictor depends on all other predictors, interactions might be present. By considering only a random subset of input variables for the split at every node and ultimately combining all resulting trees, we consider an average over various contexts of a predictor (J. D. Malley, K. G. Malley, et al. 2011). A random selection of covariates furthermore reduces correlations between single trees and with this further increases the benefit of the initial bagging.

To determine optimal parameters for the analyzed PBCG data, Figure 29 shows AUC values for RFs with different tuning parameters and a constant number of bootstrap samples of $B = 1000$. The considered RFs achieve a maximal AUC value of 76.2% for $mtry = 2$,

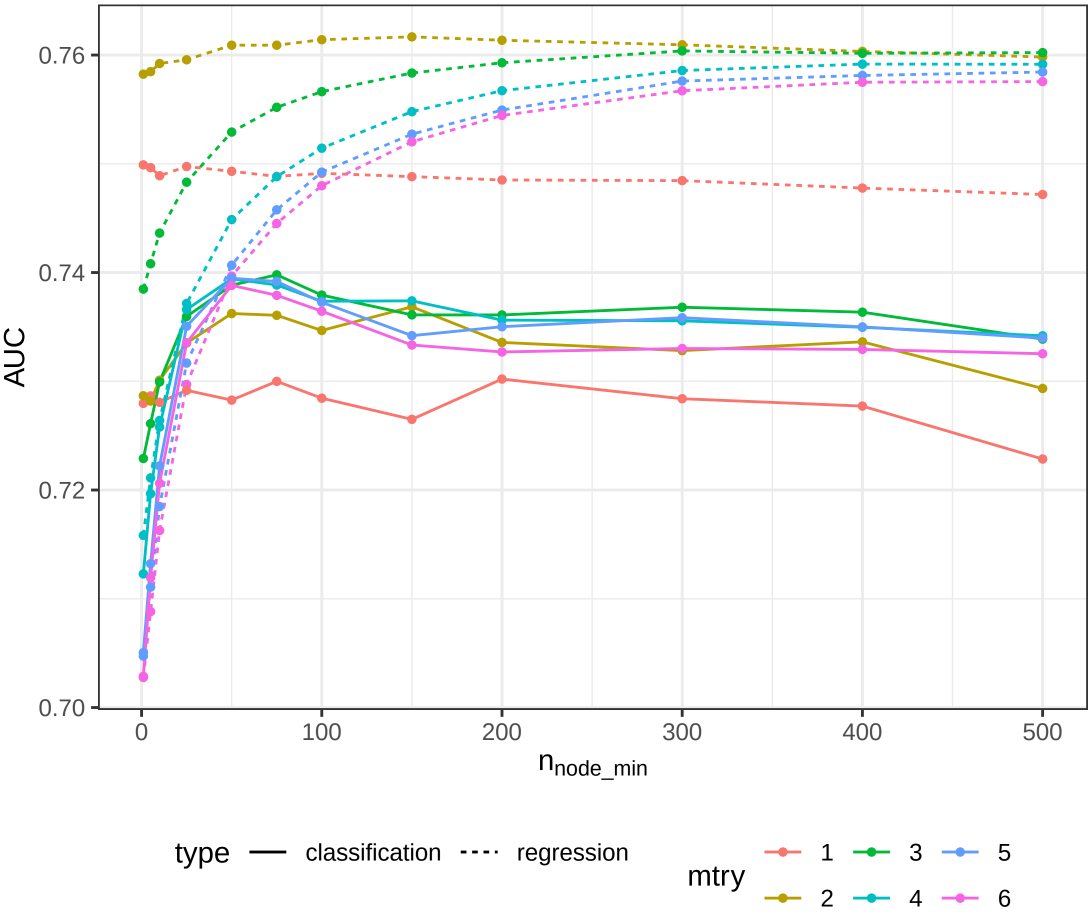


Figure 29 : AUC values of leave-one-cohort-out cross validation for RFs with different minimal node sizes n_{node_min} and different number of considered covariates at each node, $mtry$. Number of bootstrap samples kept constant with $B = 1000$.

$n_{node\_min} = 150$ and underlying regression instead of classification trees. The default value in the statistical software R for $mtry$ in regression RFs is the number of possible predictors divided by three, $mtry = \frac{p}{3} = \frac{6}{3} = 2$, so coinciding with the optimal choice. Thereby the AUC depends only slightly on the minimal size of terminal nodes. Using $mtry = 1$, equivalent to randomly choosing one predictor at each node, results in AUC values between 74.7% and 75.0%, and between 72.3% and 73.0% for regression and classification RFs, respectively. The performance is therefore robust to the choice of n_{node_min} as well. This is reasonable, as four of the six proposed predictors are binary. Thus, by randomly choosing one predictor at each node, the probability of getting the same binary predictor twice early in one branch is high. In this case the branch will terminate, independent of the minimal node size n_{node_min}, as one of the daughter cells would contain zero observations. Using higher values for $mtry$, the performance notably improves with increasing minimal node sizes and stagnates for regression RF for values greater than about 400. Classification RFs show best results for values between 50 and 100.

Overall, classification RFs outperform the ones based on regression trees only for $n_{node\_min} < 50$ and $mtry \in \{5, 6\}$. This is reasonable as for small terminal nodes a probability prediction based on the observed proportions of high-grade cancer becomes imprecise. In the extreme case of only one patient per terminal node regression and classification RFs even coincide. For smaller values of $mtry$, however, the chance to result in small terminal nodes is vanishing, the benefits of regression trees therefore outweigh classification RFs even for small n_{node_min}.

Furthermore, increasing the number of bootstrap samples for the optimal choices of $mtry$, n_{node_min} and type does not further improve the AUC. Since a lower amount of samples is less computational intensive, we stick to $B = 1000$.

In Figures 30 and 31 the first three splitting levels of trees grown on the PBCG data are shown, demonstrating the effect of randomization in the RF algorithm. The first set of trees, given in Figure 30, is grown on four bootstrap samples with the tuning parameter $mtry$ set to

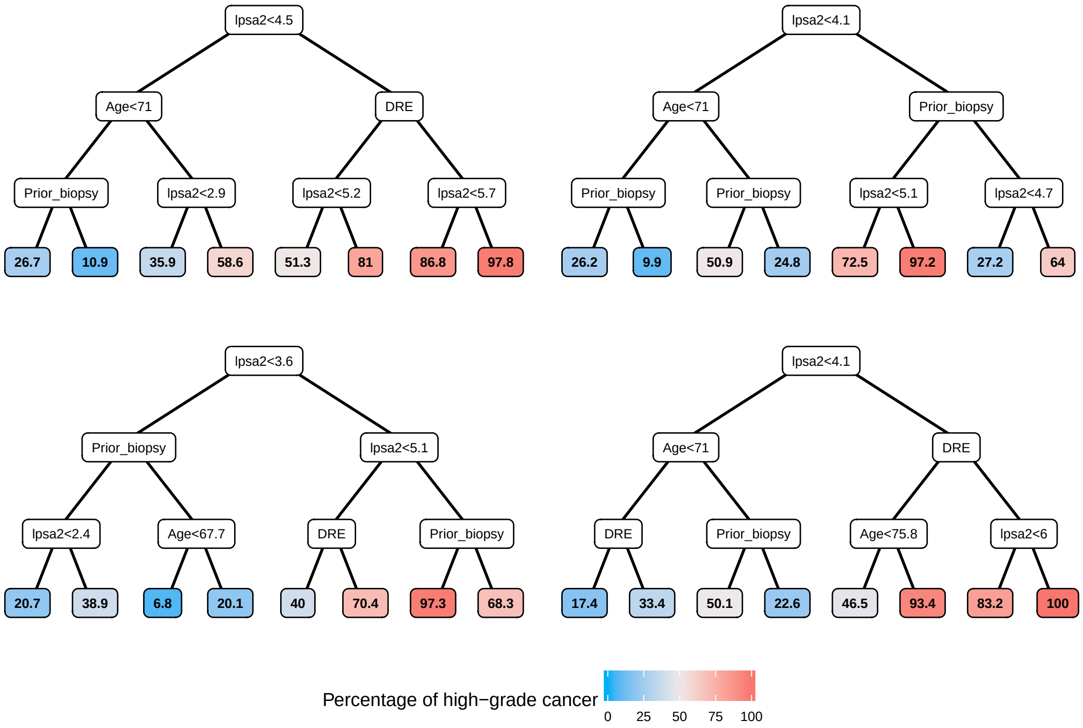


Figure 30 : Trees from four bootstrap samples in a RF with $mtry = 6$. Only the first three splitting levels are shown, along with resulting probabilities of high-grade cancer in temporary terminal nodes.

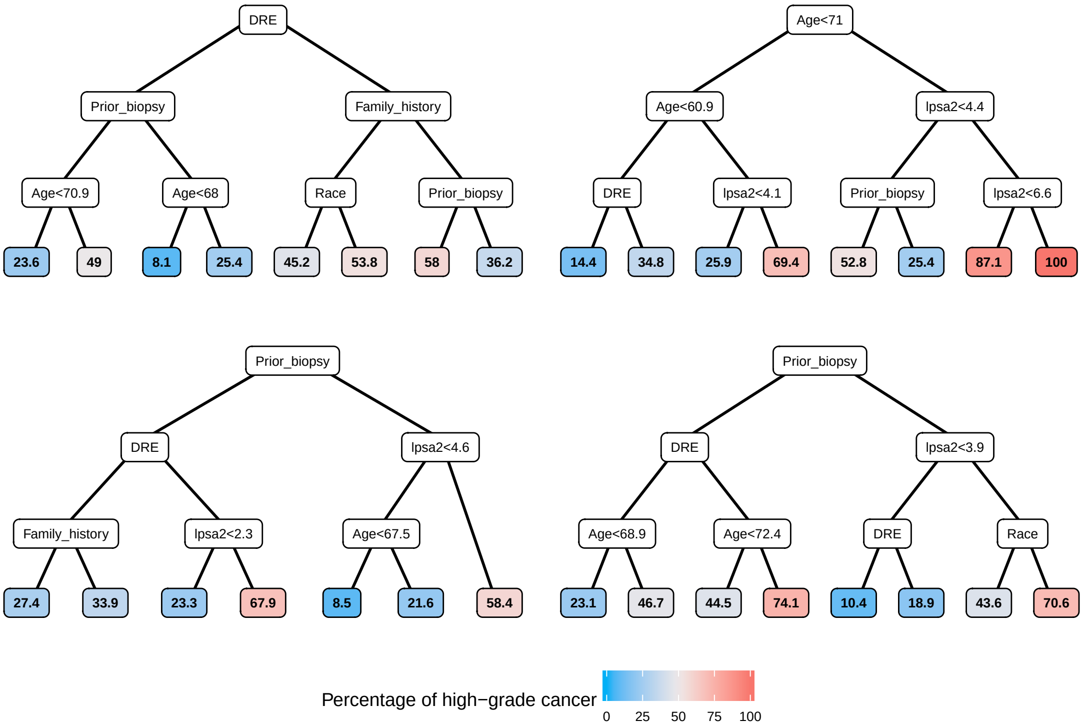


Figure 31 : Trees from four bootstrap samples in a RF with $mtry = 2$. Only the first three splitting levels are shown, along with resulting probabilities of high-grade cancer in temporary terminal nodes.

six. With this choice all predictors can be chosen at each node, resulting in similar structures for all trees. The start is given by the predictor lpsa2, along with a splitting point close to four. Furthermore, only the predictors lpsa2, age, DRE and prior negative biopsy are used for the first three splits. In contrast, the trees of Figure 31 are grown for $mtry = 2$. Therefore both levels of randomization in a RF are used, the bootstrap replicates and the random choice of covariates at each node. As a result more diverse trees are constructed, in particular all predictors are considered in the first three steps and the trees are started with diverse covariates.

Also the size of the resulting trees can vary considerably in a RF model. Supplementary Figure A.20 depicts the smallest tree of a RF with $mtry = 2$ and $B = 1000$, consisting of six splitting levels and 13 terminal nodes. Whereas the largest tree, depicted in Figure A.21, comprises 108 terminal nodes from 15 splitting levels.

5.2.2 (Bagged) k-nearest neighbors

In the KNN and bagged KNN approaches the prediction for a new observation is based on observations in the training set similar to the new one (Kuhn and Johnson 2013). Summaries, along with tuning parameters chosen according to the following considerations, are given in Algorithm 13 and 14. Implementation of both methods is done with the R package caret, which is based on the package class (Venables and Ripley 2002, Kuhn 2008).

Algorithm 13 K-nearest neighbors (KNN)

```
procedure KNN(training data, $x$)
    $n \leftarrow$ size of training data
    $k \leftarrow 160$
    training data $\leftarrow$ standardized training data
    $x \leftarrow$ standardized $x$
    for $i$ in $1$ to $n$ do
        $d_i \leftarrow$ euclidean distance between $x$ and observation $i$ in training set
    $K \leftarrow$ neighborhood of $x$: subset of $k$ observations from training set with smallest $d_i$,
including ties for $k$th observation
    $|K| \leftarrow$ number of observations in $K$
    $risk \leftarrow \frac{1}{|K|} \sum_K y_i$ proportion of high-grade cancer patients in neighborhood of $x$
    return $risk$
```

Algorithm 14 Bagged k-nearest neighbors (BNN)

```
procedure BNN(training data, $x$)
    $n \leftarrow$ size of training data
    $k \leftarrow 160$
    $B \leftarrow 100$ number of required bootstrap samples
    for i in 1 to B do
        $b_i \leftarrow$ bootstrap sample of size $n$ drawn with replacement from training data
        $p_i \leftarrow$ KNN($b_i$, $x$)
    $risk \leftarrow \frac{1}{B} \sum_{i=1}^{B} p_i$
    return $risk$
```

First the distance between the covariates of a new observation $x = (x_1, ..., x_p) \in \mathbb{R}^p$ and the predictors $x_i = (x_{i,1}, ..., x_{i,p}) \in \mathbb{R}^p,\ i = 1, ..., n$ of each observation in the training set is computed. The most commonly used measure for the distance between x_i of the training set and the new observation x is given by the Euclidean distance d_i (Ripley 1996, Kuhn and Johnson 2013):

$$d_i = \sqrt{\sum_{j=1}^{p} (x_j - x_{i,j})^2}. \tag{5.4}$$

The k observations with smallest distance to the new value are included in the subset K, named the KNN of x. In case of ties for the kth nearest observation, all tied candidates are included in the subset K. The risk prediction $p(x)$ is now given by the average response in the subset K: $p(x) = \frac{1}{|K|} \sum_K y_i$, with $|K|$ the number of observations in K, which might exceed k in case of ties.

As this algorithm fundamentally depends on the distance between observations, the scale of the predictors gets very influential. Covariates on large scales contribute more to the distance calculation and get therefore more weight. To ensure similar contribution of all predictors, we center and scale them beforehand:

$$x_{i,j,standardized} = \frac{x_{i,j} - \bar{x}_{.,j}}{s_{.,j}}, \tag{5.5}$$

$$x_{j,standardized} = \frac{x_j - \bar{x}_{.,j}}{s_{.,j}}, \tag{5.6}$$

with $\bar{x}_{.,j}$ and $s_{.,j}$ the mean and standard deviation of the predictor j in the training set. Equation (5.6) describes the standardization of a new observation x. We perform the same transformation for the binary predictors, resulting in a higher influence of the variables race and family history compared to DRE and prior biopsy. This is due to the small proportion of patients with African ancestry and a family history of prostate cancer.

This weighting of the binary predictors is quite arbitrary. In fact, a weighting based on the influence on high-grade prostate cancer for both, continuous and binary predictors, would be more reasonable. Finding a suitable weighting, however, is not part of the standard KNN approach and also not implemented in the enabled R command.

Similarly, the Euclidean distance measure of Equation (5.4) might not be optimal, even though most commonly used. Hu et al. 2016 compare its performance with KNNs based on the cosine, Minkowsky and chi-square distance function for medical data sets containing merely categorical or numerical predictors, as well as both types combined. Thereby none of the measures performed consistently best over all data sets, suggesting that a selection should be based on the data at hand. A review by Prasath et al. 2017 of 54 distance measures on 28 data sets confirms this observation.

The KNN method can be modified with bootstrap aggregating, introduced by Breiman 1996 as "bagging". For B bootstrap samples the KNN algorithm is implemented and the resulting risk predictions are averaged to get the bagged KNN estimate. As the bagging step in the implementation of RFs, this modification is used to reduce variance in the estimation.

The main tuning parameters of a standard KNN approach with Euclidean distance and without weighting are the size of the neighborhood k, along with the bootstrap parameter B for the bagged KNN. Figure 32 depicts resulting AUC values for KNNs, as well as bagged KNNs with 5, 20 and 100 bootstrap replicates, for several values of the tuning parameter k. For small neighborhoods the performance in terms of AUC improves with the neighborhood size k. For values larger than 50 the effect levels off and only small variations are present. The maximal AUC of 75.5% results within the range of 150-170 for KNN as well as bagged KNN with B=20 and B=100. Further increasing the number of boostrap replicates in this range did not improve the AUC. Due to additional computational effort for larger B, we use $B = 100$ in the following analysis. Whereas the bagged KNN with only 5 bootstrap replicates performs worst for most k, the difference between KNN and bagged KNN with $B = 20$ as well as $B = 100$ decreases with growing k and becomes negligible for values greater than 50. For the further analysis we use $k = 160$ along with $B = 100$.

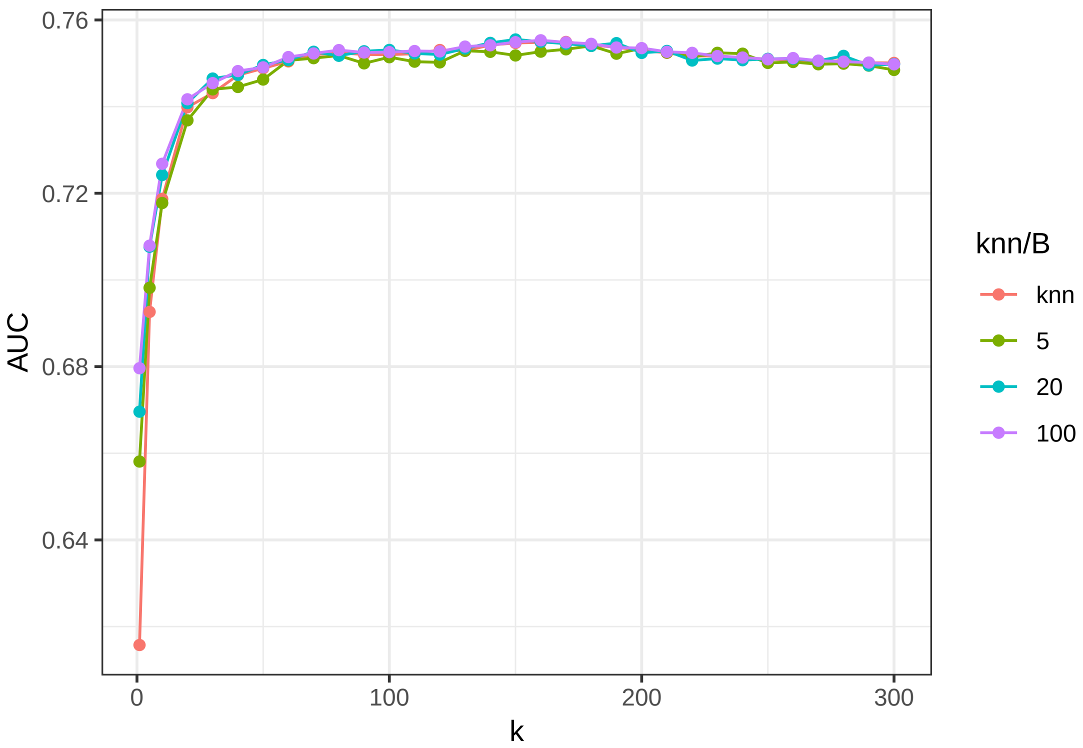


Figure 32 : AUC values of leave-one-cohort-out cross validation for KNNs with different neighborhood sizes k and number of bootstrap replicates B for bagged KNNs.

5.2.3 Artificial neural network

ANNs, often called neural nets, are models inspired by biological neural networks of the human brain. In the last decade the use of ANNs expanded with a broad variety of applications and modifications (Jensen and Bateman 2011, T. Huang et al. 2018). Due to the vast amount of slightly different approaches, it is reasonable to first consider the use of simpler and earlier models. Later developments in particular enable the use of a large amount of input variables, necessary for instance in image or speech recognition. However, the models developed in this thesis merely utilize six covariates. We therefore consider simple networks with only a few hidden layers and neurons as sufficient.

Algorithm 15 summarizes the procedure of the considered ANNs, with the tuning parameters chosen according to the subsequent discussion. In R we perform the implementation with the package neuralnet (Günther and Fritsch 2010).

In the following the detailed mathematical background for an ANN with two hidden layers, as depicted in Figure 33, is described. The p covariates $x_{i,1}, ..., x_{i,p}$ of patient i and a bias term with the value 1 are termed input layer and the output $\hat{y}_i$ is given by the predicted risk of high-grade prostate cancer. The final weights between the layers are estimated with an iterative process. After initially assigning the weights random values, they get updated within every step. The following considerations are given with respect to current weights in order to

Algorithm 15 Artificial Neural Network (ANN)

```
procedure ANN(training data, $x$)
    $n \leftarrow$ size of training data
    $H \leftarrow$ number of hidden layers
    $p^{(h)} \leftarrow$ number of neurons in hidden layer $h = 1, ..., H$
    $\alpha \leftarrow$ learning rate
    $it_{max} \leftarrow$ maximum number of iterations
    $\mu \leftarrow$ lower threshold for partial derivatives
    training data $\leftarrow$ standardized training data
    $x \leftarrow$ standardized $x$
    randomly initialize all weights
    for $i = 1, ..., n$ calculate all hidden neurons and the output
    loop
        calculate all partial derivatives for the error function cross entropy
        $STOP$ if the partial derivatives $< \mu$ or number of iterations $> it_{max}$
        update all weights
        for $i = 1, ..., n$ update all hidden neurons and the output
    $risk \leftarrow$ output for the input $x$ using the current weights
    return $risk$
```

illustrate the updating process.

For patient i, the values of the $p^{(1)}$ neurons of the first layer, $h_{i,k}^{(1)},\ k = 1, ..., p^{(1)}$, are calculated by

$$h_{i,k}^{(1)} = g\left(u_{i,k}^{(1)}\right), \tag{5.7}$$

$$u_{i,k}^{(1)} = w_{0,k}^{(I-1)} + \sum_{j=1}^{p} w_{j,k}^{(I-1)} x_{i,j}, \tag{5.8}$$

with $w_{0,k}^{(I-1)}$ the weight corresponding to the bias term 1 and $w_{j,k}^{(I-1)}$ the weights from the input j to the kth neuron of the first hidden layer. The function $g(.)$ is called activation function and a common choice is the logistic link, given by

$$g(u) = \frac{1}{1 + e^{-u}}. \tag{5.9}$$

The $p^{(2)}$ neurons of the second layer, $h_{i,l}^{(2)},\ l = 1, ..., p^{(2)}$, are calculated analogously:

$$h_{i,l}^{(2)} = g\left(u_{i,l}^{(2)}\right), \tag{5.10}$$

$$u_{i,l}^{(2)} = w_{0,l}^{(1-2)} + \sum_{k=1}^{p^{(1)}} w_{k,l}^{(1-2)} h_{i,k}^{(1)}, \tag{5.11}$$

with $w_{0,l}^{(1-2)}$ and $w_{k,l}^{(1-2)}$ the weights between the first and second layer. The output, in this application a single value indicating the probability of high-grade cancer, is then calculated to

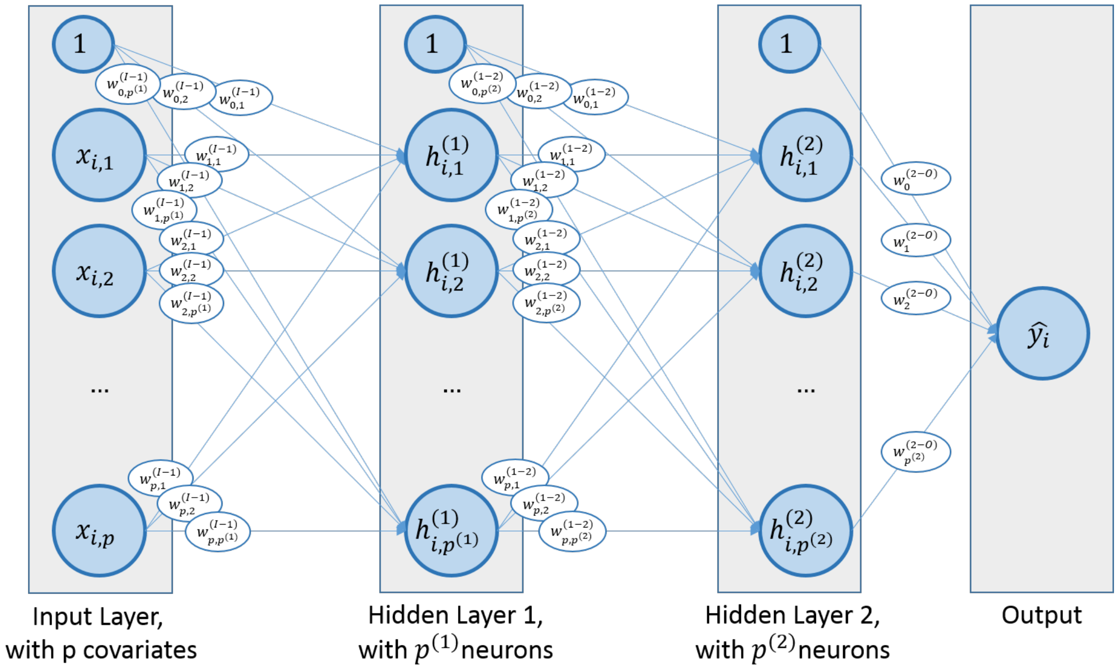


Figure 33 : Schematic representation of a feedforward ANN with two hidden layers.

be

$$\hat{y}_i = g(u_i), \tag{5.12}$$

$$u_i = w_0^{(2-O)} + \sum_{l=1}^{p^{(2)}} w_l^{(2-O)} h_{i,l}^{(2)}, \tag{5.13}$$

whereby $w_0^{(2-O)}$ and $w_l^{(2-O)}$ are set to be the weights between the second layer and the output.

For the error function evaluating the output, the sum of squared errors,

$$E = \frac{1}{2} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2, \tag{5.14}$$

might be employed. The current weights get now adjusted according to their influence on the error. This is done through backpropagation of the error by considering the derivative of E with respect to all individual weights. The new weights are then given by the old weights minus the derivative multiplied by a predefined learning rate α:

$$w_{j,k}^{(I-1,new)} = w_{j,k}^{(I-1)} - \alpha \frac{\partial E}{\partial w_{j,k}^{(I-1)}}, \quad j = 0, ..., p, \quad k = 1, ..., p^{(1)} \tag{5.15}$$

$$w_{k,l}^{(1-2,new)} = w_{k,l}^{(1-2)} - \alpha \frac{\partial E}{\partial w_{k,l}^{(1-2)}}, \quad k = 0, ..., p^{(1)}, \quad l = 1, ..., p^{(2)} \tag{5.16}$$

$$w_{l}^{(2-O,new)} = w_{l}^{(2-O)} - \alpha \frac{\partial E}{\partial w_{l}^{(2-O)}}, \quad l = 0, ..., p^{(2)}. \tag{5.17}$$

For the calculation of the derivatives we start with the weights closest to the output and use the chain rule to get for $l = 0, ..., p^{(2)}$:

$$\begin{aligned}
\frac{\partial E}{\partial w_l^{(2-O)}} &= \frac{\partial \frac{1}{2}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\partial w_l^{(2-O)}} \\
&= \sum_{i=1}^{n} \frac{\partial \frac{1}{2}(\hat{y}_i - y_i)^2}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial u_i} \frac{\partial u_i}{\partial w_l^{(2-O)}} \\
&= \sum_{i=1}^{n} \frac{1}{2} 2(\hat{y}_i - y_i) \frac{\partial g(u_i)}{\partial u_i} \frac{\partial u_i}{\partial w_l^{(2-O)}} \\
&= \sum_{i=1}^{n} (\hat{y}_i - y_i) \frac{\partial \left(\frac{1}{1+e^{-u_i}}\right)}{\partial u_i} \frac{\partial u_i}{\partial w_l^{(2-O)}} \\
&= \sum_{i=1}^{n} (\hat{y}_i - y_i) \frac{1}{1+e^{-u_i}} \frac{e^{-u_i}}{1+e^{-u_i}} \frac{\partial u_i}{\partial w_l^{(2-O)}} \\
&= \sum_{i=1}^{n} (\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i) \frac{\partial u_i}{\partial w_l^{(2-O)}} \\
&= \sum_{i=1}^{n} (\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i) \frac{\partial \left(w_0^{(2-O)} + \sum_{l=1}^{p^{(2)}} w_l^{(2-O)} h_{i,l}^{(2)}\right)}{\partial w_l^{(2-O)}} \\
&= \begin{cases} \sum_{i=1}^{n}(\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i)h_{i,l}^{(2)} & l \geq 1 \\ \sum_{i=1}^{n}(\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i) & l = 0. \end{cases}
\end{aligned} \tag{5.18}$$

We continue with the derivatives with respect to the weights of the preceding layer and get for $k = 0, ..., p^{(1)}$ and $l = 1, ..., p^{(2)}$:

$$\begin{aligned}
\frac{\partial E}{\partial w_{k,l}^{(1-2)}} &= \sum_{i=1}^{n} \frac{\partial \frac{1}{2}(\hat{y}_i - y_i)^2}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial u_i} \frac{\partial u_i}{\partial h_{i,l}^{(2)}} \frac{\partial h_{i,l}^{(2)}}{\partial u_{i,l}^{(2)}} \frac{\partial u_{i,l}^{(2)}}{\partial w_{k,l}^{(1-2)}} \\
&= \sum_{i=1}^{n} (\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i) \frac{\partial \left(w_0^{(2-O)} + \sum_{l=1}^{p^{(2)}} w_l^{(2-O)} h_{i,l}^{(2)}\right)}{\partial h_{i,l}^{(2)}} \frac{\partial h_{i,l}^{(2)}}{\partial u_{i,l}^{(2)}} \frac{\partial u_{i,l}^{(2)}}{\partial w_{k,l}^{(1-2)}} \\
&= \sum_{i=1}^{n} (\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i) w_l^{(2-O)} \frac{\partial h_{i,l}^{(2)}}{\partial u_{i,l}^{(2)}} \frac{\partial u_{i,l}^{(2)}}{\partial w_{k,l}^{(1-2)}}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} (\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i) w_l^{(2-O)} \frac{\partial \left(\frac{1}{1+e^{-u_{i,l}^{(2)}}} \right)}{\partial u_{i,l}^{(2)}} \frac{\partial u_{i,l}^{(2)}}{\partial w_{k,l}^{(1-2)}} \\
&= \sum_{i=1}^{n} (\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i) w_l^{(2-O)} \frac{1}{1+e^{-u_{i,l}^{(2)}}} \frac{e^{-u_{i,l}^{(2)}}}{1+e^{-u_{i,l}^{(2)}}} \frac{\partial u_{i,l}^{(2)}}{\partial w_{k,l}^{(1-2)}} \\
&= \sum_{i=1}^{n} (\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i) w_l^{(2-O)} h_{i,l}^{(2)}(1-h_{i,l}^{(2)}) \frac{\partial u_{i,l}^{(2)}}{\partial w_{k,l}^{(1-2)}} \\
&= \sum_{i=1}^{n} (\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i) w_l^{(2-O)} h_{i,l}^{(2)}(1-h_{i,l}^{(2)}) \frac{\partial \left(w_{0,l}^{(1-2)} + \sum_{k=1}^{p^{(1)}} w_{k,l}^{(1-2)} h_{i,k}^{(1)} \right)}{\partial w_{k,l}^{(1-2)}} \\
&= \begin{cases} \sum_{i=1}^{n} (\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i) w_l^{(2-O)} h_l^{(2)}(1-h_l^{(2)}) h_{i,k}^{(1)} & k \geq 1 \\ \sum_{i=1}^{n} (\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i) w_l^{(2-O)} h_l^{(2)}(1-h_l^{(2)}) & k = 0. \end{cases}
\end{aligned}
\tag{5.19}
$$

Similarly we get for the weights from the input to the first layer the derivative for $j = 0, ..., p,\ k = 0, ..., p^{(1)}$ and $l = 0, ..., p^{(2)}$ by

$$
\begin{aligned}
\frac{\partial E}{\partial w_{k,l}^{(I-1)}} &= \sum_{i=1}^{n} \frac{\partial \frac{1}{2}(\hat{y}_i - y_i)^2}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial u_i} \frac{\partial u_i}{\partial h_{i,l}^{(2)}} \frac{\partial h_{i,l}^{(2)}}{\partial u_{i,l}^{(2)}} \frac{\partial u_{i,l}^{(2)}}{\partial h_{i,k}^{(1)}} \frac{\partial h_{i,k}^{(1)}}{\partial u_{i,k}^{(1)}} \frac{\partial u_{i,k}^{(1)}}{\partial w_{j,k}^{(I-1)}} \\
&= \begin{cases} \sum_{i=1}^{n} (\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i) w_l^{(2-O)} h_{i,l}^{(2)}(1-h_{i,l}^{(2)}) w_{k,l}^{(1-2)} h_{i,k}^{(1)}(1-h_{i,k}^{(1)}) x_{i,j} & j \geq 1 \\ \sum_{i=1}^{n} (\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i) w_l^{(2-O)} h_{i,l}^{(2)}(1-h_{i,l}^{(2)}) w_{k,l}^{(1-2)} h_{i,k}^{(1)}(1-h_{i,k}^{(1)}). & j = 0. \end{cases}
\end{aligned}
\tag{5.20}
$$

Inserting Equations (5.18)-(5.20) in the updating rules given in Equations (5.15)-(5.17), we can calculate new weights for every iteration. As a stopping criterion we define a maximum number of iterations, it_{max}, combined with a lower threshold for the partial derivatives, μ. These are tuning parameters along with the number of hidden layers with their respective numbers of neurons, as well as the learning rate α.

