Vorhersage der Überlebenswahrscheinlichkeit für Patientenuntergruppen mit hochdimensionalen Daten am Beispiel zweier Lungenkrebskohorten

Dissertation
zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
Dr. rer. nat.

vorgelegt
der Fakultät für Statistik der
Technischen Universität Dortmund

von
Christian Netzer

Dortmund im Juli 2013

Gutachter:
Prof. Dr. Jörg Rahnenführer
Prof. Dr. Katja Ickstadt

Tag der mündlichen Prüfung:
21. Oktober 2013

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die hier vorliegende zur Promotion eingereichte Arbeit mit dem Titel *„Vorhersage der Überlebenswahrscheinlichkeit für Patientenuntergruppen mit hochdimensionalen Daten am Beispiel zweier Lungenkrebskohorten“* selbstständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe.

Christian Netzer

Danksagung

Meinem Doktorvater Professor Dr. Jörg Rahnenführer danke ich für seine exzellente Betreuung. Er hat mir den nötigen Freiraum für meine Arbeit ermöglicht und hatte stets ein offenes Ohr für neue Ideen. Die vielen anregenden Diskussionen mit ihm waren äußerst wertvoll und für seine entgegengebrachte Zeit und Unterstützung bin ich ihm sehr dankbar.

Ich danke außerdem Professor Dr. Katja Ickstadt, die sich bereiterklärt hat diese Arbeit zu begutachten und ebenso wertvolle Anregungen gegeben hat.

Auch bin ich dem Bundesministerium für Bildung und Forschung, dem Förderer des Nationalen Genomforschungsnetzes (NGFN), zu Dank verpflichtet, welches durch die Finanzierung des Projektes einen erfolgreichen Abschluss meiner Arbeit möglich gemacht hat. Ebenfalls bedanke ich mich bei allen Kooperationspartnern. Stellvertretend möchte ich hier Laura Toloşi und Jasmina Bogojeska vom Max-Planck-Institut für Informatik, Thomas Zander (Uniklinik Köln), Jürgen Wolf (Centrum für Integrierte Onkologie Köln Bonn) sowie Roman Thomas (Max-Planck-Institut für neurologische Forschung) nennen.

Ich danke meinen Kollegen und Mitarbeitern, die mir bei allen Fragen mit Rat und Tat zur Seite standen. Besonders hervorheben möchte ich Michel Lang. Durch ihn habe ich viel beim Programmieren mit R dazugelernt. Ebenso danke ich der TU Dortmund für das Bereitstellen des LiDO Hochleistungsrechners.

Mein besonderer Dank gilt meiner Frau Tanja, die mir stets Mut zugesprochen und mich in meiner Arbeit bestärkt hat. Hätte sie mir nicht den Rücken freigehalten, wäre meine Arbeit in dieser Form nicht möglich gewesen.

Inhaltsverzeichnis

1 Einleitung

Die Vorhersage der Überlebenswahrscheinlichkeit spielt eine entscheidende Rolle in der Krebsforschung. Die Überlebenszeit steht in direktem Zusammenhang mit der Malignität der Tumoren. Präzise Vorhersagen können dabei helfen, die Tumoren zu differenzieren und eine bestmögliche Therapie und Nachsorge der Patienten sicherzustellen.

Meist wird die Überlebenszeit durch ein statistisches Modell prognostiziert, welches anhand der Ausprägungen eines oder mehrerer Merkmale (Variablen) versucht Rückschlüsse auf die Überlebenszeit zu ziehen. Ein solches Modell muss zuvor auf einer Datenmenge gelernt bzw. trainiert werden. Neben klinischen Merkmalen, wie beispielsweise dem Alter oder der Histologie sind vor allem genetische Variablen von Interesse. Diese können entweder durch die Mutationsstatus einzelner Gene, Expressionswerte, Einzelnukleotid-Polymorphismen (SNPs) oder Kopienzahlvariationen gegeben sein. Letztere, auch CNVs abgekürzt (kurz für Copy Number Variations) werden in dieser Arbeit betrachtet. Aufgrund der immer feiner auflösenden Messverfahren, mit denen solche genetischen Variablen erhoben werden, nimmt die Dimension der Daten mit der Zeit weiter zu. Das dadurch resultierende Problem, das Modell aus Daten mit deutlich mehr Variablen als Beobachtungen trainieren zu müssen, wird auch als das $p \gg n$ Problem bezeichnet. Das Ziel einer derartigen Modellbildung sollte es daher sein, das entsprechende Modell auf Basis so vieler Beobachtungen wie möglich zu trainieren. Die Beobachtungen müssen indessen zwangsläufig homogen im Sinne der Effekte der Variablen auf die zu untersuchende Zielvariable (hier die Überlebenszeit) sein. Allerdings sind genetische sowie auch klinische Variablen extrem heterogen über alle Patienten verteilt. Dieses Bild wird in vielen Fachartikeln bestätigt und zeigt sich in gleicher Weise bei Betrachtung diverser Krebsarten (Chiaretti u. a. (2004), Mavaddat u. a. (2010), Bhatia u. a. (2012), Turner und Reis-Filho (2012), Yap u. a. (2012), Prat u. a. (2013)) und im Besonderen im Bereich Lungenkrebs (Russell u. a. (2011), Sun u. a. (2007), Garber u. a. (2001)). Es ist somit in Frage zu stellen, ob und inwieweit ein bestimmter Zusammenhang zwischen einem bzw. mehreren Merkmalen und der Zielvariablen innerhalb der gesamten Patientenkohorte Gültigkeit besitzt. Vielmehr liegt die Vermutung nahe, dass gewisse Effekte lediglich in einer Teilmenge (Untergruppe) der Kohorte vorzufinden sind. Sollte die Überlebenszeit gemeinsam auf Basis aller Patienten in einer Kohorte modelliert werden, kann dies in solchen Situationen zu den folgenden Problemen führen:

- Eine in der Untergruppe prognostische Variable wird nicht erkannt. Ursache ist eine zu kleine Untergruppe im Verhältnis zur Gesamtkohorte
- Hat eine Variable in der Untergruppe einen entgegengesetzten Effekt auf die Überlebenszeit, so kann bei Betrachtung aller Patienten eine falsche Schlussfolgerung

für diese Untergruppe getroffen werden, indem in der Gesamtkohorte der Effekt der restlichen Patienten überwiegt

- Eine Variable ist nur für die Untergruppe prognostisch, sonst jedoch nicht. Trotzdem wird als Folge diese Variable in der Gesamtkohorte als signifikant eingestuft. Der Effekt wird dann in der Untergruppe möglicherweise zu schwach eingeschätzt. Zudem werden falsche Schlussfolgerungen für die nicht zu der Untergruppe gehörenden Patienten gezogen

In all diesen Szenarien ist ein Zusammenhang zwischen Merkmal und Überlebenszeit spezifisch für eine Untergruppe. Bekannte Verfahren oder Ansätze wie etwa Survival Bäume oder eine stratifizierte Analyse im Cox-Modell können keine der oben skizzierten Zusammenhänge korrekt erkennen. Natürlich kann eine Variable ebenso innerhalb der gesamten Kohorte denselben Einfluss auf die Zielvariable ausüben. Die Kohorte sollte dementsprechend stets komplett analysiert werden, um für solche Effekte die Fallzahl nicht zu verkleinern. Auf eine entsprechende Adjustierung bezüglich multipler Vergleiche muss dabei geachtet werden. Allgemein können Variablen global oder niemals oder nur in der Untergruppe relevant sein.

Untergruppen können beispielsweise durch klinische Variablen, einzelne Mutationen, Cluster, Batches, o.ä. charakterisiert sein. In allen Fällen werden diese Gruppen über eine Untergruppenvariable definiert. Die Variable *Geschlecht* etwa definiert die beiden Untergruppen *Frauen* und *Männer*. Ob die entscheidende Untergruppenvariable beobachtet wurde ist allerdings nicht bekannt. Ferner ist nicht gesagt, dass die Werte einer beobachteten Untergruppenvariablen richtig gemessen wurden. Beispielsweise kann durch einen simplen Dateneingabefehler das Geschlecht falsch zugeordnet sein. Die Anzahl unterschiedlicher, nur in Untergruppen vorzufindender Zusammenhänge ist ebenfalls nicht bekannt.

In dieser Arbeit werden potentielle Untergruppenvariablen daraufhin untersucht, ob ein auffälliger Unterschied bezüglich der Überlebenszeitmodelle erkennbar ist, wenn im Vergleich dazu die gesamte Kohorte herangezogen wird. Im Folgenden werden somit ausschließlich beobachtete Variablen als Untergruppenvariablen aufgefasst und analysiert. Eine mögliche Herangehensweise ist zuvor definierte Untergruppen separat zu betrachten. Die Modelle werden dann lediglich unter Verwendung der zu der Untergruppe gehörenden Patienten gebildet. Bei sehr kleinen Untergruppen besteht jedoch die Gefahr, dass die tatsächlichen Zusammenhänge nicht präzise geschätzt werden oder womöglich nicht im resultierenden Modell wiederzufinden sind. Dazu wird nachfolgend ein Modell vorgestellt, welches alle Patienten einer Kohorte nutzt, um ein Überlebenszeitmodell für eine bestimmte Untergruppe zu schätzen. Das Problem einer zu kleinen Stichprobengröße der Untergruppe wird damit umgangen. Die Modellbildung beruht hierbei auf Stichprobengewichten. Die Beobachtungen mit ihren jeweiligen Merkmalsausprägungen tragen dadurch im unterschiedlichen Maße zur Modellierung bei. Eine vergleichbare gewichtete Modellbildung findet ebenfalls bei sogenannten lokalen Regressionsverfahren Anwendung. Die in dieser Arbeit vorgestellte Gewichtung basiert auf einer theoretischen Herleitung auf Basis der Annahme unterschiedlicher Verteilungen der Daten in

den einzelnen Untergruppen und unterscheidet sich grundlegend von den bisherigen lokalen Verfahren. Die Grundidee dabei ist die Höhergewichtung von nicht zu der Untergruppe gehörenden Beobachtungen, wenn diese von ihren Merkmalen her denen der Untergruppe ähneln. Jede Beobachtung erhält somit ein individuelles Gewicht, welches der Wahrscheinlichkeit der jeweiligen Untergruppe anzugehören gleichkommt. Diese Idee wurde ursprünglich von Bickel (2009) vorgeschlagen und zur Modellierungen von Therapieerfolgen bei HIV Patienten genutzt (Bickel u. a., 2008). Die Methodik wird in dieser Arbeit auf Überlebenszeitmodelle für Patientenuntergruppen übertragen. Ferner wird ein Modellbildungsprozess evaluiert, bei dem sich die Stichprobengewichte lediglich zwischen der Untergruppe und der jeweiligen Restgruppe unterscheiden. Eine aufwendige Bestimmung individueller Gewichte entfällt in diesem Fall.

Auf diese Weise ergeben sich mehrere Modelle, die demselben Zweck dienen; der Modellierung der Überlebenszeit in einer Untergruppe mittels klinischer und genetischer Variablen. Ziel der Arbeit ist ein Vergleich dieser Modelle entlang aller betrachteten Untergruppen. Damit sollen Unterschiede zwischen einem speziell für eine Untergruppe angepassten Modell und dem auf Basis der gesamten Kohorte gelernten Modell aufgedeckt werden. Dadurch können bestenfalls für eine Untergruppe spezifische Effekte erkannt werden. Die Modelle werden dabei anhand ihrer Vorhersagegenauigkeit miteinander verglichen.

Gliederung

Kapitel 2 beginnt mit einer kurzen Übersicht über epidemiologische Kennzahlen für Lungenkrebs. Des Weiteren werden histologische Subtypen und die Einteilung in Krankheitsstadien beschrieben. Hierbei wird zudem auf bereits bekannte Unterschiede der Überlebenswahrscheinlichkeiten nach diesen Unterteilungen eingegangen. Als nächstes werden die Verfahren zur Gewinnung der genetischen Information der Patienten vorgestellt. Zentraler Aspekt dieser Arbeit ist die Schätzung der Überlebenswahrscheinlichkeiten aus klinischen sowie genetischen Variablen. Es werden nötige Vorverarbeitungsschritte vorgestellt, um die genetischen Variablen zur Vorhersage nutzen zu können. Abschließend werden in Kapitel 2 die beiden Patientenkohorten vorgestellt, in denen die Untergruppen untersucht werden. Auf die Datenqualität wird ebenso eingegangen, wie auch auf Unterschiede im Hinblick auf die Überlebensraten in den Kohorten.

Kapitel 3 stellt die grundlegenden Konzepte der Überlebenszeitanalyse vor. Zunächst werden elementare Grundbegriffe und Modelle wie die Überlebensfunktion und die Cox Regression aufgeführt. Danach werden Algorithmen zur Modellbildung diskutiert, die für hochdimensionale Daten geeignet sind. Die hohe Dimension ist eine der größten Herausforderungen im Umgang mit genetischen Daten. Schließlich wird eine weit verbreitete Methode zur Bewertung von Überlebenszeitvorhersagen aufgeführt. Damit werden die vorgestellten Modelle miteinander verglichen. Wie auch bei der Modellbildung ist diese Methode darauf ausgerichtet zensierte Beobachtungen zu berücksichtigen.

Kapitel 4 stellt Modellbildungsverfahren gegenüber, womit Überlebenszeitmodelle für eine bestimmte Teilmenge aller Beobachtungen (Untergruppe) angepasst werden. Da-

bei wird insbesondere das Verfahren vorgestellt, welches es ermöglicht über individuelle Stichprobengewichte alle zur Verfügung stehenden Beobachtungen zu verwenden, so dass gezielt die Zusammenhänge in der jeweiligen Untergruppe modelliert werden können. Der erste Teil aus Kapitel 4 befasst sich im Detail mit der Schätzung dieser Stichprobengewichte und rechtfertigt die Vorgehensweise über eine theoretische Herleitung. Ferner wird eine alternative Gewichtung über einfachere Stichprobengewichte diskutiert. Im zweiten Teil wird der genaue Versuchsaufbau zur Schätzung und Evaluierung der Überlebenszeitmodelle auf allen zu untersuchenden Untergruppen der beiden Patientenkohorten beschrieben. Zudem wird an dieser Stelle beispielhaft auf künstlich erzeugten Daten gezeigt, in welchen Situationen die Modellbildung mittels individueller Stichprobengewichte gegenüber einer Modellierung ohne diese Gewichte zu besseren Überlebenszeitvorhersagen führt.

Kapitel 5 gibt zunächst einen Überblick über alle untersuchten Untergruppen der beiden Patientenkohorten. Anschließend wird die Schätzung der individuellen Stichprobengewichte analysiert. Zuletzt werden die Vorhersagen der angewandten Modelle für jede Untergruppe berechnet und diesbezügliche Unterschiede zwischen den Modellen entlang der Untergruppen aufgezeigt und erörtert.

Kapitel 6 fasst abschließend die Ergebnisse der Untergruppenanalyse zusammen und bietet eine ausführliche Diskussion der angewandten Verfahren. Ebenso werden mögliche Erweiterungen sowie alternative Vorgehensweisen diskutiert.

2 Lungenkrebs: Übersicht und Datenmaterial

Abschnitt 2.1 gibt zunächst einen kurzen Überblick über die Klassifikation und die Stadieneinteilung von Bronchialkarzinomen. Neben der Prävalenz der einzelnen Untertypen, die durch histopathologische Befunde oder Stadien charakterisiert werden, wird in diesem Abschnitt die allgemeine sowie die subtypen-spezifische Überlebensrate angegeben und es werden bereits bekannte Unterschiede aufgezeigt und diskutiert. In Abschnitt 2.2 wird das Verfahren zur Detektion genetischer Veränderungen auf chromosomaler Ebene der Tumoren erläutert. Diese genetische Information bildet die Basis, um Rückschlüsse von genetischen Veränderungen auf die Überlebenszeit schließen zu können. Der letzte Abschnitt 2.3 gibt Auskunft über die beiden in dieser Arbeit verwendeten Patientenkohorten.

2.1 Epidemiologie, Klassifikation und Überlebensraten

Die Lunge ist die dritthäufigste Tumorlokalisation an allen Krebsneuerkrankungen in Deutschland bei Männern (13,8 %) sowie bei Frauen (7,0 %). Aufgrund der schlechten Prognose ist Lungenkrebs jedoch die häufigste Krebstodesursache bei Männern (25,5 %) und hinter Brustkrebs die zweithäufigste Krebstodesursache bei Frauen (16,0 %) (Quelle: `www.dkfz.de`, Stand: 2010). Das relative 5-Jahres Überleben beträgt gerade 15 % bei Männern und 19 % bei Frauen. Die Mortalitätsrate ist bei Männern seit Mitte der 1980er-Jahre rückläufig, wohingegen sie bei Frauen einen stetigen Aufwärtstrend zeigt. Dieser gegenläufige Trend wird oft auf eine unterschiedliche Entwicklung des Rauchverhaltens in den beiden Geschlechtergruppen zurückgeführt. Diese und weitere Angaben über die Erkrankungshäufigkeit und Sterblichkeit sind in der aktuellen Ausgabe "Krebs in Deutschland" vom Robert Koch-Institut (2012) zu finden.

Die wichtigste histologische Unterscheidung erfolgt im Wesentlichen in kleinzellige und nicht-kleinzellige Lungenkrebskarzinome. Dabei macht die Form des kleinzelligen Lungenkrebs, mit SCLC (small cell lung cancer) bezeichnet, etwa 15 bis 20 % aller Lungenkrebsfälle aus. Die übrigen nicht-kleinzelligen Karzinome, mit NSCLC (non-small cell lung cancer) bezeichnet, lassen sich histologisch in weitere Gruppen unterteilen. Hierzu zählt das mit einem Anteil von 30 bis 40 % am häufigsten auftretende Plattenepithelkarzinom. Dieses geht von Schleimhautdeckzellen aus und ist durch eine spindelzellige (squamöse) Erscheinung charakterisiert. Es wird hier mit SQ (squamous) abgekürzt. Mit

Tabelle 2.1: TNM-Klassifikation von Bronchialkarzinomen (siehe Pschyrembel, 2007, Seite 282, Tabelle 2)

TNM	Tumorwachstum
TX	Primärtumor kann nicht beurteilt werden oder positive Zytologie
T0	kein Anhalt für Primärtumor
Tis	Carcinoma in situ
T1	Tumor ≤ 3 cm, Hauptbronchus und viszerale Pleura frei
T2	Tumor > 3 cm oder ≥ 2 cm distal der Carina oder Invasion der viszeralen Pleura oder tumorassoziierte Atelektase oder Ausbreitung in Hilusregion
T3	Infiltration von Brustwand, Zwerchfell, Perikard oder mediastinaler Pleura oder < 2 cm distal der Carina oder totale Atelektase einer Lunge
T4	Infiltration von Mediastinum, Herz, großen Gefäßen, Trachea, Speiseröhre, Wirbelkörper oder Carina oder maligner Erguss oder Satellitenmetastasen im vom Primärtumor befallenen Lungenlappen
N0	keine regionären Lymphknotenmetastasen
N1	ipsilaterale peribronchiale und/oder hiläre Lymphknoten
N2	ipsilaterale mediastinale und/oder subkarinale Lymphnoten
N3	kontralaterale mediastinale oder hiläre, ipsi- oder kontralaterale Skalenus- oder supraklavikuläre Lymphknoten
M0	keine Fernmetastasierung
M1	Fernmetastasierung

einer Häufigkeit von 25 bis 30 % tritt das von drüsenartigen Zellen abstammende Adenokarzinom, kurz AD, auf. Das großzellige Karzinom, kurz LC (large cell), ist mit einem Anteil von weniger als 10 % deutlich seltener. Ebenfalls seltener sind adenosquamöse Karzinome oder Karzinoidtumoren (kurz: CA). Die hier genannten Subtypen lassen sich oft noch weiter differenzieren (siehe Hammerschmidt und Wirtz, 2009).

Die Einteilung der Tumoren in verschiedene Stadien beschreibt die anatomische Ausdehnung des Tumors. Grundlage hierfür bildet die von der Union for International Cancer Control (UICC) festgelegte TNM-Klassifikation (Sobin und Compton, 2010). Dabei beschreibt die Kategorie T (Tumor) die Ausdehnung des Tumors, N (Nodulus) das Fehlen oder Vorhandensein von regionären Lymphknotenmetastasen und M (Metastase) das Fehlen oder Vorhandensein von Fernmetastasen. Der entsprechende Grad der Ausdehnung bezüglich einer Kategorie wird durch eine Ziffer gekennzeichnet. Zur Einteilung in

Tabelle 2.2: Stadieneinteilung von Bronchialkarzinomen (siehe Pschyrembel, 2007, Seite 282, Tabelle 2). T: Primärtumor; N: regionäre Lymphknoten; M: Fernmetastasen (vgl. Tabelle 2.1).

Stadium	T	N	M
0	TX	N0	M0
	Tis	N0	M0
I A	T1	N0	M0
I B	T2	N0	M0
II A	T1	N1	M0
II B	T2	N1	M0
	T3	N0	M0
III A	T3	N1	M0
	T1-3	N2	M0
III B	jedes T	N3	M0
	T4	jedes N	M0
IV	jedes T	jedes N	M1

Stadien (Staging) schreibt die UICC ebenfalls vor, welche Kombination der T-, N- und M-Kategorien jeweils einem bestimmten Stadium zugeordnet werden (siehe Tabellen 2.1 und 2.2). Daraus ergeben sich die Stadien I bis IV, die gegebenenfalls weiter unterteilt werden können (z.B. in I A und I B). Stadium I entspricht beispielsweise einem kleinen Tumor, Stadium IV beschreibt unabhängig vom Grad der Kategorien T und M einen Tumor mit Fernmetastasierung.

Im Hinblick auf die Überlebenszeiten der Patienten stellt insbesondere das Stadium des Tumors einen wichtigen Faktor in Bezug auf die Prognose dar. Mit aufsteigendem Stadium nimmt die zu erwartende Überlebenszeit ab. Ebenso kann die Histologie als wichtiger Indikator zur Vorhersage der Überlebenszeit eines Patienten angesehen werden. Beispielsweise weisen SCLC Patienten eine relativ kurze Überlebenszeit auf. So beträgt die Überlebenszeit in diesem Fall kaum mehr als fünf Jahre (Rosti u. a., 2006). In der Literatur finden sich vermehrt Arbeiten, in denen Kombinationen dieser klinischen Merkmale, wie die Histologie oder das Stadium, in Bezug auf Zusammenhänge mit der Überlebenszeit der Patienten untersucht werden. So stellen beispielsweise Cetin u. a. (2011) fest, dass die Überlebenswahrscheinlichkeit von Patienten mit einem Stadium IV Bronchialkarzinom in Abhängigkeit des Histologietyps variiert. Adenokarzinome (AD) weisen hierbei die höchste Überlebensrate auf, wobei sich ein großzelliges Karzinom (LC) negativ auf die Überlebenswahrscheinlichkeit auswirkt. Solche Resultate sind jedoch kritisch zu hinterfragen. Die simple Betrachtung mehrerer Subtypen oder gar Kombinationen von Subtypen ohne eine geeignete Validierung der Ergebnisse zieht einen erhöhten Fehler 1. Art nach sich (Dijkman u. a., 2009), so dass fälschlicherweise von einem Einfluss auf die Überlebenszeit ausgegangen wird. Auf eine Kontrolle der `FWER` (familywise error rate) sollte daher geachtet werden.