Furthermore, variations in the activation and error functions are possible. To evaluate the error, the cross entropy, given by

$$
C = -\sum_{i=1}^{n} \big(y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i) \big), \tag{5.21}
$$

is an alternative to the sum of squared errors for risk predictions of binary outcomes (Hastie et al. 2009, Kuhn and Johnson 2013, Bishop 1995). Figure 34 displays both error functions in the interval $[0, 1]$ for y and $\hat{y}$.

Minimizing the sum squared errors is equivalent to maximizing the log-likelihood by assuming

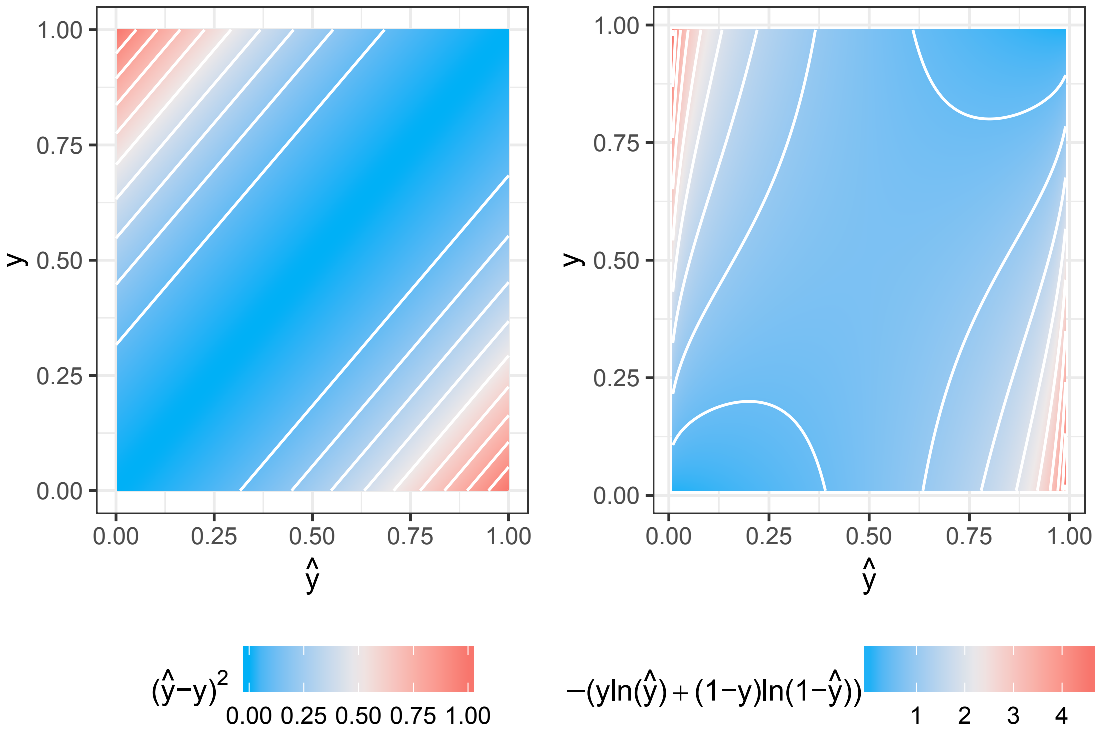


Figure 34 : Contour map of sum squared errors on the left and cross entropy on the right.

a Gaussian distribution with variance σ^2 for the outcome y:

$$\begin{aligned}
min(E) &= min\left(\frac{1}{2}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2\right) \\
&= min\left(\frac{n}{2}log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2\right) \qquad (5.22) \\
&= min\left(\sum_{i=1}^{n}\left(\frac{1}{2}log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}(\hat{y}_i - y_i)^2\right)\right) \\
&= min\left(\sum_{i=1}^{n} -log\left(\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}\right) - log\left(exp\left(-\frac{1}{2\sigma^2}(\hat{y}_i - y_i)^2\right)\right)\right) \\
&= min\left(-\left(\sum_{i=1}^{n} log\left(\frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}}exp\left(-\frac{1}{2\sigma^2}(\hat{y}_i - y_i)^2\right)\right)\right)\right) \\
&= min\left(-(log - likelihood)\right) = max(log - likelihood), \qquad (5.23)
\end{aligned}$$

whereby we used that the constant term $\frac{n}{2}log(2\pi\sigma^2)$ and factor $\frac{1}{\sigma^2} > 0$ do not influence the minimization. Bishop 1995 notes that using the sum of squared errors does not necessarily require the outcome to be Gaussian and can be also applied for classification. Nevertheless, for binary outcomes it might be a disadvantage compared to the cross entropy, which is equivalent to the log-likelihood of the Bernoulli distribution and therefore directly relates to the parameter estimation in logistic regression as introduced in Section 3.3.1:

$$\begin{aligned} min(C) &= min\left(-\sum_{i=1}^{n}\big(y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)\big)\right) \\ &= min\left(-\sum_{i=1}^{n} log\left((\hat{y}_i)^{y_i}(1-\hat{y}_i)^{1-y_i}\right)\right) \\ &= min\left(-(log - likelihood)\right) = max(log - likelihood). \end{aligned} \tag{5.24}$$

Kline and Berardi 2005 point to early results finding comparable estimation accuracy of squared errors and cross entropy. The authors suggest that as a result many researches do not report on the employed error function in ANNs and assume that they primarily use squared errors. A broad availability of squared errors in commercial packages might further strengthen this practice. However, cross entropy might result in more accurate estimations of small probabilities and might be superior to squared errors in binary classifications (Bishop 1995, Kline and Berardi 2005).

We further investigate differences in the derivatives of the two error functions, suggesting computational advantages of the cross entropy C. Consider therefore the derivatives of C with respect to the weights between the last hidden layer and the output for $l = 0, ..., p^{(2)}$:

$$\begin{aligned} \frac{\partial C}{\partial w_l^{(2-O)}} &= -\sum_{i=1}^{n} \frac{\partial\big(y_i ln(\hat{y}_i) + (1-y_i) ln(1-\hat{y}_i)\big)}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial u_i} \frac{\partial u_i}{\partial w_l^{(2-O)}} \\ &= -\sum_{i=1}^{n}\left(\frac{y_i}{\hat{y}_i} - \frac{1-y_i}{1-\hat{y}_i}\right) \frac{\partial \hat{y}_i}{\partial u_i} \frac{\partial u_i}{\partial w_l^{(2-O)}} \\ &= -\sum_{i=1}^{n} \frac{y_i - \hat{y}_i}{\hat{y}_i(1-\hat{y}_i)} \frac{\partial \hat{y}_i}{\partial u_i} \frac{\partial u_i}{\partial w_l^{(2-O)}} \\ &= \begin{cases} -\sum_{i=1}^{n} \frac{y_i-\hat{y}_i}{\hat{y}_i(1-\hat{y}_i)} \hat{y}_i(1-\hat{y}_i) h_{i,l}^{(2)} & l \geq 1 \\ -\sum_{i=1}^{n} \frac{y_i-\hat{y}_i}{\hat{y}_i(1-\hat{y}_i)} \hat{y}_i(1-\hat{y}_i) & l = 0. \end{cases} \\ &= \begin{cases} \sum_{i=1}^{n}(\hat{y}_i - y_i) h_{i,l}^{(2)} & l \geq 1 \\ \sum_{i=1}^{n}(\hat{y}_i - y_i) & l = 0. \end{cases} \end{aligned} \tag{5.25}$$

The difference between these derivatives and the respective ones for the sum squared error function, given in Equation (5.18), is the missing term $\hat{y}_i(1-\hat{y}_i)$. Figure 35 shows both derivatives with respect to the bias term, as this is the basis for all other derivatives. The same holds for the partial derivatives with respect to the remaining weights. Neglecting the term $\hat{y}_i(1-\hat{y}_i)$ enables faster learning, as the term gets close to zero for the output $\hat{y}_i$ approaching its boundaries zero or one. In the contour plots in Figure 35 this disadvantage of the sum squared error can be clearly seen. For the extreme combinations $(\hat{y}, y) \in \{(0,1);(1,0)\}$ we get a derivative of zero, compared to minus one and one for the cross entropy. In case the randomly selected initial weights result in an output close to these extremes, an ANN based

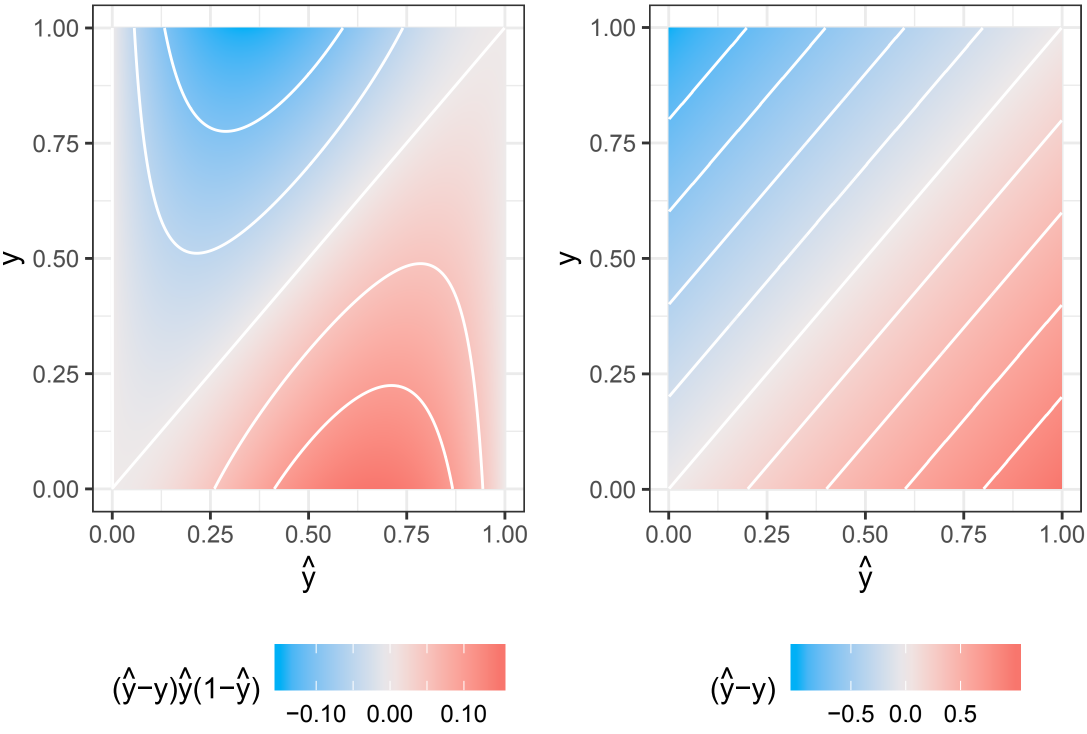


Figure 35 : Contour map of derivatives of sum squared errors on the left and cross entropy on the right.

on the sum squared error adjusts the weights only gradually at the beginning.

As for KNN, it is important to scale the data before an ANN is implemented, whereby we can utilize the standardization given in Equations (5.5) and (5.6).

In a preliminary analysis we found the learning rate ranges $0.000025 - 0.0002$ for cross entropy and $0.000025 - 0.000925$ for the sum squared error function and a threshold of $\mu = 1$ reasonable for further considerations. We furthermore assume simpler ANNs to be sufficient as we analyze only six risk factors as input neurons. We therefore consider models with one and two hidden layers, whereby the first layer can include up to six neurons and the second layer up to two neurons. Figure 36 shows resulting AUC values for all possible combinations. For every training set of the leave-one-cohort-out cross validation we enabled another set of random starting weights. Furthermore, we trained every combination five times in order to visualize remaining variability resulting from the choice of starting weights. Noting the different AUC scale for the three options of layer two, we identify the ANNs with only one hidden layer to have the smallest range of AUCs. Comparing the error functions cross entropy and sum squared error, the latter shows more variation and overall lower AUC values. However, training of ANNs with cross entropy did often not converge within $it_{max} = 100000$ iteration steps for the considered learning rates and a second hidden layer. For an easier comparison we consider the mean values in Figure 37, whereby the mean is only shown if all ANNs converged. We consistently obtain best results for ANNs with cross entropy as error function and only one hidden layer containing one neuron for considered learning rates between 0.000125

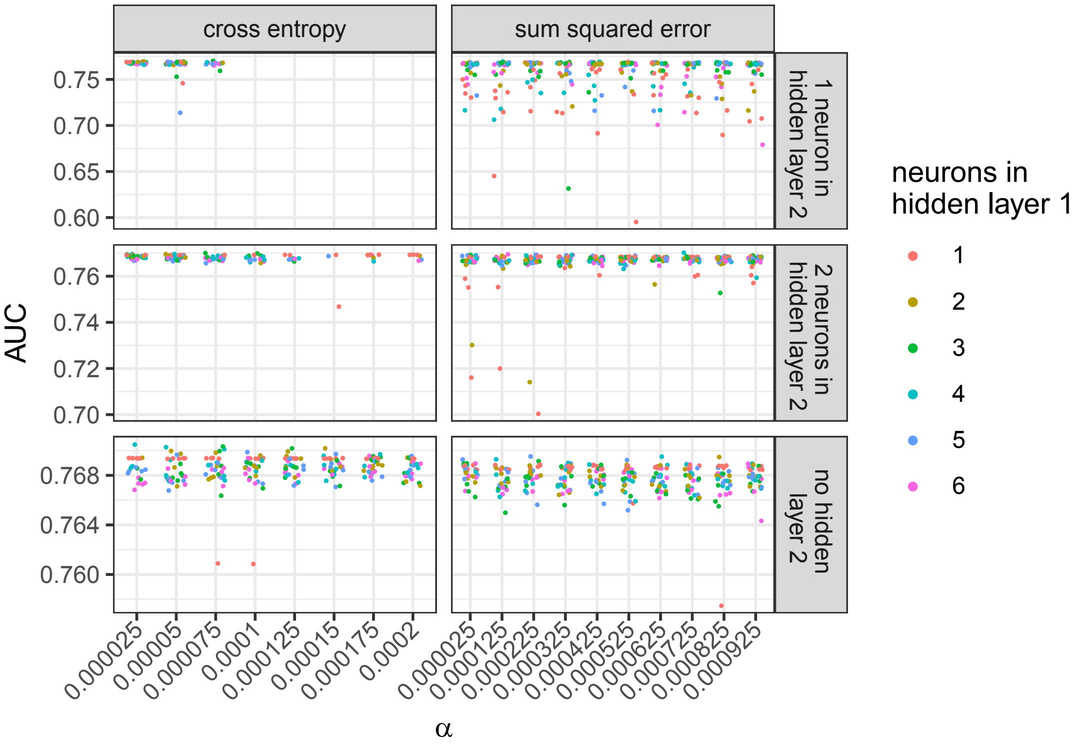


Figure 36 : AUC values of leave-one-cohort-out cross validation for ANNs with different learning rates α and number of neurons in the first and second hidden layers. For every combination five repetitions with random starting weights for every training set in the leave-one-cohort-out cross validation are shown. AUC values are missing in cases of no convergence for $it_{max} = 100000$ iterations and a lower threshold of $mu = 1$.

and 0.0002, whereby we choose the learning rate $\alpha = 0.0002$ for further analysis.

Using the same random starting weights for all training sets in the leave-one-cohort-out cross validation, we trained 20 repetition of this ANN combination and chose the starting weights with the highest resulting AUC. Note that the AUCs thereby range between 76.939% and 76.941%, the difference of starting weights is therefore negligible for this combination.

Finally we trained an ANN on all cohorts combined with the tuning parameters and starting values determined in the previous considerations. Figure 38 shows the final model with all trained weights. We obtained convergence after 8961 steps.

5.3 Results

We compare the considered machine learning approaches, along with their previously discussed optimal tuning parameters, with each other and in addition with a standard logistic regression as discussed in Chapter 3. Figure 39 shows evaluation curves comparing all methods. KNN and bagged KNN perform almost identical across all validation metrics, therefore only one of their curves is visible. In terms of net benefit all methods perform as well as or better than the strategies of referring all or no patients to biopsies for all considered thresh-

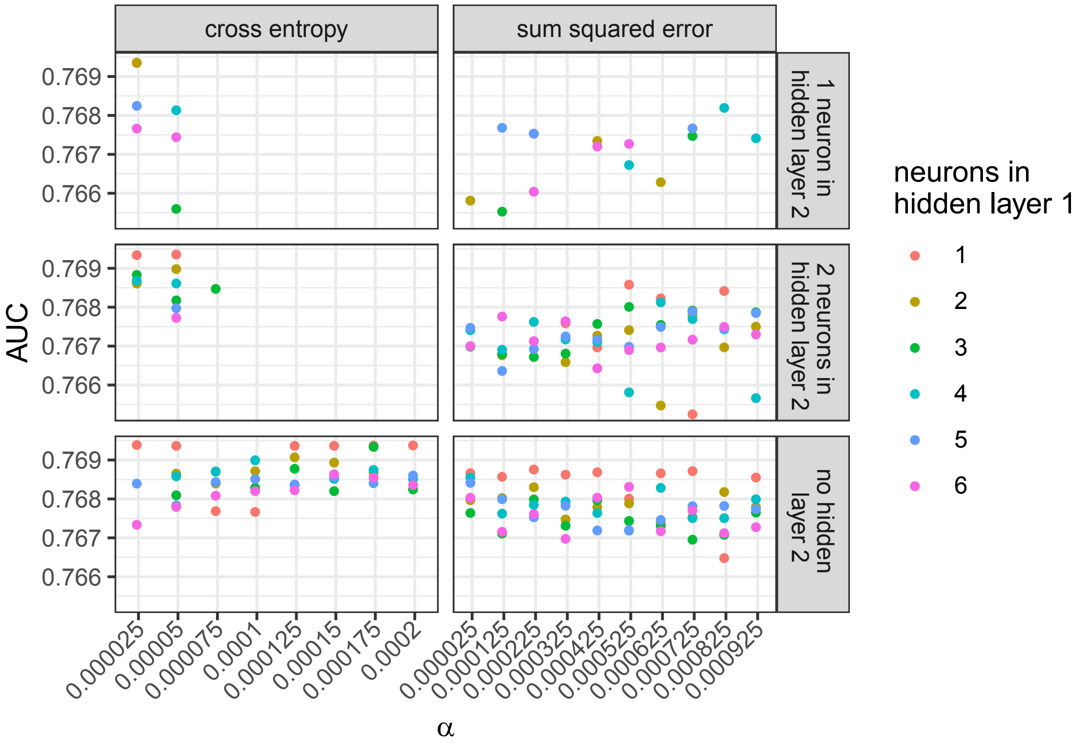


Figure 37 : Mean AUC values across five repetitions with random starting weights of leave-one-cohort-out cross validation for ANNs with different learning rates α and number of neurons in first and second hidden layer. AUC values are missing in case of no convergence for $it_{max} = 100000$ iterations and a lower threshold of $mu = 1$ for one of the repetitions.

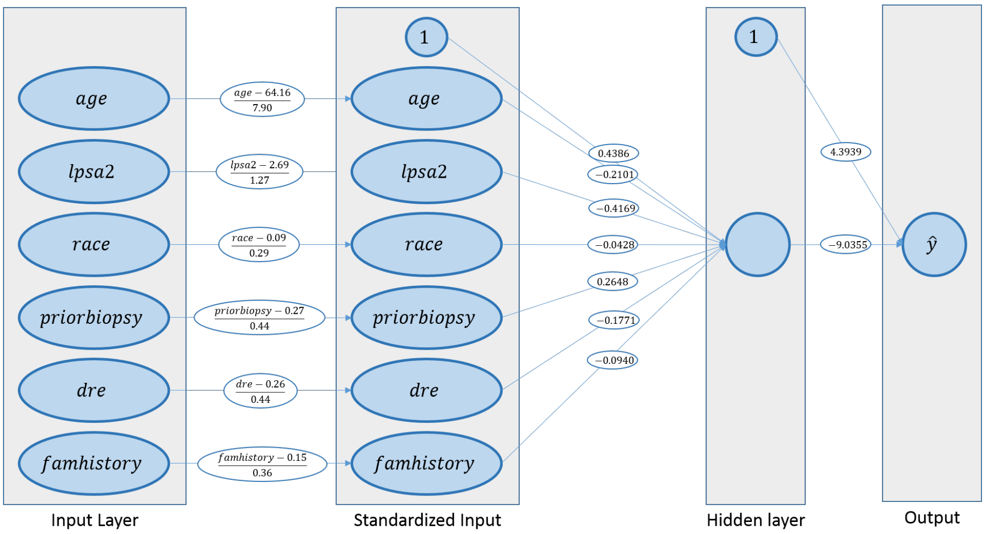


Figure 38 : Schematic representation of trained ANN with one neuron in one hidden layer, learning rate $\alpha = 0.0002$, cross entropy as error function, maximum number of iterations $it_{max} = 100000$ and lower threshold $\mu = 1$.

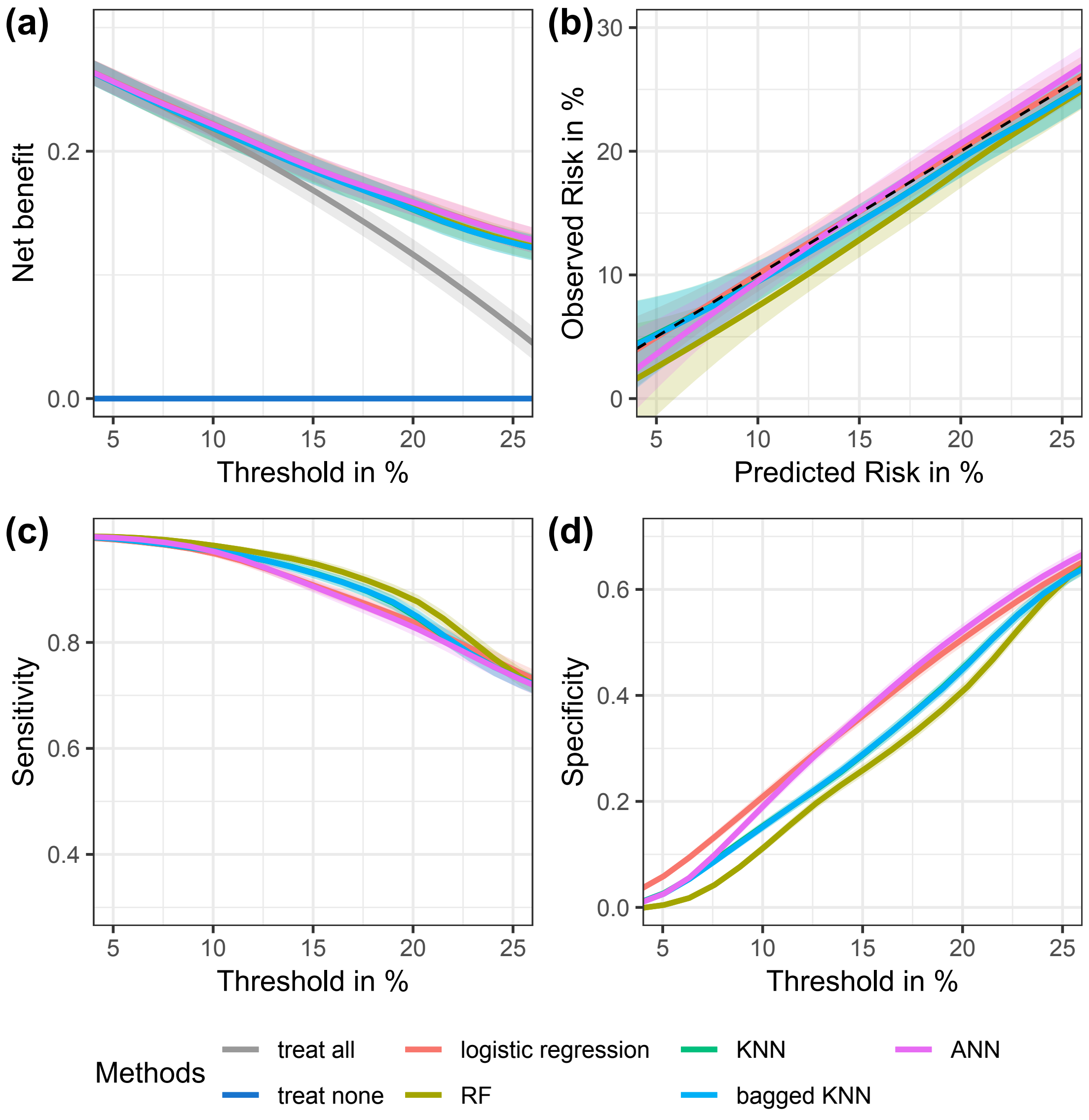


Figure 39 : (a) Net benefit, (b) calibration, (c) sensitivity, and (d) specificity cuves for high-grade cancer comparing different methods. The strategies of referring all men or none to biopsy are given in (a) for comparison. Curves for KNN not visible, as they are almost identical with curves for bagged KNN.

olds between 5 and 25%. Logistic regression and ANN perform very similar and slightly better than RF and the KNN approaches, CIs, however, are overlapping. Also for all test cohorts considered individually in the supplementary Figure A.22, the regression and machine learning methods perform as good as the reference strategies and show overlapping CIs.

We identify the RF model to have worst calibration and the best calibrated model is the logistic regression, however, all CIs are overlapping. Whereas the ANN slightly overpredicts small risks and underpredicts high risks, the KNN approaches overpredict higher risks. Also for the single cohort analysis, shown in Figure A.23, logistic regression is always amongst the best calibrated models, even though CIs are wider and overlapping.

In terms of discrimination, RF has highest sensitivity values and lowest specificity, followed by the KNN approaches. ANN and logistic regression have lowest sensitivity, whereby the regression shows slightly better sensitivity for threshold values about 20% and higher. However, these two approaches have better specificity results than RF and KNN methods. Thereby, logistic regression shows higher specificity than ANN with non-overlapping CIs for thresholds below 10%. Figures A.24 and A.25 for individual cohorts report similar results. Whereby the difference in specificity of logistic regression and ANNs is present in most individual cohorts, the CIs are not overlapping only for the large cohort Zurich.

Figure 40 shows the summary statistics AUC and HLS for all test cohorts individually. They confirm a similar performance of logistic regression and ANN, as well as of KNN and bagged KNN. In all cohorts, except SanRaffaele, the logistic regression and ANN achieve highest AUC values. The differences are, however, small and CIs are overlapping. In terms of the

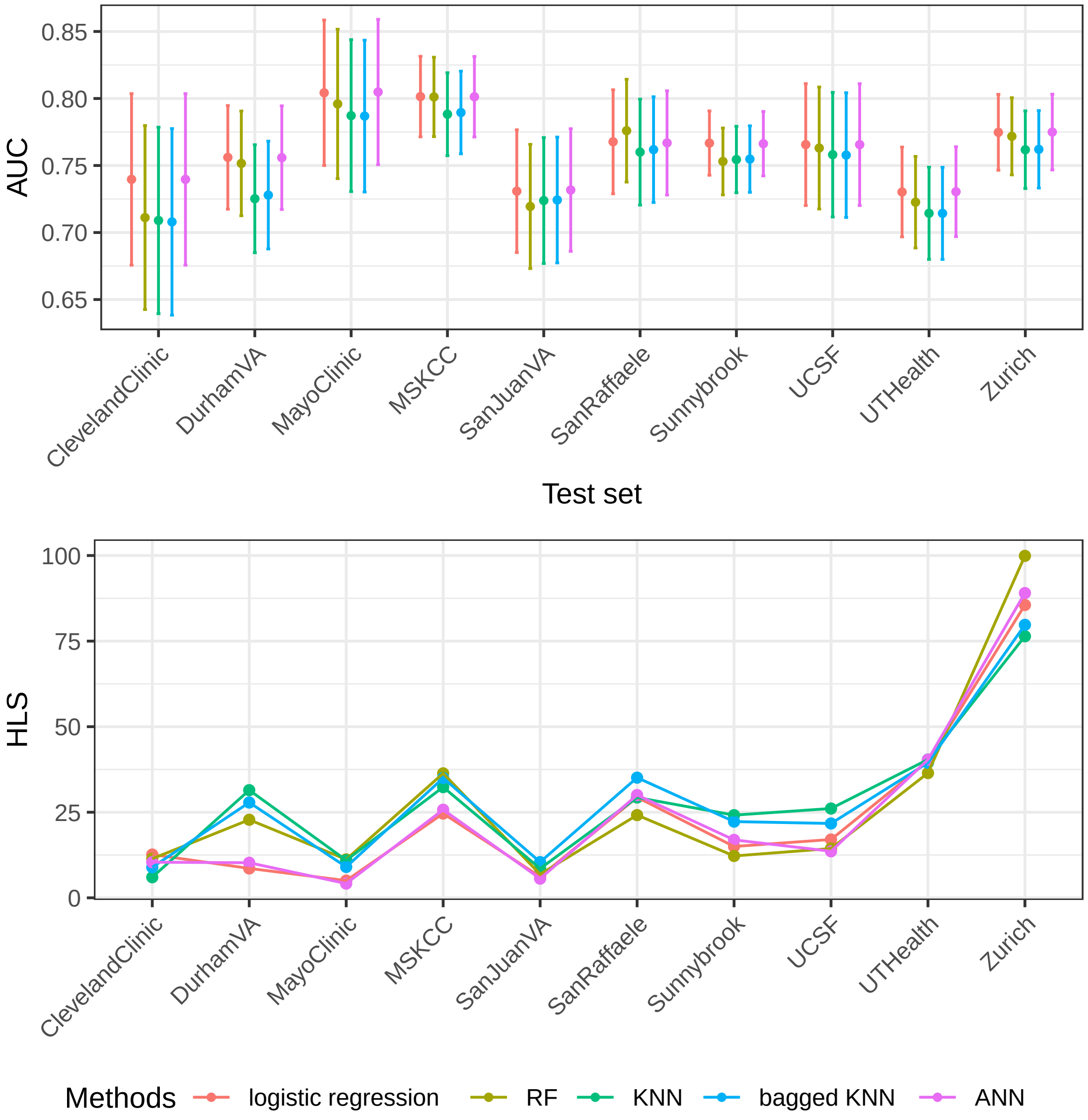


Figure 40 : AUC and HLS values for each of the ten cohorts to compare machine learning methods. For the AUC values 95%-CIs are included.

calibration summary HLS, no method outperforms the logistic regression and ANN for all test cohorts and differences are small.

At last Figure 41 depicts predicted risks and the corresponding true biopsy result for every patient. We observe that logistic regression and the ANN predict more extreme values than the RF and KNN approaches. This becomes particularly evident in combinations of the binary predictors with smaller amounts of patients. Consider, for instance, the combination of having African ancestry, a family history, no prior negative biopsy and an abnormal DRE result, coded as 1101. For this high-risk group the patient with highest observed log based 2 PSA value of 7.99 and an age of 67 has predicted values of 99.0%, 83.1%, 65.0%, 65.6% and 96.7% for logistic regression, RF, KNN, bagged KNN and ANN, respectively. In contrast, the lowest risk predictions are present for a patient with log based 2 PSA value of -0.62 and age 57. The corresponding risks comprise 12.1%, 30.3%, 31.9%, 32.7% and 11.9%.

5.4 Discussion

We compared a standard logistic regression model with diverse machine learning methods in the case of a small set of six predictors. As expected, we found merely small differences between the approaches in this setting. These results align with a literature review of clinical prediction models for binary outcomes by Christodoulou et al. 2019, finding no superior performance of machine learning methods compared to logistic regression. The authors report classification trees, RFs, ANNs and support vector machines as the most common machine learning techniques with the number of included predictors ranging from 5 to 563 with a median of 19.

Also in the field of prostate cancer prediction several researchers compared logistic regression with machine learning methods, thereby resulting in diverse conclusions. Virtanen et al. 1999, Chun et al. 2007 and Takeuchi et al. 2018, for instance, found superior performance of logistic regression compared to ANNs. They used data from 974, 3980 and 334 patients and enabled 7, 5 and 22 variables, respectively. Analysis by Gülkesen et al. 2010 on 750 patients and 5 predictors similarly resulted in lower AUC values for decision trees compared to logistic regression. Interestingly, Strobl, Vickers, et al. 2015 describe a worse performance of RFs for a data set similar to the one used in the analysis of this chapter, which differs from our observation of similar discrimination and calibration of both methods. We might attribute this discrepancy to the fact that Strobl, Vickers, et al. 2015 used classification instead of regression RFs, which have shown low AUC values in Section 5.2.1 as well.

Alternatively, in Finne et al. 2000 an ANN with two neurons in one hidden layer, which was trained on six predictors of 656 observations, showed better accuracy at high sensitivity levels than a logistic regression. Similarly Boegemann et al. 2016 report ANNs containing one hidden layer with two and three neurons to slightly outperform logistic regression with respect to AUC and net benefit. Training data for these models were seven predictors of 769 patients.

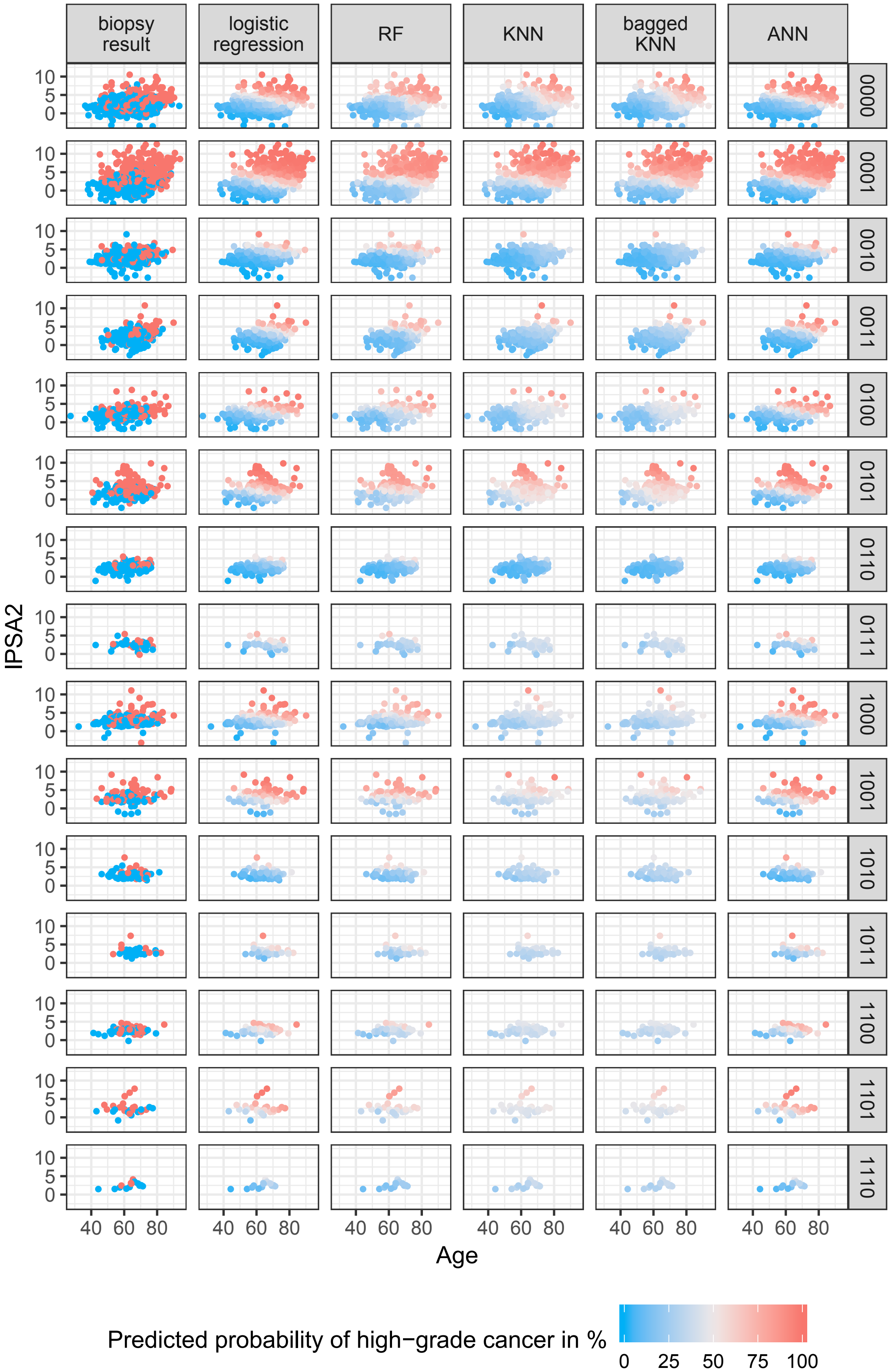


Figure 41 : Predicted probabilities for high-grade cancer for different machine learning methods, with true biopsy results given for comparison. Predictions are separated by possible combinations of the binary covariates: African ancestry, family history, prior negative biopsy and abnormal DRE result, whereby yes and no are coded as 1 and 0, respectively.

The two studies support the decision to use merely one hidden layer for a small set of variables as analyzed in this chapter.

At last, Garzotto et al. 2005 and Barnholtz-Sloan et al. 2011 discuss comparable results for classification and regression tree analysis compared to logistic regression, using 1433 and 2025 patients, respectively. Whereas the former authors analyzed seven variables, the latter ones enabled five variables in combination with several genotype predictors. Similar to these results, also the previous comparisons show mainly small differences between the methods. Furthermore most of them enabled small sets of predictors, ranging from five to seven, as well. These studies, as well as our results, therefore confirmed our initial expectations of merely small differences between the considered approaches in the setting of only few preselected variables for the prediction of prostate cancer biopsy results.

However, the analysis performed in this chapter illustrated the underlying mechanism and revealed advantages as well as disadvantages to be considered in future work. We in particular assess their implementations with respect to additional risk factors, as these might be needed to further enhance prostate cancer risk prediction. Additional predictors, as further described in Chapter 6, are, for instance, MRI markers. Since these markers were not included in the primary PBCG protocol, only a few cohorts collected this information and analysis is delayed to future work when data collection has ended. ANNs and RFs are already broadly implemented for prostate cancer detection based on MRI and might also constitute a suitable framework to combine MRI and standard risk factors (Lay et al. 2017, Qian et al. 2016, Zhu et al. 2017, Rampun et al. 2016, Song et al. 2018, Z. Wang et al. 2018).

Whereas preselection of risk factors might not be required for RFs and ANNs, we have to correctly specify all relevant variables, interactions and higher order terms to use a logistic regression. Therefore preprocessing the data becomes essential as the number of possible covariates increases. In this thesis we used stepwise model selection algorithms as introduced in Section 3.2 and more sophisticated methods might be necessary for data sets with many possible predictors.

Variables that are not relevant for the outcome can also undesirably influence KNN methods. Predictions of KNNs rely on observations most similar to the patient at hand. Distance measures assess the similarity of observations and usually do not account for the variables' relation to the outcome. As previously discussed, weights might therefore be necessary within the distance measure to handle irrelevant or noisy predictors. Otherwise, this noise may cause observations to not be included in the neighborhood, even though they are similar with respect to important features, but diverse in terms of additional covariates. However, it is not straightforward to find optimal weights and we therefore discard this approach for further considerations.

A disadvantage of ANNs is the presence of several tuning parameters. For instance, the use of a simple neural net requires optimal values for the number of hidden layers with their respective number of neurons, as well as for the learning rate, error function, number of maximal iterations and lower threshold for the partial derivatives. Further variations include the choice of an activation function and the general type of the ANN. This vast amount of options exacerbate finding the best possible model and trying several combinations is furthermore computational expensive. Even for the considered application with a small set of predictors and a predefined selection of the general structure of the ANN, we found the optimal choice of tuning parameters to be time consuming and not straightforward. RFs require specification of less tuning parameters and logistic regression do not have tuning parameters at all.

Of the considered approaches we identified ANNs and RFs as the most promising methods for future work, whereby ANNs showed better calibration. We in particular advocate their use in cases where the set of variables becomes large. Whereas RFs are easier to implement, ANNs enable an individual adjustment to diverse data due to their flexible structure. However, we emphasize the resulting specification challenges for ANNs and encourage future work to assess the characteristics of activation and error functions with respect to the data at hand.

6 Conclusions

Through visualization, statistical modeling and cross validation, this thesis addressed the three aims stated in Chapter 1, thus providing a novel comprehensive framework for developing global risk tools incorporating data from multiple heterogeneous centers.

6.1 Aim 1: Will incorporation of cohort heterogeneity improve validation of a global risk prediction model?

Large consortiums that aggregate data for unified purposes, such as for the development of a global risk tool as performed here, presuppose that such data are homogeneous in terms of distribution of risk factors, outcomes, and their relations. In the spirit of interdisciplinary data science, we therefore developed novel visualization graphs for emphasizing heterogeneity in all aspects of data across the institutions in Chapter 1. This led to new revelations on specific aberrations from institutions and fruitful dialogue among all clinicians. For example, Figure 7 revealed that SanJuanVA had an unusually high abnormal DRE rate (51.3%), but with only 28.7% of the abnormal DREs positive for cancer. This brought to light that SanJuanVA was over-diagnosing lumps in the prostate, and should consider retraining urologists. Interestingly, only by visualization and comparison to other institutions was this problem revealed.

Despite heterogeneity revealed by visualization, comparisons of five alternative methods competing adjustment for versus ignoring center heterogeneity using a novel permutation validation strategy in Chapter 3 found no differences in prediction performance. Reasons for this could be the binary nature of the outcomes. When modeling discrete outcomes compared to continuous ones, intuition would state that there is less power to detect clusters, because in the discrete case the clusters affect only the latent normal random effects distribution, which has only an indirect link to the binary outcome.

These results therefore support the common practice of simply pooling data across diverse cohorts for valid risk prediction models. Additionally, meta-analysis of IPD performed as well as other approaches, suggesting that centralized data storage, with all its compliance and ethics regulation concerns, is not necessary. This enables the possibility to reduce organizational workload. The answer to Aim 1 for the PBCG is no.

6.2 Aim 2: Should existing risk models be updated as soon as contemporary data are available?

Chapter 4 compared prediction models based on contemporary data from 2006 to 2017 of the PBCG with the risk calculator of the PCPT based on biopsies from the 1990s, using internal cross validation within the North American cohorts and external validation on the remaining European sites of the PBCG. This analysis showed a small but significant improvement in AUC for the contemporary prediction tool, however, in terms of calibration and clinical net benefit, the PBCG model clearly outperformed the PCPTRC.

The enhancement in performance occurred due to differences in patient population, as described in Section 1.2. Cohort designs and inclusion criteria led to a predominantly healthy population included in the PCPT, as this study required a PSA$\leq$4 ng/ml and a normal DRE for patients to enter the study. In contrast, in the PBCG, clinical referral preceded prostate biopsy. In addition, changes in clinical practice necessitated contemporary data and resulting model updates. These changes included an increase in number of biopsy cores and a shift in prostate cancer grading towards high-grade assessments.

We therefore advocate the use of contemporary data if available, as existing risk prediction tools might not be appropriate for current clinical practice. It is in particular desirable to continually update risk models with new data as described by Strobl, Vickers, et al. 2015 and Strobl, I. M. Thompson, et al. 2015. The answer to Aim 2 for the PBCG is yes.

6.3 Aim 3: Can more flexible machine learning methods improve traditional regression approaches for small sets of established risk factors?

Comparing more flexible machine learning methods with a standard logistic regression in Chapter 5 showed only small differences in the PBCG setting of only six covariates. However, we identified ANNs and RFs as suitable approaches to handle a larger number of predictors that may arise in future work of the PBCG. Whereas RFs are easier to implement, ANNs tend to show better calibration. The answer to Aim 3 for the current PBCG is no, but for future analyses of more variables, potentially.

6.4 Future work incorporating new biomarkers into established risk tools

In this thesis we only considered the variables PSA, family history, prior negative biopsy, race, age and DRE as these are the established and routinely reported risk factors. However, Strobl, Vickers, et al. 2015 note that new markers might be necessary to further improve discrimination of risk predictions as models based on the currently employed standard covariates might have reached their limit. The PBCG collects several additional variables, listed in Figure 42. Previous studies on prostate cancer suggest that inclusion of some of these variables might improve model performance (Grill, Fallah, Leach, I. M. Thompson, Freedland, et al. 2015, Kranse et al. 2008, Kuo et al. 2013, Lamy et al. 2018, Beebe-Dimmer et al. 2006, Nun-

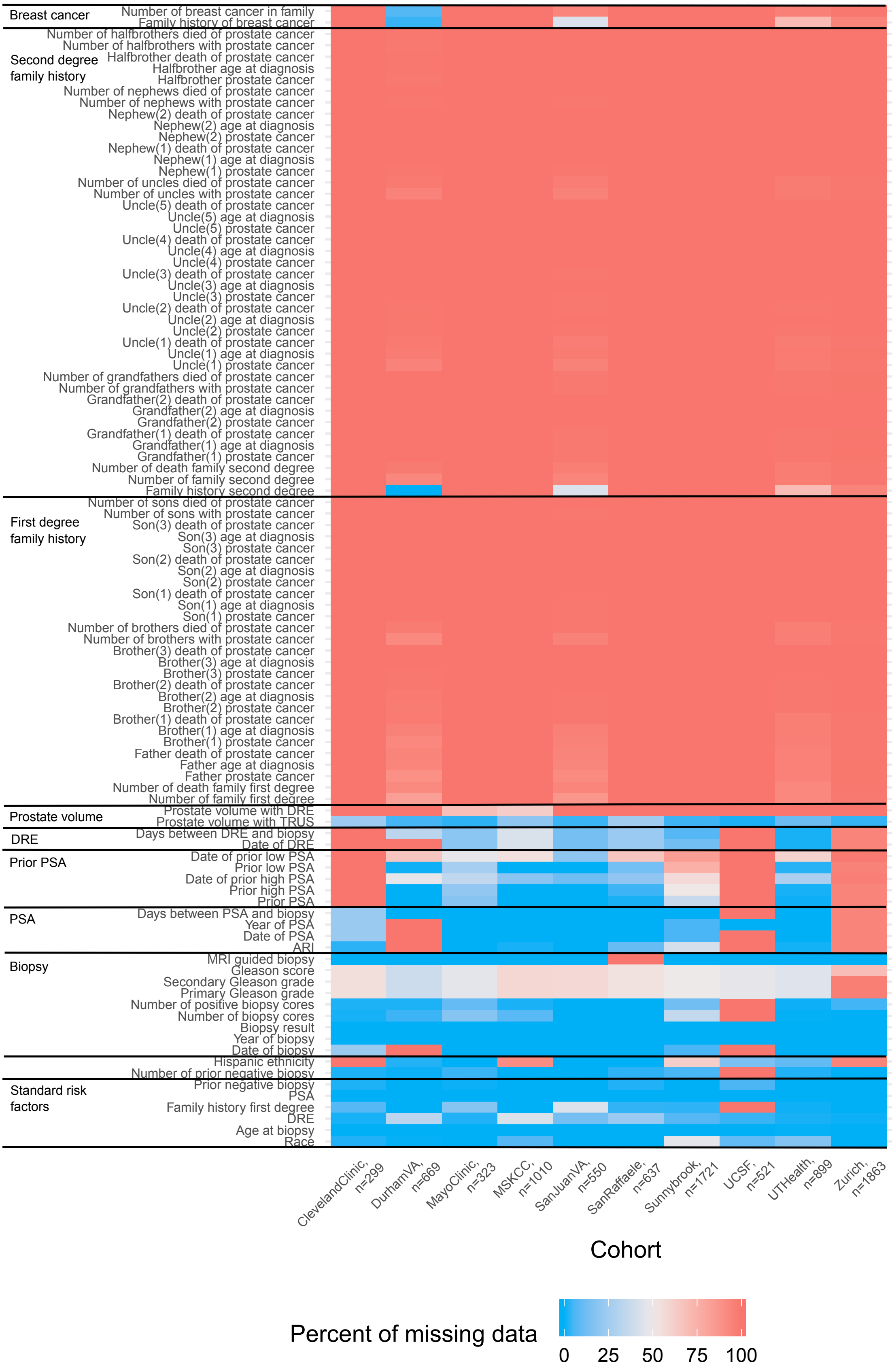


Figure 42 : Percent of missing data for all available variables, stratified by cohort. Gleason score, primary and secondary Gleason grade are only available in case of a positive biopsy. Acronyms: DRE, digitial rectal examination; TRUS, transrectal ultrasound; PSA, prostate-specific antigen; ARI, 5-alphaspacereductase inhibitors; MRI, magnetic resonance imaging.

zio et al. 2018). However, information on more rarely collected variables suffers from higher proportions of missing values. We therefore delayed analysis of these special variables until data acquisition has ended and the overall number of observation is maximized.

A further concern for future work is that external validation becomes problematic. Most other studies might not collect the considered new markers and hence not be able to perform a retrospective collection. Shariat et al. 2009 discuss that despite adopting smart medical records, the use of some variables remain impractical. Future work should take these considerations into account in case of adjusting existing risk tools to include more variables.

Many of the additional variables of the PBCG comprise detailed information on cancer history of the patient's family (Figure 42). Whereas all cohorts except UCSF provide information on whether or not a first degree family member experienced prostate cancer, only the four cohorts, DurhamVA, SanJuanVA, UTHealth and Zurich report more details. Most questions concerning detailed family history are only relevant for patients with a present prostate cancer history within their family, such as the age of diagnosis of the affected members. These data are missing by design for patients with no family history and thus do not count towards the missing data issue. The graph in Figure 42 does not reflect this distinction. However, other variables recording presence of prostate cancer in second degree family members as well as history of breast cancer suffer from high missingness among the four sites.

Similarly, most of the other additional variables provide more detailed information about the risk factors already included in the model. A date and resulting days before biopsy complement PSA and DRE values. Some cohorts also provide information on whether the patient used 5-alphaspacereductase inhibitors at PSA blood draw and information about prior PSA measurements. Detailed race and Hispanic or Latino ethnicity supplement the binary variable race of being African American or not. The number of prior negative biopsies specifies the respective binary assessment of having at least one. An additional marker is prostate volume, more accurately measured by transrectal ultrasound instead of DRE.

Finally, cohorts that performed MRI-guided biopsies additionally collected MRI markers. These will be available at the end of data acquisition and are therefore not part of Figure 42.

For parametric methods, such as logistic regression, it is important to correctly specify all risk predictors. This becomes particularly complicated for increasing sets of possible covariates with unknown correlation structures. Suitable algorithms to select variables are necessary. In this thesis we used a stepwise selection algorithm based on the BIC to determine relevant terms. Alternatives are, for instance, the least absolute shrinkage and selection operator (LASSO) introduced by Tibshirani 1996. The process selects a subset of the provided variables by shrinking the regression coefficients. The LASSO minimizes the basic regression least squares as described in Chapters 3 and 4, while simultaneously constraining the sum of absolute regression coefficients. Least angle regression is another approach developed by

Efron et al. 2004 to select variables, which is similar but less greedy than traditional forward selection methods. These methods were not used in this thesis due to the small number of six covariates, but will represent suitable alternatives for analysis of the expanded variable set in Figure 42.

Alternatively, nonparametric methods might be able to meet the challenge of many potential predictors with unknown correlation structure. Chapter 5 introduced some common approaches and identified ANNs and RFs to be promising as they do not rely on extensive preprocessing of the data. However, the implementation, in particular of ANNs, is more complex than standard logistic regression.

6.5 Dealing with missing values

Even for the small number of variables used in the models of Chapter 3 - 5, some observations were missing. We imputed these with their corresponding median. With increasing variable numbers the amount of missing values also rises (Figure 42). We in particular can not assume the pattern of missing values to be random and dependencies to other variables are possible. Furthermore, heterogeneities across cohorts might become problematic, as some variables are, for instance, only available for specific sites. Overall, more sophisticated methods to deal with missing data might improve model performance and their advantage might become even more evident for increasing numbers of possible risk factors and increased prevalence of structured missingness.

Schafer and Graham 2002 review general methods to deal with missing data and recommend use of maximum likelihood and Bayesian multiple imputation approaches, but also point to methods based on weighting such as weighted regression by Robins et al. 1994. Maximum likelihood estimates might utilize the iterative algorithm described by Dempster et al. 1977. This algorithm consists of an expectation, followed by a maximization step, therefore called the EM algorithm. Multiple imputation, proposed by Rubin 1987, first creates several complete data sets by estimating plausible values for the missing data. Estimates from individual but identical analyses on these sets are finally combined. A frequently implemented extension is the method of multivariate imputation by chained equation (Azur et al. 2011, van Buuren and Groothuis-Oudshoorn 2011, White et al. 2011). It relies on the assumption that missing values of one patient can be estimated by patients with similar characteristics for the remaining variables. This approach specifies separate imputation models for each variable with missing values based on the information of all other variables. This enables an iterative imputation of the variables to get multiple complete data sets, which are then combined into an overall estimate.

However, for the PBCG it remains unclear how to incorporate the clustered nature of the data, as it is possible to either use all patients or only patients from the same cohort for imputation. Restricting the data to a single cohort might be useful in cases where a specific variable is diverse across cohorts. For the covariate African ancestry, for instance, it is advisable to separately impute missing values for patients of European cohorts, as these have an almost exclusively Caucasian patient mix, compared to DurhamVA with over 60% of patients with African ancestry.

At the same time it is possible that a risk factor is missing for all patients from one cohort, such as family history for UCSF, so that imputations have to rely on patients from other cohorts. This is a common scenario in meta-analysis and Koopman et al. 2008 discourage imputation over trials while ignoring heterogeneity. Resche-Rigon et al. 2013 and Burgess et al. 2013 discuss combinations of multiple imputations and IPD meta-analyses, which address the problem of systematic missing values and allows for heterogeneity between studies. Jolani et al. 2015 introduce an extension to this approach, including imputation of non-continuous predictors. Instead of imputation, Jackson, White, Kostis, et al. 2009 propose to model estimates of regression coefficients from studies including all covariates and studies with less predictors in a bivariate meta-analysis. Kovačić and Varnai 2016 extend this approach with a graphical model to enable estimations in complex missing structures.

Furthermore, Kondofersky et al. 2016 suggest an ensemble of models to handle the situation of missing values for a whole cohort. This approach trains different models on all possible subsets of cohorts and afterwards averages the resulting predictions. This has the advantage that covariates, which are not reported for a specific cohort, might be excluded from the models which contain this cohort in the training set, but can be included in the remaining ones.

In case only some cohorts report a specific variable, for instance information about family history of breast cancer, alternative approaches might be appropriate. This missingness structure resembles the situation of an existing model built on a large data set, which gets updated by a new, potentially small, study including an additional predictor. Grill, Ankerst, et al. 2017 and Cheng, Taylor, T. Gu, et al. 2019 assess diverse methods for this purpose using both frequentist and bayesian approaches.

6.6 Model on demand

The described techniques for missing data use imputation, ensemble or updating methods to use as much data as possible to get an overall model including all relevant risk predictors. Their focus is thereby on missing variables of the training set, however, also new observations might not have collected all features. In general, current risk tools apply to single endpoints and specified sets of risk factors, and do not allow missing covariates or adjust to specific interests of the user.

In light of the growing interest in shared decision making, requiring precise predictions for diverse settings, it might therefore become an appealing goal for future work in prostate cancer biopsy prediction to implement a model on demand. This model should be able to exactly scale to the particular needs and available data of individual patients. It should be possible for the user of a risk prediction tool to choose a specific outcome among a range of possibilities. In particular diverse classifications of high-grade prostate cancer exist, for instance based on the overall Gleason score, the combination of primary and secondary Gleason grade and number or percentage of positive cores. As shown in Figure 42, the PBCG collects several types of information on the biopsy and diverse responses might be modeled. Furthermore, selection of predictors should be individualized, based on the available data of the patient.

We might achieve this universal approach by combining data sets of all available cohorts of the PBCG and ideally also further institutions, even though these may report different outcomes as well as risk factors. We could then build individual models for every requested use of the prediction tool, enabling all data available for the required constellation and ignoring possible heterogeneity among institutions. As we have shown in Chapter 3, simple data pooling is suitable in this context. The PCPTRC, and the analogously built risk calculator based on the PBCG data in Chapter 4 have already implemented this kind of model on demand, as these risk tools offer separate models for diverse missingness structures. These independent logistic regressions, built on different sets of included covariates, were straightforward to implement and showed good performance.

A concern is, however, that for unique constellations of outcome and covariates only small sets of training data might be available. Resulting predictions could be inaccurate and misleading. In these cases it becomes particularly important to report the sample size and to state a clear warning of the potential unreliability of the prediction tool.

This dynamic approach has the further advantage of a continuous updating process. It is possible to implement an upper limit for the time span of included data such that every individual model includes only the most recent data. This is appealing as Chapter 4 revealed the necessity of contemporary data, and is destined to become state-of-the-art, as access to and quality of electronic health records improves.

Bibliography

Abo-Zaid, G. M. A. et al. (2013). “Individual participant data meta-analyses should not ignore clustering”. In: *Journal of Clinical Epidemiology* 66.8, 865–873.e4. ISSN: 1878-5921. DOI: 10.1016/j.jclinepi.2012.12.017.

Allyn, J., N. Allou, et al. (2017). “A Comparison of a Machine Learning Model with EuroSCORE II in Predicting Mortality after Elective Cardiac Surgery: A Decision Curve Analysis”. In: *PloS One* 12.1, e0169772. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0169772.

Allyn, J., C. Ferdynus, et al. (2016). “Simplified Acute Physiology Score II as Predictor of Mortality in Intensive Care Units: A Decision Curve Analysis”. In: *PloS One* 11.10, e0164828. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0164828.

Anderson, K. M. et al. (1991). “An updated coronary risk profile. A statement for health professionals”. In: *Circulation* 83.1, pp. 356–362. ISSN: 0009-7322. DOI: 10.1161/01.CIR.83.1.356.

Anderson, S. J. et al. (1992). *NSABP Breast Cancer Prevention Trial risk assessment program, version 2. NSABP Biostatistical Center Technical Report.*

Ankerst, D. P., A. Boeck, et al. (2012). “Evaluating the PCPT risk calculator in ten international biopsy cohorts: Results from the Prostate Biopsy Collaborative Group”. In: *World Journal of Urology* 30.2, pp. 181–187. DOI: 10.1007/s00345-011-0818-5.

Ankerst, D. P., J. Groskopf, et al. (2008). “Predicting prostate cancer risk through incorporation of prostate cancer gene 3”. In: *The Journal of Urology* 180.4, pp. 1303–1308. ISSN: 1527-3792. DOI: 10.1016/j.juro.2008.06.038.

Ankerst, D. P., J. Hoefler, et al. (2014). “Prostate Cancer Prevention Trial risk calculator 2.0 for the prediction of low- vs high-grade prostate cancer”. In: *Urology* 83.6, pp. 1362–1367. ISSN: 1527-9995. DOI: 10.1016/j.urology.2014.02.035.

Ankerst, D. P., T. Koniarski, et al. (2012). “Updating risk prediction tools: A case study in prostate cancer”. In: *Biometrical Journal* 54.1, pp. 127–142. DOI: 10.1002/bimj.201100062.

Ankerst, D. P., J. Straubinger, et al. (2018). “A Contemporary Prostate Biopsy Risk Calculator Based on Multiple Heterogeneous Cohorts”. In: *European Urology* 74.2, pp. 197–203. ISSN: 1873-7560. DOI: 10.1016/j.eururo.2018.05.003.

Ankerst, D. P., C. Till, A. Boeck, P. J. Goodman, C. M. Tangen, Z. Feng, et al. (2013). “The impact of prostate volume, number of biopsy cores and American Urological Association symptom score on the sensitivity of cancer detection using the Prostate Cancer Prevention Trial risk calculator”. In: *The Journal of Urology* 190.1, pp. 70–76. ISSN: 1527-3792. DOI: 10.1016/j.juro.2012.12.108.

Ankerst, D. P., C. Till, A. Boeck, P. J. Goodman, C. M. Tangen, and I. M. Thompson (2013). “Predicting risk of prostate cancer in men receiving finasteride: Effect of prostate volume, number of biopsy cores, and American Urological Association symptom score”. In: *Urology* 82.5, pp. 1076–1081. ISSN: 1527-9995. DOI: 10.1016/j.urology.2013.07.041.

Augustin, H. et al. (2012). “Decision curve analysis to compare 3 versions of Partin Tables to predict final pathologic stage”. In: *Urologic Oncology* 30.4, pp. 396–401. DOI: 10.1016/j.urolonc.2010.07.003.

Austin, P. C. and E. W. Steyerberg (2014). “Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers”. In: *Statistics in Medicine* 33.3, pp. 517–535. ISSN: 1097-0258. DOI: 10.1002/sim.5941.

Austin, P. C., D. van Klaveren, et al. (2016). “Geographic and temporal validity of prediction models: Different approaches were useful to examine model performance”. In: *Journal of Clinical Epidemiology* 79, pp. 76–85. ISSN: 1878-5921. DOI: 10.1016/j.jclinepi.2016.05.007.

Austin, P. C., D. van Klaveren, et al. (2017). “Validation of prediction models: Examining temporal and geographic stability of baseline risk and estimated covariate effects”. In: *Diagnostic and Prognostic Research* 1:12. DOI: 10.1186/s41512-017-0012-3.

Azur, M. J. et al. (2011). “Multiple imputation by chained equations: What is it and how does it work?” In: *International Journal of Methods in Psychiatric Research* 20.1, pp. 40–49. DOI: 10.1002/mpr.329.

Babayan, R. K. and M. H. Katz (2016). “Biopsy Prophylaxis, Technique, Complications, and Repeat Biopsies”. In: *Prostate Cancer*. Elsevier, pp. 77–82. ISBN: 9780128000779. DOI: 10.1016/B978-0-12-800077-9.00009-8.

Baker, S. G. et al. (2009). “Using relative utility curves to evaluate risk prediction”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172.4, pp. 729–748. ISSN: 09641998. DOI: 10.1111/j.1467-985X.2009.00592.x.

Bamber, D. (1975). “The area above the ordinal dominance graph and the area below the receiver operating characteristic graph”. In: *Journal of Mathematical Psychology* 12.4, pp. 387–415. ISSN: 00222496. DOI: 10.1016/0022-2496(75)90001-2.

Banegas, M. P. et al. (2012). “Evaluating breast cancer risk projections for Hispanic women”. In: *Breast Cancer Research and Treatment* 132.1, pp. 347–353. DOI: 10.1007/s10549-011-1900-9.

Barnes, B. et al. (2016). *Bericht zum Krebsgeschehen in Deutschland 2016*. DOI: 10.17886/rkipubl-2016-014.

Barnholtz-Sloan, J. S. et al. (2011). “Decision tree-based modeling of androgen pathway genes and prostate cancer risk”. In: *Cancer Epidemiology, Biomarkers & Prevention* 20.6, pp. 1146–1155. ISSN: 1538-7755. DOI: 10.1158/1055-9965.EPI-10-0996.

Bates, D. et al. (2015). “Fitting Linear Mixed-Effects Models Using lme4”. In: *Journal of Statistical Software* 67.1, pp. 1–48. ISSN: 1548-7660. DOI: 10.18637/jss.v067.i01.

Beebe-Dimmer, J. L. et al. (2006). “Association between family history of prostate and breast cancer among African-American men with prostate cancer”. In: *Urology* 68.5, pp. 1072–1076. ISSN: 1527-9995. DOI: 10.1016/j.urology.2006.06.028.

Bhat, H. S. and N. Kumar (2010). *On the Derivation of the Bayesian Information Criterion.* School of Natural Sciences, University of California.

Biau, G. et al. (2008). “Consistency of Random Forests and Other Averaging Classifiers”. In: *Journal of Machine Learning Research* 9, pp. 2015–2033.

Billheimer, D. et al. (2014). “Combined benefit of prediction and treatment: A criterion for evaluating clinical prediction models”. In: *Cancer Informatics* 13.Suppl 2, pp. 93–103. ISSN: 1176-9351. DOI: 10.4137/CIN.S13780.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Clarendon Press. ISBN: 0198538642.

Bjurlin, M. A. et al. (2014). “Optimization of prostate biopsy: Review of technique and complications”. In: *The Urologic Clinics of North America* 41.2, pp. 299–313. DOI: 10.1016/j.ucl.2014.01.011.

Boegemann, M. et al. (2016). “The percentage of prostate-specific antigen (PSA) isoform -2proPSA and the Prostate Health Index improve the diagnostic accuracy for clinically relevant prostate cancer at initial and repeat biopsy compared with total PSA and percentage free PSA in men aged $\leq$65 years”. In: *BJU International* 117.1, pp. 72–79. DOI: 10.1111/bju.13139.

Bokhorst, L. P. et al. (2016). “Complications after prostate biopsies in men on active surveillance and its effects on receiving further biopsies in the Prostate cancer Research International: Active Surveillance (PRIAS) study”. In: *BJU International* 118.3, pp. 366–371. DOI: 10.1111/bju.13410.

Boos, D. D. and L. A. Stefanski (2013). *Essential Statistical Inference*. Vol. 120. New York, NY: Springer New York. ISBN: 978-1-4614-4817-4. DOI: 10.1007/978-1-4614-4818-1.

Bouwmeester, W., J. W. R. Twisk, et al. (2013). “Prediction models for clustered data: Comparison of a random intercept and standard regression model”. In: *BMC Medical Research Methodology* 13:19. ISSN: 1471-2288. DOI: 10.1186/1471-2288-13-19.

Bouwmeester, W., N. P. A. Zuithoff, et al. (2012). “Reporting and methods in clinical prediction research: A systematic review”. In: *PLoS Medicine* 9.5, pp. 1–12. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001221.

Breiman, L. (1996). “Bagging Predictors”. In: *Machine Learning* 24.2, pp. 123–140. ISSN: 1573-0565. DOI: 10.1023/A:1018054314350.

Breiman, L. (2001). “Random Forests”. In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324.

Breiman, L. (2004). *Consistency for a simple model of random forests: Technical Report 670, Statistics Department University of California at Berkeley*.

Brier, G. W. (1950). “Verification of forecasts expressed in terms of probability”. In: *Monthly Weather Review* 78.1, pp. 1–3. ISSN: 0027-0644.

Burgess, S. et al. (2013). “Combining multiple imputation and meta-analysis with individual participant data”. In: *Statistics in Medicine* 32.26, pp. 4499–4514. ISSN: 1097-0258. DOI: 10.1002/sim.5844.

Burke, D. L. et al. (2017). “Meta-analysis using individual participant data: One-stage and two-stage approaches, and why they may differ”. In: *Statistics in Medicine* 36.5, pp. 855–875. ISSN: 1097-0258. DOI: 10.1002/sim.7141.

Carlsson, S. V. et al. (2011). “No excess mortality after prostate biopsy: Results from the European Randomized Study of Screening for Prostate Cancer”. In: *BJU International* 107.12, pp. 1912–1917. DOI: 10.1111/j.1464-410X.2010.09712.x.

Carter, H. B. (2012). “Active surveillance for prostate cancer: An underutilized opportunity for reducing harm”. In: *Journal of the National Cancer Institute. Monographs* 2012.45, pp. 175–183. DOI: 10.1093/jncimonographs/lgs036.

Ceylan, C. et al. (2014). “Comparison of 8, 10, 12, 16, 20 cores prostate biopsies in the determination of prostate cancer and the importance of prostate volume”. In: *Canadian Urological Association Journal = Journal de l’Association des urologues du Canada* 8.1-2, E81–E85. ISSN: 1911-6470. DOI: 10.5489/cuaj.510.