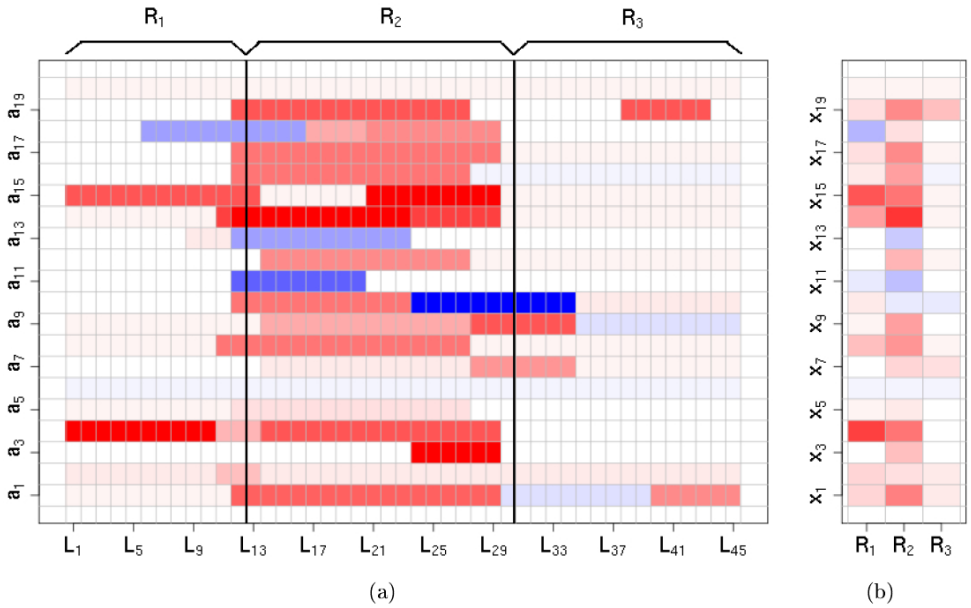
2.2 Detektion chromosomaler Aberrationen

Die Erkennung von DNA-Regionen mit veränderter Kopienzahl ist ein wichtiger Bestandteil in der Erforschung der Onkogenese. Ein Zugewinn (Amplifikation) oder ein Verlust (Deletion) bestimmter Regionen auf dem Chromosom kann in Verbindung mit der Krankheit gebracht werden, um so beispielsweise Hinweise auf die Prognose oder die Therapie zu geben. So kann das Fehlen von DNA-Stücken in einem Tumor auf ein Suppressorgen hinweisen oder umgekehrt bei Amplifikation auf ein Onkogen hindeuten.

Die Array-CGH Technik (Microarray-basierte komparative genomische Hybridisierung) ist neben der Genexpressionsanalyse ein weit verbreitetes Verfahren zur Auffindung chromosomaler Veränderungen. Erstmals wurde diese Technik von Solinas-Toldo u. a. (1997) angewandt. Beim Array-CGH Verfahren werden die Tumor-DNA und eine Kontroll-DNA aus gesundem Gewebe jeweils unterschiedlich farblich markiert und gemeinsam auf dem Microarray-Chip hybridisiert. Ein Scanner misst anschließend die jeweilige Farbintensität. Das Verhältnis der beiden Farbintensitäten gibt Aufschluss über veränderte Kopienzahlen im Tumorgewebe. Üblicherweise wird das Verhältnis der beiden Intensitäten logarithmiert. Sei beispielsweise R (Rot) die Intensität im Tumorgewebe und B (Blau) die Intensität entsprechend in der Referenzprobe. Dann wird $\log_2(R/B)$ als das *log-ratio* bezeichnet. Liegt keine Veränderung der Kopienzahl in der Tumor-DNA Probe vor, so entspricht dies einem log-ratio von Null. Gewöhnlich sind in diesem Fall zwei Kopien in jeder Probe enthalten. Sind doppelt so viele Kopien der DNA im Tumor im Vergleich zur Referenz vorhanden, so ergibt sich ein log-ratio von Eins. Im umgekehrten Fall ist das log-ratio gerade gleich -1. Da die log-ratios technologiebedingt einen Bias aufweisen, werden diese stets normalisiert (siehe Toloşi, 2012, Seite 30).

Die einzelnen Positionen (loci) auf dem Chromosom, für die jeweils ein log-ratio ermittelt wurde, werden anschließend zu Regionen zusammengefasst. Dabei werden benachbarte Positionen mit annähernd gleichem log-ratio gesucht. Dieser Schritt der Vorverarbeitung wird *Segmentation* genannt. Für jede Region wird stellvertretend ein log-ratio ermittelt (Level). Die Stellen auf dem Chromosom, die zwei Regionen voneinander trennen, werden *Breakpoints* genannt. Der verbreitetste Algorithmus zur Bestimmung der Breakpoints sowie der jeweiligen Level ist der CBS (Circular Binary Segmentation) Algorithmus von Olshen u. a. (2004). Mit Hilfe einer Likelihood Ratio Teststatistik wird die Hypothese überprüft, ob sich die mittleren log-ratios zweier disjunkter Intervalle auf dem Chromosom unterscheiden. Der CBS wurde später von Venkatraman und Olshen (2007) in Bezug auf die Laufzeit optimiert.

Die Segmentierung durch CBS findet lediglich auf Basis eines Arrays bzw. eines Patienten statt. Eine Herausforderung ist die Erkennung und Festsetzung einheitlicher Regionen über mehrere Arrays hinweg. Abbildung 2.1 (a) skizziert die Notwendigkeit gemeinsamer Breakpoints bzw. gemeinsamer Regionen ausschnittsweise an den Arrays $a_1, \ldots, a_{20}$ und den Positionen $L_1, \ldots, L_{45}$. Der von Beroukhim u. a. (2007) vorgestellte GISTIC Algorithmus (Genomic Identification of Significant Targets in Cancer) ist ein gebräuchliches Verfahren zur Findung von einheitlichen Regionen. Durch die Mittelung der Level jeder einzelnen Position über alle Arrays hinweg geht jedoch Information verloren. Toloşi (2012) schlägt drei neue, unter dem Begriff *Consensus Segmentation* zu-



(a) (b)

Abbildung 2.1: (a): Schematische Darstellung von 20 segmentierten Arrays (Ordinatenachse) an 45 Positionen (Abszissenachse). Die Farbintensität gibt das log-ratio an, wobei rot eine Amplifikation und blau eine Deletion beschreibt. (b): Ergebnis des Consensus Segmentation Algorithmus. $R_1 = \{L_1, \ldots, L_{12}\}$, $R_2 = \{L_{13}, \ldots, L_{30}\}$ und $R_3 = \{L_{31}, \ldots, L_{45}\}$ sind gemeinsame Regionen. Quelle: Tolo̧si (2012), Seite 54.

sammengefasste Algorithmen vor (Tolo̧si, 2012, Kapitel 4). Diese sind vom Prinzip her eine multivariate Erweiterung der oben beschriebenen Segmentierung. Ohne über die Positionen zu mitteln, werden in einem ersten Schritt entweder die gemeinsamen Breakpoints (Algorithmen CB-MUG und CB-KeS) oder direkt gemeinsame Regionen (CR-FC Algorithmus) gefunden, wobei für diese Arbeit der zuletzt genannte CR-FC (Consensus Regions via Feature Clustering) Algorithmus angewandt wurde. Hierbei werden die einzelnen Positionen entlang aller Arrays über ein Hierarchisches Clusterverfahren zusammengefasst. Diese Vereinigung von einzelnen Positionen bzw. Clustern wird mit dem Complete Linkage Verfahren vorgenommen (siehe Tolo̧si, 2012, Seite 70). Die optimale Anzahl an Clustern wird dabei über die *weighted clustering balance* von Jung u. a. (2003) bestimmt. Diese beschreibt den Kompromiss zwischen dem Inter- und dem Intra-Cluster-Abstand. Die resultierenden Cluster definieren allerdings noch keine Regionen. Eine Region darf lediglich aus unmittelbar angrenzenden Positionen bestehen. Um letztendlich die gemeinsamen Regionen zu erhalten, wird jedes Cluster als Vereinigung von Regionen aufgefasst. Alle Regionen, die auf diese Weise ein Cluster definieren, stellen die endgültige Menge der gemeinsamen Regionen dar. Im zweiten Schritt (Abbildung 2.1 (b)) wird für jedes Array das mittlere Level über die Region hinweg berechnet. Die resultierende Matrix gibt nun für jeden Patienten (Array) und für jede Region R_1, R_2, R_3 ein mittleres log-ratio an. Als Nebeneffekt wird gleichzeitig die Dimension der Daten reduziert.

2.3 Beschreibung der Patientenkohorten

Die Analyse von Untergruppen in Bezug auf Einflüsse genetischer Faktoren auf die Überlebenszeit wird auf zwei unabhängigen Patientenkohorten durchgeführt. Die erste Kohorte (Köln) enthält Datensätze von 833 Patienten. Die Daten wurden im Rahmen des Clinical Lung Cancer Genome Project (CLCGP) gesammelt, welches über das Nationale Genomforschungsnetz (NGFN) gefördert wurde. Die Daten wurden dabei von mehreren Kliniken weltweit zusammengeführt. Die zweite, kleinere Kohorte (Uppsala) enthält Datensätze von 192 Lungenkrebspatienten. Die Daten wurden der nationalen Biobank sowie dem Lungenkrebsregister Schwedens entnommen und stammen von Patienten aus der Region Uppsala-Örebro (Edlund, 2012).

Die Köln Kohorte umfasst Datensätze von sowohl nicht-kleinzelligen (NSCLC) als auch von kleinzelligen (SCLC) Karzinomen. In der Uppsala Kohorte sind lediglich Datensätze von NSCLC Patienten enthalten. Beide Kohorten beinhalten Informationen zu mehreren klinischen Merkmalen sowie Messungen von Genkopienzahlvariationen (copy number variation, kurz: CNV). Eine detaillierte Übersicht über sämtliche klinischen Merkmale ist in Kapitel 5.1 gegeben.

Die CNVs wurden auf Basis von Affymetrix Chips gemessen. Dafür kamen in den Kohorten unterschiedliche Chips zum Einsatz. Die Messungen der Uppsala Kohorte stammen von einem GeneChip Human Mapping 500K Array (Affymetrix, 2006). Die CNVs in der Köln Kohorte wurden mit dem neueren Genome-Wide Human SNP Array 6.0 (Affymetrix, 2009) erfasst, welcher eine wesentlich höhere Auflösung besitzt. In beiden Kohorten wurden jeweils mit dem in Abschnitt 2.2 skizzierten *consensus segmentation*

Tabelle 2.3: Anzahl der Patienten und der Ereignisse (Todesfälle) sowie die mediane Überlebenszeit $\tilde{t}_{0,5}$ und die Anzahl vorausgewählter Regionen getrennt nach den Daten aus Köln und Uppsala. Für eine Betrachtung dieser Maßzahlen getrennt nach klinischen Merkmalen sei hier auf die Tabellen 5.1 und 5.2 verwiesen.

Kohorte	Patienten	Todesfälle	$\tilde{t}_{0,5}$	Regionen
Köln	833	277 (33%)	7,0 Jahre	733
Uppsala	192	141 (73%)	3,5 Jahre	576

Verfahren potentiell relevante Regionen extrahiert. Durch dieses Verfahren resultieren aus der Köln Kohorte 733 und aus der Uppsala Kohorte 576 Regionen (siehe Tabelle 2.3).

Die Zielvariable beider Kohorten ist die sogenannte *overall survival* Zeit. Sie gibt für jeden Patienten die Zeit vom Zeitpunkt der Diagnose des Tumors bis zum Todeszeitpunkt des Patienten an. Im Folgenden wird hiervon stets von der Überlebenszeit eines Patienten gesprochen. Abbildung 2.2 zeigt jeweils den Kaplan-Meier-Schätzer getrennt nach den beiden Kohorten. Die Kurve gibt für jeden Zeitpunkt die geschätzte Überlebenswahrscheinlichkeit an (siehe Kapitel 3.1.1, Gleichung (3.1)). Die Abbildung veranschaulicht indirekt die Qualität der zugrundeliegenden Daten bezüglich der Erhebung der Überlebenszeit. Für Patienten aus der Köln Kohorte ist der tatsächliche Überlebensstatus oft nicht bekannt. In diesem Fall wird von einer Zensierung gesprochen (siehe Kapitel 3). Je niedriger der Anteil an Zensierungen in den Daten ist, desto genauer lässt sich die Überlebenswahrscheinlichkeit der Patienten schätzen. In Abbildung 2.2 ist zu erkennen, dass für Zeitpunkte unter fünf Jahren der Überlebensstatus aller Patienten der Uppsala Kohorte bekannt ist. Hingegen ist der Status der meisten Patienten aus der Köln Kohorte unbekannt. Tabelle 2.3 gibt die exakten Zensierungsanteile bzw. Todesfälle in den Daten an. Der Anteil an tatsächlich beobachteten Fällen und damit die genaue Kenntnis der Überlebenszeit ist in der Uppsala Kohorte mit 73 % deutlich über dem entsprechenden Anteil der Köln Kohorte (33 %). Die beobachtete mediane Überlebenszeit bzw. die Überlebensfunktion differiert zwischen den beiden Kohorten (siehe Tabelle 2.3 bzw. Abbildung 2.2). Die mediane Überlebenszeit $\tilde{t}_{0,5}$ entspricht dem Zeitpunkt, an dem geschätzt 50 % der Patienten noch nicht verstorben sind (siehe Abschnitt 3.1.1).

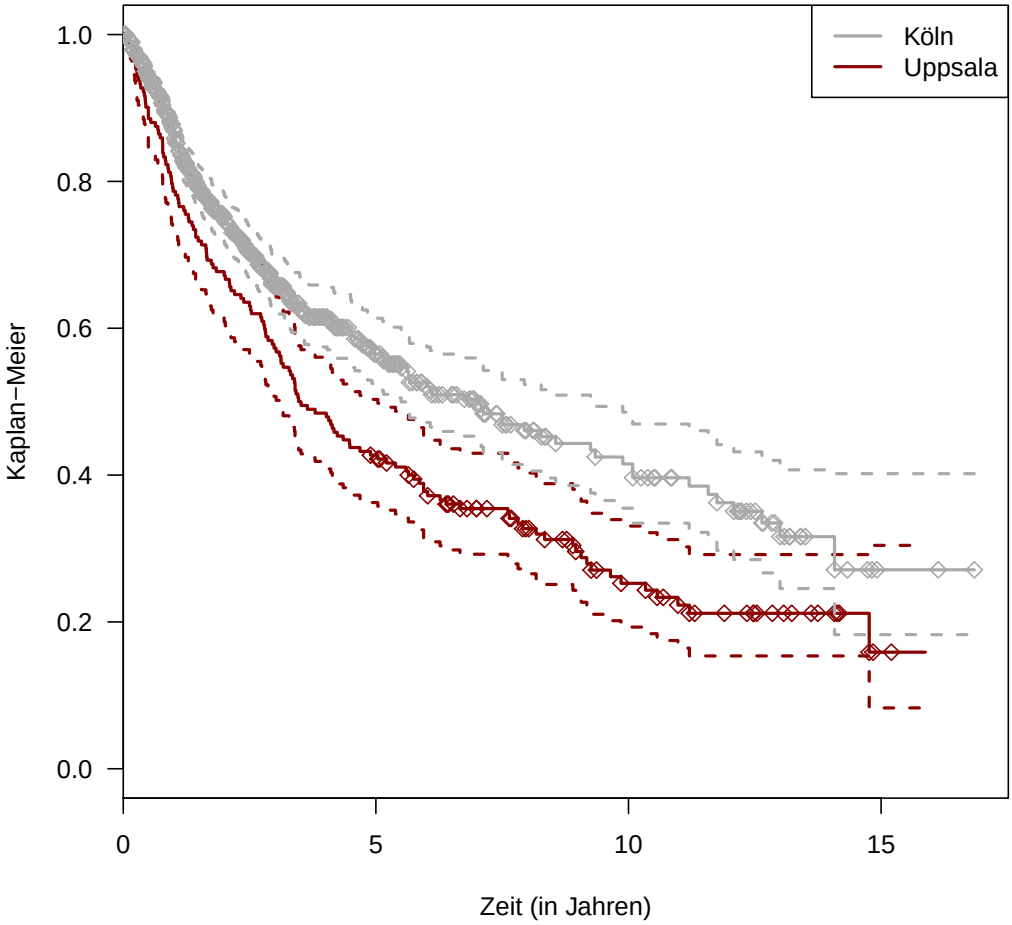


Abbildung 2.2: Kaplan-Meier-Schätzer getrennt nach den Überlebenszeitdaten aus Köln und Uppsala. Eine Raute kennzeichnet eine Zensierung. Das jeweilige 95 % Konfidenzintervall wird über die gestrichelten Linien dargestellt.

3 Überlebenszeitanalyse bei hochdimensionalen Daten

Die Überlebenszeitanalyse findet in der Krebsforschung breite Anwendung. Generell wird bei der Analyse von Überlebenszeiten die Zeit bis zum Auftreten eines Ereignisses modelliert. Oft wird deshalb auch von Ereigniszeitanalyse gesprochen. Im Kontext der Krebsforschung werden vorwiegend zwei Ereignisse unterschieden. Zum einen lässt sich die Zeit bis zum Eintritt des Todes der Patienten analysieren. Zum anderen kann die Zeit bis die Patienten ein Rezidiv erleiden betrachtet werden. Für den ersten Fall wird überwiegend die Zeit seit dem Tag der Diagnose gemessen. Ist das rezidivfreie Überleben von Interesse, wird ab dem Tag der operativen Entfernung des Tumors gemessen. Wichtig für beide Fälle ist, dass ein Patient womöglich nicht bis zum Zeitpunkt des interessierenden Ereignisses beobachtet wird oder unter Umständen das Ereignis nie eintritt. In Kapitel 3.1 werden die Grundlagen der Überlebenszeitanalyse vorgestellt.

Neben der Zensierung der Überlebenszeiten ist die Anzahl p der möglichen Einflussfaktoren auf die Überlebenszeit von entscheidender Bedeutung bei der Wahl geeigneter Regressionsverfahren. Werden genetische Einflussfaktoren untersucht, so gehen diese in der Regel aus Microarray-Experimenten hervor, wodurch mehrere tausend Merkmale (Proben) gleichzeitig untersucht werden können. Die Anzahl p dieser genetischen Merkmale ist üblicherweise deutlich größer als die Anzahl der untersuchten Patienten n. Dies wird in der Literatur oft mit dem Problem $p \gg n$ ausgedrückt. Für diesen Fall sind einige sogenannte regularisierte Verfahren entwickelt worden. Diese Verfahren begünstigen eine kleine Anzahl von Regressionsparametern bei der Modellwahl. In Kapitel 3.2 werden die nach diesem Prinzip arbeitenden Verfahren vorgestellt.

Im Hinblick auf die zu untersuchenden Untergruppen verstärkt sich die zuvor angesprochene Problematik einer zu kleinen Fallzahl n. Um dennoch ein geeignetes Modell schätzen zu können, werden in Kapitel 4 zwei Verfahren vorgestellt, mit denen die nicht in der jeweiligen Untergruppe enthaltenen Patienten mit in die Berechnung der in 3.2 beschriebenen Modelle eingeschlossen werden. Auf diese Weise wird die Fallzahl erhöht und es kann gegebenenfalls ein passenderes Modell gefunden werden, als es rein mit der Betrachtung der Patienten aus der Untergruppe möglich gewesen wäre. Wichtig ist hierbei, dass stets ein Modell ausschließlich für die Untergruppe angepasst wird. Ziel ist es, die Modelle für Untergruppen mit Modellen für die gesamte Patientenkohorte zu vergleichen.

Für den Vergleich der Modelle im Sinne der Vorhersagegenauigkeit für die Überlebenszeit werden in Kapitel 3.3 der auf die Anwendung von zensierten Ereigniszeiten angepasste Brier Score von Brier (1950) sowie Vorhersage-Fehler-Kurven definiert. In Abschnitt 3.3 wird zudem die Validierung mit Hilfe von Bootstrap Stichproben erläutert

und die für diese Arbeit genutzte Technik beschrieben.

Am Ende eines jeden Unterkapitels finden sich Angaben zur Umsetzung der Methoden mit der Software R (R Core Team, 2012). Zu jedem Verfahren wird auf das entsprechend verwendete R Paket hingewiesen und es werden nähere Erläuterungen zur eigenständigen Implementierung einiger Methoden gegeben.

3.1 Grundlagen der Überlebenszeitanalyse

Dieses Kapitel stellt die elementaren Kenngrößen und Modelle der Überlebensfunktion vor. Zunächst werden in Abschnitt 3.1.1 grundlegende Begriffe und das Schätzverfahren nach Kaplan-Meier eingeführt. Auch wird in diesem Kapitel auf die Besonderheit der Überlebensdaten, die Zensierung, eingegangen, und diesbezügliche Annahmen an die Daten werden diskutiert. Abschnitt 3.1.2 erläutert das Regressionsmodell nach Cox (1972), welches die elementare Grundlage der in dieser Arbeit durchgeführten Analysen der Überlebenszeit darstellt. Das Modell beschreibt die Überlebenszeiten der Patienten in Abhängigkeit von bestimmten Einflussfaktoren. Im Kontext der Krebsforschung sind diese Einflussvariablen durch das gemessene genetische Material der Patienten charakterisiert. Abschnitt 3.1.3 stellt den Breslow-Schätzer für das Grundrisiko, die sogenannte Baselinerate, eines angepassten Cox Modells vor. Diese Schätzung ist Voraussetzung für Vorhersagen mittels eines Cox Modells und findet in den Kapiteln 3.3 und 4.1.5 Anwendung.

Die in den folgenden Abschnitten vorgestellten Methoden sind in detaillierter Form zum Beispiel in den Werken von Klein und Moeschberger (2003) oder Kalbfleisch und Prentice (2002) nachzulesen.

3.1.1 Überlebensfunktion und Kaplan-Meier-Schätzer

Sei zunächst T eine nicht negative stetige Zufallsvariable, die die Zeit bis zum Auftreten eines Ereignisses, meist Tod oder Rezidiv, beschreibt. Die Verteilung von T wird von der sogenannten *Survival-* oder *Überlebensfunktion* eindeutig charakterisiert. Die Überlebensfunktion gibt die Wahrscheinlichkeit an, dass ein interessierendes Ereignis erst nach dem Zeitpunkt t eintritt und ist definiert als:

$$S(t) = P(T > t),\ 0 \leq t < \infty$$

$S(t)$ spiegelt die rechtsseitige Wahrscheinlichkeitsmasse der Verteilung $f(t)$ wider, wohingegen die Verteilungsfunktion $F(t) = P(T \leq t)$ die linksseitige Masse erklärt. Somit ist $S(t)$ eine nicht steigende, rechtsseitig stetige Funktion von t und es gilt $S(0) = 1$ und $\lim_{t\to\infty} S(t) = 0$. Die *mediane Überlebenszeit* wird mit $\tilde{t}_{0,5}$ bezeichnet. Diese löst die Gleichung:

$$S(\tilde{t}_{0,5}) = 0{,}5$$

Die größte Herausforderung bei der Analyse von Überlebenszeiten ist die Zensierung der beobachteten Daten. Informationen über den genauen Zeitpunkt wann das Ereignis

eingetreten ist sind oft unvollständig. Häufigstes Beispiel ist die Rechtszensierung. In diesem Fall ist lediglich bekannt, dass ein Patient mindestens bis zu einem Zeitpunkt t_c überlebt hat bzw. rezidivfrei war. Die Zeit bis zur Zensierung wird durch die Zufallsvariable C beschrieben, mit analogen Eigenschaften der Zufallsvariable T. Für rechtszensierte Daten gilt, dass sich die tatsächlich beobachtete Zeit $\tilde{t}$ eines Patienten, im Weiteren auch Risikozeit genannt, aus dem Minimum der jeweiligen Realisationen von T und C ergibt, kurz: $\tilde{T} = \min(T, C)$. Nachfolgend wird durch einen Zensierungsindikator δ angegeben, ob das Ereignis vor der Zensierungszeit eingetreten ist. Es gilt somit $\delta = \mathcal{I}(T < C)$, wobei $\mathcal{I}(\cdot)$ die Indikatorfunktion ist und den Wert 1 annimmt, wenn der Ausdruck wahr ist und Null sonst. Im Allgemeinen wird angenommen, dass die Ereigniszeit von der Zensierungszeit stochastisch unabhängig ist. Diese Annahme ermöglicht die folgende Schätzung der Überlebensfunktion aus einer Stichprobe vom Umfang n.

Zu jedem der n Patienten ist die Risikozeit $\tilde{T}$ sowie der Zensierungsindikator δ bekannt. Sei nun D die Anzahl der (geordneten) Zeitpunkte $t_{(1)} < t_{(2)} < \cdots < t_{(D)}$ an denen tatsächlich mindestens ein Ereignis eingetreten ist. d_i gibt die Anzahl Ereignisse zum Zeitpunkt $t_{(i)}$ an. Die Risikomenge

$$R(t_{(i)}) = \left\{ j \colon \tilde{t}_j \geq t_{(i)} \right\}$$

enthält zu jeder Ereigniszeit $t_{(i)}$ die Indizes derer Patienten, welche entweder das Ereignis oder eine Zensierung ebenfalls zum Zeitpunkt $t_{(i)}$ oder später erfahren, d.h. die Patienten die zum Zeitpunkt $t_{(i)}$ unter Risiko stehen. Dann ist $|R(t_{(i)})|$ die Anzahl der unter Risiko stehenden Patienten zum Zeitpunkt $t_{(i)}$. Eine Schätzung der Überlebensfunktion $S(t)$ aus einer unabhängigen Stichprobe von (T, C) ist nach Kaplan und Meier (1958) wie folgt gegeben:

$$\hat{S}(t) = \begin{cases} 1 & \text{falls } t < t_{(1)} \\ \prod_{t_{(1)} \leq t} \left[1 - \frac{d_i}{|R(t_{(i)})|} \right] & \text{falls } t \geq t_{(1)} \end{cases} \tag{3.1}$$

Dieser Schätzer wird fortan als *Kaplan-Meier-Schätzer* bezeichnet und ist der am häufigsten verwendete nichtparametrische Schätzer der Überlebensfunktion (Klein und Moeschberger, 2003). Ist zum letzten Zeitpunkt $t_{(D)}$ mindestens ein Patient zensiert, so kann $\hat{S}(t)$ nicht den Wert Null annehmen. Für $t > t_{(D)}$ ist $\hat{S}(t)$ dann nicht definiert. Jedoch gibt es theoretische Lösungen den Schätzer in geeigneter Weise fortzuführen. In den Abschnitten 3.3.1 und 3.3.2 wird auf diese Möglichkeit genauer eingegangen.