Chan, J. M. et al. (2016). “Selenium- or Vitamin E-Related Gene Variants, Interaction with Supplementation, and Risk of High-Grade Prostate Cancer in SELECT”. In: *Cancer Epidemiology, Biomarkers & Prevention* 25.7, pp. 1050–1058. ISSN: 1538-7755. DOI: 10.1158/1055-9965.EPI-16-0104.

Chatterjee, N. et al. (2016). “Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-level Information from External Big Data Sources”. In: *Journal of the American Statistical Association* 111.513, pp. 107–117. ISSN: 01621459. DOI: 10.1080/01621459.2015.1123157.

Chen, J. et al. (2006). “Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density”. In: *Journal of the National Cancer Institute* 98.17, pp. 1215–1226. ISSN: 1460-2105. DOI: 10.1093/jnci/djj332.

Cheng, W., J. M. G. Taylor, T. Gu, et al. (2019). “Informing a risk prediction model for binary outcomes with external coefficient information”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68.1, pp. 121–139. ISSN: 00359254. DOI: 10.1111/rssc.12306.

Cheng, W., J. M. G. Taylor, P. S. Vokonas, et al. (2018). “Improving estimation and prediction in linear regression incorporating external information from an established reduced model”. In: *Statistics in Medicine* 37.9, pp. 1515–1530. ISSN: 1097-0258. DOI: 10.1002/sim.7600.

Chiu, P. K. et al. (2017). “Additional benefit of using a risk-based selection for prostate biopsy: An analysis of biopsy complications in the Rotterdam section of the European Randomized Study of Screening for Prostate Cancer”. In: *BJU International* 120.3, pp. 394–400. DOI: 10.1111/bju.13913.

Christodoulou, E. et al. (2019). “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models”. In: *Journal of Clinical Epidemiology*. ISSN: 1878-5921. DOI: 10.1016/j.jclinepi.2019.02.004.

Chun, F. K.-H. et al. (2007). “Initial biopsy outcome prediction–head-to-head comparison of a logistic regression-based nomogram versus artificial neural network”. In: *European Urology* 51.5, pp. 1236–1243. ISSN: 1873-7560. DOI: 10.1016/j.eururo.2006.07.021.

Claus, E. B. et al. (1994). “Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction”. In: *Cancer* 73.3, pp. 643–651.

Claxton, K. (1999). “The irrelevance of inference: A decision-making approach to the stochastic evaluation of health care technologies”. In: *Journal of Health Economics* 18.3, pp. 341–364. ISSN: 01676296. DOI: 10.1016/S0167-6296(98)00039-3.

Cleveland, W. S. (1979). “Robust Locally Weighted Regression and Smoothing Scatterplots”. In: *Journal of the American Statistical Association* 74.368, pp. 829–836. ISSN: 01621459. DOI: 10.2307/2286407.

Cleveland, W. S. and S. J. Devlin (1988). “Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting”. In: *Journal of the American Statistical Association* 83.403, pp. 596–610. ISSN: 01621459. DOI: 10.1080/01621459.1988.10478639.

Cook, N. R. (2007). “Use and misuse of the receiver operating characteristic curve in risk prediction”. In: *Circulation* 115.7, pp. 928–935. ISSN: 0009-7322. DOI: 10.1161/CIRCULATIONAHA.106.672402.

Cook, N. R. (2008). “Comments on ’Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond’ by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929)”. In: *Statistics in Medicine* 27.2, pp. 191–195. ISSN: 1097-0258. DOI: 10.1002/sim.2987.

Cook, N. R. (2010). “Methods for evaluating novel biomarkers - a new paradigm”. In: *International Journal of Clinical Practice* 64.13, pp. 1723–1727. DOI: 10.1111/j.1742-1241.2010.02469.x.

Cooperberg, M. R. et al. (2011). “Active surveillance for prostate cancer: Progress and promise”. In: *Journal of Clinical Oncology* 29.27, pp. 3669–3676. ISSN: 1527-7755. DOI: 10.1200/JCO.2011.34.9738.

Cox, D. R. (1958). “Two Further Applications of a Model for Binary Regression”. In: *Biometrika* 45.3/4, pp. 562–565. ISSN: 00063444. DOI: 10.2307/2333203.

D’Agostino, R. B. (2006). “Risk prediction and finding new independent prognostic factors”. In: *Journal of Hypertension* 24.4, pp. 643–645. ISSN: 0263-6352. DOI: 10.1097/01.hjh.0000217845.57466.cc.

D’Agostino, R. B. et al. (2001). “Validation of the Framingham Coronary Heart Disease Prediction Scores: Results of a Multiple Ethnic Groups Investigation.” In: *JAMA* 286.2, pp. 180–187. DOI: 10.1001/jama.286.2.180.

Danneman, D. et al. (2015). “Gleason inflation 1998-2011: A registry study of 97,168 men”. In: *BJU International* 115.2, pp. 248–255. DOI: 10.1111/bju.12671.

Debray, T. P. A., K. G. M. Moons, G. M. A. Abo-Zaid, et al. (2013). “Individual participant data meta-analysis for a binary outcome: One-stage or two-stage?” In: *PloS One* 8.4, e60650. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0060650.

Debray, T. P. A., K. G. M. Moons, I. Ahmed, et al. (2013). “A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis”. In: *Statistics in Medicine* 32.18, pp. 3158–3180. ISSN: 1097-0258. DOI: 10.1002/sim.5732.

Debray, T. P. A., R. D. Riley, et al. (2015). “Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: Guidance on their use”. In: *PLoS Medicine* 12.10, e1001886. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001886.

DeFilippis, A. P. et al. (2015). “An analysis of calibration and discrimination among multiple cardiovascular risk scores in a modern multiethnic cohort”. In: *Annals of Internal Medicine* 162.4, pp. 266–275. ISSN: 1539-3704. DOI: 10.7326/M14-1281.

DeLong, E. R. et al. (1988). “Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach”. In: *Biometrics* 44.3, pp. 837–845. DOI: 10.2307/2531595.

Demidenko, E. (2012). “Confidence intervals and bands for the binormal ROC curve revisited”. In: *Journal of Applied Statistics* 39.1, pp. 67–79. ISSN: 0266-4763. DOI: 10.1080/02664763.2011.578616.

Dempster, A. P. et al. (1977). “Maximum Likelihood from Incomplete Data via the EM Algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–38.

DerSimonian, R. and N. Laird (1986). “Meta-analysis in clinical trials”. In: *Controlled Clinical Trials* 7.3, pp. 177–188. ISSN: 0197-2456. DOI: 10.1016/0197-2456(86)90046-2.

DeSantis, C. E. et al. (2019). “Cancer statistics for African Americans, 2019”. In: *CA: A Cancer Journal for Clinicians* 69.3, pp. 211–233. DOI: 10.3322/caac.21555.

Devroye, L. et al. (1996). *A Probabilistic Theory of Pattern Recognition*. Vol. 31. New York, NY: Springer New York. ISBN: 978-1-4612-6877-2. DOI: 10.1007/978-1-4612-0711-5.

Diamond, G. A. and J. S. Forrester (1979). “Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease”. In: *The New England Journal of Medicine* 300.24, pp. 1350–1358. DOI: 10.1056/NEJM197906143002402.

Eastham, J. A. et al. (1999). “Development of a nomogram that predicts the probability of a positive prostate biopsy in men with an abnormal digital rectal examination and a prostate-specific antigen between 0 and 4 ng/mL”. In: *Urology* 54.4, pp. 709–713. ISSN: 1527-9995.

Efron, B. et al. (2004). “Least angle regression”. In: *The Annals of Statistics* 32.2, pp. 407–499. DOI: 10.1214/009053604000000067.

Elwyn, G. et al. (2017). “A three-talk model for shared decision making: Multistage consultation process”. In: *BMJ (Clinical Research Ed.)* 359, j4891. ISSN: 1756-1833. DOI: 10.1136/bmj.j4891.

Epstein, J. I., W. C. Allsbrook, et al. (2005). “The 2005 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma”. In: *The American Journal of Surgical Pathology* 29.9, pp. 1228–1242. DOI: 10.1097/01.pas.0000173646.99337.b1.

Epstein, J. I., L. Egevad, et al. (2016). “The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System”. In: *The American Journal of Surgical Pathology* 40.2, pp. 244–252. DOI: 10.1097/PAS.0000000000000530.

Farago, A. and G. Lugosi (1993). “Strong universal consistency of neural network classifiers”. In: *IEEE Transactions on Information Theory* 39.4, pp. 1146–1151. ISSN: 00189448. DOI: 10.1109/18.243433.

Fenton, J. J. et al. (2018). “Prostate-Specific Antigen-Based Screening for Prostate Cancer: Evidence Report and Systematic Review for the US Preventive Services Task Force”. In: *JAMA* 319.18, pp. 1914–1931. DOI: 10.1001/jama.2018.3712.

Finne, P. et al. (2000). “Predicting the outcome of prostate biopsy in screen-positive men by a multilayer perceptron network”. In: *Urology* 56.3, pp. 418–422. ISSN: 1527-9995. DOI: 10.1016/S0090-4295(00)00672-5.

Gail, M. H., L. A. Brinton, et al. (1989). “Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually”. In: *Journal of the National Cancer Institute* 81.24, pp. 1879–1886. ISSN: 1460-2105. DOI: 10.1093/jnci/81.24.1879.

Gail, M. H. (2009). “Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model”. In: *Journal of the National Cancer Institute* 101.13, pp. 959–963. ISSN: 1460-2105. DOI: 10.1093/jnci/djp130.

Gail, M. H., J. P. Costantino, et al. (2007). “Projecting individualized absolute invasive breast cancer risk in African American women”. In: *Journal of the National Cancer Institute* 99.23, pp. 1782–1792. ISSN: 1460-2105. DOI: 10.1093/jnci/djm223.

Garzotto, M. et al. (2005). “Improved detection of prostate cancer using classification and regression tree analysis”. In: *Journal of Clinical Oncology* 23.19, pp. 4322–4329. ISSN: 1527-7755. DOI: 10.1200/JCO.2005.11.136.

Genders, T. S. S. et al. (2011). “A clinical prediction rule for the diagnosis of coronary artery disease: Validation, updating, and extension”. In: *European Heart Journal* 32.11, pp. 1316–1330. ISSN: 1522-9645. DOI: 10.1093/eurheartj/ehr014.

Gengsheng, Q. and L. Hotilovac (2008). “Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test”. In: *Statistical Methods in Medical Research* 17.2, pp. 207–221. ISSN: 1477-0334. DOI: 10.1177/0962280207087173.

Genuer, R. (2012). “Variance reduction in purely random forests”. In: *Journal of Nonparametric Statistics* 24.3, pp. 543–562. DOI: 10.1080/10485252.2012.677843.

Ghani, K. R. et al. (2005). “Trends in reporting Gleason score 1991 to 2001: Changes in the pathologist’s practice”. In: *European Urology* 47.2, pp. 196–201. ISSN: 1873-7560. DOI: 10.1016/j.eururo.2004.07.029.

Gibbons, R. J. et al. (2002). “ACC/AHA 2002 guideline update for exercise testing: Summary article. A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to Update the 1997 Exercise Testing Guidelines)”. In: *Journal of the American College of Cardiology* 40.8, pp. 1531–1540. ISSN: 0735-1097. DOI: 10.1016/S0735-1097(02)02164-2.

Gleason, D. F. (1966). “Classification of prostatic carcinomas”. In: *Cancer Chemotherapy Reports* 50.3, pp. 125–128.

Gordetsky, J. and J. I. Epstein (2016). “Grading of prostatic adenocarcinoma: Current state and prognostic implications”. In: *Diagnostic Pathology* 11, p. 25. DOI: 10.1186/s13000-016-0478-2.

Greenland, P. (2008). “Comments on ’Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond’ by M. J. Pencina, R. B. D’Agostino Sr, R. B. D’Agostino Jr, R. S. Vasan, Statistics in Medicine (DOI: 10.1002/sim.2929)”. In: *Statistics in Medicine* 27.2, pp. 188–190. ISSN: 1097-0258. DOI: 10.1002/sim.2976.

Greenland, P. and P. G. O'Malley (2005). "When is a new prediction marker useful? A consideration of lipoprotein-associated phospholipase A2 and C-reactive protein for stroke risk". In: *Archives of Internal Medicine* 165.21, pp. 2454–2456. ISSN: 0003-9926. DOI: 10.1001/archinte.165.21.2454.

Gregorio, E. P. et al. (2007). "Comparison between PSA density, free PSA percentage and PSA density in the transition zone in the detection of prostate cancer in patients with serum PSA between 4 and 10 ng/mL". In: *International Brazilian Journal of Urology* 33.2, pp. 151–160. ISSN: 1677-5538. DOI: 10.1590/S1677-55382007000200004.

Grill, S., D. P. Ankerst, et al. (2017). "Comparison of approaches for incorporating new information into existing risk prediction models". In: *Statistics in Medicine* 36.7, pp. 1134–1156. ISSN: 1097-0258. DOI: 10.1002/sim.7190.

Grill, S., M. Fallah, R. J. Leach, I. M. Thompson, S. Freedland, et al. (2015). "Incorporation of detailed family history from the Swedish Family Cancer Database into the PCPT risk calculator". In: *The Journal of Urology* 193.2, pp. 460–465. ISSN: 1527-3792. DOI: 10.1016/j.juro.2014.09.018.

Grill, S., M. Fallah, R. J. Leach, I. M. Thompson, K. Hemminki, et al. (2015). "A simple-to-use method incorporating genomic markers into prostate cancer risk prediction tools facilitated future validation". In: *Journal of Clinical Epidemiology* 68.5, pp. 563–573. ISSN: 1878-5921. DOI: 10.1016/j.jclinepi.2015.01.006.

Grossman, D. C. et al. (2018). "Screening for Prostate Cancer: US Preventive Services Task Force Recommendation Statement". In: *JAMA* 319.18, pp. 1901–1913. DOI: 10.1001/jama.2018.3710.

Gu, W. and M. S. Pepe (2009). "Estimating the capacity for improvement in risk prediction with a marker". In: *Biostatistics* 10.1, pp. 172–186. DOI: 10.1093/biostatistics/kxn025.

Gülkesen, K. H. et al. (2010). "Prediction of Prostate Cancer Using Decision Tree Algorithm". In: *Turkish Journal of Medical Sciences* 40.5, pp. 681–686. DOI: 10.3906/sag-0906-47.

Günther, F. and S. Fritsch (2010). "neuralnet: Training of Neural Networks". In: *The R Journal* 2.1, pp. 30–38. ISSN: 2073-4859. DOI: 10.32614/RJ-2010-006.

Gupta, S. et al. (2017). "Prostate Cancer: How Young is too Young?" In: *Current Urology* 9.4, pp. 212–215. ISSN: 1661-7649. DOI: 10.1159/000447143.

Györfi, L. et al. (2002). *A Distribution-Free Theory of Nonparametric Regression*. New York, NY: Springer New York. ISBN: 978-0-387-95441-7. DOI: 10.1007/b97848.

Hall, P. et al. (2004). "Nonparametric confidence intervals for receiver operating characteristic curves". In: *Biometrika* 91.3, pp. 743–750. ISSN: 00063444. DOI: 10.1093/biomet/91.3.743.

Hamdy, F. C. et al. (2016). "10-Year Outcomes after Monitoring, Surgery, or Radiotherapy for Localized Prostate Cancer". In: *The New England Journal of Medicine* 375.15, pp. 1415–1424. DOI: 10.1056/NEJMoa1606220.

Hasan, A. et al. (2016). "Fast Estimation of Multinomial Logit Models: R Package mnlogit". In: *Journal of Statistical Software* 75.3. ISSN: 1548-7660. DOI: 10.18637/jss.v075.i03.

Hastie, T. et al. (2009). *The Elements of Statistical Learning*. 2nd ed. New York, NY: Springer New York. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7.

Hendel, R. C., D. S. Berman, et al. (2009). “ACCF/ASNC/ACR/AHA/ASE/SCCT/SCMR/SNM 2009 appropriate use criteria for cardiac radionuclide imaging: A report of the American College of Cardiology Foundation Appropriate Use Criteria Task Force, the American Society of Nuclear Cardiology, the American College of Radiology, the American Heart Association, the American Society of Echocardiography, the Society of Cardiovascular Computed Tomography, the Society for Cardiovascular Magnetic Resonance, and the Society of Nuclear Medicine”. In: *Circulation* 119.22, e561–e587. ISSN: 0009-7322. DOI: 10.1161/CIRCULATIONAHA.109.192519.

Hendel, R. C., M. R. Patel, et al. (2006). “ACCF/ACR/SCCT/SCMR/ASNC/NASCI/SCAI/SIR 2006 appropriateness criteria for cardiac computed tomography and cardiac magnetic resonance imaging: A report of the American College of Cardiology Foundation Quality Strategic Directions Committee Appropriateness Criteria Working Group, American College of Radiology, Society of Cardiovascular Computed Tomography, Society for Cardiovascular Magnetic Resonance, American Society of Nuclear Cardiology, North American Society for Cardiac Imaging, Society for Cardiovascular Angiography and Interventions, and Society of Interventional Radiology”. In: *Journal of the American College of Cardiology* 48.7, pp. 1475–1497. ISSN: 0735-1097. DOI: 10.1016/j.jacc.2006.07.003.

Hense, H. (2003). “Framingham risk function overestimates risk of coronary heart disease in men and women from Germany—results from the MONICA Augsburg and the PROCAM cohorts”. In: *European Heart Journal* 24.10, pp. 937–945. ISSN: 1522-9645. DOI: 10.1016/S0195-668X(03)00081-2.

Herden, J. and L. Weissbach (2018). “Utilization of Active Surveillance and Watchful Waiting for localized prostate cancer in the daily practice”. In: *World Journal of Urology* 36.3, pp. 383–391. DOI: 10.1007/s00345-018-2175-0.

Heron, M. (2018). “Deaths: Leading Causes for 2016”. In: *National Vital Statistics Reports* 67.6, pp. 1–77.

Hilden, J. (1991). “The area under the ROC curve and its competitors”. In: *Medical Decision Making* 11.2, pp. 95–101. DOI: 10.1177/0272989X9101100204.

Hodge, K. K. et al. (1989). “Random systematic versus directed ultrasound guided transrectal core biopsies of the prostate”. In: *Journal of Urology* 142.1, pp. 71–74. ISSN: 0022-5347. DOI: 10.1016/S0022-5347(17)38664-0.

Hoeffding, W. (1948). “A Class of Statistics with Asymptotically Normal Distribution”. In: *The Annals of Mathematical Statistics* 19.3, pp. 293–325. URL: www.jstor.org/stable/2235637.

Hosmer, D. W. and S. Lemeshow (1980). “Goodness of fit tests for the multiple logistic regression model”. In: *Communications in Statistics - Theory and Methods* 9.10, pp. 1043–1069. ISSN: 0361-0926. DOI: 10.1080/03610928008827941.

Hosmer, D. W., S. Lemeshow, and R. X. Sturdivant (2013). *Applied logistic regression*. 3rd ed. Wiley series in probability and statistics. Hoboken: Wiley. ISBN: 9780470582473.

Hu, L.-Y. et al. (2016). “The distance function effect on k-nearest neighbor classification for medical datasets”. In: *SpringerPlus* 5.1, p. 1304. ISSN: 2193-1801. DOI: 10.1186/s40064-016-2941-7.

Huang, T. et al., eds. (2018). *Advances in neural networks – ISNN 2018: 15th International Symposium on Neural Networks, ISNN 2018, Minsk, Belarus, June 25-28, 2018, Proceedings*. Vol. 10878. LNCS sublibrary. SL 1, Theoretical computer science and general issues. Cham, Switzerland: Springer. ISBN: 978-3-319-92536-3.

Huang, Y. et al. (2007). “Evaluating the predictiveness of a continuous marker”. In: *Biometrics* 63.4, pp. 1181–1188. DOI: 10.1111/j.1541-0420.2007.00814.x.

Huang, Z. et al. (2016). “Bayesian reclassification statistics for assessing improvements in diagnostic accuracy”. In: *Statistics in Medicine* 35.15, pp. 2574–2592. ISSN: 1097-0258. DOI: 10.1002/sim.6899.

Jackson, D., I. R. White, J. B. Kostis, et al. (2009). “Systematically missing confounders in individual participant data meta-analysis of observational cohort studies”. In: *Statistics in Medicine* 28.8, pp. 1218–1237. ISSN: 1097-0258. DOI: 10.1002/sim.3540.

Jackson, D., I. R. White, and S. G. Thompson (2010). “Extending DerSimonian and Laird’s methodology to perform multivariate random effects meta-analyses”. In: *Statistics in Medicine* 29.12, pp. 1282–1297. ISSN: 1097-0258. DOI: 10.1002/sim.3602.

Jain, V. et al. (2012). “Doing meta-analysis in research: A systematic approach”. In: *Indian Journal of Dermatology, Venereology and Leprology* 78.3, pp. 242–250. DOI: 10.4103/0378-6323.95438.

Janssen, K. J. M., K. G. M. Moons, et al. (2008). “Updating methods improved the performance of a clinical prediction model in new patients”. In: *Journal of Clinical Epidemiology* 61.1, pp. 76–86. ISSN: 1878-5921. DOI: 10.1016/j.jclinepi.2007.04.018.

Janssen, K. J. M., Y. Vergouwe, et al. (2009). “A simple method to adjust clinical prediction models to local circumstances”. In: *Canadian Journal of Anaesthesia = Journal canadien d’anesthesie* 56.3, pp. 194–201. DOI: 10.1007/s12630-009-9041-x.

Jensen, L. J. and A. Bateman (2011). “The rise and fall of supervised machine learning techniques”. In: *Bioinformatics* 27.24, pp. 3331–3332. DOI: 10.1093/bioinformatics/btr585.

Jolani, S. et al. (2015). “Imputation of systematically missing predictors in an individual participant data meta-analysis: A generalized approach using MICE”. In: *Statistics in Medicine* 34.11, pp. 1841–1863. ISSN: 1097-0258. DOI: 10.1002/sim.6451.

Kacker, R. N. (2004). “Combining information from interlaboratory evaluations using a random effects model”. In: *Metrologia* 41.3, pp. 132–136. ISSN: 0026-1394. DOI: 10.1088/0026-1394/41/3/004.

Kang, D. D. et al. (2012). “MetaQC: Objective quality control and inclusion/exclusion criteria for genomic meta-analysis”. In: *Nucleic Acids Research* 40.2, e15. DOI: 10.1093/nar/gkr1071.

Kaur, J. S. et al. (2004). “Can the Gail model be useful in American Indian and Alaska Native populations?” In: *Cancer* 100.5, pp. 906–912. DOI: 10.1002/cncr.20047.

Kerr, K. F., M. D. Brown, et al. (2016). “Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use”. In: *Journal of Clinical Oncology* 34.21, pp. 2534–2540. ISSN: 1527-7755. DOI: 10.1200/JCO.2015.65.5654.

Kerr, K. F., R. L. McClelland, et al. (2011). “Evaluating the incremental value of new biomarkers with integrated discrimination improvement”. In: *American Journal of Epidemiology* 174.3, pp. 364–374. DOI: 10.1093/aje/kwr086.

Kibbe, W. et al. (2017). “Cancer Informatics: New Tools for a Data-Driven Age in Cancer Research”. In: *Cancer Research* 77.21, e1–e2. DOI: 10.1158/0008-5472.CAN-17-2212.

Kiciński, M. et al. (2011). “An epidemiological reappraisal of the familial aggregation of prostate cancer: A meta-analysis”. In: *PloS One* 6.10, e27130. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0027130.

Kline, D. M. and V. L. Berardi (2005). “Revisiting squared-error and cross-entropy functions for training neural network classifiers”. In: *Neural Computing and Applications* 14.4, pp. 310–318. ISSN: 0941-0643. DOI: 10.1007/s00521-005-0467-y.

Klotz, L. et al. (2015). “Long-term follow-up of a large active surveillance cohort of patients with prostate cancer”. In: *Journal of Clinical Oncology* 33.3, pp. 272–277. ISSN: 1527-7755. DOI: 10.1200/JCO.2014.55.1192.

Kondo, M. et al. (2018). “Factors predicting early postpartum glucose intolerance in Japanese women with gestational diabetes mellitus: Decision-curve analysis”. In: *Diabetic Medicine* 35.8, pp. 1111–1117. DOI: 10.1111/dme.13657.

Kondofersky, I. et al. (2016). “Three general concepts to improve risk prediction: Good data, wisdom of the crowd, recalibration”. In: *F1000Research* 5, p. 2671. ISSN: 2046-1402. DOI: 10.12688/f1000research.8680.1.

Koopman, L. et al. (2008). “Comparison of methods of handling missing data in individual patient data meta-analyses: An empirical example on antibiotics in children with acute otitis media”. In: *American Journal of Epidemiology* 167.5, pp. 540–545. DOI: 10.1093/aje/kwm341.

Kovačić, J. and V. M. Varnai (2016). “A graphical model approach to systematically missing data in meta-analysis of observational studies”. In: *Statistics in Medicine* 35.24, pp. 4443–4458. ISSN: 1097-0258. DOI: 10.1002/sim.7010.

Kowall, B. et al. (2013). “Use of areas under the receiver operating curve (AROCs) and some caveats”. In: *International Journal of Public Health* 58.3, pp. 485–488. DOI: 10.1007/s00038-012-0401-x.

Kramer, A. A. and J. E. Zimmerman (2007). “Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited”. In: *Critical Care Medicine* 35.9, pp. 2052–2056. ISSN: 0090-3493. DOI: 10.1097/01.CCM.0000275267.64078.B0.

Kranse, R. et al. (2008). “A graphical device to represent the outcomes of a logistic regression analysis”. In: *The Prostate* 68.15, pp. 1674–1680. ISSN: 1097-0045. DOI: 10.1002/pros.20840.

Kruppa, J., Y. Liu, et al. (2014). “Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory”. In: *Biometrical Journal* 56.4, pp. 534–563. DOI: 10.1002/bimj.201300068.

Kruppa, J., A. Ziegler, et al. (2012). “Risk estimation and risk prediction using machine-learning methods”. In: *Human Genetics* 131.10, pp. 1639–1654. DOI: 10.1007/s00439-012-1194-y.

Kryvenko, O. N. and J. I. Epstein (2016). “Prostate Cancer Grading: A Decade After the 2005 Modified Gleason Grading System”. In: *Archives of Pathology & Laboratory Medicine* 140.10, pp. 1140–1152. ISSN: 1543-2165. DOI: 10.5858/arpa.2015-0487-SA.

Kuhn, M. (2008). “Building Predictive Models in R Using the caret Package”. In: *Journal of Statistical Software* 28.5. ISSN: 1548-7660. DOI: 10.18637/jss.v028.i05.

Kuhn, M. and K. Johnson (2013). *Applied Predictive Modeling*. New York, NY: Springer New York. ISBN: 978-1-4614-6848-6. DOI: 10.1007/978-1-4614-6849-3.

Kuo, S.-C. et al. (2013). “Chinese nomogram to predict probability of positive initial prostate biopsy: A study in Taiwan region”. In: *Asian Journal of Andrology* 15.6, pp. 780–784. ISSN: 1745-7262. DOI: 10.1038/aja.2013.100.

Lahiri, K. and L. Yang (2017). “Confidence Bands for ROC Curves With Serially Dependent Data”. In: *Journal of Business & Economic Statistics* 36.1, pp. 115–130. ISSN: 0735-0015. DOI: 10.1080/07350015.2015.1073593.

Lamy, P.-J. et al. (2018). “Family history of breast cancer increases the risk of prostate cancer: Results from the EPICAP study”. In: *Oncotarget* 9.34, pp. 23661–23669. DOI: 10.18632/oncotarget.25320.

Lay, N. et al. (2017). “Detection of prostate cancer in multiparametric MRI using random forest with instance weighting”. In: *Journal of Medical Imaging* 4.2, p. 024506. ISSN: 2329-4302. DOI: 10.1117/1.JMI.4.2.024506.

Li, J. et al. (2013). “Multicategory reclassification statistics for assessing improvements in diagnostic accuracy”. In: *Biostatistics* 14.2, pp. 382–394. DOI: 10.1093/biostatistics/kxs047.

Liang, Y. et al. (2011). “Prospective evaluation of operating characteristics of prostate cancer detection biomarkers”. In: *The Journal of Urology* 185.1, pp. 104–110. ISSN: 1527-3792. DOI: 10.1016/j.juro.2010.08.088.

Lilja, H. et al. (2011). “Prediction of significant prostate cancer diagnosed 20 to 30 years later with a single measure of prostate-specific antigen at or before age 50”. In: *Cancer* 117.6, pp. 1210–1219. DOI: 10.1002/cncr.25568.

Lin, Y. and Y. Jeon (2006). “Random Forests and Adaptive Nearest Neighbors”. In: *Journal of the American Statistical Association* 101.474, pp. 578–590. ISSN: 01621459. DOI: 10.1198/016214505000001230.

Lindström, S. et al. (2012). “Common genetic variants in prostate cancer risk prediction–results from the NCI Breast and Prostate Cancer Cohort Consortium (BPC3)”. In: *Cancer Epidemiology, Biomarkers & Prevention* 21.3, pp. 437–444. ISSN: 1538-7755. DOI: 10.1158/1055-9965.EPI-11-1038.

Liu, J. et al. (2004). “Predictive value for the Chinese population of the Framingham CHD risk assessment tool compared with the Chinese Multi-Provincial Cohort Study”. In: *JAMA* 291.21, pp. 2591–2599. DOI: 10.1001/jama.291.21.2591.

Loeb, S. et al. (2013). “Systematic review of complications of prostate biopsy”. In: *European Urology* 64.6, pp. 876–892. ISSN: 1873-7560. DOI: 10.1016/j.eururo.2013.05.049.

Macinnis, R. J. et al. (2011). "A risk prediction algorithm based on family history and common genetic variants: Application to prostate cancer with potential clinical impact". In: *Genetic Epidemiology* 35.6, pp. 549–556. ISSN: 1098-2272. DOI: 10.1002/gepi.20605.

Malley, J. D., J. Kruppa, et al. (2012). "Probability machines: Consistent probability estimation using nonparametric learning machines". In: *Methods of Information in Medicine* 51.1, pp. 74–81. DOI: 10.3414/ME00-01-0052.

Malley, J. D., K. G. Malley, et al. (2011). *Statistical Learning for Biomedical Data*. Cambridge: Cambridge University Press. ISBN: 9780511975820. DOI: 10.1017/CBO9780511975820.

Marrugat, J. et al. (2003). "An adaptation of the Framingham coronary heart disease risk function to European Mediterranean areas". In: *Journal of Epidemiology and Community Health* 57.8, pp. 634–638. DOI: 10.1136/jech.57.8.634.

Martínez-Camblor, P. et al. (2018). "Efficient nonparametric confidence bands for receiver operating-characteristic curves". In: *Statistical Methods in Medical Research* 27.6, pp. 1892–1908. ISSN: 1477-0334. DOI: 10.1177/0962280216672490.

Matsuno, R. K. et al. (2011). "Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women". In: *Journal of the National Cancer Institute* 103.12, pp. 951–961. ISSN: 1460-2105. DOI: 10.1093/jnci/djr154.

Mease, D. and A. J. Wyner (2008). "Evidence Contrary to the Statistical View of Boosting". In: *Journal of Machine Learning Research* 9, pp. 131–156.

Meinshausen, N. (2006). "Quantile Regression Forests". In: *Journal of Machine Learning Research* 7, pp. 983–999.

Minne, L. et al. (2012a). "Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment". In: *Intensive Care Medicine* 38.1, pp. 40–46. ISSN: 1432-1238. DOI: 10.1007/s00134-011-2390-2.

Minne, L. et al. (2012b). "Statistical process control for validating a classification tree model for predicting mortality–a novel approach towards temporal validation". In: *Journal of Biomedical Informatics* 45.1, pp. 37–44. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2011.08.015.

Moher, D. et al. (2009). "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement". In: *PLoS Medicine* 6.7, e1000097. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1000097.

Moyer, V. A. (2012). "Screening for prostate cancer: U.S. Preventive Services Task Force recommendation statement". In: *Annals of Internal Medicine* 157.2, pp. 120–134. ISSN: 1539-3704. DOI: 10.7326/0003-4819-157-2-201207170-00459.

Nagelkerke, N. (1991). "A Note on a General Definition of the Coefficient of Determination". In: *Biometrika* 78.3, pp. 691–692. ISSN: 00063444. DOI: 10.1093/biomet/78.3.691.

Nam, R. K. et al. (2007). "Assessing individual risk for prostate cancer". In: *Journal of Clinical Oncology* 25.24, pp. 3582–3588. ISSN: 1527-7755. DOI: 10.1200/JCO.2007.10.6450.

Neath, A. A. and J. E. Cavanaugh (2012). "The Bayesian information criterion: Background, derivation, and applications". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.2, pp. 199–203. ISSN: 19395108. DOI: 10.1002/wics.199.

Nomura, M. et al. (2012). "Development and external validation of a nomogram for predicting cancer probability at initial prostate biopsy using the life expectancy- and prostate volume-adjusted biopsy scheme". In: *Prostate Cancer and Prostatic Diseases* 15.2, pp. 202–209. ISSN: 1476-5608. DOI: 10.1038/pcan.2011.62.

Noone, A. M. et al. (2018). *SEER Cancer Statistics Review, 1975-2015: Based on November 2017 SEER data submission, posted to the SEER web site, April 2018*. National Cancer Institute. Bethesda, MD. URL: https://seer.cancer.gov/csr/1975_2015/ (visited on 02/25/2019).

Nunzio, C. de et al. (2018). "Repeat prostate-specific antigen (PSA) test before prostate biopsy: A 20% decrease in PSA values is associated with a reduced risk of cancer and particularly of high-grade cancer". In: *BJU International* 122.1, pp. 83–88. DOI: 10.1111/bju.14197.

Optenberg, S. A. et al. (1997). "Development of a decision-making tool to predict risk of prostate cancer: The cancer of the prostate risk index (CAPRI) test". In: *Urology* 50.5, pp. 665–672. ISSN: 1527-9995. DOI: 10.1016/S0090-4295(97)00451-2.

Parker, C. (2004). "Active surveillance: Towards a new paradigm in the management of early prostate cancer". In: *The Lancet Oncology* 5.2, pp. 101–106. ISSN: 14702045. DOI: 10.1016/S1470-2045(04)01384-1.

Parmigiani, G. et al. (1998). "Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2". In: *American Journal of Human Genetics* 62.1, pp. 145–158. ISSN: 0002-9297. DOI: 10.1086/301670.

Pastor-Barriuso, R. et al. (2013). "Recalibration of the Gail model for predicting invasive breast cancer risk in Spanish women: A population-based cohort study". In: *Breast Cancer Research and Treatment* 138.1, pp. 249–259. DOI: 10.1007/s10549-013-2428-y.

Paul, P. et al. (2013). "Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets". In: *Statistics in Medicine* 32.1, pp. 67–80. ISSN: 1097-0258. DOI: 10.1002/sim.5525.

Pavlou, M. et al. (2015). "A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes". In: *BMC Medical Research Methodology* 15, p. 59. ISSN: 1471-2288. DOI: 10.1186/s12874-015-0046-6.

Pencina, M. J. and R. B. D'Agostino (2015). "Evaluating Discrimination of Risk Prediction Models: The C Statistic". In: *JAMA* 314.10, pp. 1063–1064. DOI: 10.1001/jama.2015.11082.

Pencina, M. J., R. B. D'Agostino, and O. V. Demler (2012). "Novel metrics for evaluating improvement in discrimination: Net reclassification and integrated discrimination improvement for normal variables and nested models". In: *Statistics in Medicine* 31.2, pp. 101–113. ISSN: 1097-0258. DOI: 10.1002/sim.4348.

Pencina, M. J., R. B. D'Agostino, and E. W. Steyerberg (2011). "Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers". In: *Statistics in Medicine* 30.1, pp. 11–21. ISSN: 1097-0258. DOI: 10.1002/sim.4085.

Pencina, M. J., R. B. D'Agostino, and R. S. Vasan (2008). "Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond". In: *Statistics in Medicine* 27.2, pp. 157–172. ISSN: 1097-0258. DOI: 10.1002/sim.2929.

Pennello, G. et al. (2016). “Comparing diagnostic tests on benefit-risk”. In: *Journal of Biopharmaceutical Statistics* 26.6, pp. 1083–1097. DOI: 10.1080/10543406.2016.1226335.

Pepe, M. S., Z. Feng, and W. Gu (2008). “Comments on ’Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond’ by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929)”. In: *Statistics in Medicine* 27.2, pp. 173–181. ISSN: 1097-0258. DOI: 10.1002/sim.2991.

Pepe, M. S., Z. Feng, Y. Huang, et al. (2008). “Integrating the predictiveness of a marker with its performance as a classifier”. In: *American Journal of Epidemiology* 167.3, pp. 362–368. DOI: 10.1093/aje/kwm305.

Pepe, M. S. and H. E. Janes (2008). “Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer”. In: *Journal of the National Cancer Institute* 100.14, pp. 978–979. ISSN: 1460-2105. DOI: 10.1093/jnci/djn215.

Petrosino, A. J. (2016). “Specifying Inclusion Criteria for a Meta-Analysis”. In: *Evaluation Review* 19.3, pp. 274–293. DOI: 10.1177/0193841X9501900303.

Pinsky, P. F. et al. (2014). “Mortality and complications after prostate biopsy in the Prostate, Lung, Colorectal and Ovarian Cancer Screening (PLCO) trial”. In: *BJU International* 113.2, pp. 254–259. DOI: 10.1111/bju.12368.

Pölsterl, S. et al. (2016). “Heterogeneous ensembles for predicting survival of metastatic, castrate-resistant prostate cancer patients”. In: *F1000Research* 5, p. 2676. ISSN: 2046-1402. DOI: 10.12688/f1000research.8231.1.

Prasath, V. B. S. et al. (2017). *Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier - A Review*. URL: https://arxiv.org/pdf/1708.04321 (visited on 06/11/2019).

Pulleyblank, R. et al. (2013). “Decision curve analysis for assessing the usefulness of tests for making decisions to treat: An application to tests for prodromal psychosis”. In: *Psychological Assessment* 25.3, pp. 730–737. DOI: 10.1037/a0032394.

Qian, C. et al. (2016). “In vivo MRI based prostate cancer localization with random forests and auto-context model”. In: *Computerized Medical Imaging and Graphics* 52, pp. 44–57. DOI: 10.1016/j.compmedimag.2016.02.001.

Raji, O. Y. et al. (2010). “Incorporation of a genetic factor into an epidemiologic model for prediction of individual risk of lung cancer: The Liverpool Lung Project”. In: *Cancer Prevention Research* 3.5, pp. 664–669. DOI: 10.1158/1940-6207.CAPR-09-0141.

Rampun, A. et al. (2016). “Computer aided diagnosis of prostate cancer: A texton based approach”. In: *Medical Physics* 43.10, pp. 5412–5425. DOI: 10.1118/1.4962031.

Resche-Rigon, M. et al. (2013). “Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data”. In: *Statistics in Medicine* 32.28, pp. 4890–4905. ISSN: 1097-0258. DOI: 10.1002/sim.5894.

Ridker, P. M. et al. (2007). “Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: The Reynolds Risk Score”. In: *JAMA* 297.6, pp. 611–619. DOI: 10.1001/jama.297.6.611.

Riley, R. D., K. R. Abrams, et al. (2007). “Bivariate random-effects meta-analysis and the estimation of between-study correlation”. In: *BMC Medical Research Methodology* 7, p. 3. ISSN: 1471-2288. DOI: 10.1186/1471-2288-7-3.

Riley, R. D., P. C. Lambert, et al. (2010). “Meta-analysis of individual participant data: Rationale, conduct, and reporting”. In: *BMJ (Clinical Research Ed.)* 340, p. c221. ISSN: 1756-1833. DOI: 10.1136/bmj.c221.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press. ISBN: 0521460867.

Robin, X. et al. (2011). “pROC: An open-source package for R and S+ to analyze and compare ROC curves”. In: *BMC Bioinformatics* 12, p. 77. DOI: 10.1186/1471-2105-12-77.

Robins, J. M. et al. (1994). “Estimation of Regression Coefficients When Some Regressors Are Not Always Observed”. In: *Journal of the American Statistical Association* 89.427, pp. 846–866. ISSN: 01621459. DOI: 10.2307/2290910.

Roobol, M. J. et al. (2013). “A Calculator for Prostate Cancer Risk 4 Years After an Initially Negative Screen: Findings from ERSPC Rotterdam”. In: *European Urology* 63.4, pp. 627–633. ISSN: 1873-7560. DOI: 10.1016/j.eururo.2012.07.029.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley series in probability and mathematical statistics. New York and Chichester: Wiley. ISBN: 0-471-08705-X.

Ryan, F. and S. Cunningham (2014). “Shared decision making in healthcare”. In: *Faculty Dental Journal* 5.3, pp. 124–127. ISSN: 2042-6852. DOI: 10.1308/204268514X14017784505970.

Samaratunga, H. et al. (2017). “The Evolution of Gleason Grading of Prostate Cancer”. In: *Journal of Diagnostic Pathology* 12.1, pp. 5–11. DOI: 10.4038/jdp.v12i1.7732.

Scarpato, K. R. and P. C. Albertsen (2016). “Prostate-Specific Antigen Screening Guidelines”. In: *Prostate Cancer*. Elsevier, pp. 111–116. ISBN: 9780128000779. DOI: 10.1016/B978-0-12-800077-9.00013-X.

Schafer, J. L. and J. W. Graham (2002). “Missing data: Our view of the state of the art”. In: *Psychological Methods* 7.2, pp. 147–177. DOI: 10.1037//1082-989X.7.2.147.

Schwarz, G. (1978). “Estimating the Dimension of a Model”. In: *The Annals of Statistics* 6.2, pp. 461–464. URL: www.jstor.org/stable/2958889.

Shao, J. (1997). “An asymptotic theory for linear model selection.” In: *Statistica Sinica* 7.2, pp. 221–242. URL: www.jstor.org/stable/24306073.

Shapiro, A. R. (1977). “The evaluation of clinical predictions. A method and initial application”. In: *The New England Journal of Medicine* 296.26, pp. 1509–1514. DOI: 10.1056/NEJM197706302962607.

Shariat, S. F. et al. (2009). “Critical review of prostate cancer predictive tools”. In: *Future Oncology* 5.10, pp. 1555–1584. DOI: 10.2217/fon.09.121.

Siegel, R. L. et al. (2019). “Cancer statistics, 2019”. In: *CA: A Cancer Journal for Clinicians* 69.1, pp. 7–34. DOI: 10.3322/caac.21551.

Sigman, M. (2011). “A meta-analysis of meta-analyses”. In: *Fertility and Sterility* 96.1, pp. 11–14. DOI: 10.1016/j.fertnstert.2011.05.029.

Skates, S. J. et al. (2001). “Screening Based on the Risk of Cancer Calculation From Bayesian Hierarchical Changepoint and Mixture Models of Longitudinal Markers”. In: *Journal of the*

American Statistical Association 96.454, pp. 429–439. ISSN: 01621459. DOI: 10.1198/016214501753168145.

Skrondal, A. and S. Rabe-Hesketh (2009). “Prediction in multilevel generalized linear models”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172.3, pp. 659–687. ISSN: 09641998. DOI: 10.1111/j.1467-985X.2009.00587.x.

Smith, E. B. et al. (2002). “Gleason scores of prostate biopsy and radical prostatectomy specimens over the past 10 years: Is there evidence for systematic upgrading?” In: *Cancer* 94.8, pp. 2282–2287. DOI: 10.1002/cncr.10457.

Song, Y. et al. (2018). “Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric MRI”. In: *Journal of Magnetic Resonance Imaging* 48.6, pp. 1570–1577. DOI: 10.1002/jmri.26047.

Sprague, S. et al. (2009). “Multicenter collaboration in observational research: Improving generalizability and efficiency”. In: *The Journal of Bone and Joint Surgery. American Volume* 91 Suppl 3, pp. 80–86. ISSN: 1535-1386. DOI: 10.2106/JBJS.H.01623.

Stamey, T. A. et al. (1987). “Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate”. In: *The New England Journal of Medicine* 317.15, pp. 909–916. DOI: 10.1056/NEJM198710083171501.

Stephan, C. et al. (2002). “Multicenter evaluation of an artificial neural network to increase the prostate cancer detection rate and reduce unnecessary biopsies”. In: *Clinical Chemistry* 48.8, pp. 1279–1287. ISSN: 0009-9147.

Steyerberg, E. W. (2009). *Clinical Prediction Models*. New York, NY: Springer New York. ISBN: 978-0-387-77243-1. DOI: 10.1007/978-0-387-77244-8.

Steyerberg, E. W. et al. (2010). “Assessing the performance of prediction models: A framework for traditional and novel measures”. In: *Epidemiology* 21.1, pp. 128–138. DOI: 10.1097/EDE.0b013e3181c30fb2.

Stiggelbout, A. M. et al. (2012). “Shared decision making: Really putting patients at the centre of healthcare”. In: *BMJ (Clinical Research Ed.)* 344, e256. ISSN: 1756-1833. DOI: 10.1136/bmj.e256.

Stone, C. J. (1977). “Consistent Nonparametric Regression.” In: *The Annals of Statistics* 5.4, pp. 595–620. URL: www.jstor.org/stable/2958783.

Strobl, A. N., I. M. Thompson, et al. (2015). “The Next Generation of Clinical Decision Making Tools: Development of a Real-Time Prediction Tool for Outcome of Prostate Biopsy in Response to a Continuously Evolving Prostate Cancer Landscape”. In: *The Journal of Urology* 194.1, pp. 58–64. ISSN: 1527-3792. DOI: 10.1016/j.juro.2015.01.092.

Strobl, A. N., A. J. Vickers, et al. (2015). “Improving patient prostate cancer risk assessment: Moving from static, globally-applied to dynamic, practice-specific risk calculators”. In: *Journal of Biomedical Informatics* 56, pp. 87–93. ISSN: 1532-0480. DOI: 10.1016/j.jbi.2015.05.001.

Su, H. et al. (2009). “Empirical Likelihood-Based Confidence Interval of ROC Curves”. In: *Statistics in Biopharmaceutical Research* 1.4, pp. 407–414. ISSN: 1946-6315. DOI: 10.1198/sbr.2009.0044.

Sun, X. and W. Xu (2014). “Fast Implementation of DeLong’s Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves”. In: *IEEE Signal Processing Letters* 21.11, pp. 1389–1393. ISSN: 1070-9908. DOI: 10.1109/LSP.2014.2337313.

Tabak, E. R. et al. (1991). “Definition and yield of inclusion criteria for a meta-analysis of patient education studies in clinical preventive services”. In: *Evaluation & the health professions* 14.4, pp. 388–411. ISSN: 0163-2787. DOI: 10.1177/016327879101400402.

Takeuchi, T. et al. (2018). “Prediction of prostate cancer by deep learning with multilayer artificial neural network”. In: *Canadian Urological Association Journal = Journal de l’Association des urologues du Canada* 13.5, E145–E150. ISSN: 1911-6470. DOI: 10.5489/cuaj.5526.

Talluri, R. and S. Shete (2016). “Using the weighted area under the net benefit curve for decision curve analysis”. In: *BMC Medical Informatics and Decision Making* 16, p. 94. ISSN: 1472-6947. DOI: 10.1186/s12911-016-0336-x.

Thompson, I. M., D. P. Ankerst, et al. (2006). “Assessing prostate cancer risk: Results from the Prostate Cancer Prevention Trial”. In: *Journal of the National Cancer Institute* 98.8, pp. 529–534. ISSN: 1460-2105. DOI: 10.1093/jnci/djj131.

Thompson, I. M., P. J. Goodman, C. M. Tangen, M. S. Lucia, et al. (2003). “The influence of finasteride on the development of prostate cancer”. In: *The New England Journal of Medicine* 349.3, pp. 215–224. DOI: 10.1056/NEJMoa030660.

Thompson, I. M., P. J. Goodman, C. M. Tangen, H. L. Parnes, et al. (2013). “Long-term survival of participants in the prostate cancer prevention trial”. In: *The New England Journal of Medicine* 369.7, pp. 603–610. DOI: 10.1056/NEJMoa1215932.

Thompson, I. M., R. J. Leach, et al. (2014). “Focusing PSA testing on detection of high-risk prostate cancers by incorporating patient preferences into decision making”. In: *JAMA* 312.10, pp. 995–996. DOI: 10.1001/jama.2014.9680.

Tibshirani, R. (1996). “Regression Shrinkage and Selection via the Lasso.” In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288. URL: http://www.jstor.org/stable/2346178.

Tierney, J. F. et al. (2015). “Individual Participant Data (IPD) Meta-analyses of Randomised Controlled Trials: Guidance on Their Use”. In: *PLoS Medicine* 12.7, e1001855. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001855.

Tjur, T. (2009). “Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination”. In: *The American Statistician* 63.4, pp. 366–372. ISSN: 00031305. DOI: 10.1198/tast.2009.08210.

Torre, L. A. et al. (2016). “Global Cancer Incidence and Mortality Rates and Trends–An Update”. In: *Cancer Epidemiology, Biomarkers & Prevention* 25.1, pp. 16–27. ISSN: 1538-7755. DOI: 10.1158/1055-9965.EPI-15-0578.

Trottier, G. et al. (2011). “Comparison of risk calculators from the Prostate Cancer Prevention Trial and the European Randomized Study of Screening for Prostate Cancer in a contemporary Canadian cohort”. In: *BJU International* 108.8b, E237–E244. DOI: 10.1111/j.1464-410X.2011.10207.x.

Tyrer, J. et al. (2004). “A breast cancer prediction model incorporating familial and personal risk factors”. In: *Statistics in Medicine* 23.7, pp. 1111–1130. ISSN: 1097-0258. DOI: 10.1002/sim.1668.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). “mice: Multivariate Imputation by Chained Equations in R”. In: *Journal of Statistical Software* 45.3. ISSN: 1548-7660. DOI: 10.18637/jss.v045.i03.

van den Bergh, R. C. N. et al. (2008). “The Prostate Cancer Prevention Trial and European Randomized Study of Screening for Prostate Cancer risk calculators indicating a positive prostate biopsy: A comparison”. In: *BJU International* 102.9, pp. 1068–1073. DOI: 10.1111/j.1464-410X.2008.07940.x.

van Houwelingen, H. C. et al. (2002). “Advanced methods in meta-analysis: Multivariate approach and meta-regression”. In: *Statistics in Medicine* 21.4, pp. 589–624. ISSN: 1097-0258. DOI: 10.1002/sim.1040.

Venables, W. N. and B. D. Ripley (2002). *Modern applied statistics with S*. 4th ed. Statistics and computing. New York, NY: Springer New York. ISBN: 0-387-95457-0.

Vickers, A. J., A. M. Cronin, et al. (2008). “Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers”. In: *BMC Medical Informatics and Decision Making* 8, p. 53. ISSN: 1472-6947. DOI: 10.1186/1472-6947-8-53.

Vickers, A. J. and E. B. Elkin (2006). “Decision curve analysis: A novel method for evaluating prediction models”. In: *Medical Decision Making* 26.6, pp. 565–574. DOI: 10.1177/0272989X06295361.

Viechtbauer, W. (2010). “Conducting meta-analyses in R with the metafor package”. In: *Journal of Statistical Software* 36.3, pp. 1–48. ISSN: 1548-7660. DOI: 10.18637/jss.v036.i03.

Virtanen, A. et al. (1999). “Estimation of prostate cancer probability by logistic regression: Free and total prostate-specific antigen, digital rectal examination, and heredity are significant variables”. In: *Clinical Chemistry* 45.7, pp. 987–994. ISSN: 0009-9147.

Vrieze, S. I. (2012). “Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC)”. In: *Psychological Methods* 17.2, pp. 228–243. DOI: 10.1037/a0027127.

Wacholder, S. et al. (2010). “Performance of common genetic variants in breast-cancer risk models”. In: *The New England Journal of Medicine* 362.11, pp. 986–993. DOI: 10.1056/NEJMoa0907727.

Wang, T. J. et al. (2006). “Multiple biomarkers for the prediction of first major cardiovascular events and death”. In: *The New England Journal of Medicine* 355.25, pp. 2631–2639. DOI: 10.1056/NEJMoa055373.

Wang, Z. et al. (2018). “Automated Detection of Clinically Significant Prostate Cancer in mp-MRI Images Based on an End-to-End Deep Neural Network”. In: *IEEE Transactions on Medical Imaging* 37.5, pp. 1127–1139. DOI: 10.1109/TMI.2017.2789181.

Ware, J. H. and T. Cai (2008). “Comments on ’Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond’ by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929)". In: *Statistics in Medicine* 27.2, pp. 185–187. ISSN: 1097-0258. DOI: 10.1002/sim.2985.

Welch, H. G. and P. C. Albertsen (2009). "Prostate cancer diagnosis and treatment after the introduction of prostate-specific antigen screening: 1986-2005". In: *Journal of the National Cancer Institute* 101.19, pp. 1325–1329. ISSN: 1460-2105. DOI: 10.1093/jnci/djp278.

White, I. R. et al. (2011). "Multiple imputation using chained equations: Issues and guidance for practice". In: *Statistics in Medicine* 30.4, pp. 377–399. ISSN: 1097-0258. DOI: 10.1002/sim.4067.

Williams, S. B. et al. (2012). "Selective detection of histologically aggressive prostate cancer: An Early Detection Research Network Prediction model to reduce unnecessary prostate biopsies with validation in the Prostate Cancer Prevention Trial". In: *Cancer* 118.10, pp. 2651–2658. DOI: 10.1002/cncr.26396.

Wilson, P. W. et al. (1998). "Prediction of coronary heart disease using risk factor categories". In: *Circulation* 97.18, pp. 1837–1847. ISSN: 0009-7322.

Wu, Y. et al. (2015). "Developing a utility decision framework to evaluate predictive models in breast cancer risk estimation". In: *Journal of Medical Imaging* 2.4, p. 041005. ISSN: 2329-4302. DOI: 10.1117/1.JMI.2.4.041005.

Wynants, L. et al. (2016). "Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study". In: *Statistical Methods in Medical Research*. ISSN: 1477-0334. DOI: 10.1177/0962280216668555.

Xu, J. et al. (2009). "Estimation of absolute risk for prostate cancer using genetic markers and family history". In: *The Prostate* 69.14, pp. 1565–1572. ISSN: 1097-0045. DOI: 10.1002/pros.21002.

Yanke, B. V. et al. (2006). "African-American race is a predictor of prostate cancer detection: Incorporation into a pre-biopsy nomogram". In: *BJU International* 98.4, pp. 783–787. DOI: 10.1111/j.1464-410X.2006.06388.x.

Yates, J. (1982). "External correspondence: Decompositions of the mean probability score". In: *Organizational Behavior and Human Performance* 30.1, pp. 132–156. ISSN: 00305073. DOI: 10.1016/0030-5073(82)90237-9.

Zareba, P. et al. (2009). "The impact of the 2005 International Society of Urological Pathology (ISUP) consensus on Gleason grading in contemporary practice". In: *Histopathology* 55.4, pp. 384–391. ISSN: 1365-2559. DOI: 10.1111/j.1365-2559.2009.03405.x.

Zastrow, S. et al. (2015). "Decision curve analysis and external validation of the postoperative Karakiewicz nomogram for renal cell carcinoma based on a large single-center study cohort". In: *World Journal of Urology* 33.3, pp. 381–388. DOI: 10.1007/s00345-014-1321-6.

Zheng, J. et al. (2012). "Predictive performance of prostate cancer risk in Chinese men using 33 reported prostate cancer risk-associated SNPs". In: *The Prostate* 72.5, pp. 577–583. ISSN: 1097-0045. DOI: 10.1002/pros.21462.

Zhou, X.-H. and G. Qin (2005). "Improved confidence intervals for the sensitivity at a fixed level of specificity of a continuous-scale diagnostic test". In: *Statistics in Medicine* 24.3, pp. 465–477. ISSN: 1097-0258. DOI: 10.1002/sim.1563.

Zhu, Y. et al. (2017). “MRI-based prostate cancer detection with high-level representation and hierarchical classification”. In: *Medical Physics* 44.3, pp. 1028–1039. DOI: 10.1002/mp.12116.

Appendix

A.1 Supporting tables

Variables	SanRaffaele	Zurich	Hamburg	PCPT
Number of biopsies	637	1863	7877	6664
Age median (quartiles)	65.0 (60.0; 71.0)	63.8 (58.5; 68.7)	66.0 (60.0; 71.0)	69.0 (65.0; 73.0)
PSA median** (quartiles)	6.6 (4.8; 9.7)	5.3 (3.5; 8.5)	7.3 (5.1; 11.0)	1.2 (0.7; 2.2)
<=4	85 (13.3%)	635 (34.1%)	963 (12.2%)	5748 (86.3%)
(4,20]	515 (80.9%)	1099 (59.0%)	6282 (79.8%)	916 (13.7%)
>20	37 (5.8%)	129 (6.9%)	632 (8.0%)	0 (0.0%)
DRE result	-	-	-	-
Normal	386 (60.6%)	1433 (76.9%)	5872 (74.6%)	5647 (84.7%)
Abnormal	106 (16.6%)	425 (22.8%)	1114 (14.1%)	1017 (15.3%)
Unknown	145 (22.8%)	5 (0.3%)	891 (11.3%)	0 (0.0%)
1st degree family history	-	-	-	-
No	496 (77.8%)	1810 (97.2%)	2760 (35.0%)	5562 (83.5%)
Yes	124 (19.5%)	53 (2.8%)	614 (7.8%)	1102 (16.5%)
Unknown	17 (2.7%)	0 (0.0%)	4503 (57.2%)	0 (0.0%)
Race	-	-	-	-
Black or African American	0 (0.0%)	0 (0.0%)	33 (0.4%)	219 (3.3%)
Others	637 (100.0%)	1863 (100.0%)	2902 (36.9%)	6445 (96.7%)
Unknown	0 (0.0%)	0 (0.0%)	4942 (62.7%)	0 (0.0%)
Prior negative biopsy	-	-	-	-
Yes	189 (29.7%)	734 (39.4%)	605 (7.7%)	904 (13.6%)
No	440 (69.1%)	1129 (60.6%)	180 (2.3%)	5760 (86.4%)
Unknown	8 (1.2%)	0 (0.0%)	7092 (90.0%)	0 (0.0%)
Biopsy result	-	-	-	-
Positive	297 (46.6%)	561 (30.1%)	4466 (56.7%)	1196 (17.9%)
Negative	340 (53.4%)	1302 (69.9%)	3411 (43.3%)	5468 (82.1%)
Gleason score*	-	-	-	-
<=6	83 (28.0%)	231 (41.2%)	1559 (34.9%)	942 (78.8%)
7	151 (50.8%)	190 (33.9%)	1938 (43.4%)	201 (16.8%)
>=8	63 (21.2%)	140 (24.9%)	969 (21.7%)	53 (4.4%)

Table A.1 : Description of the European PBCG cohorts SanRaffaele and Zurich, and the additional data sets of Hamburg and the PCPT. *for positive diagnosis only **PSA>10 excluded in PCPT data set

Variables	ClevelandClinic	DurhamVA	MayoClinic	MSKCC	SanJuanVA	Sunnybrook	UCSF	UTHealth
Number of biopsies	299	669	323	1010	550	1721	521	899
Age median (quartiles)	64.5 (58.2; 69.0)	66.0 (62.0; 69.0)	64.0 (58.0; 70.0)	62.5 (57.2; 67.8)	67.0 (63.0; 71.0)	64.0 (58.1; 70.0)	65.0 (60.0; 69.0)	64.0 (59.0; 69.0)
PSA median (quartiles)	5.0 (3.6; 7.2)	6.5 (5.0; 9.7)	5.4 (3.8; 8.0)	5.8 (4.3; 8.4)	5.4 (4.2; 7.9)	6.4 (4.6; 9.7)	6.5 (4.7; 10.0)	6.0 (4.4; 9.1)
<=4	98 (32.8%)	52 (7.8%)	87 (26.9%)	192 (19.0%)	125 (22.7%)	284 (16.5%)	87 (16.7%)	173 (19.2%)
(4,20]	191 (63.9%)	553 (82.6%)	213 (66.0%)	766 (75.8%)	402 (73.1%)	1308 (76.0%)	403 (77.4%)	667 (74.2%)
>20	10 (3.3%)	64 (9.6%)	23 (7.1%)	52 (5.2%)	23 (4.2%)	129 (7.5%)	31 (5.9%)	59 (6.6%)
DRE result	-	-	-	-	-	-	-	-
Normal	226 (75.6%)	314 (47.0%)	197 (61.0%)	448 (44.3%)	194 (35.3%)	1126 (65.4%)	371 (71.2%)	540 (60.1%)
Abnormal	71 (23.7%)	136 (20.3%)	122 (37.8%)	131 (13.0%)	282 (51.3%)	471 (27.4%)	133 (25.5%)	352 (39.1%)
Unknown	2 (0.7%)	219 (32.7%)	4 (1.2%)	431 (42.7%)	74 (13.4%)	124 (7.2%)	17 (3.3%)	7 (0.8%)
1st degree family history	-	-	-	-	-	-	-	-
No	224 (74.9%)	557 (83.3%)	178 (55.1%)	697 (69.0%)	237 (43.1%)	1398 (81.2%)	0 (0.0%)	723 (80.4%)
Yes	53 (17.7%)	112 (16.7%)	88 (27.2%)	313 (31.0%)	73 (13.3%)	284 (16.5%)	0 (0.0%)	172 (19.1%)
Unknown	22 (7.4%)	0 (0.0%)	57 (17.6%)	0 (0.0%)	240 (43.6%)	39 (2.3%)	521 (100.0%)	4 (0.5%)
Race	-	-	-	-	-	-	-	-
Black or African American	48 (16.1%)	422 (63.1%)	2 (0.6%)	38 (3.8%)	68 (12.4%)	72 (4.2%)	14 (2.7%)	116 (12.9%)
Others	247 (82.6%)	246 (36.8%)	321 (99.4%)	899 (89.0%)	482 (87.6%)	849 (49.3%)	457 (87.7%)	631 (70.2%)
Unknown	4 (1.3%)	1 (0.1%)	0 (0.0%)	73 (7.2%)	0 (0.0%)	800 (46.5%)	50 (9.6%)	152 (16.9%)
Prior negative biopsy	-	-	-	-	-	-	-	-
Yes	87 (29.1%)	96 (14.3%)	54 (16.7%)	302 (29.9%)	180 (32.7%)	276 (16.0%)	192 (36.8%)	159 (17.7%)
No	209 (69.9%)	573 (85.7%)	268 (83.0%)	708 (70.1%)	370 (67.3%)	1445 (84.0%)	299 (57.4%)	740 (82.3%)
Unknown	3 (1.0%)	0 (0.0%)	1 (0.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	30 (5.8%)	0 (0.0%)
Biopsy result	-	-	-	-	-	-	-	-
Positive	134 (44.8%)	411 (61.4%)	173 (53.6%)	418 (41.4%)	232 (42.2%)	867 (50.4%)	277 (53.2%)	497 (55.3%)
Negative	165 (55.2%)	258 (38.6%)	150 (46.4%)	592 (58.6%)	318 (57.8%)	854 (49.6%)	244 (46.8%)	402 (44.7%)
Gleason score*	-	-	-	-	-	-	-	-
<=6	53 (39.6%)	172 (41.8%)	69 (39.9%)	123 (29.4%)	63 (27.1%)	316 (36.5%)	118 (42.6%)	149 (30.0%)
7	59 (44.0%)	157 (38.2%)	61 (35.3%)	213 (51.0%)	122 (52.6%)	407 (46.9%)	113 (40.8%)	234 (47.1%)
>=8	22 (16.4%)	82 (20.0%)	43 (24.8%)	82 (19.6%)	47 (20.3%)	144 (16.6%)	46 (16.6%)	114 (22.9%)

Table A.2 : Description of North American cohorts in the PBCG. *for positive diagnosis only

	PCPTRC			PBCG		
Risk factor	Odds Ratio	95%-CI	P-value	Odds Ratio	95%-CI	P-value
High- versus low-grade cancer						
Prior neg. biopsy	1.27	0.85-1.90	0.2	0.65	0.50-0.84	0.001
PSA (log base 2)	1.57	1.38-1.77	<0.0001	1.94	1.75-2.15	<0.0001
Family history	0.95	0.67-1.36	0.8	1.32	1.06-1.65	0.01
DRE	1.55	1.09-2.21	0.01	2.25	1.84-2.75	<0.0001
Age (in 10 years)	1.36	1.07-1.74	0.01	1.46	1.29-1.65	<0.0001
African ancestry	2.51	1.39-4.52	0.002	0.84	0.66-1.07	0.2
High-grade versus no cancer						
Prior neg. biopsy	0.81	0.57-1.15	0.2	0.28	0.23-0.34	<0.0001
PSA (log base 2)	2.02	1.80-2.27	<0.0001	2.23	2.04-2.43	<0.0001
Family history	1.25	0.91-1.73	0.2	1.84	1.52-2.22	<0.0001
DRE	1.49	1.09-2.05	0.01	2.36	2.00-2.77	<0.0001
Age (in 10 years)	1.61	1.29-2.01	<0.0001	1.74	1.57-1.94	<0.0001
African ancestry	2.83	1.71-4.68	<0.0001	1.85	1.48-2.32	<0.0001
Low-grade versus no cancer						
Prior neg. biopsy	0.63	0.51-0.79	<0.0001	0.43	0.34-0.54	<0.0001
PSA (log base 2)	1.29	1.22-1.37	<0.0001	1.15	1.05-1.26	0.004
Family history	1.31	1.10-1.57	0.003	1.39	1.13-1.72	0.002
DRE	0.96	0.79-1.17	0.7	1.05	0.86-1.27	0.6
Age (in 10 years)	1.18	1.04-1.34	0.01	1.19	1.07-1.34	0.002
African ancestry	1.13	0.77-1.67	0.5	2.20	1.73-2.79	<0.0001

Table A.3 : Comparison between odds ratios of PCPTRC and PBCG for multinomial logistic regression models with no missing data. PBCG model is built on all eight North American cohorts pooled together. Odds ratios, CIs and p-values are given for the reference level reported to the left versus the endpoint to the right.

	AUC (CI) PCPTRC	AUC (CI) PBCG	P-value
North American cohorts (n=5,992)	69.9% (68.6-71.2%)	71.8% (70.5-73.1%)	<0.0001
European cohorts (n=10,377)	66.4% (65.4-67.4%)	68.8% (67.8-69.8%)	<0.0001
P-value	<0.0001	0.0003	

Table A.4 : AUC values for overall versus no cancer with corresponding 95% Delong CI for the PCPTRC and PBCG models. P-values by the Delong test for two correlated ROC-curves are used to compare the PCPTRC and PBCG models (bottom row), and for two uncorrelated ROC-curves to compare the North American and European cohorts (last column). Risk predictions for the North American cohorts by the PBCG model are calculated by leave-one-cohort-out cross validation.

A.2 Selected R code

The R calculations are based on R version R-3.4.2.