Der Kaplan-Meier-Schätzer lässt sich über das Produkt bedingter Wahrscheinlichkeiten herleiten und wird daher in der Literatur auch als Produkt-Limit-Schätzer bezeichnet. Anschaulich gesehen wird die Wahrscheinlichkeitsmasse eines Ereignisses einer zensierten Beobachtung gleichmäßig auf alle nachfolgenden Ereigniszeitpunkte aufgeteilt.

Die Varianz des Kaplan-Meier-Schätzers kann mit *Greenwood's Formel*

$$\hat{V}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum_{t_{(i)} < t} \frac{d_i}{n_i(n_i - d_i)}$$

geschätzt werden (siehe Klein und Moeschberger, 2003, Kapitel 4.2). Ein einfaches ap-

proximatives Konfidenzintervall für $\hat{S}(t)$ zum Niveau 5% kann mit

$$\hat{S}(t) \pm 1{,}96 \left[\hat{V}\left(\hat{S}(t)\right)\right]^{1/2}$$

angegeben werden.

Das Standardpaket in R zur Darstellung der Risikozeiten und des Zensierungsindikators sowie zur Berechnung der Punkt- und Varianzschätzung der Überlebensfunktion über den Kaplan-Meier-Schätzer ist das Paket *survival* von Therneau und Grambsch (2000), aktuell in Version 2.37-2 (Therneau, 2012).

3.1.2 Cox-Regression

Eine weitere fundamentale Größe der Überlebenszeitanalyse ist die *Hazardrate*. Sie beschreibt das Risiko eines Patienten im nächsten Augenblick das betrachtete Ereignis zu widerfahren, falls dieser bis dahin das Ereignis nicht erfahren hat. Die Hazardrate $h(t)$ eines Patienten zum Zeitpunkt t steht in eindeutigem Zusammenhang mit der Überlebensfunktion $S(t)$ und ist definiert durch:

$$h(t) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T < t + \Delta t \mid T \geq t\right)}{\Delta t} = \frac{f(t)}{S(t)}$$

$h(t)$ beschreibt das Risiko, dass zum nächstmöglichen Zeitpunkt das Ereignis eintritt, wenn es bis dahin noch nicht stattgefunden hat. Für die Hazardrate gilt die Einschränkung $h(t) \geq 0$ (siehe Klein und Moeschberger, 2003, Seite 27). Die Hazardrate der Zufallsvariablen T ist aufgrund ihrer Darstellungsform meist einfacher zu interpretieren als die Überlebensfunktion $S(t)$ (siehe Klein und Moeschberger, 2003, Seite 31). Die kumulierte Hazardrate, die ebenso direkt mit der Überlebensfunktion in Zusammenhang steht, ist definiert durch:

$$H(t) = \int_0^t h(u)\,\mathrm{d}u = \int_0^t \frac{f(u)}{S(u)}\,\mathrm{d}u = -\ln S(t)$$

Die Relevanz einzelner oder mehrerer unabhängiger erklärender Variablen auf die Hazardrate und damit auf die Überlebenszeit T kann mit dem semi-parametrischen Regressionsmodell von Cox (1972) modelliert werden. Es ist das am weitesten verbreitete Regressionsmodell zur Modellierung der Hazardrate durch einen Kovariablenvektor $Z = (Z_1, \ldots, Z_p)$ und wird meist kurz als das *Cox-Modell* bezeichnet. Das Cox-Modell beruht auf der Annahme, dass sich die Hazardrate aus einer Baseline-Rate (Grundrisiko) und einem zeitunabhängigen Faktor zusammensetzt. Die Modellgleichung sieht wie folgt aus:

$$h\left(t \mid Z\right) = h_0(t) \exp\left(\beta' Z\right) = h_0(t) \exp\left(\sum_{k=1}^{p} \beta_k Z_k\right) \tag{3.2}$$

Dabei spiegelt der Parametervektor $\beta = (\beta_1, \ldots, \beta_p)$ die Stärke des Einflusses der Kovariablen auf die Hazardrate wider. Die Baseline-Rate $h_0(t)$ ist nichtparametrischer Natur. Aus diesem Grund wird das Cox-Modell auch als semi-parametrisches Modell bezeichnet. Der Faktor $\beta' Z$ wird auch als Risiko-Score (risk score) oder Prognostischer Index (prognostic index) bezeichnet. Eine Erhöhung dieses Faktors geht mit einem Anstieg der Hazardrate einher und das Risiko das (negative) Ereignis zu erfahren steigt.

Die Besonderheit des Cox-Modells liegt in der folgenden Eigenschaft: Für zwei Patienten mit unterschiedlichen Kovariablenvektoren Z und Z^* gilt:

$$\frac{h\left(t \mid Z\right)}{h\left(t \mid Z^*\right)} = \frac{h_0(t) \exp\left[\sum_{k=1}^{p} \beta_k Z_k\right]}{h_0(t) \exp\left[\sum_{k=1}^{p} \beta_k Z_k^*\right]} = \exp\left[\sum_{k=1}^{p} \beta_k \left(Z_k - Z_k^*\right)\right]$$

Der Quotient der Hazardraten zweier Patienten, das *Hazardratio*, ist konstant über die Zeit. Die Hazardraten sind demnach proportional zueinander.

Der Parametervektor β wird über ein Maximum-Likelihood-Verfahren geschätzt. Sei dafür zunächst wieder eine Stichprobe der Größe n verfügbar. Darin enthalten ist für jeden Patienten $j = 1, \ldots, n$ die individuell beobachtete Risikozeit $\tilde{t}_j$, der dazugehörige Indikator δ_j sowie der Kovariablenvektor Z_j. Analog zu 3.1.1 bezeichne mit $t_{(i)}$, $i = 1 \ldots D$ die geordneten Ereigniszeiten. Sei $Z_{(i)k}$, die k-te Kovariable desjenigen Patienten, bei welchem exakt zum Zeitpunkt $t_{(i)}$ das Ereignis eintritt. Zunächst wird dementsprechend angenommen, dass keine Bindungen der Ereigniszeiten auftreten. Dann ist nach Klein und Moeschberger (2003), Seite 253, die *partielle Likelihood* gegeben durch:

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp\left(\sum_{k=1}^{p} \beta_k Z_{(i)k}\right)}{\sum_{j \in R(t_{(i)})} \exp\left(\sum_{k=1}^{p} \beta_k Z_{jk}\right)} \tag{3.3}$$

Der Zähler wird lediglich durch den Risiko-Score des zum Zeitpunkt $t_{(i)}$ ausgefallenen Patienten bestimmt. Der Nenner spiegelt das Gesamtrisiko ab diesem Zeitpunkt, die Summe der Risiko-Scores aller übrig gebliebenen Patienten, wider. Die partielle Likelihood kann als Quotient zweier bedingter Wahrscheinlichkeiten interpretiert werden. Die Gleichung hängt dabei nicht von der Baseline-Rate ab.

Bezeichne nun $LL(\beta) = \ln\left(L(\beta)\right)$. Die Maximierung der aus (3.3) resultierenden partiellen Log-Likelihood

$$LL(\beta) = \sum_{i=1}^{D} \sum_{k=1}^{p} \beta_k Z_{(i)k} - \sum_{i=1}^{D} \ln\left[\sum_{j \in R(t_{(i)})} \exp\left(\sum_{k=1}^{p} \beta_k Z_{jk}\right)\right] \tag{3.4}$$

führt zur ML-Schätzung der Parameter β. Dieses numerische Maximierungsproblem von $LL(\beta)$ lässt sich iterativ beispielsweise mit einem Newton-Raphson-Verfahren lösen. Aber auch andere iterative Algorithmen sind denkbar.

Bisher wurde der Zähler der partiellen Likelihood lediglich durch den Risiko-Score eines einzelnen Patienten bestimmt. Werden nun Bindungen in den Daten zugelassen, d.h. dürfen zu einem Zeitpunkt mehrere Ereignisse auftreten, so gibt es nach Klein und

Moeschberger (2003) mehrere Adaptionen der partiellen Likelihood aus Gleichung (3.3), von denen die Verfahren nach Breslow (1974) sowie nach Efron (1977) am verbreitetsten sind. Seien wiederum $t_{(i)}, i = 1, \ldots, D$ die geordneten Ereigniszeiten, an denen nun jedoch mehrere Ereignisse aufgetreten sein können. Analog zu Abschnitt 3.1.1 bezeichnet d_i die Anzahl der Ereignisse zum Zeitpunkt $t_{(i)}$. $\mathbb{D}_i$ sei der Index aller Patienten, die zu diesem Zeitpunkt ein Ereignis haben. Die partielle Likelihood nach Breslow (1974) ist wie folgt definiert:

$$L_{Breslow}(\beta) = \prod_{i=1}^{D} \frac{\exp\left(\beta' \left(\sum_{j \in \mathbb{D}_i} Z_j\right)\right)}{\left[\sum_{j \in R(t_{(i)})} \exp\left(\beta' Z_j\right)\right]^{d_i}}$$

Etwas genauer und damit bei vielen Zensierungen geeigneter ist die partielle Likelihood nach Efron (1977):

$$L_{Efron}(\beta) = \prod_{i=1}^{D} \frac{\exp\left(\beta' \left(\sum_{j \in \mathbb{D}_i} Z_j\right)\right)}{\prod_{j=1}^{d_i} \left[\sum_{k \in R(t_{(i)})} \exp\left(\beta' Z_k\right) - \frac{j-1}{d_i} \sum_{k \in \mathbb{D}_i} \exp\left(\beta' Z_k\right)\right]}$$

Sind keine Bindungen der Ereigniszeiten in den Daten vorhanden, so gilt $L_{Breslow}(\beta) = L_{Efron}(\beta) = L(\beta)$.

Das Schätzen eines Cox-Modells ist ebenfalls im R Paket *survival* von Therneau und Grambsch (2000) implementiert. Standardmäßig wird die partielle Likelihood nach Efron (1977) maximiert. Seit den neusten Versionen werden zudem regularisierte Modelle, wie z.B. die Ridge Regression unterstützt. Die Strafterme lassen sich jedoch nicht mit dem *survival* Paket optimieren. Hierfür empfiehlt sich beispielsweise das Paket *glmnet* von Simon u. a. (2011) (siehe dazu Abschnitt 3.2.2).

3.1.3 Vorhersage der Überlebensfunktion

Die wichtigste Voraussetzung für einen validen Vergleich von Vorhersagefehlern, wie er in dieser Arbeit vorgenommen werden soll, ist die Möglichkeit ein geschätztes Modell anhand einer unabhängigen Testmenge zu überprüfen. Konkret soll ein geschätztes Cox-Modell nicht nur auf der Stichprobe auf dem es angepasst wurde (Trainingsmenge) überprüft werden, sondern auch auf einer von der Anpassung des Modells unabhängigen zweiten Stichprobe (Testmenge). Die Parameterschätzung aus Abschnitt 3.1.2 erlaubt zunächst nur eine Überprüfung der Parameter und des Risiko-Scores, nicht aber der Überlebensfunktion. Soll die Überlebensfunktion für eine zweite unabhängige Stichprobe geschätzt werden und mit der tatsächlich beobachteten Überlebenszeit verglichen werden, so muss die Überlebenswahrscheinlichkeit eines Patienten aus der neuen Stichprobe über dessen Risiko-Score vorhergesagt werden. Liegt das Cox-Modell zugrunde, ist die Kenntnis der Baseline-Rate $h_0(t)$ erforderlich. Anders als etwa bei Tests über die geschätzten Parameter, muss für eine Vorhersage der Überlebenswahrscheinlichkeit das Grundrisiko bekannt sein bzw. ebenfalls geschätzt werden.

Dafür sei zunächst die Likelihood der Daten in folgender Form gegeben (siehe Klein und Moeschberger, 2003, Kapitel 3.5 und Kapitel 8.3):

$$\begin{aligned} L &= \prod_{j=1}^{n} f(\tilde{t}_j|Z_j)^{\delta_j} S(\tilde{t}_j|Z_j)^{1-\delta_j} \\ &= \prod_{j=1}^{n} h(\tilde{t}_j|Z_j)^{\delta_j} \exp\left(-H(\tilde{t}_j|Z_j)\right) \end{aligned}$$

Wird nun angenommen, dass den Daten ein Cox-Modell zugrunde liegt, so lässt sich mit Gleichung (3.2) die Likelihood wie folgt schreiben:

$$L = \prod_{j=1}^{n} h_0(\tilde{t}_j)^{\delta_j} \left[\exp(\beta' Z_j)\right]^{\delta_j} \exp\left(-H_0(\tilde{t}_j)\exp(\beta' Z_j)\right) \tag{3.5}$$

Dabei ist $H_0(t) = \sum_{t_{(i)} \le t} h_0(t_{(i)})$ die diskrete kumulierte Baseline-Rate zum Zeitpunkt t (vgl. Klein und Moeschberger, 2003, Seite 32, Theoretical Note 1). Für die diskrete Baseline-Rate gilt $h_0(t) = 0$ für alle Zeitpunkte $t \notin \{t_{(1)}, \ldots, t_{(D)}\}$.

Sei nun bereits eine Maximum-Likelihood-Schätzung $\hat{\beta}$ für den Parametervektor β aus einem Cox-Modell gegeben. Für diese Schätzung bedarf es nicht der Kenntnis der Baseline-Rate. Die Baseline-Rate an den Zeitpunkten $t \in \{t_{(1)}, \ldots, t_{(D)}\}$ lässt sich dann mit gegebenem $\hat{\beta}$ durch Maximieren der Likelihood (3.5) bzgl. $h_0(t)$ schätzen. Die zu maximierende Likelihood hat damit folgende Gestalt:

$$\begin{aligned} L_{\hat{\beta}}(h_0(t)) &= \left[\prod_{i=1}^{D} h_0(t_{(i)}) \exp(\hat{\beta}' Z_{(i)})\right] \exp\left[-\sum_{j=1}^{n} H_0(\tilde{t}_j)\exp(\hat{\beta}' Z_j)\right] \\ &= \left[\prod_{i=1}^{D} h_0(t_{(i)}) \exp(\hat{\beta}' Z_{(i)})\right] \exp\left[-\sum_{j=1}^{n} \sum_{t_{(i)} \le \tilde{t}_j} h_0(t_{(i)})\exp(\hat{\beta}' Z_j)\right] \\ &= \left[\prod_{i=1}^{D} h_0(t_{(i)}) \exp(\hat{\beta}' Z_{(i)})\right] \exp\left[-\sum_{i=1}^{D} \sum_{j \in R(t_{(i)})} h_0(t_{(i)})\exp(\hat{\beta}' Z_j)\right] \qquad (3.6) \\ &= \prod_{i=1}^{D} h_0(t_{(i)}) \exp(\hat{\beta}' Z_{(i)}) \exp\left[-\sum_{j \in R(t_{(i)})} h_0(t_{(i)})\exp(\hat{\beta}' Z_j)\right] \\ &= \prod_{i=1}^{D} h_0(t_{(i)}) \exp(\hat{\beta}' Z_{(i)}) \exp\left[-h_0(t_{(i)}) \sum_{j \in R(t_{(i)})} \exp(\hat{\beta}' Z_j)\right] \qquad (3.7) \\ &\propto \prod_{i=1}^{D} h_0(t_{(i)}) \exp\left[-h_0(t_{(i)}) \sum_{j \in R(t_{(i)})} \exp(\hat{\beta}' Z_j)\right] \qquad (3.8) \end{aligned}$$

In Gleichung (3.6) wird die Doppelsumme umgestellt, so dass zunächst die Baseline-Rate $h_0(t_{(i)})$ mit der Anzahl der zum Zeitpunkt $t_{(i)}$ unter Risiko stehenden Patienten multipliziert wird. Dabei ist $R(t_{(i)}) = \left\{j \colon \tilde{t}_j \geq t_{(i)}\right\}$ die Risikomenge (siehe Abschnitt 3.1.1). Die Baseline-Rate $h_0(t_{(i)})$ zu einem festen Zeitpunkt $t_{(i)}$ hängt nicht von den Beobachtungen $j \in R(t_{(i)})$ ab. Somit kann in Gleichung (3.7) die Baseline-Rate vor die Summe geschrieben werden.

Sei $\eta = \sum_{j \in R(t_{(i)})} \exp(\hat{\beta}' Z_j)$. Es gilt:

$$\frac{\partial}{\partial h_0(t_{(i)})} \left(h_0(t_{(i)}) \exp\left[-\eta \cdot h_0(t_{(i)})\right]\right) = \left(1 - \eta \cdot h_0(t_{(i)})\right) \exp\left(-\eta \cdot h_0(t_{(i)})\right)$$

Damit ist der ML-Schätzer von (3.8) an den Zeitpunkten $t_{(1)}, \ldots, t_{(D)}$ gegeben durch

$$\hat{h}_0(t_{(i)}) = \frac{1}{\eta} = \frac{1}{\sum_{j \in R(t_{(i)})} \exp(\hat{\beta}' Z_j)}$$

(vgl. Klein und Moeschberger, 2003, Kapitel 8.3, Theoretical Note 2). Mit $\hat{H}_0(t) = \sum_{t_{(i)} \leq t} \hat{h}_0(t_{(i)})$ lassen sich diese Schätzer zu einem Schätzer der kumulierten Baseline-Rate kombinieren. Treten mindestens zwei Ereignisse gleichzeitig auf, lassen sich diese als in einem infinitesimal kleinen Abstand einzeln auftretende Ereignisse auffassen. Daraus ergibt sich der sogenannte *Breslow-Schätzer* für $H_0(t)$ (siehe Klein und Moeschberger, 2003, Kapitel 8.8) wie folgt:

$$\hat{H}_0(t) = \sum_{t_{(i)} \leq t} \left(\frac{d_i}{\sum_{j \in R(t_{(i)})} \exp\left(\sum_{k=1}^{p} \hat{\beta}_k Z_{jk}\right)} \right) \tag{3.9}$$

Über den Breslow-Schätzer kann ein Schätzer für die Baseline-Überlebensfunktion $S_0(t) = \exp\left[-H_0(t)\right]$ angeben werden:

$$\hat{S}_0(t) = \exp\left[-\hat{H}_0(t)\right]$$

$\hat{S}_0(t)$ ist die geschätzte Überlebensfunktion für einen Patienten mit einem Kovariablenvektor $Z = 0$. Für einen Patienten mit einem beliebigen Kovariablenvektor $Z = z$ ergibt sich hieraus und aus der Modellgleichung (3.2) die für diesen Patienten geschätzte individuelle Überlebensfunktion zu:

$$\hat{S}(t|Z = z) = \hat{S}_0(t)^{\exp\left(\sum_{k=1}^{p} \hat{\beta}_k z_k\right)} \tag{3.10}$$

Der Ausdruck $\sum_{k=1}^{p} \hat{\beta}_k z_k$ wird auch als Risiko-Score (Risk Score) bezeichnet. Mit dem Breslow-Schätzer ist es zusammen mit dem Risiko-Score demnach möglich für einen Patienten eine Vorhersage seiner individuellen Überlebenswahrscheinlichkeit für einen beliebigen Zeitpunkt zu treffen. Ist die tatsächliche Überlebenszeit dieses Patienten bekannt, so kann diese mit der Vorhersage verglichen werden (siehe Kapitel 3.3).

Die R Funktionen `basehaz` und `predict.coxph` aus dem *survival* Paket bieten lediglich die Möglichkeit die Baseline-Rate bzw. die Überlebensfunktion direkt aus einem `coxph`-Objekt zu schätzen. Die Option einen beliebigen Risiko-Score-Vektor einzugeben sowie die Option die Baseline-Rate und die Überlebenswahrscheinlichkeiten für frei wählbare Zeitpunkte auszugeben fehlt. Aus diesem Grund wurde die Funktion `PREDmat` geschrieben, welche im Anhang B zu finden ist. Für Zeitpunkte $t > t_{(D)}$ wird die Baseline-Rate konstant fortgesetzt. Die Funktion erlaubt ebenfalls die Verwendung einer gewichteten Version der Likelihood aus (3.8). Eine Herleitung des gewichteten Breslow-Schätzers ist in Abschnitt 4.1.5 gegeben.

3.2 Regularisierte Modellbildung

Das Maximierungsproblem der partiellen (Log-)Likelihood aus Abschnitt 3.1.2 ist nur lösbar, wenn die Anzahl der betrachteten Kovariablen p kleiner ist als die Anzahl der zur Verfügung stehenden Beobachtungen (Patienten) n. Ansonsten kann kein eindeutiges Maximum gefunden werden. Jedoch ist dieses Szenario in der genetischen Krebsforschung selten vorzufinden. Stattdessen steigt mit besseren technischen Voraussetzungen und immer feiner auflösenden Messverfahren die Zahl der verfügbaren (genetischen) Kovariablen weiter an. Die Anzahlen belaufen sich bei ArrayCGH Chips für die Analyse von Copy Number Variationen derzeit auf dem Niveau von ca. 6 000 Regionen. Im Bereich der SNP (Single Nucleotide Polymorphism) Analyse sind es je nach Experiment bzw. Technologie 500 000 bis hin zu 1 000 000 mögliche SNPs, die als Kovariablen in Betracht kommen können (Affymetrix, 2009).

Für solche $p \gg n$ Szenarien wurden bereits einige Verfahren entwickelt. Die Grundidee bei diesen Methoden ist die zusätzliche Einführung eines weiteren Parameters in die zu maximierende Likelihood. Dieser (Regularisierungs-)Parameter, fortan mit λ bezeichnet, reguliert die absolute Größe der jeweiligen Regressionskoeffizienten β. Hierzu wird bei der ML-Schätzung ein Strafterm der Form $\lambda \sum_{k=1}^{p} |\beta_k|^\alpha$ von der Log-Likelihood (3.4) subtrahiert. Daraus ergibt sich folgendes Maximierungsproblem:

$$\hat{\beta}_\lambda = \text{argmax}_\beta \left(LL(\beta) - \lambda \sum_{k=1}^{p} |\beta_k|^\alpha \right), \ \lambda, \alpha \geq 0 \tag{3.11}$$

Diese Version der Likelihood wird auch als regularisierte partielle Log-Likelihoodfunktion $LL_\lambda(\beta)$ bezeichnet. Der Wert von λ wird über eine K-fache Kreuz-Validierung optimiert. Hierfür werden die Daten (Patientenindizes) in K möglichst gleich große Teilmengen $l = 1, \ldots, K$ aufgeteilt. Die kreuz-validierte partielle Log-Likelihood ist dann nach Verweij und Van Houwelingen (1993) definiert als:

$$\texttt{CVPL}(\lambda) = \sum_{l=1}^{K} LL_\lambda(\hat{\beta}_{\lambda,(-l)}) - LL_{\lambda,(-l)}(\hat{\beta}_{\lambda,(-l)}) \tag{3.12}$$

$\hat{\beta}_{\lambda,(-l)}$ entspricht dem ML-Schätzer bezüglich der Likelihood der Daten ohne die l-te

Teilmenge, welche mit $LL_{\lambda,(-l)}$ bezeichnet wird. Die Patienten aus der l-ten Teilmenge werden somit nicht zur Schätzung $\hat{\beta}_{\lambda,(-l)}$ verwendet. Der Wert der kreuz-validierten partiellen Log-Likelihood gibt an, wie gut die Likelihood der Patienten aus den Teilmengen über die restlichen Patienten geschätzt werden kann (Verweij und Van Houwelingen, 1993, Seite 2306). Die einzelnen Summanden beschreiben dabei jeweils die Verbesserung der Likelihood durch Hinzunahme der l-ten Teilmenge. Der optimale Parameter λ_{opt} ergibt sich durch Maximieren der `CVPL` nach λ. Die Algorithmen aus den Abschnitten 3.2.1 und 3.2.2 durchlaufen dazu ein Intervall für die möglichen Werte für λ.