R-code for multinomial logistic regression. The variables PSA, Age and Race are mandatory; Prior biopsy, DRE and Family history are allowed missing

```
##### explanation of input variables
# psa: enter prostate-specific antigen in ng/ml

# age: enter age in years

# race: enter 1 for African Ancestry, 0 otherwise

# priorbiopsy: enter 1 if there has been one or more prior biopsies
# (all negative for prostate cancer), 0 otherwise

# dre: enter 1 if digital rectal examination is abnormal (suspicious
# for prostate cancer), 0 otherwise

# famhistory: enter 1 if there is a first-degree family history of
# prostate cancer, 0 otherwise

# psa, age and race are mandatory, priorbiopsy, dre and famhistory
# are allowed missing


 risk = function(psa, age, race, priorbiopsy, dre, famhistory) {

##### create persons data set
data=c(1, log(psa,2), age, race)

# is priorbiopsy known?
a = as.numeric(is.na(priorbiopsy)==FALSE)
if(a==1){data=c(data, priorbiopsy)}

# is dre known?
b = as.numeric(is.na(dre)==FALSE)
if(b==1){data=c(data, dre)}

# is famhistory known?
c = as.numeric(is.na(famhistory)==FALSE)
```

```
if(c==1){data=c(data, famhistory)}

##### choose correct model
# psa, age, race, priorbiopsy, dre, famhistory
if(a==1 & b==1 & c==1){
no.low=c(-2.44052108 , 0.13617244 , 0.01780617 , 0.78721039 ,
-0.83613721 , 0.04612721 , 0.33233636)
no.high=c(-6.36851856 , 0.79996510 , 0.05566536 , 0.61596975 ,
-1.27437249 , 0.85780143 , 0.61003848)
}

# psa, age, race, priorbiopsy, dre
if(a==1 & b==1 & c==0){
no.low=c(-2.29687989 , 0.13785591 , 0.01758914 , 0.63876791 ,
-0.86200471 , 0.07193350)
no.high=c(-6.06621401 , 0.76053930 , 0.05509847 , 0.51701373 ,
-1.38390751 , 0.83442202)
}

# psa, age, race, priorbiopsy, famhistory
if(a==1 & b==0 & c==1){
no.low=c(-2.64840984 , 0.13125283 , 0.02044166 , 0.81792881 ,
-0.98610357 , 0.31447017)
no.high=c(-6.70538152 , 0.77635003 , 0.06542705 , 0.52401464 ,
-1.43681965 , 0.55443478)
}

# psa, age, race, dre, famhistory
if(a==0 & b==1 & c==1){
no.low=c(-2.16147411 , 0.07409519 , 0.01322988 , 0.76131045 ,
0.05397516 , 0.29246219)
no.high=c(-5.99897055 , 0.70727793 , 0.04992968 , 0.56485952 ,
0.89154384 , 0.56910873)
}

# psa, age, race, priorbiopsy
if(a==1 & b==0 & c==0){
no.low=c(-2.49050385 , 0.12961272 , 0.02020429 , 0.67674970 ,
-0.97275826)
no.high=c(-6.41089002 , 0.74110558 , 0.06476911 , 0.42814591 ,
-1.50274350)
}
```

```
# psa, age, race, dre
if(a==0 & b==1 & c==0){
no.low=c(-2.01851079 , 0.06745424 , 0.01263369 , 0.63938472 ,
0.08562844)
no.high=c(-5.68203352 , 0.65059244 , 0.04883786 , 0.49214793 ,
0.87421554)
}

# psa, age, race, famhistory
if(a==0 & b==0 & c==1){
no.low=c(-2.39161580 , 0.06129651 , 0.01600515 , 0.81132928 ,
0.27501639)
no.high=c(-6.42320154 , 0.67779036 , 0.06092178 , 0.50429130 ,
0.50805684)
}

# psa, age, race
if(a==0 & b==0 & c==0){
no.low=c(-2.23794923 , 0.05343098 , 0.01553627 , 0.69593716)
no.high=c(-6.13292904 , 0.62979529 , 0.06002002 , 0.43816016)
}

##### predicting probabilities
S1=no.low%*%data
S2=no.high%*%data
risk.no=1/(1+exp(S1)+exp(S2))*100
risk.low=exp(S1)/(1+exp(S1)+exp(S2))*100
risk.high=100-risk.no-risk.low

##### outcome
risk.outcome=cbind(risk.no,risk.low,risk.high)
dimnames(risk.outcome)=list(NULL, c('Chance␣of␣No␣Cancer',
'Risk␣of␣Low␣Grade␣Cancer', 'Risk␣of␣High␣Grade␣Cancer'))
return(risk.outcome)
}
```

A.3 Additional Graphics

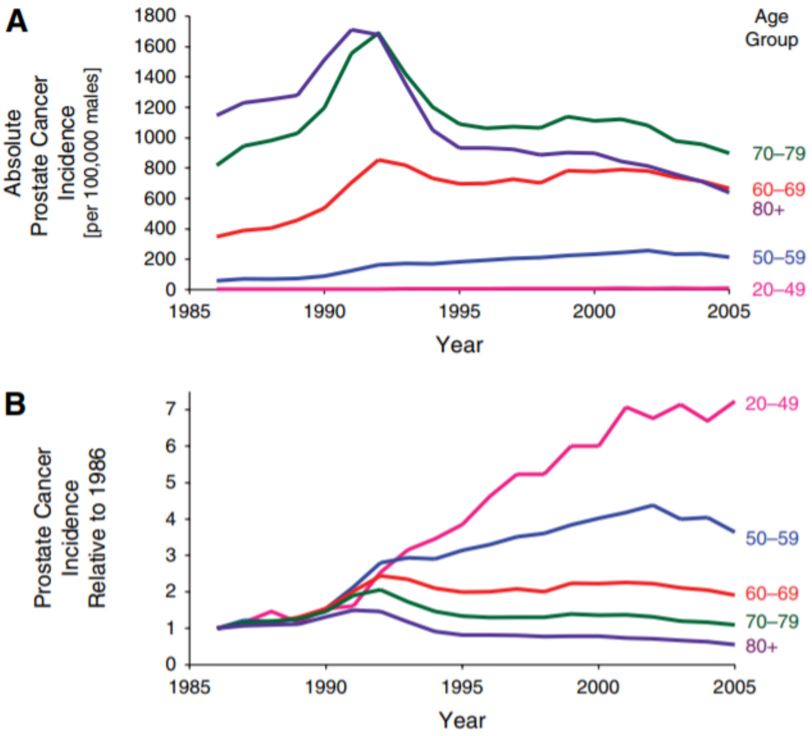


Figure A.1 : (A) absolute and (B) relative to 1986 prostate cancer incidences in the U.S. between 1986 and 2005, stratified by age group (Welch and Albertsen 2009). Data is based on SEER program.

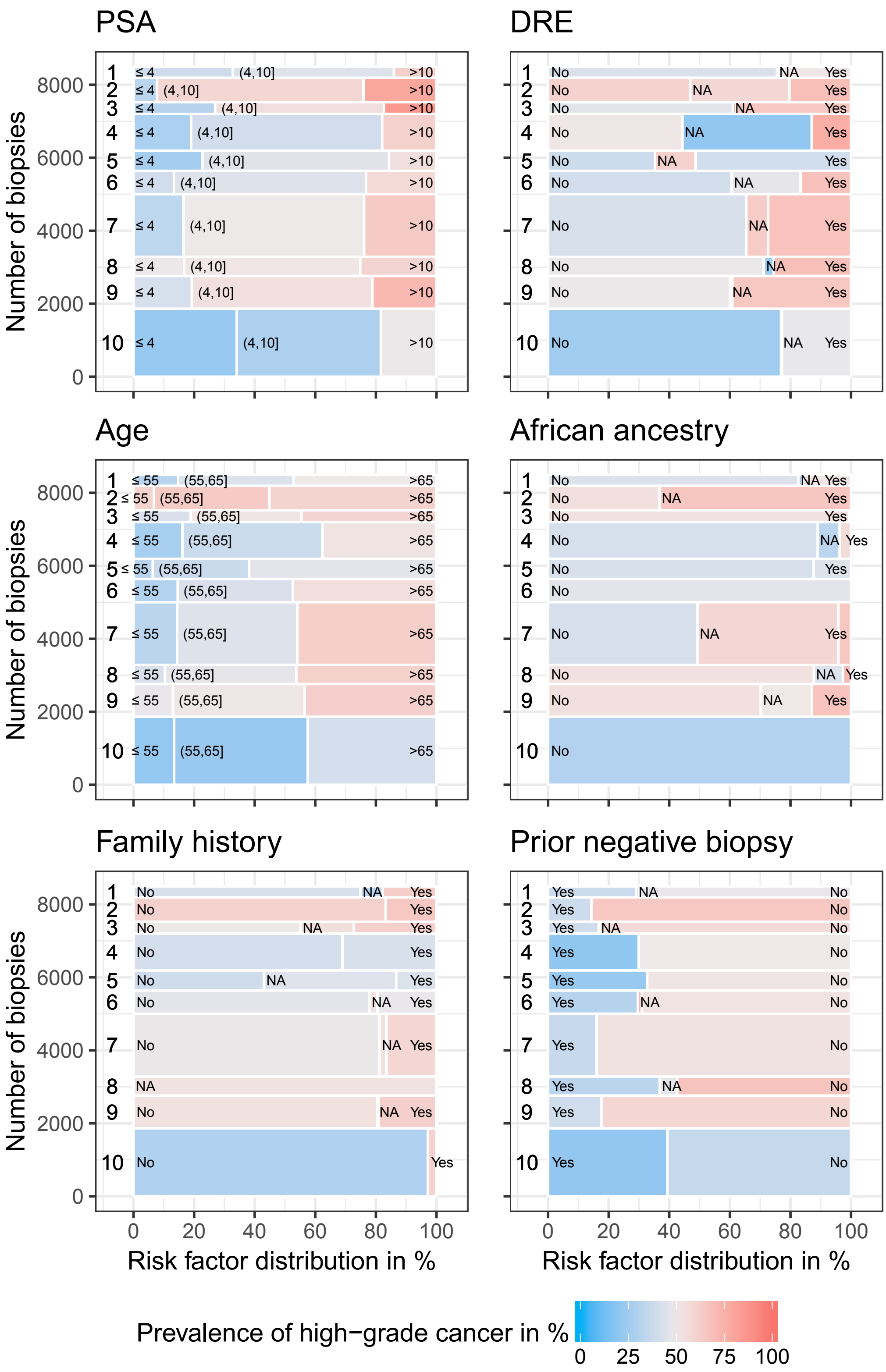


Figure A.2 : Percentage of overall cancer by risk factor (x) and number of biopsies (y); 1.ClevelandClinic, 2.DurhamVA, 3.MayoClinic, 4.MSKCC, 5.SanJuanVA, 6.SanRaffaele, 7.Sunnybrook, 8.UCSF, 9.UTHealth, 10.Zurich. NA denotes missing values.

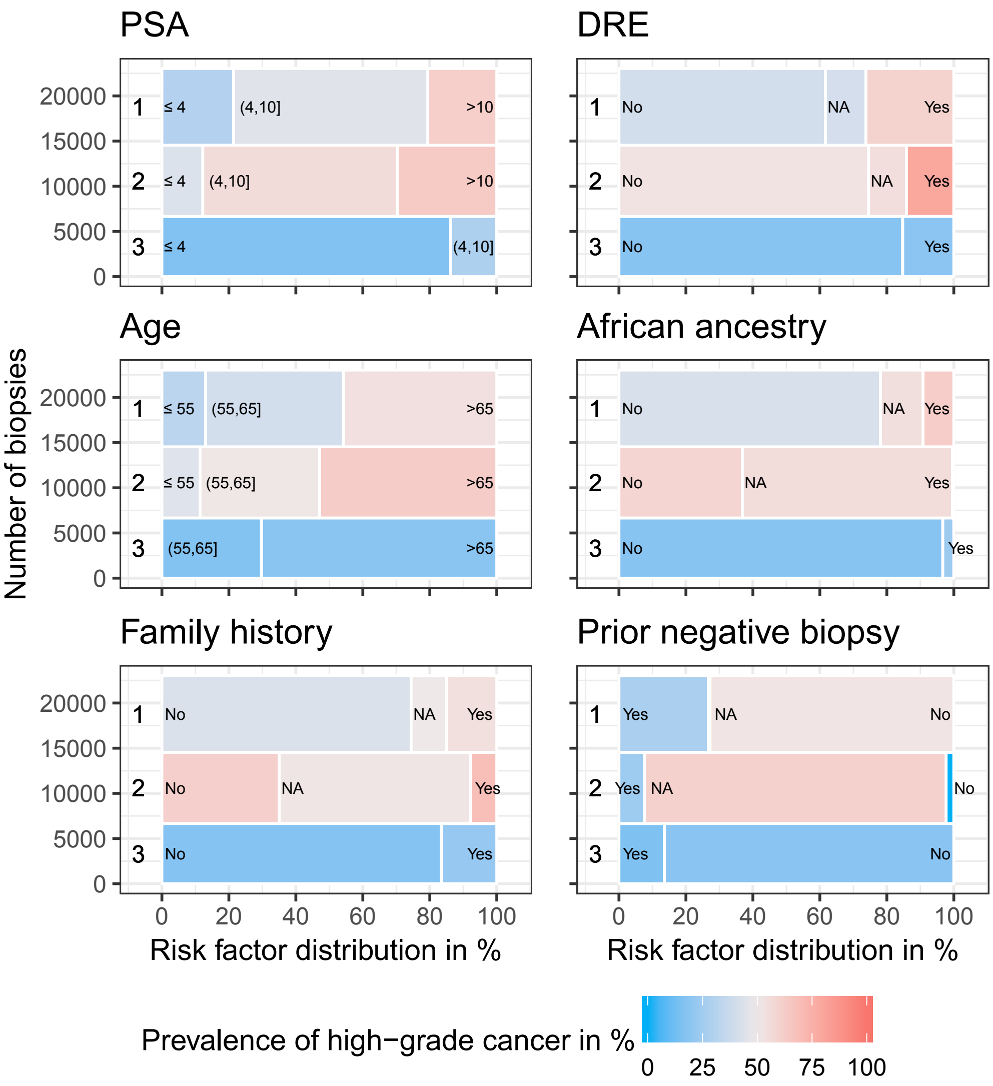


Figure A.3 : Percentage of overall cancer by risk factor (x) and number of biopsies (y); 1.PBCG, 2.Hamburg, 3.PCPT. NA denotes missing values.

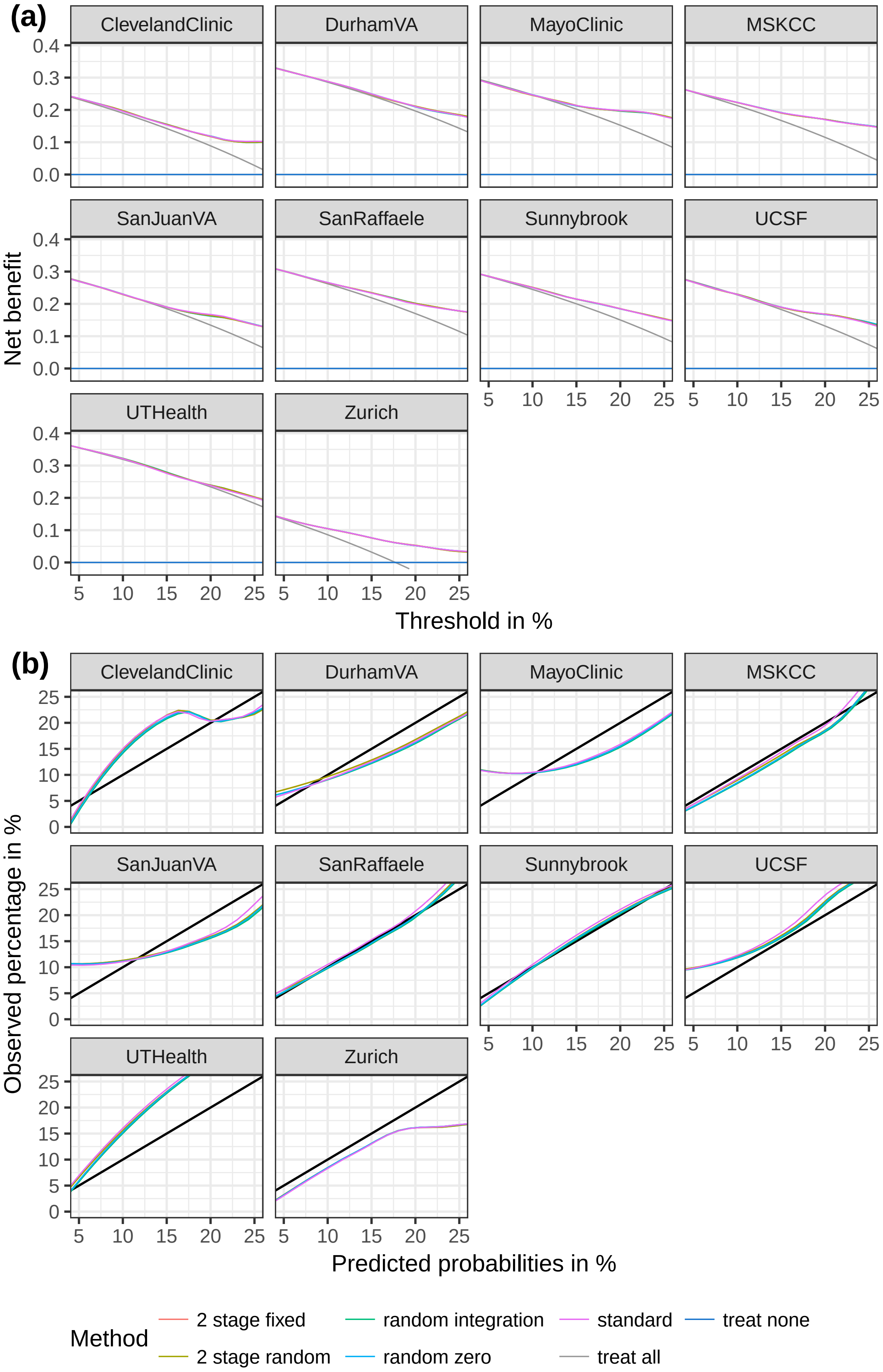


Figure A.4 : (a) Net benefit and (b) calibration curves for high-grade cancer of leave-one-cohort-out cross validation comparing regression methods. They are given for each cohort separately and the strategies of referring all men or none to biopsy are given for comparison.

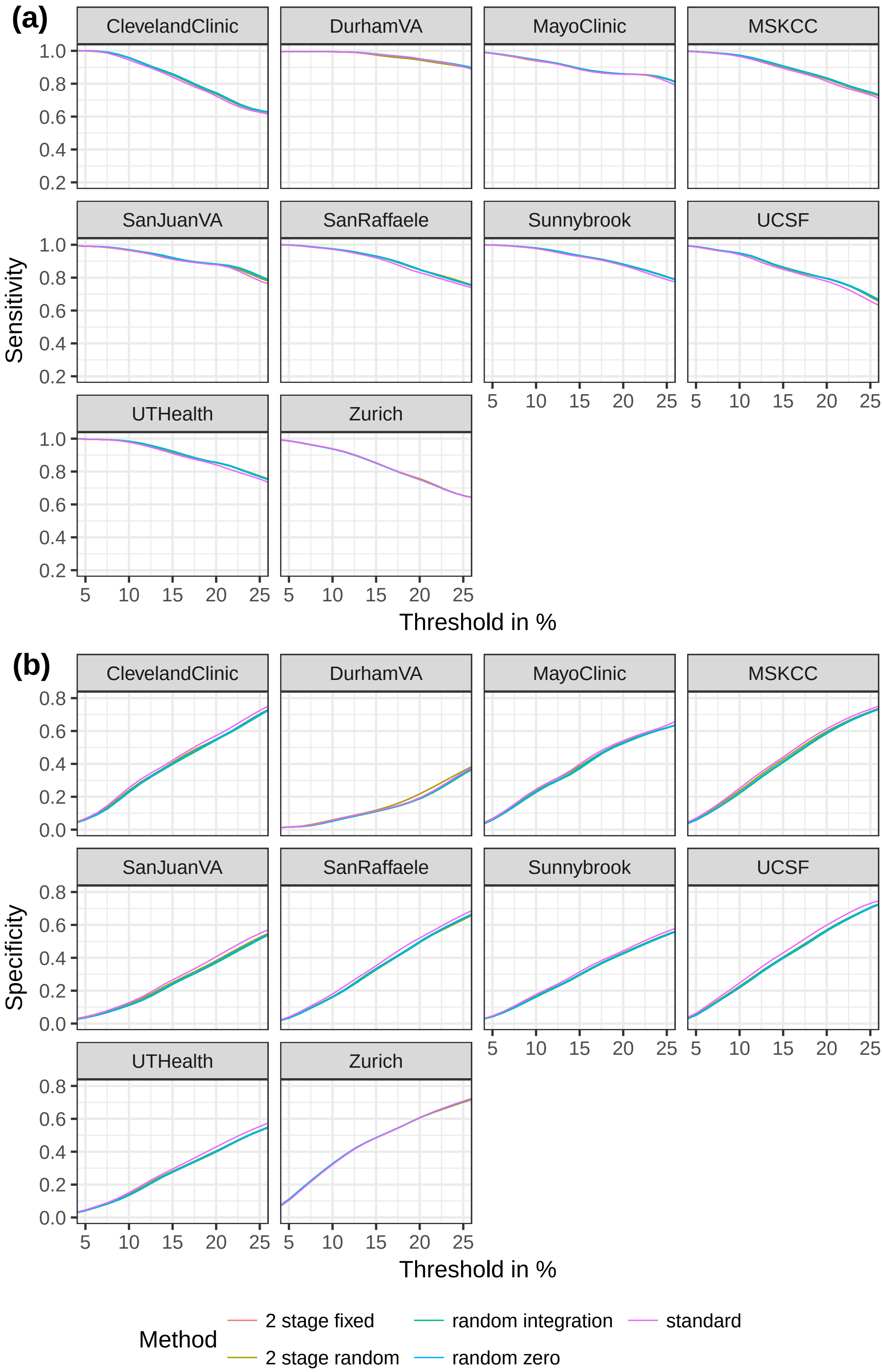


Figure A.5 : (a) Sensitivity and (b) specificity curves for high-grade cancer of leave-one-cohort-out cross validation comparing regression methods. They are given for each cohort separately.

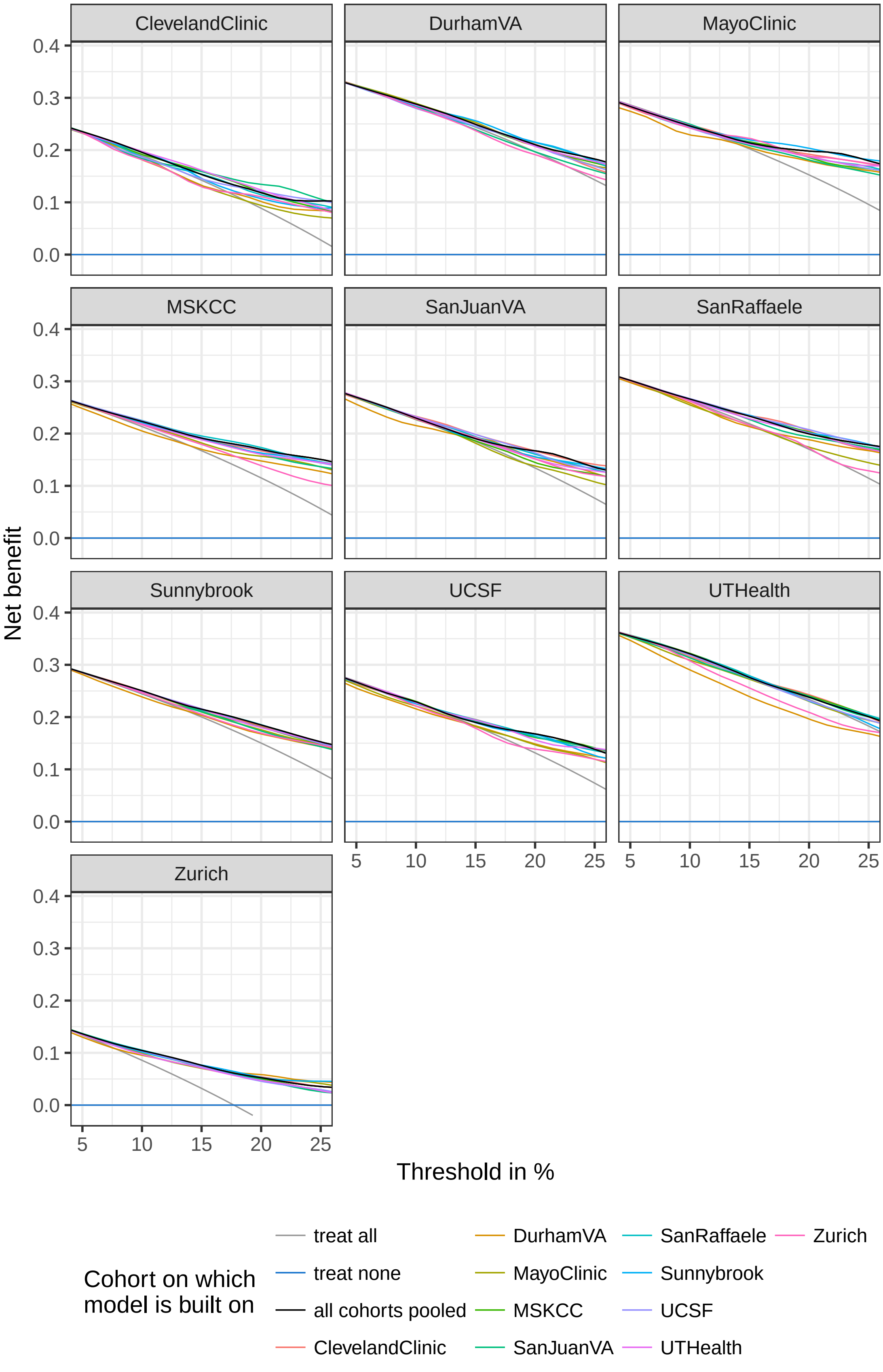


Figure A.6 : Net benefit curves for models built on individual cohorts. Leave-one-cohort-out cross validation results for standard multiple logistic regression and the strategies of referring all men or none to biopsy are given for comparison.

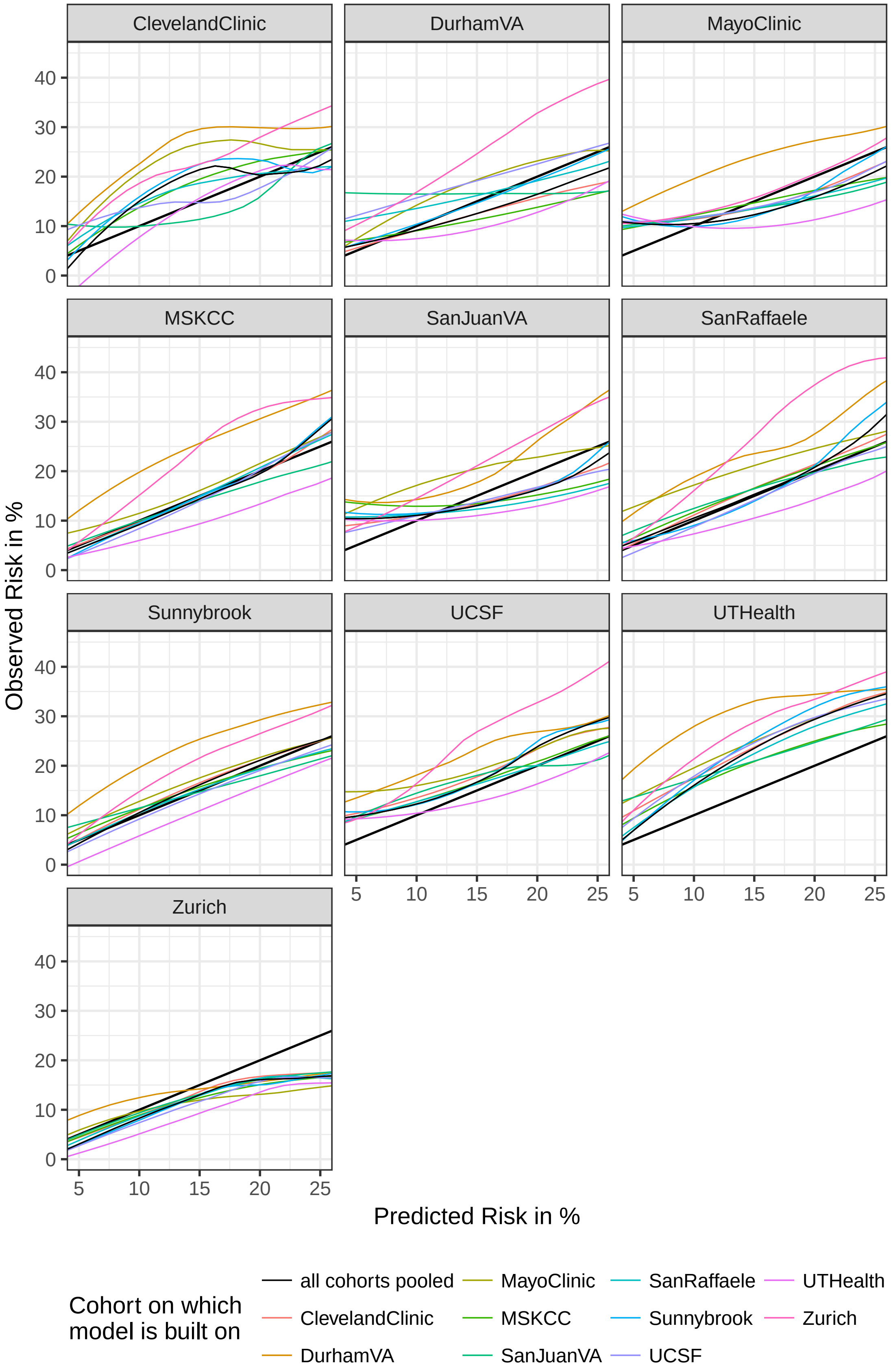


Figure A.7 : Calibration curves for models built on individual cohorts. Leave-one-cohort-out cross validation results for standard multiple logistic regression are given for comparison in black.

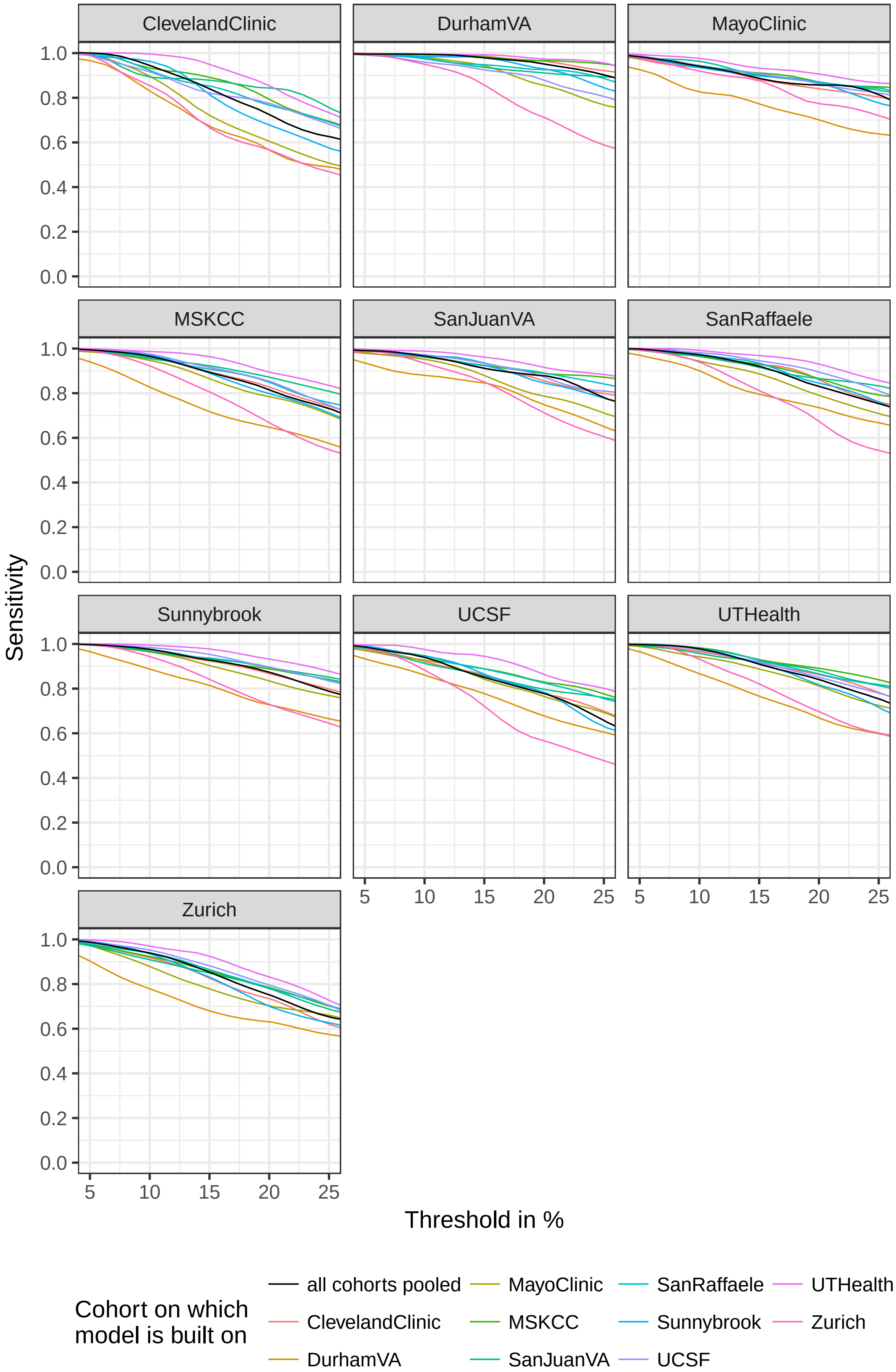


Figure A.8 : Sensitivity curves for models built on individual cohorts. Leave-one-cohort-out cross validation results for standard multiple logistic regression are given for comparison in black.

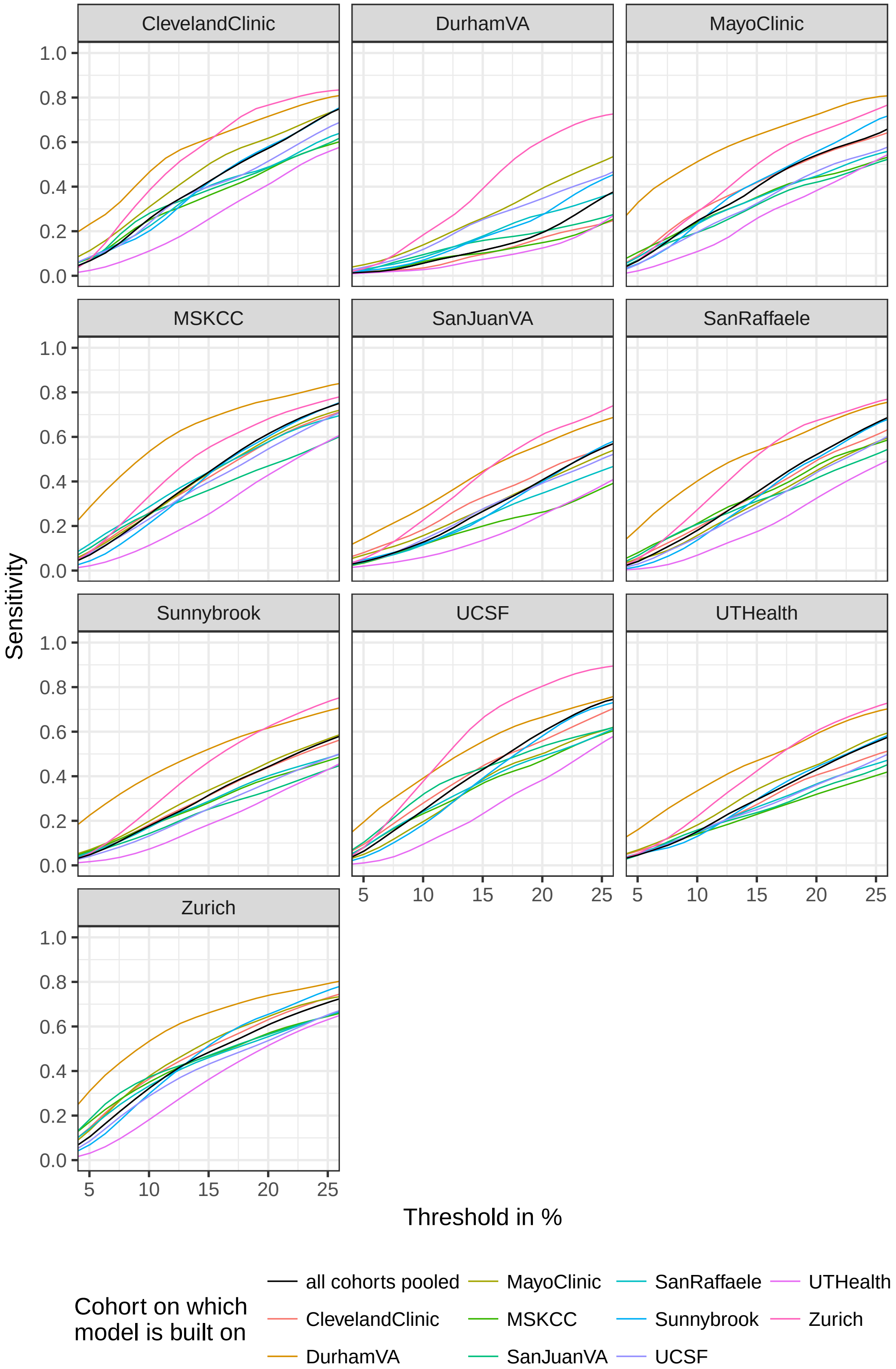


Figure A.9 : Specificity curves for models built on individual cohorts. Leave-one-cohort-out cross validation results for standard multiple logistic regression are given for comparison in black.

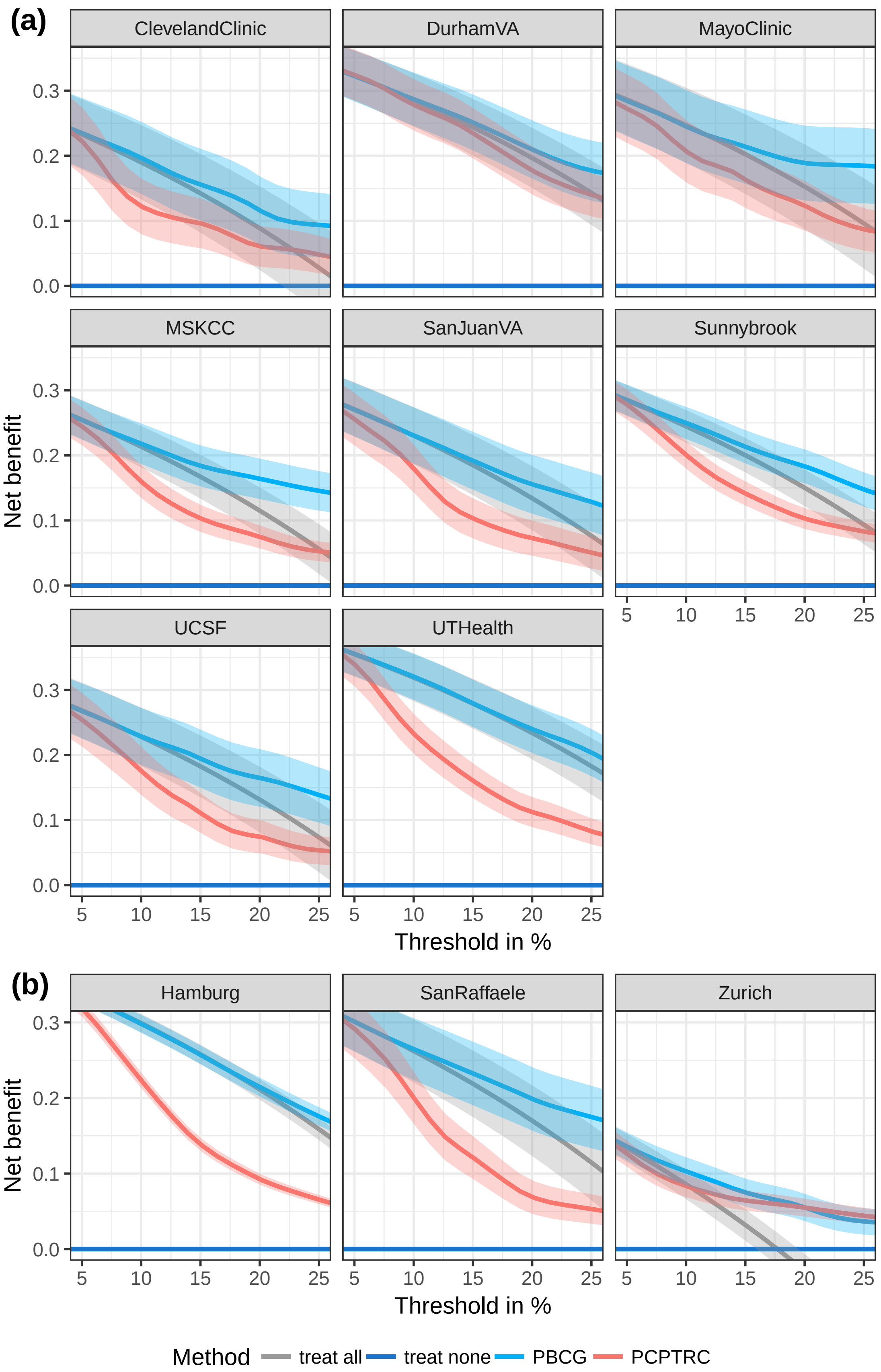


Figure A.10 : Net benefit curves for high-grade cancer comparing PBCG and PCPT models. They are given for (a) each North American and (b) European cohort separately. Strategies of referring all men or none to biopsy are provided for comparison and pointwise 95%-CIs are shown with shading.

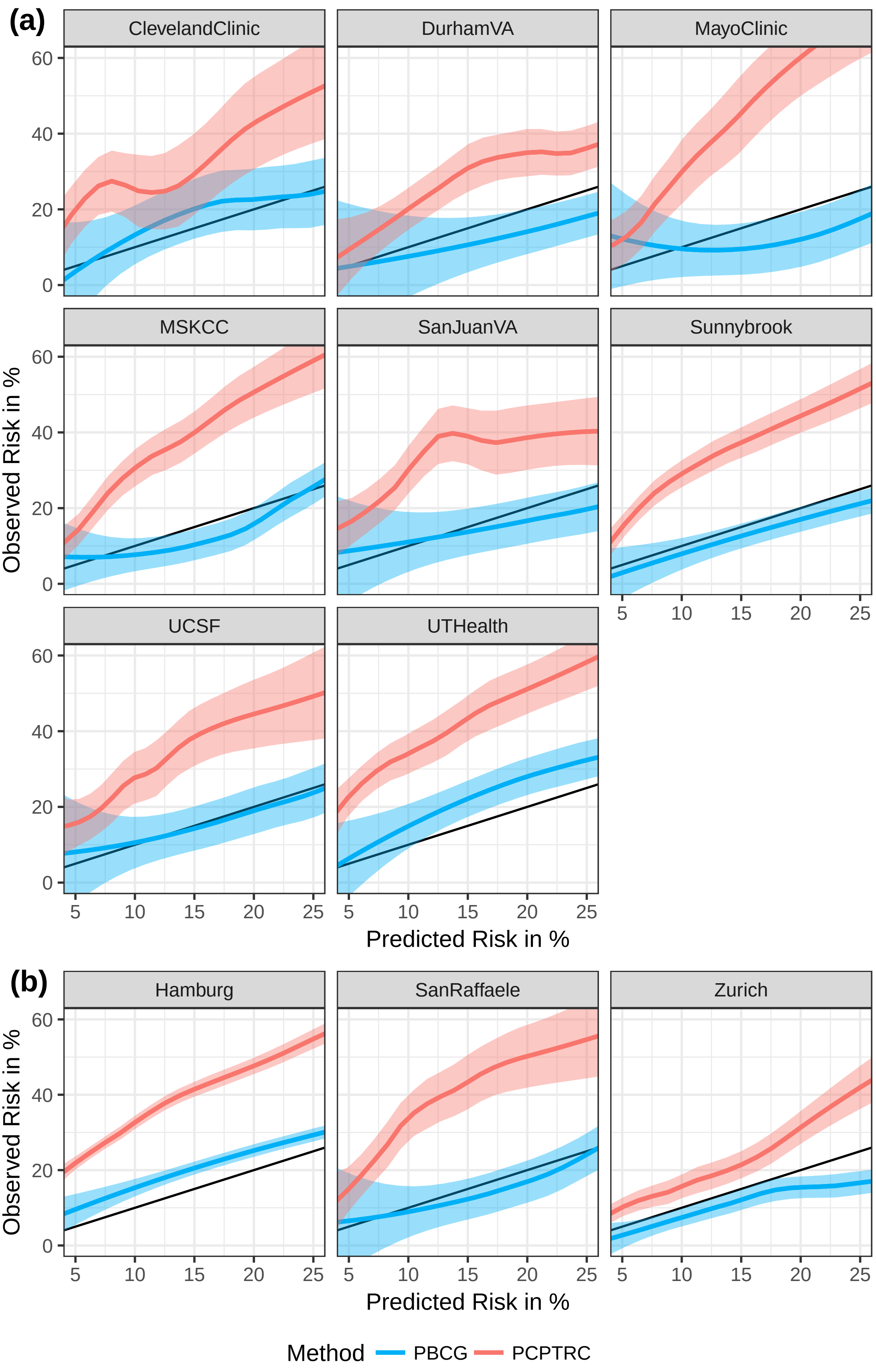


Figure A.11 : Calibration curves for high-grade cancer comparing PBCG and PCPT models. They are given for (a) each North American and (b) European cohort separately. Pointwise 95%-CIs are shown with shading and black lines show where predicted risks equal observed risks.

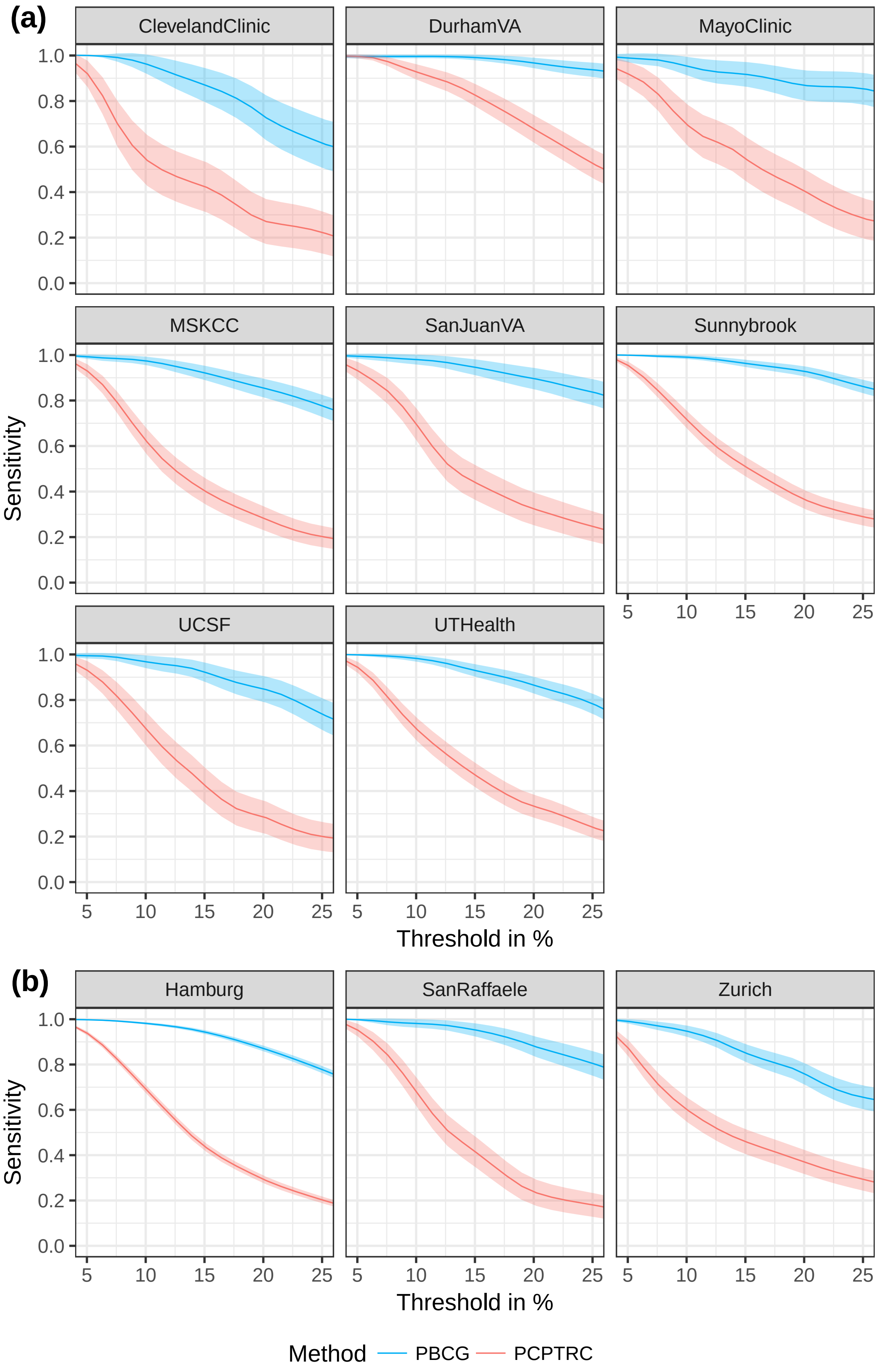


Figure A.12 : Sensitivity curves for high-grade cancer comparing PBCG and PCPT models. They are given for (a) each North American and (b) European cohort separately and pointwise 95%-CIs are shown with shading.

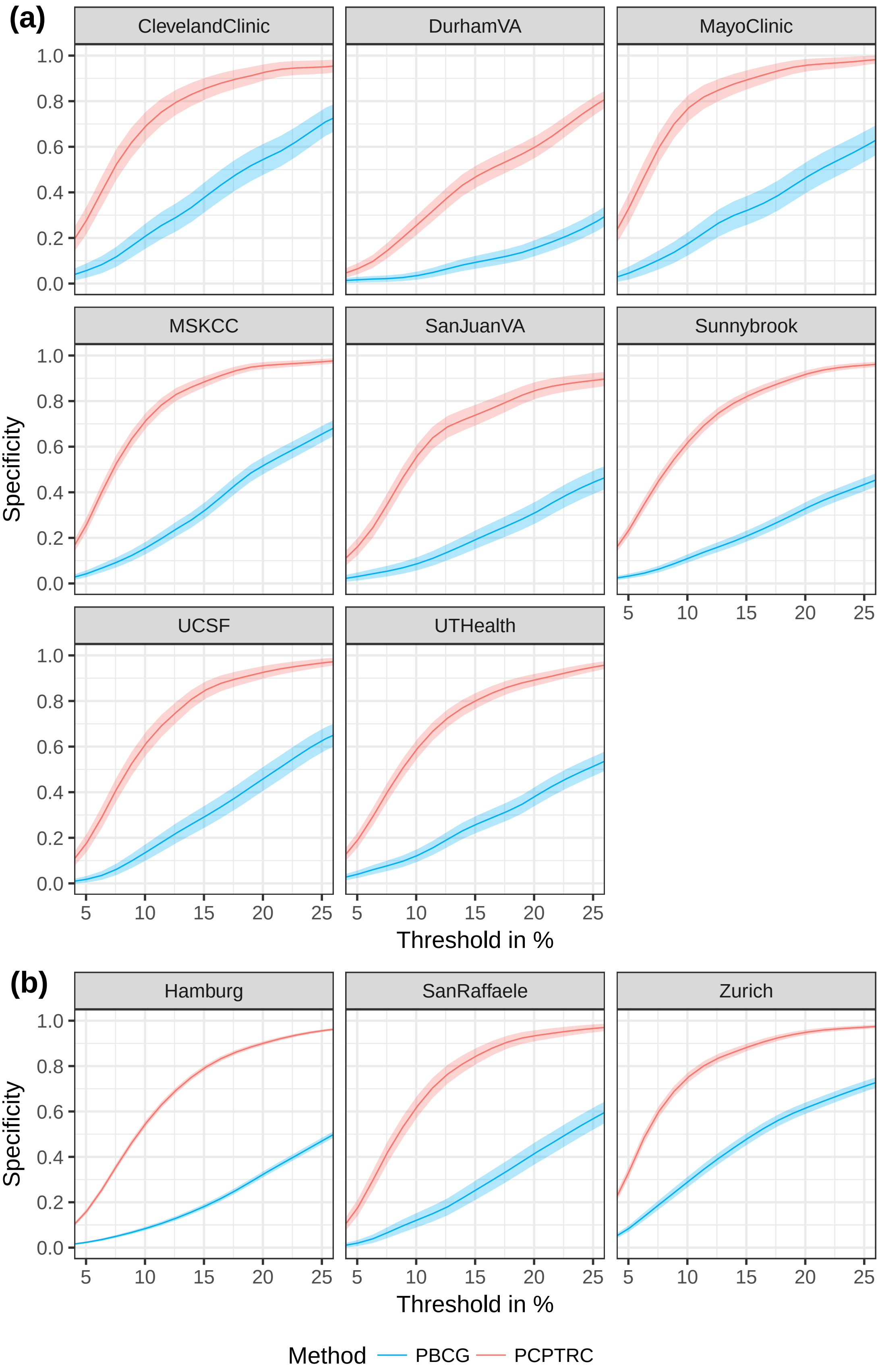


Figure A.13 : Specificity curves for high-grade cancer comparing PBCG and PCPT models. They are given for (a) each North American and (b) European cohort separately and pointwise 95%-CIs are shown with shading.

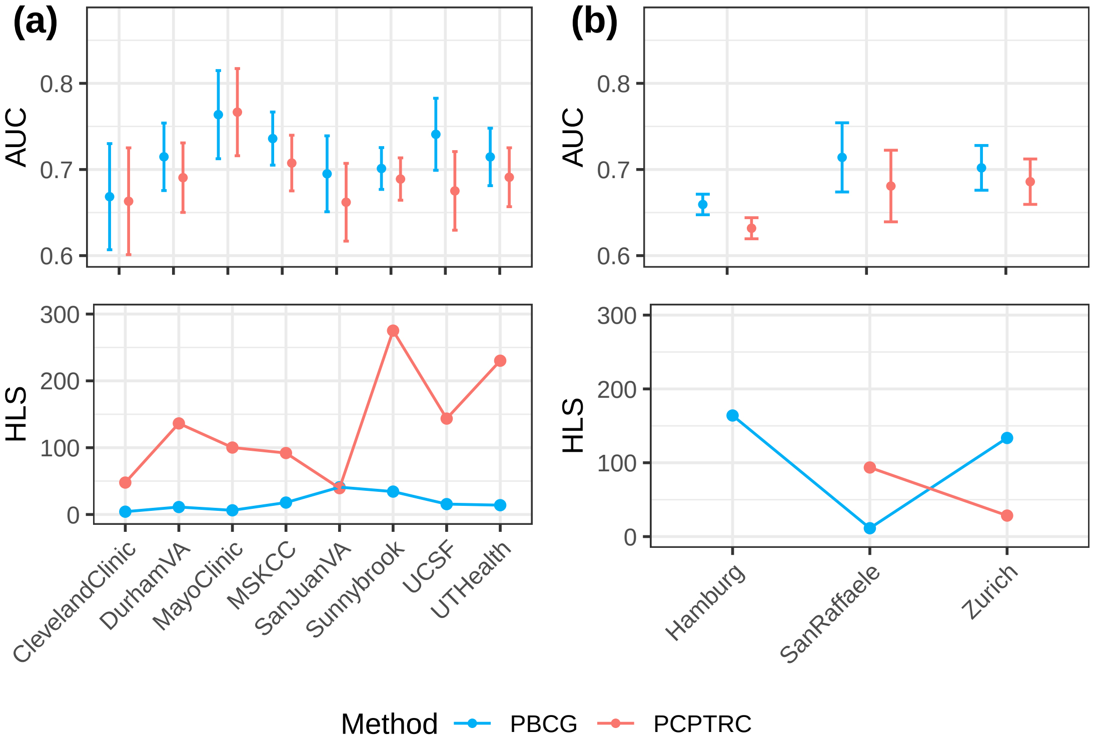


Figure A.14 : AUC and HLS values for overall cancer of PBCG and PCPT by site. (a) Results of the internal cross validation of the North American cohorts, (b) external validation on the European sites. The HLS value for the PCPTRC applied to the Hamburg cohort is neglected as it exceeds 2,000. For AUC higher values are better, while for HLS lower values are preferred. Sample sizes are given by 299 for ClevelandClinic, 669 for DurhamVA, 323 for MayoClinic, 1,010 for MSKCC, 550 for SanJuanVA, 1,721 for Sunnybrook, 521 for UCSF, 899 for UTHealth, 7,877 for Hamburg, 637 for SanRaffaele and 1,863 for Zurich.

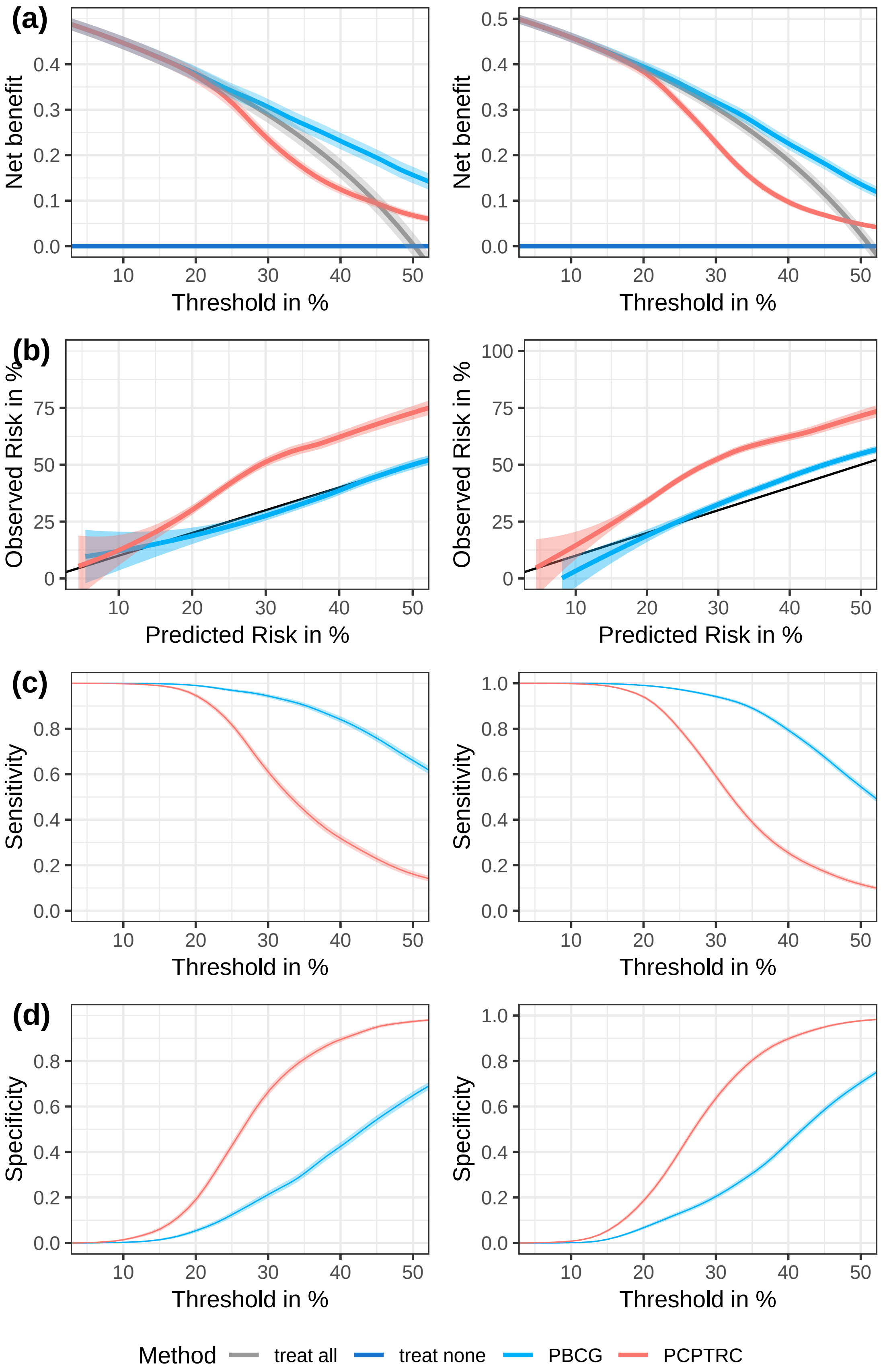


Figure A.15 : (a) Net benefit, (b) calibration, (c) sensitivity, and (d) specificity curves for overall cancer comparing the PBCG and PCPT models. Results of the internal cross validation of the North American cohorts (left) and external validation with the European sites (right). Strategies of referring all men or none to biopsy are provided in (a) for comparison, pointwise 95%-CIs are shown with shading and black lines in (b) show where predicted risks equal observed risks.

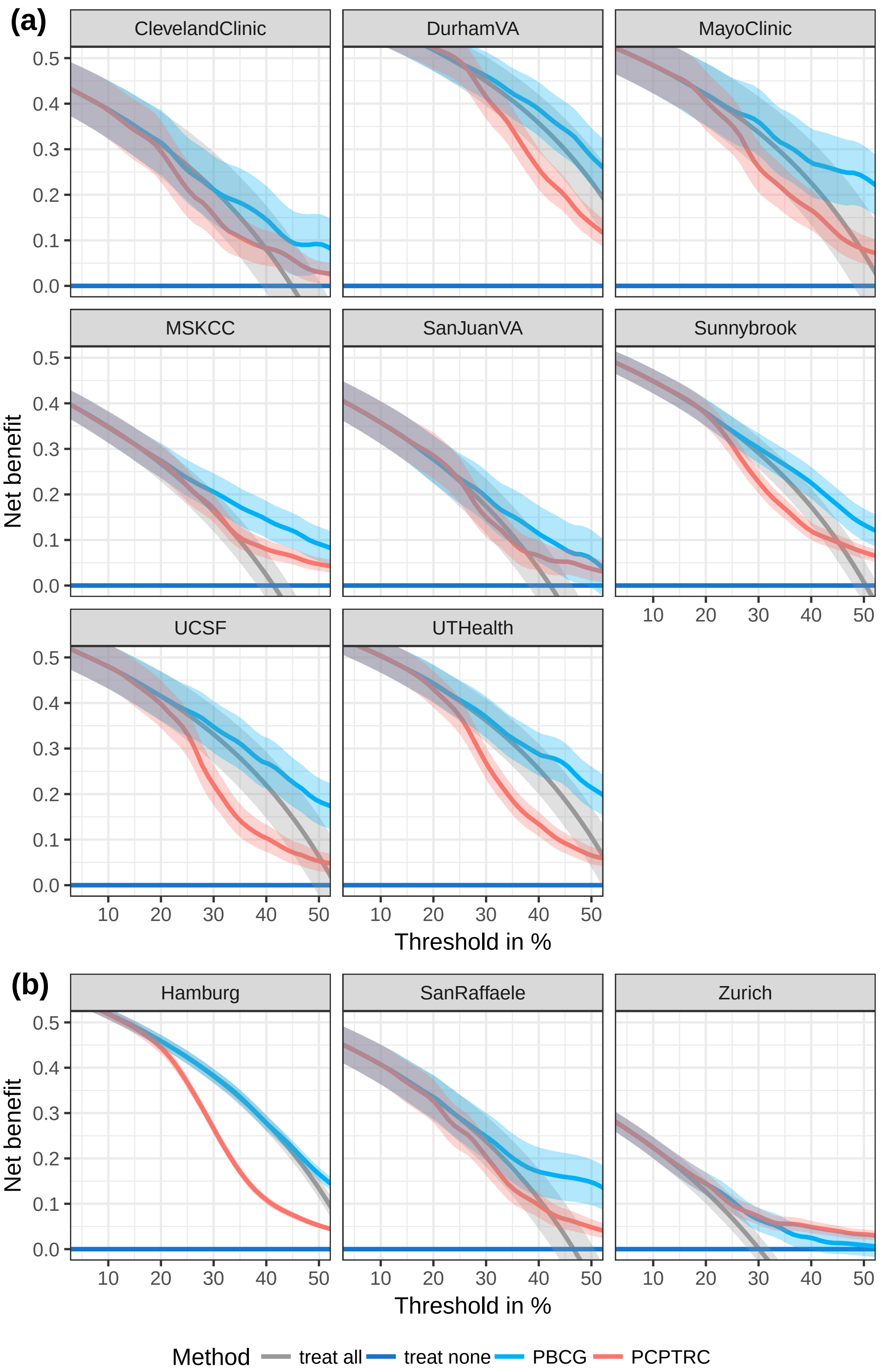


Figure A.16 : Net benefit curves for overall cancer comparing PBCG and PCPT models. They are given for (a) each North American and (b) European cohort separately. Strategies of referring all men or none to biopsy are provided for comparison and pointwise 95%-CIs are shown with shading.

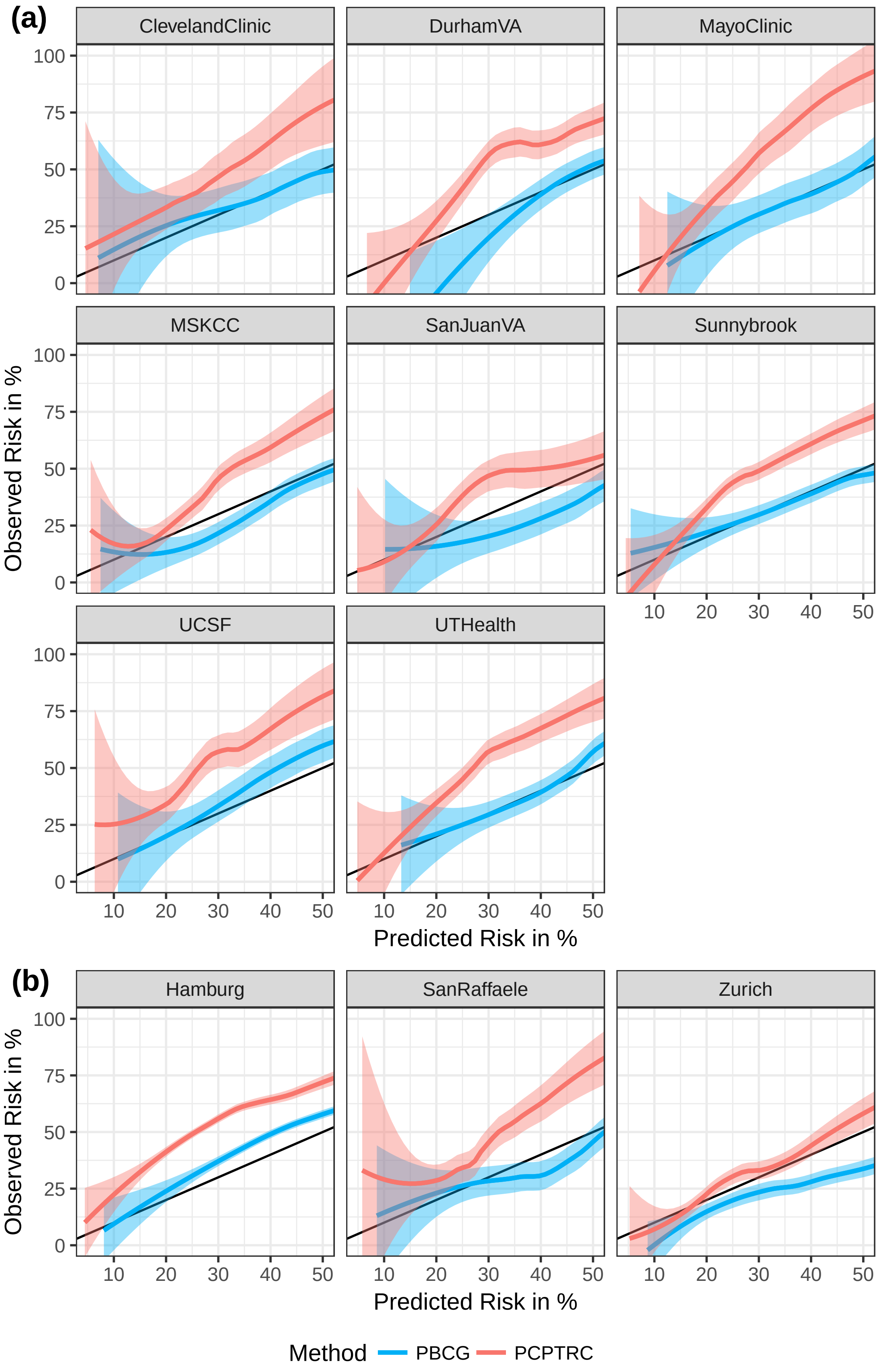


Figure A.17 : Calibration curves for overall cancer comparing PBCG and PCPT models. They are given for (a) each North American and (b) European cohort separately. Pointwise 95%-CIs are shown with shading and black lines show where predicted risks equal observed risks.