Der Parameter α ist vorab zu setzen. Für $\lambda > 0$ werden kleine absolute Werte von β begünstigt. Die Koeffizienten werden somit verkleinert (geschrumpft). Bei der Schrumpfung werden die Koeffizienten β in Richtung Null gedrückt. Auf diese Weise wird gleichzeitig jedoch ein Bias eingeführt. Die Größe des Parameter λ beschreibt den Grad der Schrumpfung. Sind wenige Koeffizienten vorhanden, genügt ein kleiner Wert. Ist hingegen die Zahl der Koeffizienten übermäßig hoch im Vergleich zur Anzahl Patienten, muss der Regularisierungsparameter für eine optimale Schätzung erhöht werden.

In den folgenden zwei Abschnitten werden die in dieser Arbeit verwendeten Verfahren zur Bestimmung der Regressionsparameter β und der Optimierung von λ für das Problem $p \gg n$ vorgestellt. Abschnitt 3.2.1 beschreibt das von Binder und Schumacher (2008) entworfene komponentenweise Boosting Verfahren *CoxBoost*. Diese Methode hat, wie das Lasso Verfahren von Tibshirani (1996) aus Abschnitt 3.2.2, den Vorteil gleichzeitig eine Variablenselektion durchzuführen. Beide Verfahren setzen $\alpha = 1$. Neben der erwähnten Lasso Regression wird in Abschnitt 3.2.2 die Ridge Regression von Hoerl und Kennard (1970) beschrieben. Die Ridge Regression ($\alpha = 2$) gehört zu den ersten Verfahren, die es aufgrund von einem neu eingeführtem Schrumpfungsparameter ermöglichen, eine stabile Schätzung für hochdimensionale Regressionsprobleme zu finden (Binder u. a., 2011).

3.2.1 Komponentenweises Boosting mit CoxBoost

Das Ziel ist erneut die Schätzung der Regressionskoeffizienten β aus der Modellgleichung eines Cox-Modells wie in Gleichung (3.2) angegeben. Sei jedoch eine Situation der Art $p \gg n$ gegeben. Eine Möglichkeit dennoch ein geeignetes Modell zu schätzen bietet sich über sogenannte Boosting Techniken. Beim Boosting wird eine Funktion von dem beobachteten Kovariablenvektor Z und den zu schätzenden Koeffizienten β optimiert. Die Verlustfunktion dieser Optimierung ist meist die negative regularisierte partielle Log-Likelihood. Im Gegensatz zu normalen Gradienten basierten Boosting-Verfahren, in denen die Koeffizienten β gleichzeitig in den Boostingschritten angepasst werden, werden beim komponentenweisen Boosting nur einzelne Koeffizienten in einem Boostingschritt adaptiert. Dieses Vorgehen impliziert, dass sukzessive einzelne Kovariablen für das Modell ausgewählt werden. Aus diesem Grund eignet sich komponentenweises Boosting besonders für hochdimensionale Daten (siehe Tutz und Binder, 2007). Genauso kann es jedoch von Nachteil sein, wenn relevante Kovariablen nicht ausgewählt werden. Das im Folgenden vorgestellte CoxBoost Verfahren von Binder und Schumacher (2008) wirkt diesem Problem entgegen, indem in jedem Boostingschritt jeweils eine Teilmenge der Kovariablen für eine Adaption in Frage kommen kann (partial boosting). So können

wiederum mehrere Koeffizienten gleichzeitig angepasst werden. Des Weiteren lassen sich mit CoxBoost Gruppen von Kovariablen unterschiedlich getrennt bestrafen, so dass einige Regressionskoeffizienten von einer möglichen Schrumpfung ausgeschlossen sind. Auf diese Möglichkeiten wird in ähnlicher Weise in Kapitel 4 eingegangen.

Genauer werden in dem Boosting Verfahren nach Tutz und Binder (2007) mehrere Mengen mit Kovariablen innerhalb eines Boostingschrittes betrachtet. Der CoxBoost Algorithmus verfährt in ähnlicher Art. In Boostingschritt $s = 1, \ldots, M$ gibt es genau q_s Indexmengen $\mathcal{I}_{sl} \subseteq \{1, \ldots, p\}, l = 1, \ldots, q_s$, die jeweils die Indizes von zuvor ausgewählten Kovariablen enthalten. Die jeweils zu diesen Indexmengen gehörenden Kovariablen werden in einem Schritt s gleichzeitig angepasst. Pro Schritt werden q_s unterschiedliche Anpassungen betrachtet, aus denen die beste ausgewählt wird.

Bezeichne nun mit $\hat{\beta}_{s-1} = (\hat{\beta}_{s1}, \ldots, \hat{\beta}_{sp})'$ den Schätzer für den Koeffizientenvektor β nach dem Boostingschritt $(s-1)$. In Schritt s gilt es nun herauszufinden, wie der Koeffizientenvektor $\hat{\beta}_{s-1}$ angepasst werden kann. Dazu wird zu jeder der q_s möglichen Anpassungen die regularisierte partielle Log-Likelihood

$$\mathrm{LL}_\lambda(\hat{\gamma}_{sl}) = \sum_{i=1}^{D} \hat{\beta}_{s-1} Z_{(i)} + \hat{\gamma}_{sl} Z_{(i),\mathcal{I}_{sl}} - \sum_{i=1}^{D} \ln \left(\sum_{j \in R(t_{(i)})} \exp\left(\hat{\beta}_{s-1} Z_j + \hat{\gamma}_{sl} Z_{j,\mathcal{I}_{sl}} \right) \right) - \lambda \hat{\gamma}_{sl}^2$$

berechnet. Der Risiko-Score $\hat{\gamma}_{sl} Z_{(i),\mathcal{I}_{sl}}$ setzt sich dabei zusammen aus den zur Indexmenge $\mathcal{I}_{sl}$ gehörenden Kovariablen und der Änderung der Koeffizienten $\hat{\gamma}_{sl}$. Der Wert der Änderung $\hat{\gamma}_{sl}$ wird über einen Newton-Raphson Algorithmus bestimmt. Hierbei ist lediglich ein Iterationsschritt nötig, da sich die Koeffizienten noch in weiteren Bootstrapschritten ändern können. Der Newton-Raphson Algorithmus wird auf die (Log-)Likelihood $LL_\lambda(\hat{\gamma}_{sl})$ ohne den Strafterm $\lambda \hat{\gamma}_{sl}^2$ angewandt.

Der CoxBoost Algorithmus läuft iterativ wie folgt ab:

(1) Starte mit den Regressionskoeffizienten $\hat{\beta}_0 = (0, \ldots, 0)'$

(2) Für jeden Boostingschritt $s = 1, \ldots, M$:

   (a) Ermittle die potentielle Änderung $\hat{\gamma}_{sl}$ je für die Koeffizienten aus den Indexmengen $\mathcal{I}_{sl}$, wobei $l = 1, \ldots, q_s$

   (b) Wähle aus den q_s Änderungen diejenige aus, die den größten Zuwachs der regularisierten partiellen Log-Likelihood zur Folge hat. Bezeichne mit $\mathcal{I}_{sl^*}$ die zu dieser Änderung gehörende Indexmenge

   (c) Aktualisiere den Koeffizientenvektor $\hat{\beta}_{(s-1)}$ wie folgt:

$$\hat{\beta}_{sk} = \begin{cases} \hat{\beta}_{s-1,k} + \hat{\gamma}_{sl^*,k}, & \text{falls } k \in \mathcal{I}_{sl^*} \\ \hat{\beta}_{s-1,k}, & \text{falls } k \notin \mathcal{I}_{sl^*} \end{cases}, \; k = 1, \ldots, p$$

Komponentenweises Boosting wird erzielt, indem die Indexmengen definiert werden zu $\mathcal{I}_s = \{\{1\}, \ldots, \{p\}\}$. Es wird so nur jeweils ein Regressionskoeffizient pro Boostingschritt

angepasst. Dies führt im Ergebnis dazu, dass, ähnlich wie bei der Lasso Methode von Tibshirani (1996) (siehe Abschnitt 3.2.2), viele der Koeffizienten mit Null geschätzt werden Binder und Schumacher (2008).

Über die Konstruktion entsprechender Indexmengen kann ferner erzwungen werden bestimmte Kovariablen in allen Iterationsschritten zu berücksichtigen. Beispielsweise wird die erste Kovariable bei Definition der Indexmengen $\mathcal{I}_s = \{\{1,2\}, \{1,3\}, \ldots, \{1,p\}\}$ stets berücksichtigt. Dies ist für die Anwendung interessant, wenn die erste Kovariable beispielsweise eine klinische Größe widerspiegelt, wobei die restlichen, meist in deutlich höherer Anzahl vertretenen Variablen, genetische Faktoren beschreiben.

Die regularisierte partielle Log-Likelihood erlaubt eine flexible Gestaltung des Strafterms $\lambda\hat{\gamma}_{sl}^2$. So können zu der im vorigen Absatz beschriebenen Begünstigung von einigen Kovariablen außerdem die Koeffizienten ausgewählter Variablen ohne Schrumpfung geschätzt werden.

Der CoxBoost Algorithmus benötigt zwei manuelle Eingaben. Der Regularisierungsparameter λ sowie die Anzahl M der Iterationen müssen übergeben werden. M sollte zwingend optimiert werden, wohingegen λ nicht von entscheidender Bedeutung für die Resultate ist (Binder u. a., 2011). Wie von Binder (2011) implementiert, wird $\lambda = 9 \cdot \sum_{j=1}^{N} \delta_j$ gesetzt, wobei δ_j dem Zensierungsstatus von Patient j entspricht. M wird mittels Maximierung der kreuz-validierten partiellen Log-Likelihood bestimmt, wobei die Methode nach Verweij und Van Houwelingen (1993) verwendet wird.

Der Algorithmus des komponentenweisen CoxBoost ist in dem gleichnamigen R Paket *CoxBoost* (Version 1.3) von Binder (2011) implementiert. Die Optimierung der Anzahl der Boostingschritte wird in der Funktion `cv.CoxBoost` standardmäßig mit einer 10-fachen Kreuz-Validierung nach der Methode von Verweij und Van Houwelingen (1993) durchgeführt. Die maximal zu untersuchende Anzahl beträgt dabei 100 Boostingschritte. Anschließend wird der CoxBoost Algorithmus mit der optimierten Anzahl M_{opt} auf allen Daten durchlaufen. Auf die Möglichkeit der Optimierung von λ (Funktion `optimCoxBoostPenalty`) wird aus den oben angegebenen Gründen verzichtet, da diese Optimierung zudem viele Ressourcen in Anspruch nimmt. Treten Bindungen bei den Ereigniszeiten auf, so wird auch hier die Korrektur nach Efron (1977), siehe Abschnitt 3.1.2, angewandt.

3.2.2 Ridge und Lasso Regression

Im Gegensatz zu dem CoxBoost Algorithmus von Binder und Schumacher (2008) (siehe Abschnitt 3.2.1), der über einen Boosting Algorithmus in jedem Iterationsschritt alle möglichen Anpassungen der Koeffizienten miteinander vergleicht und die beste Anpassung auswählt, wird bei dem Algorithmus zu den hier vorgestellten beiden Methoden über ein zyklisches Koordinatenabstiegsverfahren (siehe Bühlmann und van de Geer, 2011, Seite 38 ff.) jeweils nur ein zuvor ausgewählter Koeffizient angepasst. Bei hochdimensionalen Daten mit wenigen informativen Kovariablen arbeitet dieses Verfahren deutlich schneller als Verfahren mit exakter Liniensuche, wie beispielsweise der LARS-Algorithmus von Tibshirani u. a. (2004) (siehe Simon u. a., 2011).

Im Folgenden werden nun zwei weitere Verfahren vorgestellt, die ebenfalls über eine regularisierte partielle Log-Likelihood den Parametervektor β aus einem Cox-Modell schätzen. Die Herangehensweisen unterscheiden sich lediglich durch den Strafterm der Likelihood, d.h. durch die Wahl des festzulegenden Parameters α aus dem Maximierungsproblem (3.11).

Die sogenannte *Ridge Regression* ist durch einen quadratischen Strafterm charakterisiert. Die zu maximierende regularisierte partielle Log-Likelihood hat die folgende Form:

$$LL_{\lambda,2}(\beta) = LL(\beta) - \lambda \sum_{k=1}^{p} \beta_k^2$$

Die Ridge Regression wurde von Hoerl und Kennard (1970) vorgeschlagen und von Verweij und Van Houwelingen (1994) erstmals im Kontext der Überlebenszeiten, insbesondere unter Verwendung des Cox-Modells beschrieben. Die Regressionskoeffizienten werden durch den Strafterm klein gehalten, so dass das geschätzte Modell viele Koeffizienten nahe bei Null beinhaltet. Es gilt jedoch stets $\hat{\beta}_k > 0$ (siehe z.B. Simon u. a., 2011). Der Strafterm wird meist auch mit *L2-penalty* bezeichnet.

Die *Lasso* Methode nach Tibshirani (1996) setzt im Vergleich zum quadratischen Strafterm der Ridge Regression Absolutwerte in den Strafterm ein. Sie wurde von Tibshirani (1997) für die Schätzung von Cox-Modellen angepasst. Die Lasso Methode definiert die regularisierte Likelihood wie folgt:

$$LL_{\lambda,1}(\beta) = LL(\beta) - \lambda \sum_{k=1}^{p} |\beta_k|$$

In $p \gg n$ Szenarien werden in der Praxis die meisten Koeffizienten häufig exakt Null geschätzt (siehe Bühlmann und van de Geer, 2011, Seite 9), so dass die entsprechenden Variablen keinen Einfluss in dem geschätzten Modell haben. Der Lasso Strafterm heißt in der Literatur auch *L1-penalty*.

Generell gilt wie auch bei dem CoxBoost Algorithmus, dass die Effekte durch die Regularisierung kleiner geschätzt werden als sie möglicherweise sind (Bias). Dagegen ist jedoch die Varianz der Schätzer klein. Die Ergebnisse von Lasso lassen sich meist gut interpretieren, da nur wenige Koeffizienten von Null verschieden geschätzt werden. Bei der Ridge Regression ist die Interpretierbarkeit der Modelle aufgrund der vielen Koeffizienten im Modell oft nicht gegeben. Die Ridge Regression ist der Lasso Methode jedoch überlegen, wenn sich das wahre Modell aus vielen kleinen Effekten zusammensetzt. Gerade solche Modelle werden in der Genetik und Biologie häufig unterstellt, was die Vergleichsstudien von Bøvelstad u. a. (2007) und Kammers u. a. (2011) unterstreichen; hier hat die Ridge Regression besser abgeschnitten. Gute Eigenschaften hat die Ridge Regression zudem, wenn für das Modell relevante Kovariablen korreliert sind. Die entsprechenden Koeffizienten werden in der Regel gleich stark in das Modell aufgenommen. Lasso wählt in solchen Fällen oft nur eine Kovariable unter den korrelierten Kovariablen aus. Die Variablenauswahl von Lasso ist hingegen ein entscheidender Vorteil gegenüber der Ridge Regression. Lasso und Ridge Regression bzw. deren Strafterme können über

ein sogenanntes *elastic net* (Zou und Hastie, 2005) kombiniert werden (siehe Simon u. a., 2011).

Die Lasso Methode und die Ridge Regression sind in dem R Paket *glmnet*, aktuell in Version 1.8.5., implementiert (Simon u. a., 2011). Die Parameterschätzung ist über ein Koordinatenabstiegsverfahren realisiert. Iterativ werden die Koeffizienten $1, \ldots, p$ durchlaufen. In diesem zyklischen Algorithmus wird je Iterationsschritt der entsprechende Koeffizient angepasst, bis die Konvergenz der Koeffizienten eingetreten ist. Die Anzahl der Iterationen wird auf 10^5 Schritte begrenzt. Die Anpassung des jeweiligen Koeffizienten erfolgt über den Gradienten einer Verlustfunktion. Diese Verlustfunktion wird über eine Taylorentwicklung der regularisierten partiellen Likelihood hergeleitet (siehe dazu Simon u. a., 2011, Kapitel 2). Die Anpassung eines einzigen Koeffizienten hat den Vorteil, dass dieser Algorithmus deutlich weniger Ressourcen benötigt als der CoxBoost Algorithmus aus Abschnitt 3.2.1 (siehe Kapitel 5.4 für einen Überblick über die benötigte Rechenzeit der verschiedenen Algorithmen).

Der Parameter λ wird mit der Funktion `cv.glmnet` analog zu Abschnitt 3.2.1 mittels 10-facher Kreuz-Validierung optimiert. Dabei wird ein flexibles Gitter von λ_{max} bis λ_{min} über dem Parameterraum $\mathbb{R}_0^+$ abgesucht. Gestartet wird dabei mit λ_{max}, so dass $\hat{\beta} = 0$ geschätzt wird (siehe Simon u. a., 2011, Abschnitt 2.3). Der optimierte Wert λ_{opt} maximiert entsprechend $\texttt{CVPL}(\lambda)$ (siehe Formel (3.12)).

3.3 Vorhersagefehler zensierter Überlebenszeitdaten

Der Vergleich von geschätzten Überlebensfunktionen und den zugrundeliegenden Modellen ist elementarer Bestandteil dieser Arbeit. Eine Schätzung der Überlebensfunktion ist stets durch ein geschätztes Cox-Modell gegeben, siehe Abschnitt 3.1.3. Die Anpassung eines Cox-Modells kann dabei beispielsweise über einen der in Abschnitt 3.2 verwendeten regularisierten Algorithmen zur Modellgenerierung erfolgen. Um nun mehrere Modelle bzw. die daraus geschätzten Überlebensfunktionen für die einzelnen Patienten miteinander vergleichen zu können, ist ein Maß für die Genauigkeit, mit der die tatsächliche Überlebenswahrscheinlichkeit der Patienten vorhergesagt werden kann, erforderlich. Ein Modell wäre dem anderen vorzuziehen, wenn die Qualität der Vorhersage besser im Sinne dieses Maßes ist.

Der Brier Score nach Graf u. a. (1999) ist ein solches Maß und beschreibt den mittleren quadratischen Vorhersagefehler der geschätzten Überlebenswahrscheinlichkeit zu dem tatsächlichen Zustand des Patienten. Der Brier Score wird zunächst in Abschnitt 3.3.1 eingeführt. Abschnitt 3.3.2 beschreibt das Vorgehen zur Validierung des Brier Scores. Durch die Validierung wird vermieden, überangepasste Modelle zu bevorzugen (overfitting). Dafür soll die Vorhersagegenauigkeit der Überlebenszeiten der Patienten auf für die Schätzung der Modelle unabhängigen Testdaten überprüft werden.

3.3.1 Brier Score

Für jeden Patienten sei wie in Abschnitt 3.1.2 der Kovariablenvektor $Z_j = Z_{j1}, \ldots, Z_{jp}$ gegeben. Die Ereigniszeiten der Zufallsvariablen T seien vorerst ebenfalls bekannt. Sei $\hat{S}(t|Z_j)$ analog zu Abschnitt 3.1.3 die geschätzte individuelle Überlebensfunktion für Patient j. Dann ist dies eine Schätzung der bedingten Wahrscheinlichkeit, dass ein Patient unter Berücksichtigung seines individuellen Kovariablenvektors den Zeitpunkt t überlebt. Zu jedem Zeitpunkt t sei der tatsächliche Zustand des j-ten Patienten $Y_j(t) = \mathcal{I}(t_j > t)$ bekannt. Diese Größe kann als beobachtete individuelle Überlebensfunktion interpretiert werden und nimmt den Wert Eins an, solange der Patient lebt und fällt auf Null, sobald der Patient verstirbt.

Der Brier Score nach Graf u. a. (1999) vergleicht die geschätzte mit der beobachteten Überlebensfunktion eines jeden Patienten und ist definiert durch:

$$\texttt{BS}(t, \hat{S}) = E_Z \left(Y(t) - \hat{S}(t|Z = z) \right)^2 \tag{3.13}$$

$\texttt{BS}(t, \hat{S})$ gibt in Abhängigkeit der Zeit t den erwarteten quadratischen Abstand zwischen der geschätzten individuellen Überlebensfunktion und dem beobachteten Status eines Patienten an. Der Erwartungswert wird dabei auf Basis der für die Schätzung von $S(t|Z = z)$ betrachteten Patientenkohorte bestimmt.

Der Brier Score eines Schätzers $\hat{S}(t|Z = z)$ sollte mindestens unterhalb der folgenden beiden Bezugspunkte liegen. Die Werte der Überlebensfunktion zufällig aus dem Intervall [0,1] auszuwählen, stellt eine naive Schätzung der Überlebensfunktion $S(t)$ dar. Der erwartete Brier Score für diese Schätzung beträgt für jeden Zeitpunkt gerade 1/3 (Mogensen u. a., 2012). Dieser Wert ergibt sich aus dem Erwartungswert einer quadrierten gleichverteilten Zufallsvariable im Intervall [0,1]. Eine weitere naive Schätzung kann dadurch erzielt werden, den Wert der Überlebensfunktion für alle Zeitpunkte mit 50 % anzugeben. Unabhängig vom Zeitpunkt wird dadurch der Brier Score stets einen Wert von 1/4 aufweisen (Mogensen u. a., 2012).

Der Kaplan-Meier-Schätzer (Formel 3.1) berücksichtigt keine Kovariablen und kann ähnlich der beiden Bezugspunkte aus dem vorangegangenen Absatz ebenfalls als Referenzschätzer dienen. Die Eigenschaften des Brier Scores auf Basis der Kaplan-Meier-Schätzung der Überlebenszeit sind in Abbildung 3.1 beispielhaft skizziert. Abbildung 3.1 zeigt die Überlebensfunktion $S(t) = \exp(-(t/\lambda)^\alpha)$ einer Weibull-verteilten Zufallsvariable T, wobei $\lambda = 4$ und $\alpha = 2$ gewählt wurden. Die mediane Überlebenszeit ergibt sich zu $\tilde{t}_{0,5} = \lambda \ln(2)^{1/\alpha} \approx 3{,}33$ (siehe Klein und Moeschberger (2003), Seite 395, für nähere Informationen zu Weibull-verteilten Überlebenszeiten). Abbildung 3.1 zeigt außerdem den Brier Score für den Kaplan-Meier-Schätzer asymptotisch für $n \to \infty$. Da der Kaplan-Meier-Schätzer $\hat{S}$ nach Klein und Moeschberger (2003) (siehe Kapitel 4.2, Theoretical Note 6) konsistent ist, gilt:

$$\begin{aligned} \lim_{n\to\infty} \texttt{BS}(t, \hat{S}) &= S(t) \cdot (1 - S(t))^2 + (1 - S(t)) \cdot (0 - S(t))^2 \\ &= S(t) - S(t)^2 \end{aligned}$$

Es folgt sofort $\texttt{BS}(\tilde{t}_{0,5}) = 0{,}25$. Zum Zeitpunkt $t = 0$ beträgt auch der Brier Score Null. Der Brier Score nimmt mit der Zeit zu, bis er sein Maximum von $1/4$ zum Zeitpunkt der medianen Überlebenszeit erreicht. Anschließend fällt der Brier Score wieder ab. Interpretatorisch macht der Kaplan-Meier-Schätzer zum Zeitpunkt $\tilde{t}_{0,5}$ den größten Fehler im Sinne des Brier Scores. An früheren Zeitpunkten $t \ll \tilde{t}_{0,5}$ sind die Status der Patienten überwiegend Eins und der Kaplan-Meier-Schätzer ist ebenso nahe Eins. Zu späten Zeitpunkten $t \gg \tilde{t}_{0,5}$ sind beide Werte nahe Null. An den extremen Zeitpunkten werden somit nur kleinere Fehler im Sinne des Brier Scores gemacht.

Eine Schätzung der Überlebensfunktion kann mit Hilfe des Brier Scores über die Zeit hinweg begutachtet werden. Dies ist von Vorteil, wenn beispielsweise ein Schätzer A an frühen und ein anderer Schätzer B an späten Zeitpunkten wenig Fehler und zu den jeweils anderen Zeitpunkten viele Fehler macht. Soll hingegen die Information der Fehler in einer Kennzahl zusammengefasst werden, so kann der integrierte Brier Score bis zu einem bestimmten Zeitpunkt t^* angegeben werden. Dieser hat jedoch den Nachteil, dass ein zuvor skizzierter Unterschied zwischen zwei Schätzern nicht unbedingt zu erkennen ist. Der *integrierte Brier Score* ist definiert als:

$$\texttt{IBS}(t^*) = \frac{1}{t^*} \int_0^{t^*} \texttt{BS}(t)\,\mathrm{d}t \tag{3.14}$$

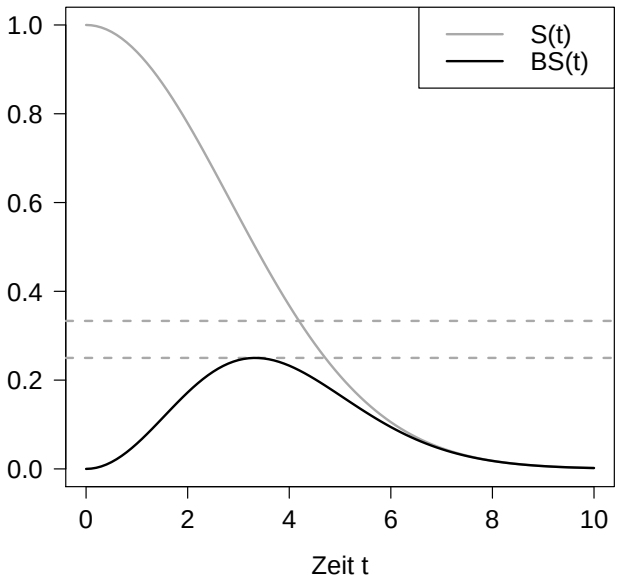


Abbildung 3.1: Grau: Überlebensfunktion einer Weibull verteilten Zufallsvariable mit Parametern $\lambda = 4$ und $\alpha = 2$. Schwarz: Brier Score für den Kaplan-Meier-Schätzer für $n \to \infty$. Die gestrichelten Linien geben die Bezugspunkte 0,33 und 0,25 an.

Seien nun wie in Abschnitt 3.1.2 Daten von $j = 1, \ldots, n$ Patienten wie folgt gegeben. Es wird von rechtszensierten Daten ausgegangen. Zu jedem Patienten j ist deshalb die beobachtete Risikozeit $\tilde{t}_j$ bekannt. Die Risikozeit ist eine Realisation der Zufallsvariable $\tilde{T} = \min(T, C)$, wobei T die Ereigniszeit und C die Zensierungszeit beschreibt. Dann ist mit $\tilde{y}_j(t) = \mathcal{I}(\tilde{t}_j > t)$ der Status bekannt, ob das Ereignis oder die Zensierung erst nach dem Zeitpunkt t eingetreten ist. Zudem seien wieder die Kovariablenvektoren $Z_j = Z_{j1}, \ldots, Z_{jp}$ gegeben. $\hat{S}(t|Z_j)$ sei die geschätzte individuelle Überlebensfunktion eines jeden Patienten.

Der von Graf u. a. (1999) vorgeschlagene *empirische Brier Score* für rechtszensierte Daten ist gegeben durch:

$$\widehat{\mathtt{BS}}(t) = \frac{1}{n} \sum_{j=1}^{n} \hat{g}_j(t) \left(\tilde{y}_j(t) - \hat{S}(t|Z_j) \right)^2 \tag{3.15}$$

Die Gewichte $g_j(t)$ werden durch

$$\hat{g}_j(t) = \frac{(1 - \tilde{y}_j(t))\delta_j}{\hat{P}(C > t_j-)} + \frac{\tilde{y}_j(t)}{\hat{P}(C > t)} \tag{3.16}$$

geschätzt. Dabei ist $\delta_j = \mathcal{I}(t_j < c_j)$ der Zensierungsindikator für Patient j. $\hat{P}(C > t)$ ist der Kaplan-Meier-Schätzer der Zensierungszeiten und gibt die Wahrscheinlichkeit dafür an, dass eine Zensierung erst nach dem Zeitpunkt t eintritt. $\hat{P}(C > t_j-)$ kennzeichnet die entsprechende Wahrscheinlichkeit unmittelbar vor dem Zeitpunkt t_j (Binder u. a., 2011, Seite 177). Um den Kaplan-Meier-Schätzer der Zensierungszeiten zu erhalten, ist in Gleichung (3.1) d_i gleich der Anzahl der Zensierungen zum Zeitpunkt $t_{(i)}$ zu setzen. Unter Risiko stehen dann diejenigen Patienten, die zu diesem Zeitpunkt noch nicht zensiert sind.

Der empirische Brier Score nach Graf u. a. (1999) ersetzt den tatsächlichen Status $y_j(t)$ durch den beobachteten Status $\tilde{y}_j(t)$. Ist ein Patient zum Zeitpunkt t bereits zensiert, so darf sein Status $\tilde{y}_j(t) = 0$ nicht zur Berechnung des Vorhersagefehlers zu diesem Zeitpunkt verwendet werden. Der Brier Score würde sonst den Vorhersagefehler überschätzen. Durch die Gewichtung mit $\hat{g}_j(t)$ bekommen bereits zensierte Patienten das Gewicht Null (erster Summand in Gleichung (3.16)). Die Gewichtung der übrigen Patienten wird so vorgenommen, dass die Summe der Gewichte für jeden Zeitpunkt stets der Anzahl der Patienten n entspricht. Patienten, die zum Zeitpunkt t noch nicht ausgefallen sind, werden am stärksten gewichtet (zweiter Summand in Gleichung (3.16)). Nach Gerds und Schumacher (2006) ist der auf diese Weise gewichtete empirische Brier Score von Graf u. a. (1999) eine konsistente Schätzung des Brier Scores aus Gleichung (3.13). Die Gewichte $\hat{g}_j(t)$ werden in der Literatur *Inverse Probability of Censoring Weights*, kurz IPCW, genannt (Mogensen u. a., 2012). Sind keine Überlebenszeiten zensiert, so gilt $\hat{g}_j(t) = 1$ für alle j und für alle t.

Der Einfluss der Gewichte $\hat{g}_j(t)$ sei am Beispiel der zuvor verwendeten Weibull-verteilten Überlebenszeiten verdeutlicht. Hierzu wurden 20 Überlebenszeiten zufällig aus der in Abbildung 3.1 skizzierten Verteilung gezogen. Dabei wurden zwei Zeiten zensiert. Ab-

bildung 3.2 zeigt den Kaplan-Meier-Schätzer dieser Daten. Die zwei Zensierungen sind mit einer Raute gekennzeichnet. Ebenfalls in Abbildung 3.2 sind zwei Varianten des empirischen Brier Score gezeichnet. Dabei wurden einmal die Zensierungen berücksichtigt und die Gewichte aus (3.16) entsprechend berechnet (schwarze Kurve) und andererseits die Zensierung ignoriert und jedes Gewicht auf den Wert Eins gesetzt (rote Kurve). Die konsistente Schätzung des Brier Scores bezweckt zum Zeitpunkt der ersten Zensierung ($t = 1{,}35$) keine Änderung, der Brier Score wird weiterhin mit 0,16 für diesen Zeitpunkt geschätzt. Die Variante ohne Berücksichtigung der Zensierung wertet hingegen einen Fehler und erhöht den Brier Score auf 0,19, so dass der Brier Score bis zur medianen Überlebenszeit überschätzt wird. Die zweite Zensierung ($t = 4{,}75$) verstärkt die Unterschätzung des Fehlers für Zeiten nach der medianen Überlebenszeit. Dieses Beispiel zeigt, dass selbst bei einer Zensierungsrate von 10 % eine konsistente Schätzung des Brier Scores das Resultat deutlich verbessern kann.

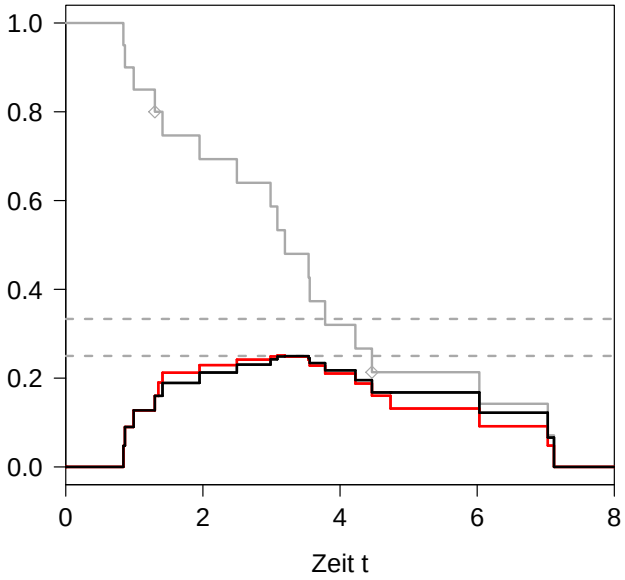


Abbildung 3.2: Empirischer Brier Score nach Graf u. a. (1999) mit (schwarz) und ohne (rot) Verwendung der Gewichte $\hat{g}_j(t)$ zur konsistenten Schätzung im Falle zensierter Daten. Der Brier Score wurde jeweils für den Kaplan-Meier-Schätzer (grau) berechnet. Dieser schätzt die Überlebensfunktion aus Abbildung 3.1 anahnd 20 zufällig gezogener Überlebenszeiten. Zwei Zensierungen (Raute) wurden den Daten hinzugefügt.

Durch die Gewichtung wird die Fehlermasse der zensierten Zeiten auf die späteren Zeitpunkte aufgeteilt. In ähnlicher Weise verfährt der Kaplan-Meier-Schätzer, der die Masse der Ausfallwahrscheinlichkeiten von zensierten Beobachtungen auf spätere Zeitpunkte gleichmäßig aufteilt.

3.3.2 Validierung der Vorhersagefehler

Im vorangegangenen Kapitel wurde der Vorhersagefehler eines Schätzers stets auf denselben Daten erhoben, auf denen er geschätzt wurde. Der so ermittelte Fehler wird als Trainingsfehler bezeichnet. Relevant ist jedoch der erwartete Testfehler, d.h. der erwartete Vorhersagefehler auf einer für die Schätzung unabhängigen Stichprobe. Der Trainingsfehler führt im Allgemeinen zu einer Unterschätzung des Testfehlers, da die zugrundeliegenden Modelle zu sehr an die erhobenen Daten angepasst sind (siehe Hastie u. a., 2009, Kapitel 7.4). Bei diesen Daten handelt es sich lediglich um eine Stichprobe aus der Gesamtpopulation, die zufälligen Schwankungen unterzogen ist.

Binder u. a. (2011) sowie Mogensen u. a. (2012) fassen mehrere Verfahren zusammen, mit denen der erwartete Testfehler im Sinne des Brier Scores ermittelt werden kann, wenn keine unabhängigen Testdaten zur Verfügung stehen. Im Folgenden wird das in dieser Arbeit angewandte Bootstrap Verfahren ohne Zurücklegen, auch *Subsampling* genannt, beschrieben. Dabei werden die Daten B-mal jeweils zufällig in zwei Teile, einen Trainingsteil und einen Testteil aufgeteilt. In jedem Bootstrap-Schritt $b = 1, \ldots, B$ wird zunächst der Index $\mathcal{I}_b \subset \{1, \ldots, n\}$ der Trainingsmenge bestimmt. Dabei werden ohne Zurücklegen zufällig $|\mathcal{I}_b| = \lfloor r \cdot n \rfloor$ Indizes ausgewählt. Der Umfang der Trainingsmenge wird über das Verhältnis $r \in (0, 1)$ angegeben und hier mit $r = 0{,}632$ festgelegt. Die übrige Menge $\{1, \ldots, n\} \backslash \mathcal{I}_b$ charakterisiert die Testdaten. Bezeichne nun mit $\hat{S}^{(b)}$ eine Schätzung für $S(t|Z = z)$, für die nur die Trainingsmenge $\mathcal{I}_b$ des b-ten Bootstrap-Schrittes herangezogen wurde. Pro Bootstrap-Schritt wird demnach ein neues Modell basierend auf den Trainingsdaten geschätzt. Danach kann jeweils der Brier Score dieses Schätzers auf der entsprechend disjunkten Testmenge bestimmt werden. Durch Mittelung dieser Brier Scores über die Bootstrap Stichproben hinweg ergibt sich der geschätzte erwartete Brier Score der Testdaten zu:

$$\texttt{BootCvBS}(t, \hat{S}) = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{|j \notin \mathcal{I}_b|} \sum_{j \notin \mathcal{I}_b} \hat{g}_j(t) \left(\tilde{y}_j(t) - \hat{S}^{(b)}(t|Z = z_j) \right)^2 \tag{3.17}$$

Es gilt zu beachten, dass der Kaplan-Meier-Schätzer der Zensierungszeiten, der für die Gewichte $\hat{g}_j(t)$ und damit für eine konsistente Schätzung entscheidend ist, in (3.17) jeweils mit der Testmenge $\{j \colon j \notin \mathcal{I}_b\}$ ermittelt wird. Das Modell zur Ermittlung von $\hat{S}^{(b)}$ wird nur auf der entsprechenden Trainingsmenge $\mathcal{I}_b$ erstellt.

Der Vorhersagefehler in (3.17) berücksichtigt jegliche Form von Zufallseinflüssen der Trainingsdaten. `BootCvBS` wird der Einfachheit halber im Weiteren *mittlerer Brier Score* genannt. Zur Veranschaulichung dient erneut das Beispiel Weibull-verteilter Überlebenszeiten aus Abschnitt 3.3.1. Aus einer Stichprobe von $n = 100$ Überlebenszeiten soll der mittlere Brier Score für den Kaplan-Meier-Schätzer bestimmt werden. Dabei werden $B = 10$ Bootstrap-Stichproben der Größe $|\mathcal{I}_b| = 66$ ohne Zurücklegen gezogen. Diese dienen jeweils als Trainingsdaten und die entsprechenden komplementären Mengen als Testdaten. Auf jedem der 10 Trainings-Stichproben wird der Kaplan-Meier-Schätzer (3.1) ermittelt. Anschließend wird jeweils der Brier Score dieser Schätzer auf den entsprechenden Testdaten errechnet. Abbildung 3.3 zeigt in grau jeweils den Brier Score

im Verlauf der Zeit für die 10 Test-Stichproben. Die schwarze Kurve zeigt den mittleren Brier Score aus Gleichung (3.17). In der Abbildung sind zudem die theoretischen Bezugspunkte von 0,25 und 0,33 für den Brier Score eingezeichnet. Während die empirischen Brier Scores auf den Trainingsdaten (nicht gezeigt) die Marke von 0,25 zu keinem Zeitpunkt überschreiten, so können diese auf den Testdaten sehr wohl Werte oberhalb dieser Marke annehmen. Schwanken die Überlebenszeiten zwischen Trainings- und Testmengen, wie hier für die Fallzahl von $n = 100$, so kann der mittlere Brier Score der Testdaten somit für manche Zeitpunkte schlechtere Werte als die Referenzwerte für eine naive bzw. zufällige Schätzung annehmen.

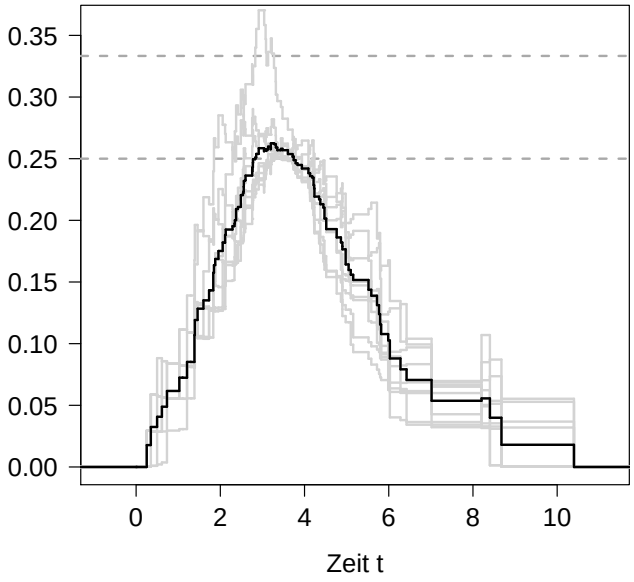


Abbildung 3.3: Mit $n = 100$ zufällig aus einer Weibull-Verteilung gezogenen Überlebenszeiten wurde der mittlere Brier Score `BootCvBS` bestimmt (schwarze Kurve). Der mittlere Wert ergibt sich in diesem Beispiel aus $B = 10$ Bootstrap-Stichproben. Die geschätzten Brier Scores auf den Testdaten der einzelnen Bootstrap-Stichproben wurden in grau eingezeichnet.

Mogensen u. a. (2012) stellen mit der Funktion `PEC` aus dem gleichnamigen R-Paket ein Werkzeug zur Verfügung, mit dem sich der mittlere Brier Score analysieren lässt. Allerdings sind die zur Wahl stehenden Schätzfunktionen eingeschränkt, so dass die Verfahren aus Abschnitt 3.2.2 sowie aus Kapitel 4 nicht verwendet werden können. Ferner ist es nicht möglich die Berechnungen auf Batch-Systemen wie dem LiDOng (**Li**nux cluster **DO**rtmund **n**ext **g**eneration) der TU Dortmund auszulagern, was hinsichtlich der benötigten Ressourcen (siehe Abschnitt 5.4) zwingend erforderlich ist. Aus diesen Gründen wurden die hier vorgestellten Verfahren zur Berechnung und Validierung des

Brier Scores eigenhändig implementiert (siehe Anhang B) und mit Hilfe der Pakete *BatchJobs* (Bischl u. a., 2012b) und *BatchExperiments* (Bischl u. a., 2012a) auf dem Rechencluster LiDOng ausgeführt. Für Details bezüglich der beiden Pakete sei hier auf die Arbeit von Bischl u. a. (2012c) verwiesen.

4 Erkennung relevanter Unterschiede in den Überlebenszeitmodellen von Patientenuntergruppen

Welche Beobachtungen einer vorliegenden Stichprobe zur Bildung eines statistischen Modells verwendet werden sollten, wird meist unter der Annahme von unabhängig identisch verteilten Stichprobenvariablen nicht hinterfragt. In diesem Fall werden alle Beobachtungen (hier Patienten) zur Modellbildung genutzt. Im Bereich der Krebsforschung ist die Rechtfertigung dieser Annahme jedoch höchst fraglich. Im Hinblick auf die Modellierung der Überlebenszeit der Patienten in Abhängigkeit genetischer und klinischer Kovariablen stellt sich die Frage, ob ein Zusammenhang zwischen bestimmten Kovariablen mit der Überlebenszeit tatsächlich in der gesamten Kohorte gleichermaßen Gültigkeit besitzt. Viele Forschungsergebnisse zeigen, dass bestimmte Zusammenhänge lediglich in charakteristischen Untergruppen zu beobachten sind. Speziell in der Lungenkrebsforschung wird gehäuft von heterogenen Subtypen berichtet (siehe beispielsweise Bianchi u. a. (2007), Cetin u. a. (2011), Micke u. a. (2011) oder Govindan u. a. (2012)). Dabei werden in diesen Arbeiten die Untergruppen meist durch klinische Variablen charakterisiert. Oft ist dies ein bestimmter Histologietyp, aber auch der Raucherstatus oder das Stadium des Tumors spielen eine Rolle.

Ziel dieser Arbeit ist zunächst die Schätzung der Überlebenszeit von Patienten einer bestimmten Untergruppe. Sei dazu formal eine Untergruppenvariable in Form einer diskreten Zufallsvariable U gegeben, deren Ausprägungen die zu untersuchenden Untergruppen darstellen. Dann ist konkret für eine Untergruppe $U = g$, $g \in \{1, \ldots, G\}$, zu entscheiden, welche Patienten zur Modellierung des zu dieser Untergruppe gehörenden Überlebenszeitmodells genutzt werden sollen. Direkt können zwei unterschiedliche Ansätze verfolgt werden. Zum einen kann mit allen Patienten unabhängig ihrer Untergruppenzugehörigkeit ein Modell gefunden werden. Dieses Modell würde demnach für alle möglichen Untergruppen dieselben Zusammenhänge zwischen Überlebenszeit und Kovariablen annehmen. Nur in bestimmten Untergruppen präsente Zusammenhänge werden durch ein auf diese Weise erstelltes Modell je nach Untergruppengröße selten bis gar nicht berücksichtigt. Zum anderen ließen sich nur diejenigen Patienten aus der interessierenden Untergruppe betrachten. Je seltener jedoch eine Untergruppe ist, desto größer ist die Varianz der Modellparameter und damit die Unsicherheit der Schätzung der Überlebenszeit. In Kapitel 4.1 werden weitere Modellbildungsverfahren vorgestellt, welche

nicht nur einen Kompromiss der beiden zuvor erläuterten Verfahren darstellen, sondern darüber hinaus in bestimmten Situationen überhaupt erst eine sinnvolle Schätzung der Überlebensfunktion von Patienten einer Untergruppe ermöglichen.

Ein weiterer Schwerpunkt dieser Arbeit liegt neben der Schätzung der Überlebensfunktion auf dem Vergleich der verwendeten Methoden zur Modellbildung. Im Hinblick auf die Erkennung relevanter Untergruppen mit spezifischen Zusammenhängen zwischen den unabhängigen Variablen und der Überlebenszeit soll die Qualität der Vorhersagen der Modelle bewertet und verglichen werden. Dazu stellt Kapitel 4.2 die Herangehensweise vor, so dass für die betrachteten Überlebenszeitmodelle jeweils deren Vorhersagefehler ermittelt und entsprechend validiert werden kann. Zudem wird die Eignung der Modelle in bestimmten simulierten Szenarien diskutiert.

4.1 Berücksichtigung unterschiedlicher Verteilungen der Daten zwischen den Untergruppen

Dieser Abschnitt erläutert im Detail die gewichteten Modellbildungsverfahren zur Generierung von Überlebenszeitmodellen spezifisch für bestimmte Untergruppen in den Daten. Die Untergruppen setzen sich aus einzelnen Beobachtungen (Patienten) zusammen. Die Modelle für eine Untergruppe werden stets mit allen verfügbaren Beobachtungen trainiert. Dabei geht jede Beobachtung mit einem entsprechenden Gewicht in die Modellberechnung ein. Eine solche Gewichtung wird ebenfalls bei sogenannten lokalen Regressionsverfahren genutzt. Eine methodische Abgrenzung solcher Verfahren zu der in dieser Arbeit vorgestellten Gewichtung findet sich in Abschnitt 4.1.1.

In Abschnitt 4.1.2 wird ein einfacher Ansatz zur Bestimmung von Stichprobengewichten diskutiert. Es wird lediglich unterschieden, ob eine Beobachtung zu der Untergruppe oder der Restgruppe gehört. Daraufhin wird ein für diese Gruppe einheitliches Gewicht vergeben. Die zu verwendenden Gewichte werden a priori festgelegt.

Anschließend wird ein Verfahren vorgestellt, welches jeder Beobachtung ein individuelles Stichprobengewicht zuweist. Die Gewichte beschreiben jeweils, wie gut eine Beobachtung zu der Verteilung der Daten in der Untergruppe passt. Die Grundidee dieser Stichprobengewichte beruht auf dem sogenannten *distribution matching* Verfahren von Bickel (2009), der dieses Prinzip zur Modellierung von Therapieerfolgen bei HIV Patienten angewandt hat (siehe Bickel u. a. (2008) und Bogojeska u. a. (2010)). Dabei entspricht das optimale Gewicht gerade dem Dichtequotienten der Verteilung der Daten bedingt auf die Untergruppe und auf die Verteilung aller Daten. Die theoretische Grundlage dieser Herangehensweise wird in Abschnitt 4.1.3 erläutert. Abschnitt 4.1.4 beschreibt die regularisierte multinomiale logistische Regression, mit der die Stichprobengewichte bzw. diese Wahrscheinlichkeiten geschätzt werden.

Ziel der Modellbildung ist es, die individuelle Überlebensfunktion eines Patienten anhand gegebener Kovariablen zu schätzen. Wie in den Abschnitten 3.1.2 und 3.1.3 gezeigt, sind hierfür zwei Schritte nötig. Durch Maximieren der partiellen Likelihood (3.3) lässt sich noch keine Überlebensfunktion eines einzelnen Patienten angeben, da nach (3.10)

hierzu die Baseline-Rate benötigt wird. Diese wird erst durch Maximieren der Likelihood (3.8) geschätzt. Die Stichprobengewichte in den hier vorgestellten Verfahren müssen in beiden Optimierungsschritten berücksichtigt werden. Abschnitt 4.1.5 zeigt, wie die Stichprobengewichte in diese beiden Likelihoods aufzunehmen sind, um zu einer gewichteten Schätzung der individuellen Überlebensfunktion zu gelangen.

Abschnitt 4.1.6 diskutiert abschließend einen alternativen Ansatz zur Schätzung der Stichprobengewichte sowie zur Modellierung der Überlebenszeit über die Formulierung eines hierarchischen Bayes Modells.