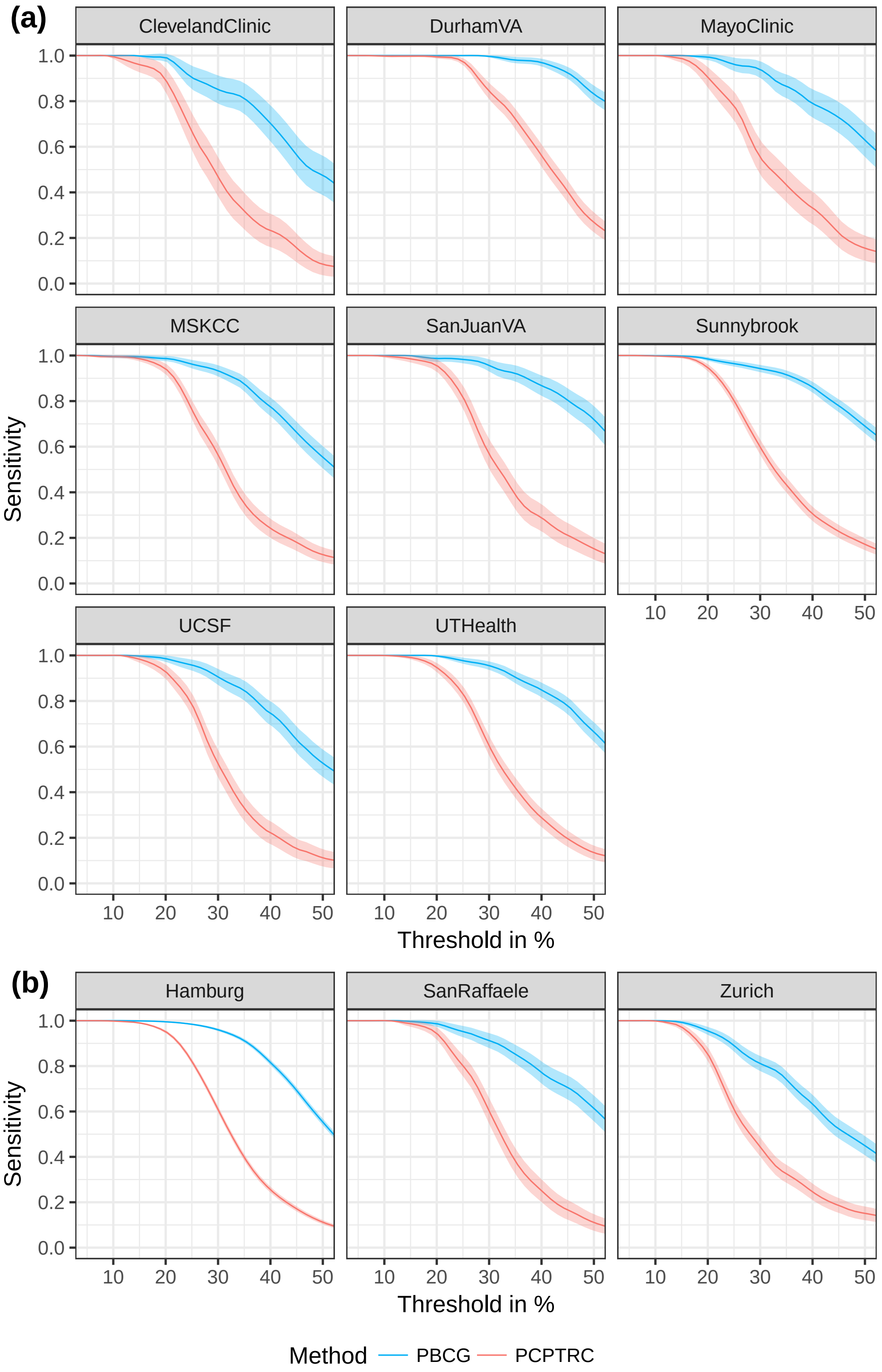


Figure A.18 : Sensitivity curves for overall cancer comparing PBCG and PCPT models. They are given for (a) each North American and (b) European cohort separately and pointwise 95%-CIs are shown with shading.

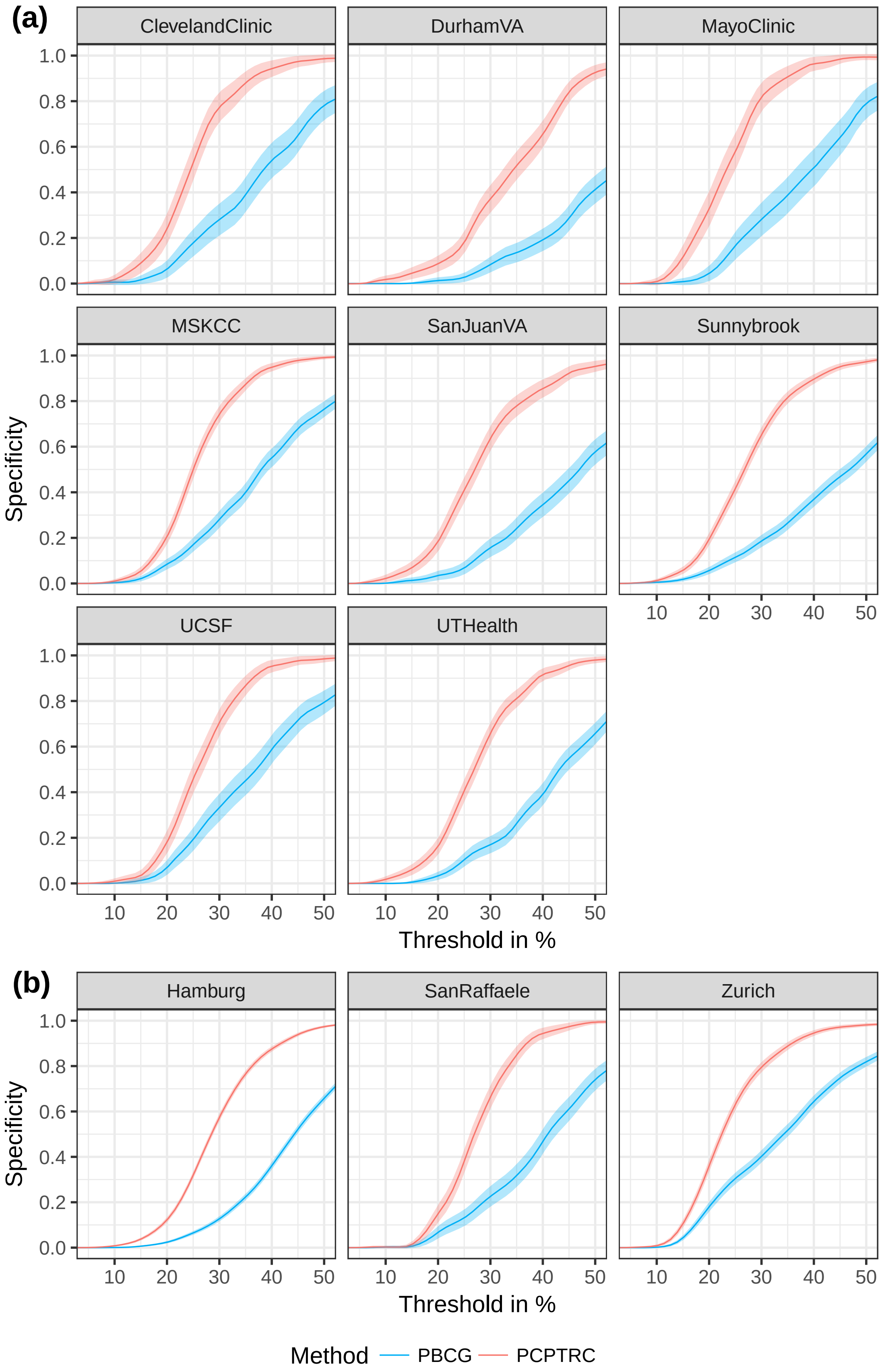


Figure A.19 : Specificity curves for overall cancer comparing PBCG and PCPT models. They are given for (a) each North American and (b) European cohort separately and pointwise 95%-CIs are shown with shading.

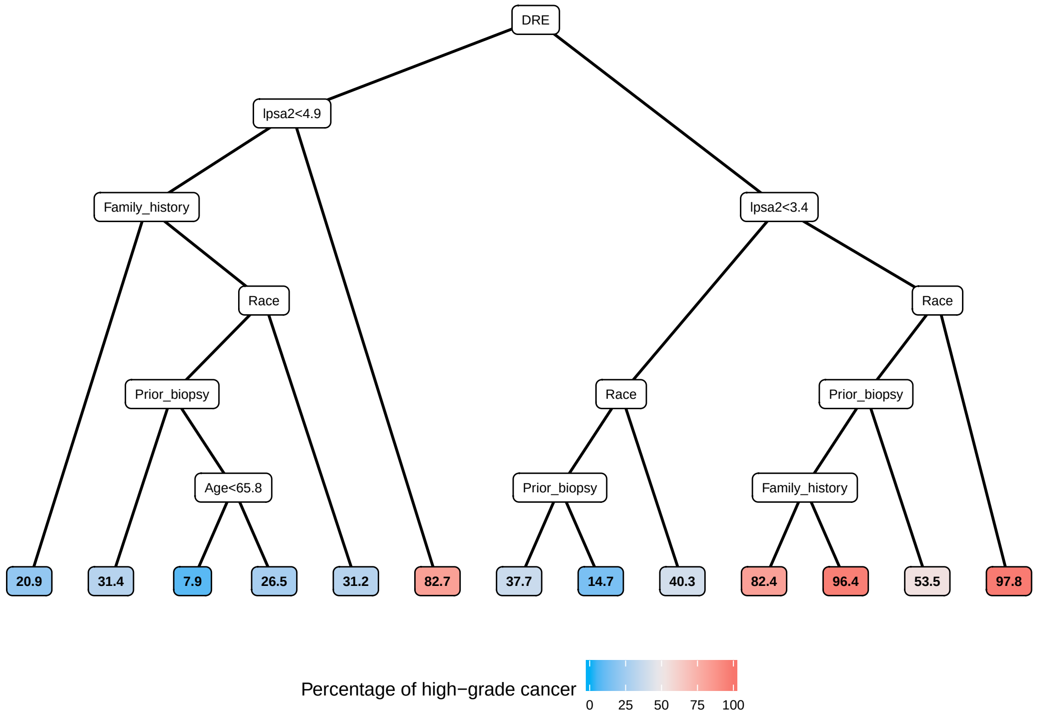


Figure A.20 : Smallest tree in final RF model with resulting percentages of high-grade cancer in the terminal nodes.

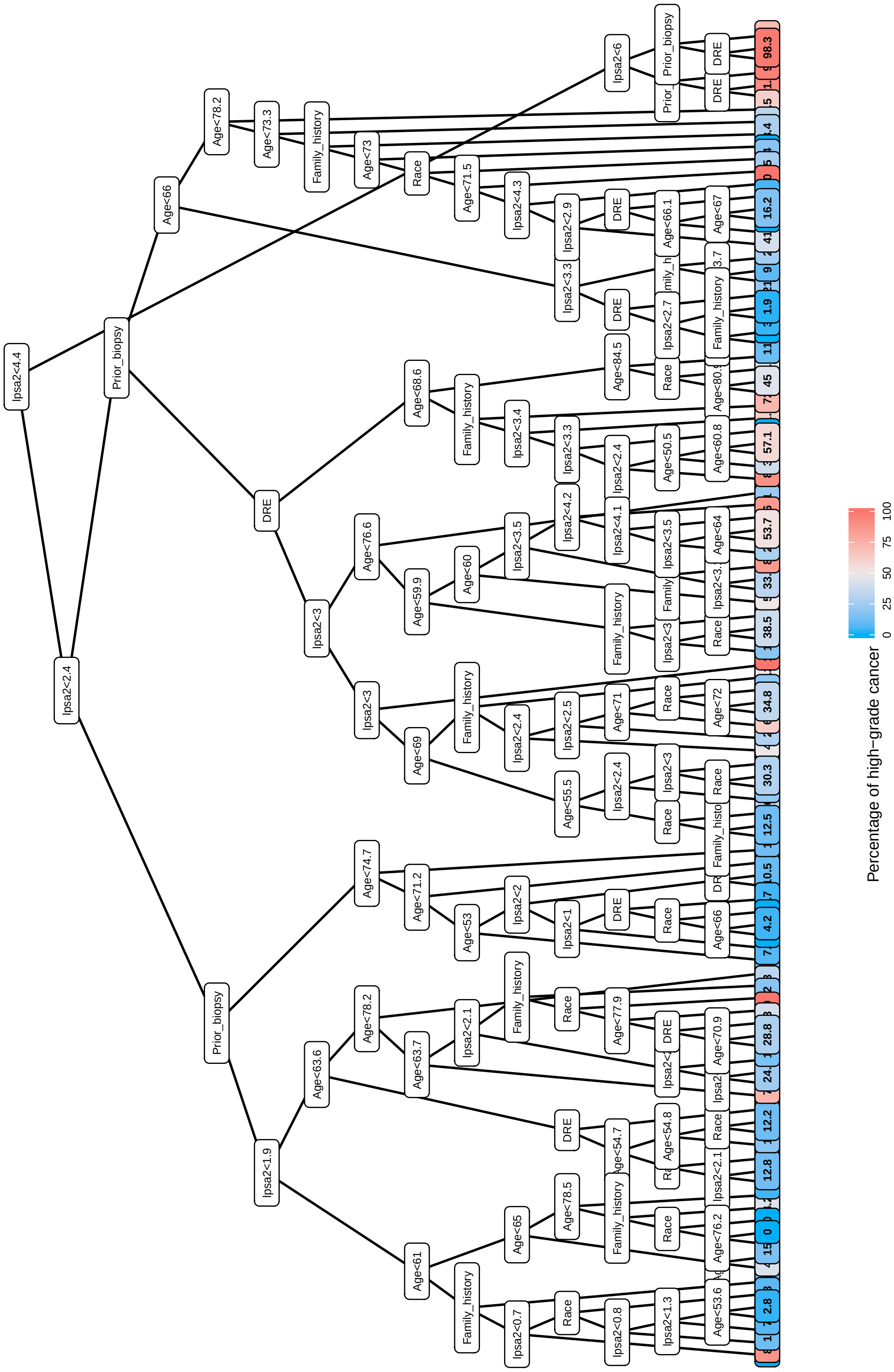


Figure A.21 : Largest tree in final RF model with resulting percentages of high-grade cancer in the terminal nodes.

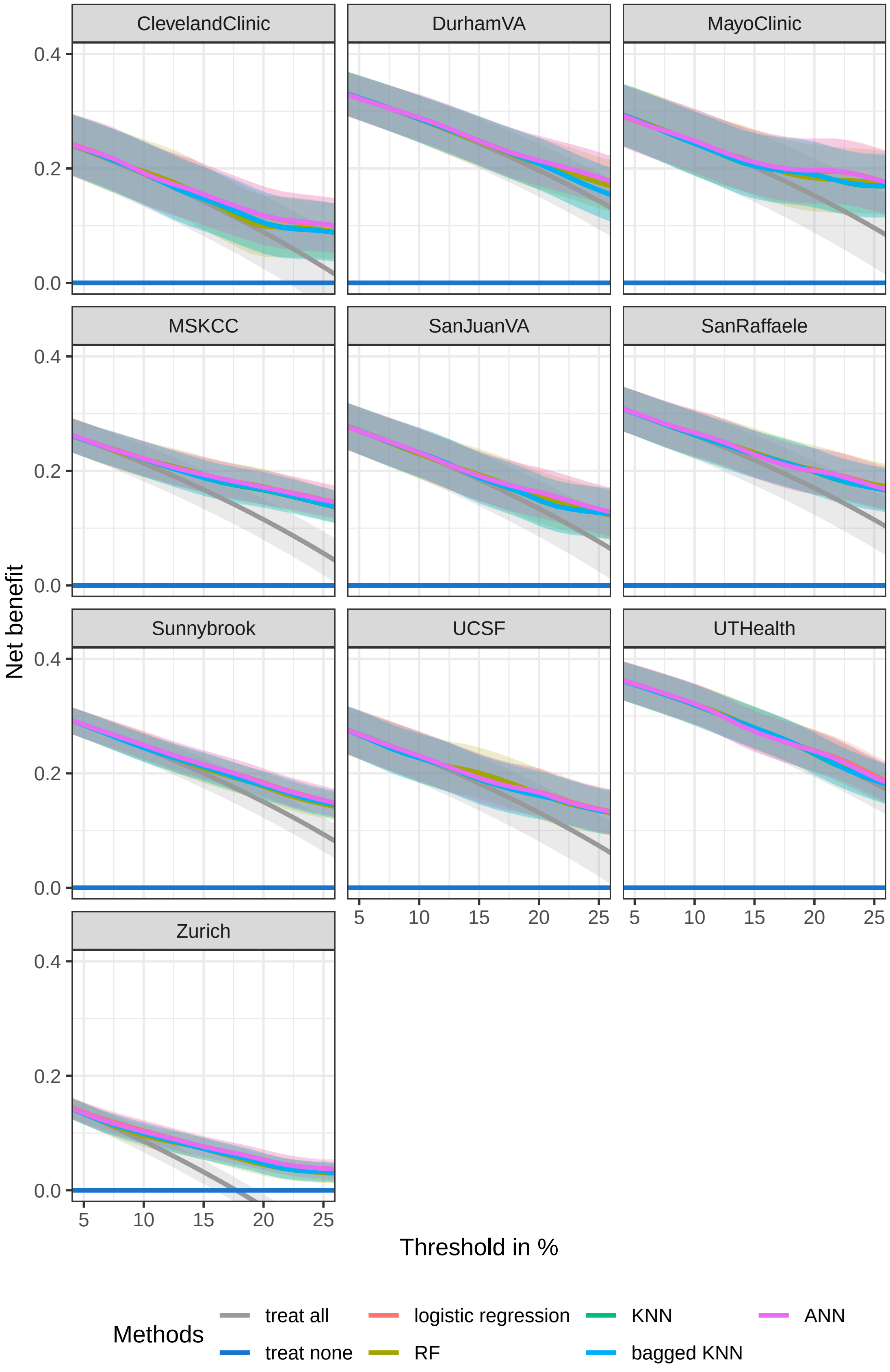


Figure A.22 : Net benefit curves for high-grade cancer of leave-one-cohort-out cross validation comparing comparing machine learning methods. They are given for each cohort separately and the strategies of referring all men or none to biopsy are given for comparison.

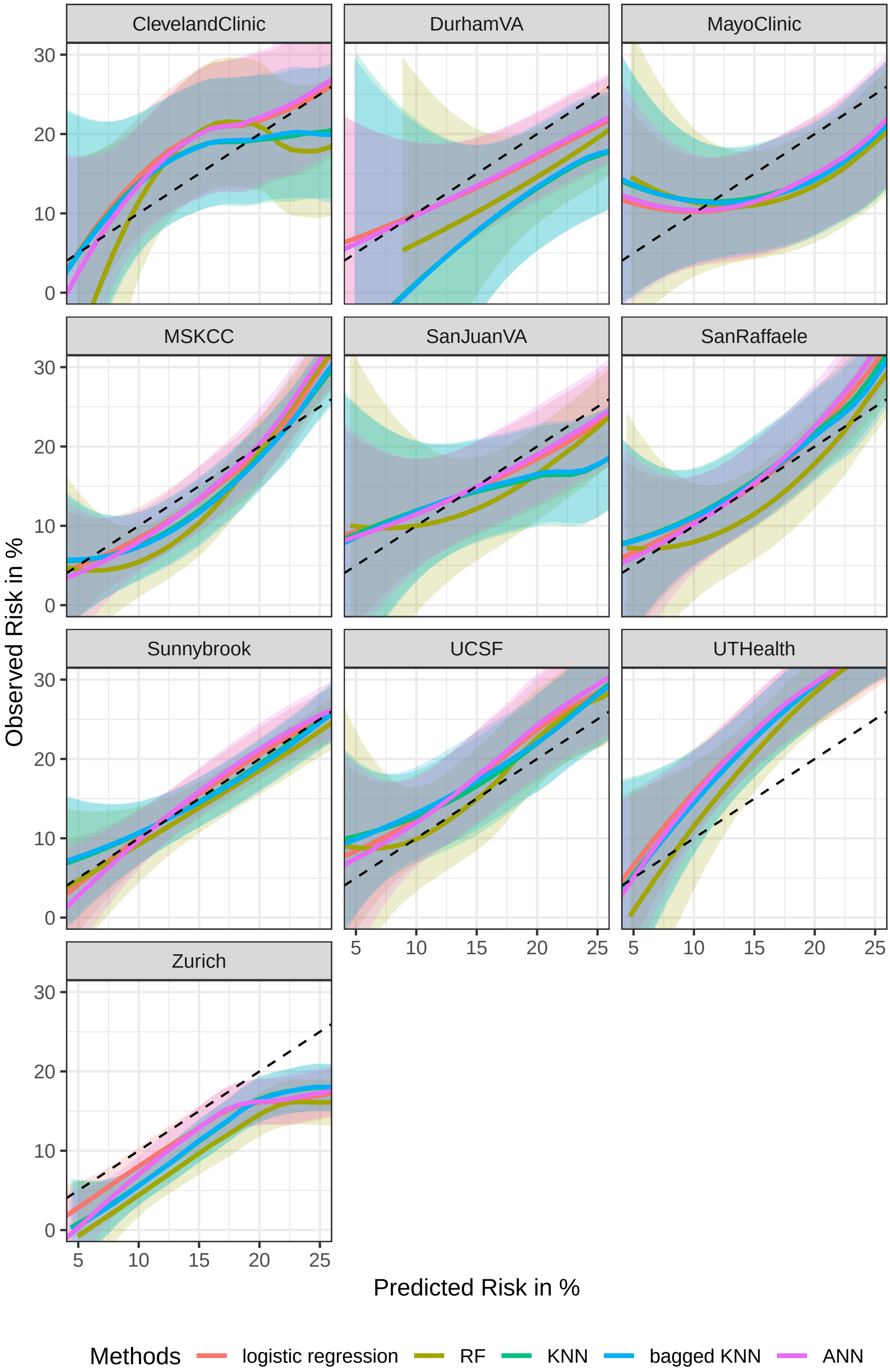


Figure A.23 : Calibration curves for high-grade cancer of leave-one-cohort-out cross validation comparing comparing machine learning methods. They are given for each cohort separately.

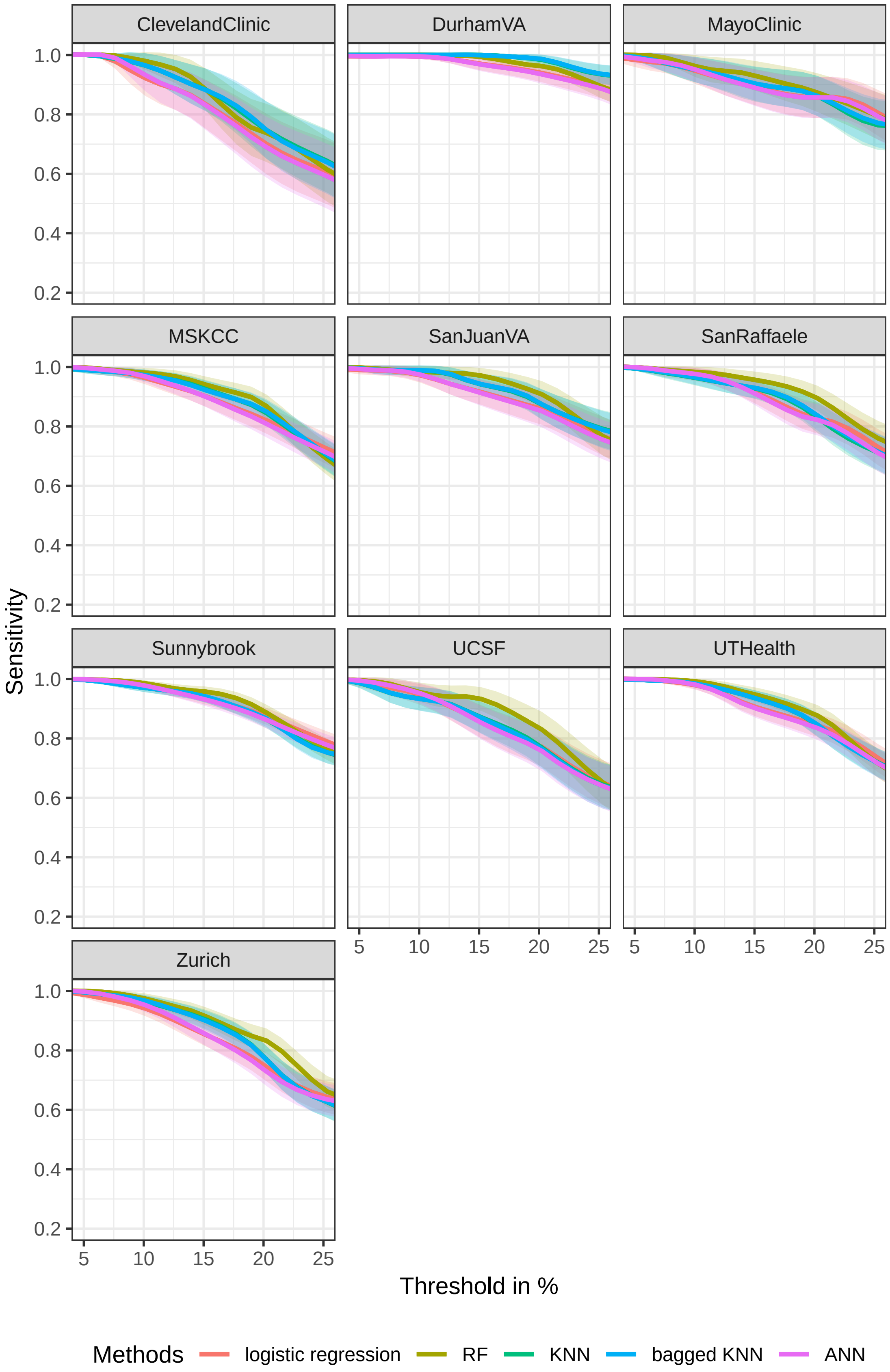


Figure A.24 : Sensitivity curves for high-grade cancer of leave-one-cohort-out cross validation comparing comparing machine learning methods. They are given for each cohort separately.

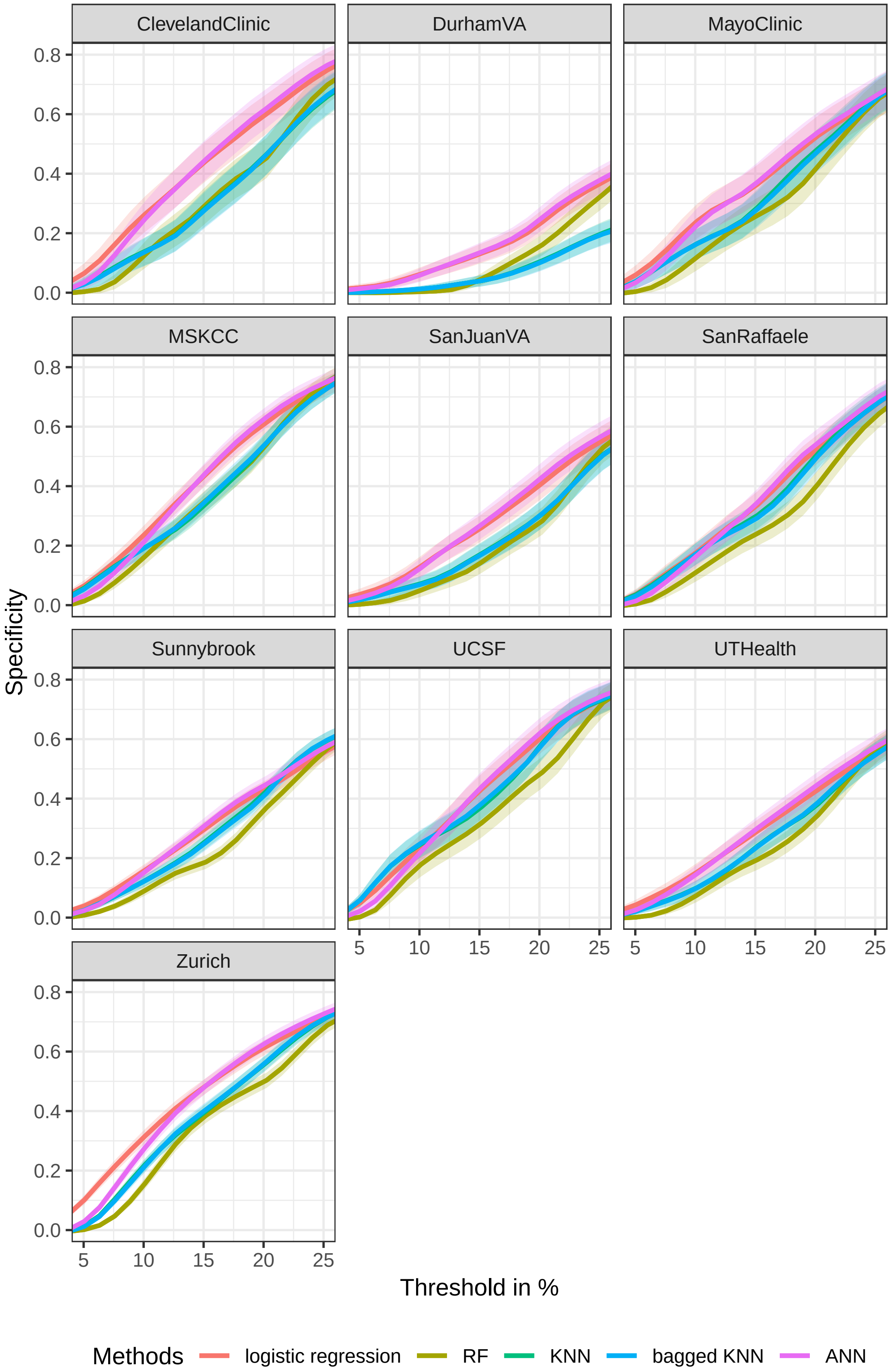


Figure A.25 : Specificity curves for high-grade cancer of leave-one-cohort-out cross validation comparing comparing machine learning methods. They are given for each cohort separately.

Acknowledgements

First of all, I would like to express my gratitude to my dedicated supervisor, Prof. Donna Ankerst. I am most grateful for her valuable input, constructive feedback, and support throughout my time as her PhD student. I consider myself very lucky that I had the opportunity to join her team and to profit from her professional knowledge as well as the great working environment. I could not have wished for a better advisor.

Thanks a lot to Prof. Aurélien Tellier and Prof. Jonathan Gelfond for serving on my PhD committee, I appreciate the time and effort you have spent. I also wish to express my gratitude to all co-authors and project partners for their scientific contribution and to the doctors from the Klinikum rechts der Isar for interesting collaborations and the chance to get some insights in their important daily work. Further thanks go to my graduate school mentor Dr. Regine Werner for keeping me in touch with the working world outside of academia.

Great thanks also go to Dr. Hannes Petermeier for all the support in teaching and organizational duties as well as great comments and conversations, both professional and personal. Furthermore I would like to thank all my colleagues at the Technical University of Munich, Lehrstuhl for Mathematical Modeling of Biological Systems of Prof. Fabian Theis: Thanks to my office mates – with some of which I could share the office for several years, with others I only had a few months - Katharina Selig, Eva Stadler, Anna Fiedler, Yiyao Chen, Bendix Koopmann, Augustine Okolie, Hanna Märkle and Thibaut Sellinger. It was great working with you, thanks a lot for the friendly atmosphere, various help and interesting discussions. Special thanks go to Silke Bauer for lifting all bureaucratic secrets and helping sort them out.

Furthermore, I would like to express special thanks to my family, in particular my parents Regina and Heinrich, who made my PhD possible in supporting me throughout my whole studies, in fact throughout my whole life.

Sincere thanks go to Lena Lermer, without her I would not have made the first year of studies and therefore would not have been able to start my PhD to begin with. But I would also like to thank her and all other friends, in particular the Stusta with all the great individuals I was fortune enough to meet during my studies and my PhD, for reminding me of the life outside the university.

Last, but by no means least, I would like to thank my husband Lukas. He encouraged me to start with the PhD and his patience in helping me face the hurdles occurring during the last few years in giving personal support as well as technical assistance was priceless. He managed to not only calm me down in nerve wracking times but to also motivate me when I got too comfortable. Furthermore, he was proofreading the whole manuscript. Thanks a lot for standing by my side this whole time!