4.1.1 Methodische Abgrenzung zu bisherigen lokalen Regressionsverfahren

Die Likelihood in einem Modellbildungsprozess zu gewichten findet ebenfalls bei sogenannten lokalen Regressionsverfahren Anwendung. Eine ausführliche Betrachtung verschiedenster lokaler Verfahren findet sich im Buch *Local Regression and Likelihood* von Loader (1999). In Kapitel 7 gibt Loader darin eine Einführung in das lokale Likelihood Modell für zensierte Überlebenszeitdaten. Generell kann jedes parametrische Modell in ein lokales Modell überführt werden, wenn die Beobachtungen in die Modellbildung gewichtet aufgenommen werden (siehe Hastie u. a., 2009, Seite 205). Der Begriff *lokal* ist dabei keinesfalls eindeutig. Folglich gibt es viele unterschiedliche lokale Verfahren. Am verbreitetsten sind sogenannte Kernel-basierte lokale Methoden. Dabei wird jeweils ein Modell für jeden möglichen Punkt im Kovariablenraum angepasst. Über ein entsprechendes Abstandsmaß fließen ausschließlich benachbarte Punkte in die Modellbildung ein. Eine weitere Variante stellen Cluster-basierte Verfahren dar. Dabei wird anstelle eines Modells für jeden Punkt jeweils ein Modell für eine zuvor über ein Clusterverfahren gefundene Teilmenge angepasst. Analog gehen auch hier die Beobachtungen im Verhältnis zu ihrem Abstand zum jeweiligen Cluster in die Modellbildung ein.

In den folgenden Charakteristika unterscheiden sich die lokalen Verfahren aus der Literatur mit der in dieser Arbeit vorgestellten Methodik. Zunächst wird die Nähe bzw. der Abstand zu der Verteilung der Daten in einer zuvor fest definierten Untergruppe gemessen. Diese Verteilung basiert auf einem Träger, der sich aus den Kovariablen und der späteren Zielvariablen (hier der Überlebenszeit) zusammensetzt. Sowohl innerhalb als auch außerhalb der Untergruppe können einzelne Beobachtungen gewichtet in die spätere Schätzung der Überlebenswahrscheinlichkeit eingehen. Des Weiteren ist hier die Schätzung des Überlebenszeitmodells für eine Untergruppe spezifisch. Dies bedeutet unter anderem, dass dieses Modell auch ausschließlich auf Testdaten der Untergruppe evaluiert wird (siehe Abschnitt 4.2.1). Außerdem ist gerade anders als bei der kernelbasierten Verfahren eine einfache Interpretation des Modells gegeben, da jeweils ein Parametervektor die Zusammenhänge für die ganze Untergruppe beschreibt.

Im Folgenden wird kurz auf die beiden bedeutendsten Varianten lokaler Regressionsverfahren eingegangen und der Unterschied zu dem in dieser Arbeit verfolgten Ansatz aufgezeigt.

Kernel-basierte lokale Regression

Bei einer Kernel-basierten lokalen Regression wird separat jeweils ein Modell an jedem Punkt z_0 im Kovariablenraum angepasst (siehe Hastie u. a., 2009, Kapitel 6). Der Punkt z_0 wird dabei auch *query point* genannt (siehe Atkeson u. a., 1997, Seit 12). Für die Schätzung eines Modells an einem solchen query point werden benachbarte Punkte (Beobachtungen) hinzugezogen. Die Nachbarschaft wird dabei im Kovariablenraum definiert. Über eine entsprechende Gewichtsfunktion $K_\lambda(z_0, z_j)$, den Kernel, wird für jeden Punkt z_j seine Nähe zum query point z_0 festgelegt. Somit wird über den Kernel die Nachbarschaft charakterisiert. Eine Beobachtung j geht damit gewichtet in Abhängigkeit zum Abstand von z_j zu z_0 in die Schätzung des Modells in z_0 ein. Der Parameter λ beschreibt dabei die Ausdehnung dieser Nachbarschaft. Durch Variieren dieses Parameters kann die Nachbarschaft eingegrenzt oder ausgedehnt werden. λ wird meist in einer Kreuzvalidierungsschleife optimiert.

Der Kernel bzw. das Abstandsmaß muss zunächst geeignet gewählt werden. Das Stichprobengewicht, welches durch den Kernel festgelegt ist, kann in einem Bereich um z_0 für alle Beobachtungen gleich sein (vgl. k-nächste Nachbarn Methode) oder mit größerem Abstand abnehmen. Dabei sind je nach Kernel die folgenden zwei Eigenschaften zu beachten (vgl. Altman, 1992). Wird der Kernel beispielsweise durch die k-nächste Nachbarn Methode beschrieben, so bleibt die Anzahl an Beobachtungen, die für die Schätzung genutzt werden, für jeden query point gleich. Über ein k-nächste Nachbarn Verfahren (Silverman und Jones, 1989) werden die k Beobachtungen, die zum Punkt z_0 am nächsten sind, gleich gewichtet zur Schätzung herangezogen. Somit bleibt die Varianz der einzelnen Schätzer konstant. Lediglich die Ausdehnung der Nachbarschaft, die Bandbreite, ändert sich über die query points. Andere Kernel, wie beispielsweise der Epanechnikov Kernel (Hastie u. a., 2009, Seite 192), sind durch eine feste Bandbreite charakterisiert. Somit ändert sich ggf. die Anzahl Beobachtungen und damit auch die Varianz über die query points hinweg.

Ähnlich wie bei einem gleitenden Durchschnitt gelangt man durch Aneinanderfügen der lokalen Schätzungen für jeden Kovariablenpunkt zu einem Gesamt-Schätzer für alle Datenpunkte (Cleveland, 1979). Die Gesamtheit aller an den query points geschätzten Modelle spiegelt letztendlich das Ergebnis der lokalen gewichteten Regression auf den gesamten Daten wider. Die lokale gewichtete Regression wird im eindimensionalen Fall oftmals auch *loess* genannt (Cleveland und Devlin, 1988). Allerdings kann bei einer lokalen Regression der Zusammenhang zwischen Kovariablen und Zielvariablen nicht über einen globalen Regressionsparameter angegeben werden. Da für alle möglichen Punkte z_0 jeweils ein separates Modell angepasst wird, ist die Interpretation schwierig. Im Gegensatz dazu wird in dieser Arbeit ein Modell für eine Untergruppe geschätzt und ist somit leicht zu interpretieren.

Die verwendeten Kernel lassen sich problemlos auf hochdimensionalen Daten verwenden (Ruppert und Wand (1994), Hastie u. a. (2009)). Allerdings ist der Hauptkritikpunkt der lokalen gewichteten Regression gegenüber einem globalen Regressionsmodell der Fallzahlverlust. Für jeden query point bleibt lediglich eine kleine Anzahl Beobachtungen, über die das Modell geschätzt wird. Im Fall hochdimensionaler Daten wirkt

dieser Nachteil besonders stark. Tutz und Binder (2004) diskutieren dazu Methoden zur lokalen Variablenselektion. Für jeden query point können andere Kovariablen für die Modelle gewählt werden. Dies entspricht theoretisch dem Gedanken der personalisierten Medizin, wenn ein query point als Patient aufgefasst wird. Jedoch führt die hohe Dimension der Daten zu keiner stabilen Schätzung. Eine Validierung gestaltet sich aufgrund der geringen Stichprobengrößen um den query point schwierig und auch die Interpretation der Modelle (für jede Beobachtung/jeden Patienten eines) ist unübersichtlich. Das in Abschnitt 4.1.3 motivierte Modell erlaubt hingegen eine einfache Interpretation des Modells für jede Untergruppe. Die Untergruppen bilden in gewissem Sinn die query points. Somit sind diese durch die Anzahl an Untergruppen beschränkt. Daneben wird die Nachbarschaft in dem Modell über den Zusammenhang zwischen Kovariablen und späterer Zielvariablen charakterisiert. Hierzu folgt ebenfalls in Abschnitt 4.1.3 eine theoretische Herleitung. Die Auswahl der Kernel bzw. der Abstandsmaße ist hingegen eher heuristischer Natur.

Cluster-basierte lokale Regression

Entgegen der Kernel-basierten lokalen Regression werden bei der Cluster-basierten lokalen Regression zunächst über ein Clusterverfahren Teilmengen der Daten identifiziert, für die dann ein Modell separat geschätzt wird. Binder u. a. (2012) haben diese Methode vorgeschlagen und ähnlich zu dieser Arbeit auf Überlebenszeitdaten angewandt. Die Beobachtungen werden zunächst anhand der genetischen Kovariablen geclustert. Hierbei kann beispielsweise ein beliebiges hierarchisches Clusterverfahren angewandt werden. Die ermittelten Cluster können ebenfalls als Untergruppen angesehen werden. Es wird jeweils ein Modell für jedes Cluster (Untergruppe) angepasst. Die Patienten aus den übrigen Clustern gehen in die Modellbildung über ein konstantes Gewicht ein. Dieser Ansatz ist vergleichbar mit der im folgenden Abschnitt 4.1.2 vorgestellten Herangehensweise, wobei Binder u. a. (2012) das entsprechende Gewicht optimieren. Neben dem konstanten Gewicht wird die Anzahl der Cluster ebenfalls über eine Kreuzvalidierungsschleife optimiert. Somit ergeben sich zwei zusätzliche Tuningparameter, was die benötigte Rechenzeit stark ansteigen lässt. Binder u. a. (2012) haben deshalb nicht den gesamten Raum der Gewichte abgesucht, sondern lediglich einzelne Gewichte exemplarisch getestet.

Durch die Vergabe eines konstanten Gewichtes wird die entscheidende Möglichkeit ausgeschlossen, dass lediglich ein kleiner Teil der Patienten aus der Restgruppe zur Schätzung der Untergruppe beiträgt. Stattdessen tragen alle Patienten gleich wenig oder stark zur Schätzung des Untergruppenmodells bei. Dies stellt neben der Definition der Untergruppen den wesentlichen Unterschied zu dem in Abschnitt 4.1.3 vorgestellten Verfahren dar. Eine Alternative wäre die Abstandsdefinition über ein Distanzmaß zur Clustermitte. Auf diese Weise könnten ebenso individuelle Stichprobengewichte generiert werden. Des Weiteren können die durch die Cluster definierten Untergruppen nicht sofort interpretiert werden. Dies ließe sich durch ein nachgeschaltetes Cluster Profiling realisieren.

4.1.2 Konstante Gewichtung der Beobachtungen innerhalb einzelner Untergruppen

Ist der Einfluss der Kovariablen auf die Überlebenszeit nicht homogen über bestimmte Untergruppen und soll ein Überlebenszeitmodell für eine dieser Untergruppen geschätzt werden, so kann es dennoch sinnvoll sein, die nicht zu der interessierenden Untergruppe gehörenden Patienten zu einem gewissen Anteil (Gewicht) mit für die Schätzung der Überlebensfunktion der Patienten der Untergruppe zu verwenden. Dies lässt sich über den Kompromiss zwischen Bias und Varianz für einen mittleren quadratischen Fehler (`MSE`) motivieren. Dabei gilt $\texttt{MSE} = \texttt{Var} + \texttt{Bias}^2$ (Hastie u. a., 2009, Seite 24, Gleichung (2.25)).

Einerseits kann eine Modellierung ausschließlich auf Basis der Patienten in der Untergruppe bei einer sehr kleinen Untergruppengröße zu keinem stabilen Modell führen (hohe Varianz). Andererseits kann bei Verwendung aller Patienten der für die Untergruppe spezifische Einfluss zwischen Kovariablen und Überlebenszeit nicht korrekt erkannt werden (Bias). Über eine entsprechende Gewichtung der Patienten kann evtl. ein Kompromiss zwischen diesen beiden Größen gefunden werden, so dass das resultierende Modell in dieser Hinsicht optimal ist.

Diese Gewichtung kann heuristisch ermittelt werden. Sei hierzu die interessierende Untergruppe mit $U = g$ bezeichnet. Jeder Patient $j = 1, \ldots, n$ erhält sein Gewicht $\hat{\omega}_{gj}$ wie folgt:

$$\hat{\omega}_{gj} = \begin{cases} 1 & \text{falls } u_j = g \\ \upsilon & \text{sonst} \end{cases}, \ \upsilon \in (0{,}1)$$

Dabei gibt u_j die Ausprägung der Untergruppenvariable U für den j-ten Patienten an. Damit gehen die Patienten aus der Untergruppe stets mit einem Gewicht von Eins in die Schätzung ein. Die übrigen Patienten erhalten ein einheitliches Gewicht υ. Dieses wird a priori aus dem Intervall (0,1) festgelegt. Wie im Detail die Modellierung der Überlebenszeit mit entsprechenden Gewichten durchgeführt wird, ist in Abschnitt 4.1.5 beschrieben.

Soll das hier vorgeschlagene Gewicht υ optimiert werden, darf dies selbstverständlich nicht auf allen Daten geschehen. Eine geeignete Wahl von υ wäre es, jenes Gewicht zu wählen, welches einen Bootstrap-Fehler in einem Trainings- und Testszenario minimiert (vgl. Abschnitt 3.3.2). Dieser Fehler kann beispielsweise durch den integrierten Brier Score aus Gleichung (3.14) angegeben werden. Entscheidend ist die Verwendung einer Bootstrap-Stichprobe mit exakt der absoluten Häufigkeit der Untergruppe in der Trainingsmenge, da das optimale Gewicht von dieser Größe abhängt. Je größer die Fallzahl der Untergruppe, desto kleiner ist das optimale Gewicht. Dies impliziert eine Bootstrap-Stichprobe mit Zurücklegen zu ziehen. Erste Tests zeigten jedoch, dass die Anzahl der benötigten Bootstrap-Stichproben zu groß für eine zeitnahe Berechnung ist ($B \gg 1\,000$). Abschnitt 5.4 geht im Detail auf die benötigten Ressourcen ein. Abschließend sei gesagt, dass sich bei mindestens drei verschiedenen Untergruppen dieses Szenario erweitern ließe, indem jede Gruppe ein separates Gewicht bekommt. Dies würde jedoch den Raum, über dem die Gewichte abgesucht werden müssten, in die Höhe treiben.

4.1.3 Motivation individueller Stichprobengewichte

Sei mit folgender vereinfachten Notation $l\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_0^+,\ (y, f_u(z)) \mapsto l\,(y, f_u(z))$ eine beliebige nicht negative Verlustfunktion zwischen einer beobachteten abhängigen Variable $y \in \mathbb{R}$ und den vorliegenden Kovariablen $z \in \mathbb{R}^p$ gegeben. $f_u(z)$ bezeichnet eine Schätzfunktion $f\colon \mathbb{R}^p \to \mathbb{R}$, die für eine bestimmte Untergruppe $U = u$ die Zielvariable y aus den beobachteten Kovariablen z vorhersagt. Für diese Untergruppe soll nun der erwartete Verlust

$$E_{p(y,z|u)}\left[l\left(y, f_u(z)\right)\right] \tag{4.1}$$

minimiert werden (vgl. Bickel u. a. (2008) und Bogojeska u. a. (2010)). Dabei charakterisiert $p(y, z|u)$ die Verteilung der Daten bedingt auf die Untergruppe u. Der Brier Score (siehe (3.13)) aus Abschnitt 3.3.1 ist beispielsweise ein erwarteter Verlust im Sinne von Gleichung (4.1). Mit der hier eingeführten Notation bezeichnet der Brier Score aus Gleichung (3.13) den erwarteten Verlust bezüglich der Verteilung aller zugrundeliegenden Daten $p(y, z)$. Gilt $p(y, z|u) = p(y, z)$, so minimiert dieser stets den erwarteten Verlust in der Untergruppe u. Im Fall heterogener Untergruppen kann jedoch im Allgemeinen nicht von der Gleichheit der Verteilungen $p(y, z|u)$ und $p(y, z)$ ausgegangen werden. Ein weiteres Beispiel für eine Verlustfunktion ist die aus Gleichung (3.3) resultierende negative (Log-)Likelihood der einzelnen Patienten.

Unterscheiden sich die Verteilungen der Daten in einer oder mehrerer Untergruppen, so führt die Verwendung aller Daten nicht zur Minimierung des Verlustes in den Untergruppen. In Gleichung (4.4) zeigt sich jedoch, dass unter Verwendung geeigneter Gewichte in Form des Dichtequotienten

$$\omega_u(y, z) = \frac{p(y, z|u)}{p(y, z)} \tag{4.2}$$

der erwartete Verlust in der Untergruppe mit Hilfe aller Daten minimiert wird. Die Herleitung dieser Aussage ergibt sich wie folgt (vgl. Bickel, 2009, Seite 57):

$$\begin{aligned}
E_{p(y,z|u)}\left[l\left(y, f_u(z)\right)\right] &= \int p(y, z|u) \cdot l\left(y, f_u(z)\right)\, \mathrm{d}y\, \mathrm{d}z \\
&= \int \frac{p(y, z|u)}{p(y, z)} p(y, z) \cdot l\left(y, f_u(z)\right)\, \mathrm{d}y\, \mathrm{d}z \qquad (4.3) \\
&= E_{p(y,z)}\left[\frac{p(y, z|u)}{p(y, z)} \cdot l\left(y, f_u(z)\right)\right] \\
&= E_{p(y,z)}\left[\omega_u(y, z) \cdot l\left(y, f_u(z)\right)\right] \qquad (4.4)
\end{aligned}$$

Der erwartete Verlust in (4.1) bezüglich der Verteilung der Daten der Untergruppe wurde in Gleichung (4.3) um einen neutralen Ausdruck erweitert. Dort, wo $p(y, z) > 0$ gilt, ist der erwartete Verlust fortan definiert. Hierdurch lässt sich der Erwartungswert nun in Bezug auf die Verteilung aller Daten der gesamten Stichprobe angeben. Es wird angenommen, dass die entsprechenden Dichten existieren. In Gleichung (4.4) wird der Zusam-

menhang der beiden Erwartungswerte deutlich: Der erwartete Verlust der Untergruppe lässt sich minimieren, indem der erwartete Verlust bezüglich aller Daten gewichtet mit dem Dichtequotienten aus Gleichung (4.2) ermittelt wird. Bickel (2009) bezeichnet diese Gewichte auch als *rescaling weights*, da die Verteilung aller Daten an die Verteilung der Untergruppe angeglichen wird. Hängt die Verteilung nicht von der Untergruppe ab, so ist $\omega_u = 1$ für alle (y, z).

$\omega_u(y, z)$ lässt sich im Allgemeinen aufgrund der hohen Dimension der Kovariablen z nicht ohne Weiteres schätzen. Über das Bayestheorem lässt sich der Zähler aus Gleichung (4.2) umformen zu:

$$p(y, z|u) = \frac{p(y, z, u)}{p(u)} = \frac{p(y, z) \cdot p(u|y, z)}{p(u)} \tag{4.5}$$

Dies setzt voraus, dass die Wahrscheinlichkeit für das Auftreten einer Untergruppe stets größer ist als Null ($p(u) > 0$). Mit (4.2) und (4.5) lässt sich ω_u wie folgt darstellen (vgl. Bickel (2009)):

$$\omega_u(y, z) = \frac{p(u|y, z)}{p(u)} \tag{4.6}$$

In Gleichung (4.6) ist nun die Abhängigkeit umgekehrt worden. Anstelle der Wahrscheinlichkeit der Daten gegeben einer Untergruppe bleibt die Wahrscheinlichkeit der Untergruppe bedingt auf die Daten zu bestimmen. $p(y, z)$ wurde im Nenner ebenfalls ersetzt durch die Wahrscheinlichkeit $p(u)$. Beide Wahrscheinlichkeiten können aus den beobachteten Daten geschätzt werden. $\hat{p}(u)$ ergibt sich aus der relativen Häufigkeit der jeweiligen Untergruppe u. Die bedingte Wahrscheinlichkeit $p(u|y, z)$ wird hier mit Hilfe einer logistischen Regression (siehe Abschnitt 4.1.4) geschätzt.

4.1.4 Logistische Regression zur Schätzung der Stichprobengewichte

In diesem Abschnitt wird das Vorgehen zur Schätzung der Gewichte $\omega_u(y, z) = \frac{p(u|y,z)}{p(u)}$ erläutert. Wie im vorangegangenen Abschnitt 4.1.3 beschrieben, wird $p(u)$ durch die relative Häufigkeit der Untergruppe u in den Daten geschätzt. Somit bleibt der Zähler $p(u|y, z)$ zu bestimmen. Diese Wahrscheinlichkeit wird hier über eine regularisierte multinomiale logistische Regression nach Zhu und Hastie (2004) modelliert.

Sei zunächst U eine diskrete Zufallsvariable mit möglichen Ausprägungen $\{1, \ldots, G\}$. Zur Vereinfachung der Schreibweise werden die Kovariablen zu $x = (y, z)$ zusammengefasst. Damit hat das Modell der multinomialen logistischen Regression die folgende Form (siehe Hastie u. a., 2009, Seite 119):

$$\log \frac{p(U=1|x)}{p(U=G|x)} = \beta_{10} + \beta_1' x$$
$$\log \frac{p(U=2|x)}{p(U=G|x)} = \beta_{20} + \beta_2' x$$
$$\vdots$$
$$\log \frac{p(U=G-1|x)}{p(U=G|x)} = \beta_{(G-1)0} + \beta_{G-1}' x$$

Die jeweiligen logarithmierten Brüche werden auch *log-odds* oder *logit-Transformation* genannt. Unabhängig davon, welche der G Ausprägungen im Nenner verwendet wird, resultieren dieselben Schätzer der Regressionsparameter.

Im Folgenden wird die von Zhu und Hastie (2004) vorgeschlagene symmetrische Darstellungsform

$$p(U=u|x) = \frac{\exp\left(\beta_{u0} + \beta_u' x\right)}{\sum_{g=1}^{G} \exp\left(\beta_{g0} + \beta_g' x\right)}$$

gewählt (vgl. Hastie u. a., 2009, Seite 657). Damit dieses Modell eindeutig geschätzt werden kann, ist die Nebenbedingung

$$\sum_{g=1}^{G} \hat{\beta}_{gk} = 0,\ k = 1, \ldots, p \tag{4.7}$$

erforderlich. Jeder Regressionskoeffizient β_g, $g = 1, \ldots, G$, ist ein p-dimensionaler Vektor.

Setze $\theta = \{\beta_{10}, \beta_1', \ldots, \beta_{G0}, \beta_G'\}$. Weiter sei $u_j \in \{1, \ldots, G\}$ die Untergruppe der j-ten Beobachtung. Dann kann das entsprechende Maximierungsproblem der regularisierten Log-Likelihood wie folgt formuliert werden (vgl. Hastie u. a., 2009, Seite 661):

$$\hat{\theta}_\lambda = \text{argmax}_\theta \left(\frac{1}{N} \sum_{j=1}^{N} \log p(U = u_j|x_j) - \lambda \sum_{g=1}^{G} \sum_{k=1}^{p} |\beta_{gk}|^\alpha \right),\ \lambda, \alpha > 0 \tag{4.8}$$

Friedman u. a. (2010) haben gezeigt, dass dieses Maximierungsproblem der Nebenbedingung aus Gleichung (4.7) genügt (siehe Friedman u. a., 2010, Theorem 1, Seite 11). $\alpha = 1$ entspricht der Lasso Methode. $\alpha = 2$ resultiert in einer Ridge-Regression. Die Parameter θ werden wie in Abschnitt 3.2.2 erläutert durch ein zyklisches Koordinatenabstiegsverfahren (siehe Hastie u. a., 2009, Seite 92 f.) geschätzt. Der Regularisierungsparameter λ wird durch eine 10-fache Kreuzvalidierung bestimmt. Der entsprechende Algorithmus ist ebenfalls im R-Paket *glmnet* implementiert und wird ausführlich von Friedman u. a. (2010) beschrieben.

Bewertung der geschätzten Modelle

Um die Qualität der Modellschätzung bzw. die Genauigkeit der geschätzten Wahrscheinlichkeiten $\hat{p}(U = u)$ zu bewerten, können die beiden folgenden Maßzahlen herangezogen werden. Die erste Maßzahl beschreibt die Exaktheit der durch

$$\hat{u}_j = \max_g \left(\hat{p}(U = g|x_j)\right),\ g \in \{1, \ldots, G\}$$

geschätzten Untergruppenzugehörigkeit. Dabei wird jedem Patienten j diejenige Untergruppe unterstellt, deren geschätzte Wahrscheinlichkeit für diese Gruppe maximal ist. Hier gehen somit alle Regressionskoeffizienten in die Entscheidung ein. Die sogenannte *accuracy* (zu Deutsch Exaktheit), hier mit `ACC` bezeichnet, vergleicht die geschätzte Untergruppe $\hat{u}$ mit der wahren Gruppenzugehörigkeit für jeden Patienten:

$$\texttt{ACC} = \frac{1}{n} \sum_{j=1}^{n} \mathcal{I}(u_j = \hat{u}_j) \tag{4.9}$$

Ein Referenzwert für die `ACC` kann über die Bayes-Regel ermittelt werden. Hierfür wird allen Patienten gerade diejenige Untergruppe zugeordnet, die in den Daten am häufigsten vertreten ist, d.h. es wird $\hat{u} = \max_g h(g)$ gesetzt, wobei $h(g)$ die relative Häufigkeit der Untergruppe g ist. Die `ACC` vergleicht nicht separat die Qualität der geschätzten Wahrscheinlichkeiten $\hat{p}(U = g)$ pro Untergruppe g sondern gibt einen einzelnen Wert für alle Untergruppen gemeinsam.

Die Vorhersage kann für jede Untergruppe einzeln wie folgt bewertet werden. Seien zunächst die folgenden beiden Kennzahlen definiert (vgl. Hastie u. a., 2009, Seite 314 ff.):

> **Sensitivität**: Wahrscheinlichkeit einen Patienten zur Untergruppe u zuzuordnen, wenn er tatsächlich zu dieser Untergruppe gehört
>
> **Spezifität**: Wahrscheinlichkeit einen Patienten nicht zur Untergruppe u zuzuordnen, wenn er tatsächlich nicht zu dieser Untergruppe gehört

Die Entscheidung einem Patienten eine bestimmte Untergruppe g zuzuordnen wird von einem Cutoff c abhängig gemacht:

$$\hat{u}_j = g, \text{ falls } \quad \hat{p}(U = g|x_j) > c$$

Die Wahrscheinlichkeit $\hat{p}(U = g|x_j)$ hängt nur von den Modellparametern $\hat{\beta}_g$ der Untergruppe g ab. Jede Untergruppe g wird demnach einzeln betrachtet. Pro Untergruppe wird die sogenannte *receiver operating characteristic curve* (kurz: ROC) betrachtet (siehe Fawcett, 2006). Diese Kurve zeigt die Trennschärfe auf, d.h. den Kompromiss zwischen der Sensitivität und der Spezifität in Abhängigkeit des Cutoffs c. Die Trennschärfe wird bewertet mit der Fläche unter dieser Kurve. Diese Maßzahl wird als `AUC` (Area under the curve) bezeichnet. Eine perfekte Trennschärfe resultiert in einer Fläche von Eins ($\texttt{AUC} = 1$). Dies bedeutet, es existiert ein Cutoff, so dass die Sensitivität sowie die Spezifität jeweils Eins ist. Gilt hingegen für alle Cutoffs Sensitivität $= 1 -$ Spezifität, ist

keine Trennschärfe gegeben und es gilt `AUC` = 0,5. Im Gegensatz zu der `ACC` können mit der `AUC` die Modelle der jeweiligen Untergruppen einzeln bewertet werden. So ist es beispielsweise möglich, dass ein logistisches Modell eine gute Trennschärfe für eine bestimmte Untergruppe aufweist, aber nicht zu den restlichen Untergruppen passt.

In diesem Abschnitt wurden sowohl das Verfahren zur Schätzung der Gewichte ω als auch Kenngrößen zur Beurteilung der Qualität dieser Schätzer beschrieben. Beides ist in der R-Funktion `EstimateWeights` (siehe Anhang B) implementiert. Die Funktion gibt unter anderem die entweder durch Lasso oder Ridge Regression auf der Trainingsmenge geschätzten Gewichte pro Untergruppe aus. Des Weiteren werden automatisch die `ACC` und `AUC` auf der Testmenge ermittelt. Die `AUC` wird mit dem R-Paket *ROCR* von Sing u. a. (2005) berechnet. Dieses Paket liegt aktuell in Version 1.0.4 vor (Sing u. a., 2012).

4.1.5 Gewichtete Schätzung der Überlebensfunktion

Wird nach einem Überlebenszeitmodell für die Patienten in einer bestimmten Untergruppe $g \in \{1, \ldots, G\}$ gesucht, können, wie in Abschnitt 4.1.3 gezeigt, hierzu alle Patienten aus den erhobenen Daten genutzt werden, wenn diese geeignet gewichtet werden (siehe Gleichung (4.4)).

Seien erneut die folgenden Daten von $j = 1, \ldots, n$ Patienten gegeben. Wie in Kapitel 3 liegt zu jedem Patienten j die beobachtete Risikozeit $\tilde{t}_j$ und der entsprechende Zensierungsindikator δ_j vor. Weiter seien die Kovariablen $z_j = z_{j1}, \ldots, z_{jp}$ beobachtet. Zudem ist für jeden Patienten die Realisation u_j der Zufallsvariablen U bekannt, die die Gruppenzugehörigkeit zu einer Untergruppe angibt. Für die Patienten der Untergruppe mit $u_j = g$ soll nun die Überlebensfunktion $S(t|Z = z_j)$ geschätzt werden. Diese wird aus allen Patienten geschätzt, wobei jeder Patient mit dem Gewicht $\hat{\omega}_{gj}$ zu dieser Schätzung beiträgt. Diese Gewichte seien entweder über das heuristische Verfahren aus Abschnitt 4.1.2 oder über die in Abschnitt 4.1.4 beschriebene multinomiale logistische Regression gegeben.

Die Überlebensfunktion wird in zwei Schritten geschätzt. Im ersten Schritt wird zunächst ein Überlebenszeitmodell nach Cox (siehe Abschnitt 3.1.2) angepasst. Anschließend wird im zweiten Schritt aus den ermittelten Modellparametern die Überlebensfunktion der Patienten geschätzt (vgl. Abschnitt 3.1.3). Damit sind nacheinander die folgenden beiden Likelihood-Funktionen zu maximieren:

1. Kreuzvalidierte regularisierte partielle Log-Likelihood (`CVPL`):

   Diese Likelihood wird im ersten Schritt maximiert, um die Koeffizienten im Cox-Modell zu schätzen

2. Likelihood der Daten bei fixiertem Parametervektor $\hat{\beta}$:

   Im zweiten Schritt wird das Maximum der Likelihood bzgl. der Baseline-Rate ermittelt. Dies führt wie in Abschnitt 3.1.3 gezeigt zu einem Schätzer der Überlebensfunktion

Die Überlebensfunktion wird gewichtetet geschätzt, indem nun jeweils die einzelnen Terme der beiden Likelihoods mit $\hat{\omega}_{gj}$ gewichtet werden. Damit wird die Überlebenszeitfunktion für Patienten der Untergruppe g geschätzt. Daraus ergibt sich für Schritt 1 die *gewichtete partielle Likelihood* (vgl. (3.3)) zu:

$$L^{\hat{\omega}_g}(\beta) = \prod_{i=1}^{D} \left[\frac{\exp\left(\sum_{k=1}^{p} \beta_k Z_{(i)k}\right)}{\sum_{j \in R(t_{(i)})} \exp\left(\sum_{k=1}^{p} \beta_k Z_{jk}\right)} \right]^{\hat{\omega}_{g(i)}} \tag{4.10}$$

Dabei ist $\hat{\omega}_{g(i)}$ das Gewicht desjenigen Patienten, welcher zum Zeitpunkt $t_{(i)}$ ausfällt. Aus der gewichteten partiellen Likelihood kann ohne Weiteres die `CVPL` ermittelt werden. Diese lässt sich dann in einem der regularisierten Modellbildungsverfahren aus Abschnitt 3.2 verwenden.

Im zweiten Schritt wird die gewichtete Version der vollständigen Likelihood der Daten bei festem, sich aus Schritt 1 ergebenem $\hat{\beta}$ herangezogen. Die gewichtete vollständige Likelihood unter der Annahme proportionaler Hazards (vgl. (3.5)) hat die folgende Gestalt:

$$\begin{aligned} L^{\hat{\omega}_{gj}} &= \prod_{j=1}^{n} \left[h_0(\tilde{t}_j)^{\delta_j} \left[\exp(\beta' Z_j)\right]^{\delta_j} \exp\left(-H_0(\tilde{t}_j)\exp(\beta' Z_j)\right) \right]^{\hat{\omega}_{gj}} \\ &= \prod_{j=1}^{n} h_0(\tilde{t}_j)^{\hat{\omega}_{gj}\delta_j} \left[\exp(\beta' Z_j)\right]^{\hat{\omega}_{gj}\delta_j} \exp\left(-H_0(\tilde{t}_j)\hat{\omega}_{gj}\exp(\beta' Z_j)\right) \end{aligned}$$

Der Maximum-Likelihood-Schätzer für $h_0(t)$ ergibt sich analog zu Abschnitt 3.1.3:

$$\begin{aligned} L_{\hat{\beta}}^{\hat{\omega}_{gj}}(h_0(t)) &= \left[\prod_{i=1}^{D} h_0(t_{(i)})^{\hat{\omega}_{g(i)}} \exp(\hat{\beta}' Z_{(i)})^{\hat{\omega}_{g(i)}} \right] \exp\left[-\sum_{j=1}^{n} H_0(\tilde{t}_j)\hat{\omega}_{gj}\exp(\hat{\beta}' Z_j) \right] \\ &= \left[\prod_{i=1}^{D} h_0(t_{(i)})^{\hat{\omega}_{g(i)}} \exp(\hat{\beta}' Z_{(i)})^{\hat{\omega}_{g(i)}} \right] \exp\left[-\sum_{j=1}^{n} \sum_{t_{(i)} \le \tilde{t}_j} h_0(t_{(i)})\hat{\omega}_{gj}\exp(\hat{\beta}' Z_j) \right] \\ &= \left[\prod_{i=1}^{D} h_0(t_{(i)})^{\hat{\omega}_{g(i)}} \exp(\hat{\beta}' Z_{(i)})^{\hat{\omega}_{g(i)}} \right] \exp\left[-\sum_{i=1}^{D} \sum_{j \in R(t_{(i)})} h_0(t_{(i)})\hat{\omega}_{gj}\exp(\hat{\beta}' Z_j) \right] \\ &= \prod_{i=1}^{D} h_0(t_{(i)})^{\hat{\omega}_{g(i)}} \exp(\hat{\beta}' Z_{(i)})^{\hat{\omega}_{g(i)}} \exp\left[-\sum_{j \in R(t_{(i)})} h_0(t_{(i)})\hat{\omega}_{gj}\exp(\hat{\beta}' Z_j) \right] \\ &= \prod_{i=1}^{D} h_0(t_{(i)})^{\hat{\omega}_{g(i)}} \exp(\hat{\beta}' Z_{(i)})^{\hat{\omega}_{g(i)}} \exp\left[-h_0(t_{(i)}) \sum_{j \in R(t_{(i)})} \hat{\omega}_{gj}\exp(\hat{\beta}' Z_j) \right] \\ &\propto \prod_{i=1}^{D} h_0(t_{(i)})^{\hat{\omega}_{g(i)}} \exp\left[-h_0(t_{(i)}) \sum_{j \in R(t_{(i)})} \hat{\omega}_{gj}\exp(\hat{\beta}' Z_j) \right] \end{aligned} \tag{4.11}$$

Sei $\eta = \sum_{j \in R(t_{(i)})} \hat{\omega}_{gj} \exp(\hat{\beta}' Z_j)$. Dann ist die Ableitung der einzelnen Faktoren aus (4.11) gegeben durch:

$$\frac{\partial \left(h_0(t_{(i)})^{\hat{\omega}_{g(i)}} \exp\left[-\eta \cdot h_0(t_{(i)}) \right] \right)}{\partial h_0(t_{(i)})} = \left(\hat{\omega}_{g(i)} - \eta \cdot h_0(t_{(i)}) \right) h_0(t_{(i)})^{\hat{\omega}_{g(i)} - 1} \exp\left(-\eta \cdot h_0(t_{(i)}) \right)$$

Damit ergibt sich der ML-Schätzer von (4.11) an den Zeitpunkten $t_{(1)}, \ldots, t_{(D)}$ zu:

$$\hat{h}_0(t_{(i)}) = \frac{\hat{\omega}_{g(i)}}{\eta} = \frac{\hat{\omega}_{g(i)}}{\sum_{j \in R(t_{(i)})} \hat{\omega}_{gj} \exp(\hat{\beta}' Z_j)}$$

Mit $\hat{H}_0(t) = \sum_{t_{(i)} \leq t} \hat{h}_0(t_{(i)})$ berechnet sich daraus analog zu Abschnitt 3.1.3 ein Schätzer der Baseline-Rate, wenn keine Bindungen in den Daten auftreten. Im Fall von auftretenden Bindungen der Ereigniszeiten, lässt sich der Schätzer ähnlich wie in Gleichung (3.9) erweitern. Damit ist der *gewichtete Breslow-Schätzer* definiert als

$$\hat{H}_0^{\hat{\omega}_g}(t) = \sum_{t_{(i)} \leq t} \left(\frac{\sum_{j:\, t_j = t_{(i)}} \hat{\omega}_{gj}}{\sum_{j \in R(t_{(i)})} \hat{\omega}_{gj} \exp\left(\sum_{k=1}^{p} \hat{\beta}_k Z_{jk} \right)} \right) \tag{4.12}$$

Der Zähler in (4.12) summiert jeweils die geschätzten Gewichte $\hat{\omega}_{gj}$ derer Patienten auf, denen gleichzeitig zum Zeitpunkt $t_{(i)}$ das betreffende Ereignis widerfährt. Mit dem gewichteten Breslow-Schätzer kann abschließend die Überlebensfunktion basierend auf dem Modell für die Untergruppe ermittelt werden (siehe hierzu Abschnitt 3.1.3).

Zur Modellbildung mittels der gewichteten partiellen Likelihood (Schritt 1) kann sowohl im R-Paket *CoxBoost* als auch im Paket *glmnet* die `weights=` Option genutzt werden. Für den zweiten Schritt der gewichteten Schätzung der Überlebensfunktion wurde in der Funktion `PREDmat` die hier vorgestellte gewichtete Version des ML-Schätzers der Baseline-Rate (siehe Gleichung (4.12)) implementiert.

4.1.6 Alternativer Ansatz über hierarchische Bayes Modelle

Um zu einer gewichteten Schätzung der Überlebenszeit zu gelangen, wird im ersten Schritt in Abschnitt 4.1.5 die gewichtete Likelihood (4.10) bezüglich β maximiert. Die entsprechenden Gewichte $\hat{\omega}$ werden dazu im Vorhinein wie in Abschnitt 4.1.4 gezeigt über eine logistische Regression geschätzt. Die Schätzung der Parameter ω und β kann alternativ jeweils über ein hierarchisches Bayes Modell (siehe Gelman u. a., 2003) erfolgen. Für beide Parameter werden im Folgenden Anregungen zur Schätzung über ein Bayes Modell gegeben.

Bezeichne erneut $y = (\tilde{t}, \delta)$ als die Zielvariable sowie die beobachteten Kovariablen mit z. Die a-posteriori Wahrscheinlichkeit der Stichprobengewichte ω kann wie folgt

angegeben werden (vgl. Bogojeska und Lengauer, 2012):

$$p(\omega, \phi_\omega | y, z) \propto p(y, z|\omega) \cdot p(\omega|\phi_\omega) \cdot p(\phi_\omega) \tag{4.13}$$

Dabei beschreibt der Hyperparameter ϕ_ω den Mittelwert der Stichprobengewichte. Bogojeska und Lengauer (2012) schlagen mit $\omega \sim \mathcal{N}(\phi_\omega, \sigma^2)$ und $\phi_\omega \sim \mathcal{N}(0, \sigma^2)$ normalverteilte Prior und Hyperprior vor. σ^2 ist im Vorfeld zu wählen. Hier kann beispielsweise mit $\sigma^2 = 1$ gestartet werden.

Mit dem Maximum-a-posteriori-Schätzer (MAP) $\hat{\omega}$ des Modells aus (4.13) kann anschließend analog zu Abschnitt 4.1.5 die gewichtete partielle Likelihood $LL^{\hat{\omega}_g}$ (4.10) aufgestellt werden. Anstelle des ML-Schätzers lässt sich der Parameter β ebenfalls über ein hierarchisches Bayes Modell ermitteln. Die entsprechende a-posteriori Wahrscheinlichkeit kann wie folgt formuliert werden:

$$p(\beta, \tau_\beta | y, z) \propto p(y, z|\beta, \tau_\beta) \cdot p(\beta|\tau_\beta) \cdot p(\tau_\beta) \tag{4.14}$$

Der Hyperparameter τ_β beschreibt die Varianz der Verteilung des Regressionskoeffizienten. Mit $\beta \sim \mathcal{N}(0, \tau_\beta^2)$ kann ein normalverteilter Prior angenommen werden. Die Varianz kann mit $\tau_\beta^2 \sim \Gamma(a, b)$ über eine Gammaverteilung modelliert werden. In diesem Fall werden häufig die Startwerte $a = b = 0{,}01$ herangezogen (Ibrahim u. a., 2001). In (4.14) wird die partielle gewichtete Likelihood gewählt, so dass auf die zusätzliche Modellierung der Baseline-Rate verzichtet werden kann (vgl. Sinha u. a., 2003). Wird die volle Likelihood zugrunde gelegt, so bedarf es eines weiteren Prior für die Baseline-Rate $h_0(t)$. Statt eines vollständigen Bayes Modells kann alternativ ein empirisches Modell aufgestellt werden, in dem die Hyperparameter entsprechend fix gewählt werden. Für weitere Informationen zur Bayesianischen Modellierung von Überlebenszeiten sei hier auf das Werk von Ibrahim u. a. (2001) verwiesen.

Ein wichtiger Aspekt bei der Analyse hochdimensionaler Daten ist die Variablenselektion. Im Kontext der Bayesschen Überlebenszeitanalyse diskutieren Ibrahim u. a. (1999) geeignete Verfahren zur Selektion bestimmter wichtiger Faktoren (Variablen), die in Zusammenhang mit der Überlebenszeit stehen. Methoden, die bereits für das Lineare Modell vorgeschlagen wurden (siehe George und McCulloch, 1993), können gegebenenfalls übertragen werden (Ibrahim u. a., 1999, Seite 17).

4.2 Bildung und Evaluierung der Überlebenszeitmodelle in den Untergruppen

Das in Abschnitt 4.1 vorgeschlagene Prinzip der gewichteten Modellbildung hängt von der Schätzung der Stichprobengewichte ω ab. In dem hypothetischen Fall diese Gewichte exakt zu kennen, generiert dieses Verfahren die kleinsten Vorhersagefehler für die betrachtete Untergruppe. Da diese Gewichte jedoch geschätzt werden müssen, kann in Situationen, in denen die Stichprobengewichte nicht sinnvoll geschätzt werden können, die Anwendung der gewichteten Likelihood zu keinem adäquaten Überlebenszeitmodell

für die entsprechende Untergruppe führen. Ferner kann ein nur auf der Untergruppe geschätztes Modell oder ein Modell auf Basis der Patienten aus allen Untergruppen bessere Vorhersagen generieren, da diese Modelle nicht auf die Zuverlässigkeit der Schätzer für die Stichprobengewichte angewiesen sind. Welches Modell am geeignetsten ist, hängt zudem von der Homogenität der Daten ab.

Abschnitt 4.2.1 stellt den Versuchsaufbau vor, um die soeben skizzierten Modelle miteinander zu vergleichen und um Rückschlüsse auf heterogene Untergruppen ziehen zu können. Alle Modelle werden jeweils auf Trainingsdaten angepasst und anschließend auf unabhängigen Testdaten der Untergruppe evaluiert. Die zu untersuchenden Untergruppenvariablen werden dabei durch klinische Variablen definiert. Als Beispiel kann die Histologie eines Tumors genannt werden. Alle möglichen Ausprägungen dieser Variable stellen dann eine Untergruppe dar. Eine Untergruppe der Lungenkarzinome bilden beispielsweise die Adenokarzinome.

In Abschnitt 4.2.2 wird der Vorhersagefehler der in 4.2.1 genauer erläuterten Modelle auf künstlich erzeugten Daten analysiert. Dabei wird aufgezeigt, dass das Modellbildungsverfahren aus Abschnitt 4.1 auf Basis der gewichteten Likelihood in vielen Szenarien vergleichbar gute Vorhersagen erzielt. Es existieren darüber hinaus Szenarien, in denen das gewichtete Verfahren den anderen Modellen überlegen ist.

4.2.1 Training/Test Szenario

Dieser Abschnitt beschreibt den Versuchsablauf zur Schätzung und Evaluierung der zu untersuchenden Modellbildungsverfahren. Der Fokus liegt dabei auf dem Verfahren mit einer Gewichtung der Likelihood, wie es im vorangegangenen Unterkapitel 4.1 ausführlich beschrieben wurde. Im Folgenden wird dieses Verfahren als `WEIGHTED` Modell bezeichnet. Dieses Modell soll mit dem `SUBGROUP` Modell sowie dem `ALL` Modell verglichen werden, welche hier konkret beschrieben werden.

Über das `WEIGHTED` Modell sollen die Überlebenszeiten der Patienten in einer bestimmten Untergruppe $U = g$ geschätzt werden. Dies rechtfertigt sofort die Betrachtung der beiden zusätzlichen Modelle `SUBGROUP` und `ALL`. Zur Schätzung des `SUBGROUP` Modells werden nur die Patienten aus der Untergruppe g verwendet. Alle übrigen Patienten gehen nicht in die Schätzung mit ein, was zu einer Reduzierung der Fallzahl führt. Dieses Modell ist besonders geeignet, wenn ein bestimmter Zusammenhang zwischen einigen Kovariablen und der Überlebenszeit nicht in der restlichen Kohorte Bestand hat und keine Information über die Verteilung der Daten innerhalb der Untergruppen im Vergleich zur Gesamtkohorte vorliegt. Im `ALL` Modell hingegen werden alle Patienten aus den Daten zur Schätzung des Modells herangezogen. Hierdurch wird die maximal zur Verfügung stehende Fallzahl genutzt. Ist ein Zusammenhang zwischen Kovariablen und Überlebenszeit unabhängig über alle Untergruppen hinweg gegeben, so wird der Schätzer der Überlebenszeit aus dem `ALL` Modell die kleinste Varianz aufweisen und die genausten Vorhersagen liefern. Abbildung 4.1 skizziert die Verwendung der unterschiedlich großen Stichproben zur Schätzung der Modelle.

Wie in Kapitel 2 beschrieben, enthalten die Daten aus Köln und Uppsala neben den genetischen Variablen jeweils fünf klinische Variablen. Die klinischen Variablen werden

zur Definition der zu untersuchenden Untergruppen herangezogen. Der hier vorgestellte Versuchsaufbau sieht vor, dass iterativ jede klinische Variable stellvertretend für die Untergruppenvariable U ausgewählt wird und jede mögliche Ausprägung $U = u$ dieser Variable eine Untergruppe darstellt. Für diese Untergruppe wird daraufhin jedes der drei im vorangegangenen Absatz diskutierten Modelle angepasst, wobei die Gewichte des `WEIGHTED` Modells entweder wie in Abschnitt 4.1.4 erläutert über eine logistische Regression geschätzt werden oder a priori wie in Abschnitt 4.1.2 gezeigt bestimmt werden.

Die Modelle werden mit Hilfe der Algorithmen zur regularisierten Modellbildung aus Abschnitt 3.2 bestimmt. In wie weit sich dabei die Resultate zwischen diesen Algorithmen auf realen Daten unterscheiden ist Inhalt des Abschnitts 5.3.1.

Neben den Modellbildungsverfahren wird auf folgende Weise ebenso die Auswahl der Kovariablen untersucht. Jedes Modell wird zunächst ausschließlich mit Hilfe der klinischen Variablen angepasst, wobei eine der klinischen Variablen die entsprechende Untergruppe definiert. Folglich gehen nur die übrigen klinischen Variablen als Kovariablen in die Modelle `SUBGROUP` und `WEIGHTED` ein. In das `ALL` Modell gehen alle klinischen Variablen ein, um für einen fairen Vergleich der drei Modelle einen möglichen Einfluss der Untergruppenvariablen zu berücksichtigen. Kategorielle klinische Variablen werden entsprechend durch Dummy-Variablen ersetzt.

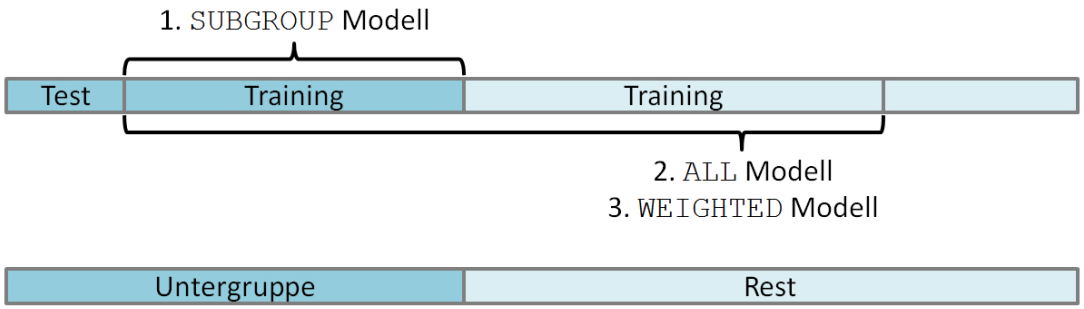


Abbildung 4.1: Schematische Darstellung des Evaluierungsprozesses der Modelle für eine bestimmte Untergruppe $U = g$. Für alle restlichen Untergruppen $U = u$, $u \neq g$ (enthalten in der Restmenge) wird auf derselben Datenaufsplittung simultan verfahren. Dadurch ergibt sich je eine Testmenge pro Untergruppe. Die Abbildung skizziert lediglich eine Testmenge für eine bestimmte Untergruppe.

Die Modelle ausschließlich mit den klinischen Variablen können als Referenzmodelle angesehen werden. Modelle mit genetischen Kovariablen sollten im Idealfall die Vorhersage dieser Referenzmodelle verbessern. In dieser Arbeit werden zwei mögliche Konstellationen betrachtet. Zum einen werden als Kovariablen ausschließlich die genetischen Variablen verwendet. Dies wird den Einfluss der genetischen Information auf die Überlebenszeit unabhängig von den klinischen Variablen zeigen. Diese Variante ist jedoch

anfällig gegenüber kreuzenden Effekten mit den klinischen Variablen. Zum anderen werden die Modelle mit beiden Typen von Kovariablen (klinisch und genetisch) gebildet. Insgesamt ergeben sich so drei mögliche Kovariablenmengen: klinische, genetische sowie klinische und genetische Variablen.

Alle stetigen Kovariablen werden über den Mittelwert und die Standardabweichung standardisiert. Dazu zählen insbesondere die genetischen Variablen. Die Dummy-Variablen der kategoriellen klinischen Variablen werden nicht standardisiert.

Zusammenfassend lassen sich die betrachteten Modelle wie folgt beschreiben:

- `SUBGROUP` Modell: Nur Patienten der zu untersuchenden Untergruppe g werden zur Modellbildung genutzt. Sind klinische Variablen erlaubt, so wird die entsprechende Untergruppenvariable nicht in das Modell aufgenommen

- `WEIGHTED` Modell: Alle Patienten werden zur Modellbildung genutzt, wobei zunächst deren Gewichte $\hat{\omega}_j$, siehe Gleichung (4.6), geschätzt bzw. bestimmt werden. Daraufhin werden im Modellbildungsprozess jeweils die gewichteten Likelihoods aus Abschnitt 4.1.5 verwendet. Sind klinische Variablen erlaubt, so wird die entsprechende Untergruppenvariable ebenfalls nicht in das Modell aufgenommen

- `ALL` Modell: Alle Patienten werden zur Modellbildung genutzt. Sind klinische Variablen erlaubt, so wird die entsprechende Untergruppenvariable mit in das Modell aufgenommen

Setzen sich die Kovariablen aus klinischen und genetischen Variablen zusammen, so geht stets die absolute Größe der Regressionskoeffizienten der klinischen Variablen bzw. deren Dummy-Variablen nicht in den Regularisierungsterm der Log-Likelihood (3.11) ein. Der Regularisierungsparameter λ_k dieser Variablen wird auf Null gesetzt. Damit sind die klinischen Variablen in jedem Fall im entsprechenden Modell enthalten. Werden nur die klinischen Variablen zur Modellbildung genutzt, wird λ_k nicht auf Null gesetzt.

Der Kaplan-Meier-Schätzer aus Gleichung (3.1) wird ebenfalls als Referenzschätzer für die Überlebenszeit betrachtet. Dieser wird sowohl auf der gesamten Kohorte als auch auf der interessierenden Untergruppe ermittelt. Der Vergleich mit dem Kaplan-Meier-Schätzer ist deutlich sinnvoller in Bezug auf die Bewertung der untersuchten Modelle als die Referenzgrößen 0,25 und 0,33, da letztere für sehr frühe sowie sehr späte Zeitpunkte keine realistischen Bezugsgrößen darstellen (siehe Kapitel 3.3).

Wie in Abbildung 4.1 skizziert, werden sämtliche Modelle auf einer zuvor zufällig bestimmten Trainingsmenge geschätzt. Der Vorhersagefehler in Form des Brier Scores aus Gleichung (3.15) wird anschließend auf der Testmenge der entsprechenden Untergruppe g ermittelt. Dies wird für dieselbe Aufsplittung der Daten jeweils für jede Untergruppe $g = 1, \ldots, G$ durchgeführt. Da aufgrund einer einzelne Aufteilung der Daten in Trainings- und Testmenge nichts über die Variabilität der Schätzer ausgesagt werden kann und zudem die Vorhersagefehler zu sehr von der Zufälligkeit der Aufsplittung der Daten beeinflusst werden (siehe z. B. Bøvelstad u. a. (2007)), werden hier, ähnlich wie in Kammers u. a. (2011), mehrere Trainings- und Testmengen betrachtet. Dabei werden die Daten wiederholt zufällig aufgeteilt. Auf jeder Testmenge wird der Vorhersagefehler

ermittelt. Anschließend werden die einzelnen Vorhersagefehler über den mittleren Brier Score (3.17) aus Abschnitt 3.3.2 zusammengefasst. Gerade bei der Betrachtung von Untergruppen mit niedriger Häufigkeit kommt es vor, dass für ein Modell keine Kovariablen ausgewählt werden. Dies ist beispielsweise bei den Algorithmen CoxBoost oder Lasso möglich. In diesem Fall resultiert ein Null-Modell als geschätztes Modell. Damit der mittlere Brier Score dennoch sinnvoll geschätzt werden kann, wird in diesem Fall der Kaplan-Meier-Schätzer zugrunde gelegt.

Zur Umsetzung des Sampling-Aufbaus wurden die folgenden R-Pakete zur Hilfe gezogen: Die Funktion `addProblem` aus dem Paket *BatchExperiments* spezifiziert das Vorgehen zur Bestimmung der jeweiligen Trainings- und Testmengen. Die Option `dynamic=` verweist auf die dazu benötigte R-Funktion mit Namen `Subsample`. Eine Beschreibung dieser Funktion ist im Anhang B zu finden. Sie führt ein stratifiziertes Subsampling (Ziehen ohne Zurücklegen) durch, so dass das jeweilige relative Häufigkeitsverhältnis der Ausprägungen der Untergruppenvariablen U dem Häufigkeitsverhältnis aus den gesamten Daten entspricht. Dadurch ist jede Untergruppe g in jeder Stichprobe vertreten. Des Weiteren wird durch einen Seed in der Funktion `addProblem` erreicht, dass jedes der Modelle jeweils auf denselben Trainings- und Testdaten geschätzt bzw. evaluiert wird. Vor- oder Nachteile einer Methode durch bestimmte Zufallsprozesse beim Ziehen der B Bootstrapstichproben werden dadurch vermieden. Insgesamt werden $B = 400$ Stichproben erzeugt.

4.2.2 Exemplarische Auswertung anhand einfacher künstlicher Daten

In den folgenden beiden Beispielen werden jeweils drei Szenarien erzeugt, in denen die im vorangegangenen Abschnitt 4.2.1 erläuterten Modelle anhand eines Datenbeispiels miteinander verglichen werden. Dabei kann in jedem Szenario die Überlegenheit eines der Modelle gegenüber den anderen Modellen gezeigt werden, wobei dabei entweder das `SUBGROUP`, `WEIGHTED` oder `ALL` Modell den kleinsten Vorhersagefehler im Sinne des Brier Scores liefert.

Dazu werden für jedes Szenario aus den beiden Beispielen künstliche Daten erzeugt. Es wird der einfachste Fall mit nur einer beobachteten Kovariable $Z = z$ angenommen. Die Untergruppenvariable U ist hier binär ($G = 2$). Von Interesse ist die Untergruppe $U = g$. Ziel ist es ein Modell zu finden, welches den Vorhersagefehler für die Überlebenszeiten dieser Untergruppe minimiert. Das Komplement der Untergruppe g, die Restgruppe, wird hier mit $\bar{g}$ bezeichnet. Mit n_g bzw. $n_{\bar{g}}$ wird jeweils der Stichprobenumfang der Untergruppe bzw. der Stichprobenumfang der Restgruppe angegeben. Es werden nun künstlich erzeugte Zusammenhänge zwischen der Überlebenszeit t der Beobachtungen aus den beiden Gruppen g und $\bar{g}$ und der Kovariablen z simuliert. Dabei wird ein Cox Modell der Form (3.2) zugrunde gelegt. Die Werte der Kovariablen z werden aus einer $\mathcal{N}(\mu, \sigma)$-Verteilung gezogen, wobei die Kovariablen der Untergruppe einer $\mathcal{N}(2, 1)$-Verteilung und die Kovariablen der Restgruppe einer $\mathcal{N}(-2, 2)$-Verteilung folgen. Der

Einfachheit halber werden die Überlebenszeiten aus einer Exponentialverteilung mit Parameter λ_j, $j = 1, \ldots, (n_g + n_{\bar{g}})$ gezogen. Damit ergibt sich eine über die Zeit konstante Hazardrate $h(t|z_j) = \lambda_j$ (siehe Klein und Moeschberger, 2003, Seite 38). Sei nun das folgende Hazardratio `HR` definiert (vgl. Netzer und Rahnenführer, 2012):

$$\texttt{HR} = \frac{h(t|z+1)}{h(t|z)} = \exp(\beta)$$

Der Regressionskoeffizient β ist in diesem Fall eindimensional. Der soeben erläuterte Zusammenhang zwischen der Kovariablen z und den Überlebenszeiten wird im ersten Beispiel konkret über die Hazardratios $\texttt{HR}_g$ und $\texttt{HR}_{\bar{g}}$ a priori festgelegt. Dabei ist $\texttt{HR}_g$ bzw. $\texttt{HR}_{\bar{g}}$ das Hazardratio in der Untergruppe g bzw. in der Restgruppe $\bar{g}$. O.B.d.A. wird $h_0(t) \equiv 1$ unterstellt. Die Überlebenszeiten können daraufhin aus einer Exponentialverteilung mit Parameter $\lambda_j = \texttt{HR}^{z_j}$ gezogen werden (siehe Netzer und Rahnenführer, 2012, Seite 5). Zensierungen werden nicht erzeugt. Somit gilt $\tilde{t}_j = t_j$ für alle Beobachtungen. Die Daten in dem zweiten hier betrachteten Beispiel hängen zudem von einem weiteren Parameter q wie folgt ab. Die Beobachtungen der Restgruppe $\bar{g}$ setzen sich zu einem bestimmten Verhältnis q zusammen aus Beobachtungen, die auch hätten zu der Untergruppe g gezählt werden können und Beobachtungen, die tatsächlich den Eigenschaften der Restgruppe folgen. q beschreibt den prozentualen Anteil der Beobachtungen der Restgruppe, welche aus der Verteilung der Untergruppe gezogen wurden. Damit setzen die beiden betrachteten Beispiele aus zwei unterschiedlichen Blickwinkeln an:

- Beispiel 1 (*Bias-Varianz Kompromiss*): Der Zusammenhang zwischen Kovariablen und Überlebenszeit ist grundsätzlich in den Gruppen verschieden. Für $n_g \to \infty$ genügt das `SUBGROUP` Modell für eine optimale Schätzung. Ist n_g hingegen klein, können die Modelle `WEIGHTED` und `ALL` eventuell aufgrund einer kleineren Varianz bessere Vorhersagen generieren. Sollte der Unterschied der beiden Gruppen zu groß sein, wiegt der Bias diesen Vorteil wieder auf

- Beispiel 2 (*Heterogene Restgruppe*): Ein Anteil q der Beobachtungen aus der Restgruppe weist denselben Zusammenhang zwischen Kovariablen und Überlebenszeit auf wie die Beobachtungen der Untergruppe. Hier gilt ebenso, dass für einen hinreichend großen Stichprobenumfang n_g das `SUBGROUP` Modell genügt. Für kleinere Stichprobenumfänge können je nach Anteil q das `WEIGHTED` oder ALL Modell die Zusammenhänge in der Untergruppe besser erkennen

Im Folgenden werden für jedes Beispiel durch Variieren der Stichprobenumfänge, der Hazardratios und (in Beispiel 2) durch Verschieben des Anteils q jeweils drei Szenarien konstruiert, in denen eines der Modelle `SUBGROUP`, `WEIGHTED` bzw. `ALL` überlegen ist. Wie diese Parameter in den jeweiligen Szenarios festgelegt sind, ist in Tabelle 4.1 angegeben. In allen Szenarios wird ebenfalls der Brier Score des Kaplan-Meier-Schätzers ermittelt. Dieser dient als Referenzschätzer der Überlebenswahrscheinlichkeit. Alle Vorhersagefehler werden auf 1 000 separaten Testdaten der Untergruppe g ermittelt. Abbildung 4.2 zeigt die mittleren Brier Scores der betrachteten Modelle in den drei Szenarien beider

Tabelle 4.1: Überblick über die in den Szenarien der beiden Beispiele gewählten Stichprobenumfänge n_g und $n_{\bar{g}}$, die Hazardratios $\texttt{HR}_g$ und $\texttt{HR}_{\bar{g}}$ sowie den in Beispiel 2 gewählten Anteil q

Beispiel	Szenario	n_g	$\texttt{HR}_g$	$n_{\bar{g}}$	$\texttt{HR}_{\bar{g}}$	q
1	1.1	10	2,8	60	1,5	–
	1.2	6	2,8	80	1,5	–
	1.3	10	2,0	60	1,9	–
2	2.1	30	1,8	100	1,0	0,01
	2.2	10	1,8	100	1,0	0,30
	2.3	10	1,8	100	1,0	0,70

Beispiele. Der mittlere Brier Score des Kaplan-Meier-Schätzers ist aufgrund einer besseren Übersicht dabei nicht gezeichnet worden.

Szenario 1.1 bezeichnet das erste Szenario des ersten Beispiels. Die Hazardrate in der Untergruppe ist mit $\texttt{HR}_g = 2{,}8$ im Vergleich zur Hazardrate der Restgruppe ($\texttt{HR}_{\bar{g}} = 1{,}5$) deutlich höher. In diesem Fall genügt der Stichprobenumfang von $n_g = 10$, um mit dem `SUBGROUP` Modell einen kleineren Vorhersagefehler im Sinne des Brier Scores zu erzielen als mit dem `ALL` Modell ($n_{\bar{g}} = 60$). Das `WEIGHTED` Modell ist hier ebenfalls leicht besser als das `ALL` Modell (siehe Tabelle 4.2). Im zweiten Szenario 1.2 wurden lediglich die Stichprobengrößen geändert ($n_g = 6$, $n_{\bar{g}} = 80$). Die Vorhersagefehler der Modelle `SUBGROUP` und `ALL` sind dann ungleich schlechter. Das `WEIGHTED` Modell findet einen guten Kompromiss der beiden anderen Modelle (siehe Abbildung 4.2). Der integrierte Brier Score liegt 18 Prozentpunkte unterhalb dem des Kaplan-Meier-Schätzers (Tabelle 4.2). In Szenario 1.3 gleichen sich die Hazardratios der Gruppen annähernd, so dass das `ALL` Modell die kleinsten Vorhersagefehler macht ($\texttt{HR}_g = 2{,}0$ und $\texttt{HR}_{\bar{g}} = 1{,}9$).

Das zweite Beispiel behandelt im Gegensatz zu Beispiel 1 eine heterogene Restgruppe. q Prozent der Patienten aus der Restgruppe basieren auf derselben Verteilung und dem gleichen Zusammenhang von T und Z wie in der Untergruppe. In allen Szenarien gilt $\texttt{HR}_g = 1{,}8$ und $\texttt{HR}_{\bar{g}} = 1$. In den Szenarien wird überwiegend der Parameter q variiert (Tabelle 4.1). Szenario 2.1 bezeichnet analog das erste Szenario des zweiten Beispiels. Eine Beobachtung ($q = 1\%$) der Restgruppe ($n_{\bar{g}} = 100$) besitzt ebenfalls ein Hazardratio von 1,8. Alle weiteren Beobachtungen unterliegen keinem Zusammenhang der Überlebenszeit mit der Kovariablen. In diesem Szenario führt die Schätzung ausschließlich basierend auf der Untergruppe ($n_g = 30$) zum kleinsten Vorhersagefehler (siehe Abbildung 4.2). Das `WEIGHTED` Modell ist annähernd gleich gut. Der mittlere integrierte Brier Score ist um 10 Prozentpunkte statt um 12 Prozentpunkte unter dem des Kaplan-Meier-Schätzers (Tabelle 4.2). Alle Beobachtungen für die Schätzung zu verwenden führt in diesem Fall sogar zu einer Verschlechterung im Vergleich zum Kaplan-Meier-Schätzer (+5%). In Szenario 2.2 ist der Anteil informativer Beobachtungen in der Restgruppe deutlich höher ($q = 30\%$ und $n_{\bar{g}} = 100$). Die Untergruppe wurde jedoch verkleinert ($n_g = 10$). Das `SUBGROUP` sowie das `ALL` Modell unterbieten den Vorhersagefehler des Kaplan-Meier-

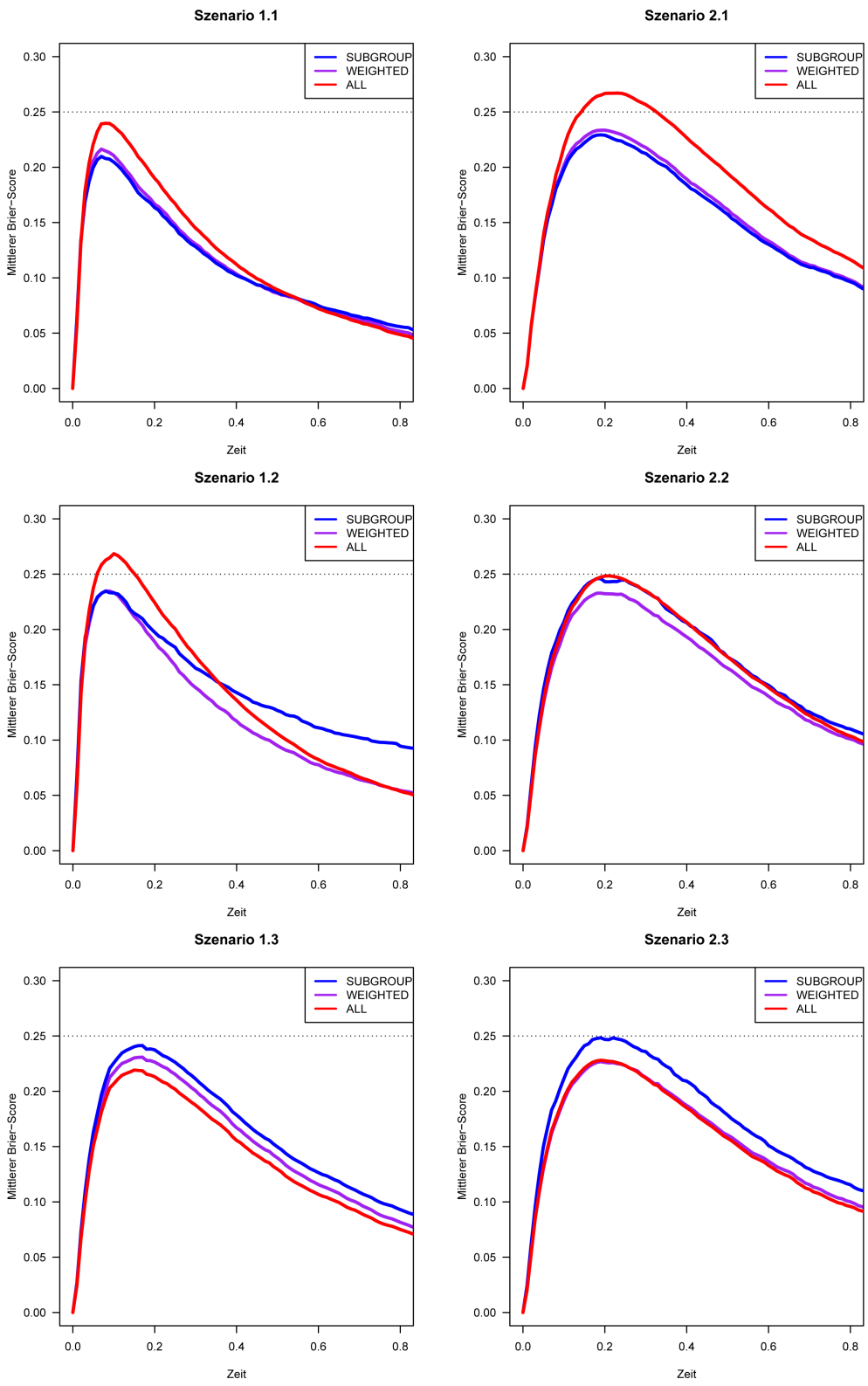


Abbildung 4.2: Mittlerer Brier Score der Modelle in den Szenarien aus Tabelle 4.1

Tabelle 4.2: Mittlerer integrierter Brier Score bis zum Zeitpunkt $t^* = 0{,}8$ der Modelle auf den simulierten Testdaten in den Szenarien aus Tabelle 4.1. Der Wert in Klammern gib die Verbesserung (in Prozentpunkten) gegenüber dem Kaplan-Meier-Schätzer an.

Beispiel	Szenario	Modell						Kaplan-Meier
		`SUBGROUP`		`WEIGHTED`		`ALL`		
1	1.1	11,43	(−22)	11,50	(−21)	12,43	(−15)	14,64
	1.2	14,93	(−3)	12,71	(−18)	14,38	(−7)	15,47
	1.3	16,44	(−11)	15,44	(−16)	14,46	(−22)	18,48
2	2.1	16,14	(−12)	16,49	(−10)	19,24	(+5)	18,38
	2.2	17,82	(−9)	16,71	(−14)	17,65	(−10)	19,51
	2.3	18,11	(−7)	16,30	(−17)	16,14	(−18)	18,76

Schätzers um 9 bis 10 %. Abbildung 4.2 zeigt jedoch, dass diese beiden Modelle im Bereich um die mediane Überlebenszeit knapp an der kritischen 0,25-Marke des Brier Scores liegen und somit wenig überzeugende Vorhersagen liefern. Das `WEIGHTED` Modell hingegen weist einen um 14 Prozentpunkte geringeren integrierten Brier Score als der Kaplan-Meier-Schätzer auf. Der `IBS` liegt nach Abbildung 4.2 zu jedem Zeitpunkt deutlich unter dem Referenzwert von 0,25. Im letzten Szenario 2.3 gilt $q = 70$ %. Somit gleicht die Verteilung von T und Z von über 2/3 der Beobachtungen aus der Restgruppe ($n_{\bar{g}} = 100$) der Verteilung aus der Untergruppe ($n_g = 10$). Wie zu erwarten, weist das `ALL` Modell den kleinsten Fehler auf. Das `WEIGHTED` Modell ist jedoch nicht merklich schlechter. Das `SUBGROUP` Modell führt hingegen zu keinem guten Ergebnis (siehe Abbildung 4.2).

Beispiel 1 macht deutlich, dass gerade für kleine Untergruppen die Vorhersage durch das `WEIGHTED` Modell deutlich verbessert werden kann, auch wenn nicht exakt die Zusammenhänge der Untergruppe in der Restgruppe herrschen. Des Weiteren unterstreicht dieses Beispiel in Szenario 1.2 die Tatsache, dass weitaus mehr Information erkannt werden kann, wenn der Brier Score über die Zeit hinweg betrachtet wird, anstatt diesen über den integrierten Brier Score zusammenzufassen. In dem betrachteten Beispiel mit heterogener Restgruppe (Beispiel 2) ist das `WEIGHTED` Modell in allen Szenarien dem jeweils besten anderen Modell mindestens ebenbürtig. Besonders hervorzuheben ist dabei das zweite Szenario 2.2. Es zeigt, dass es Konstellationen gibt, in denen das `WEIGHTED` Modell die Überlebenszeit gut vorhersagen kann, wohingegen die anderen beiden Modelle keine zufriedenstellenden Ergebnisse liefern.

Beide Beispiele zeigen schon im eindimensionalen Fall verschiedenste Szenarien auf, in denen eines der drei Modelle den anderen überlegen ist. Gerade für Untergruppen mit geringer Häufigkeit und gleichzeitig einem entsprechenden Unterschied zwischen dem Hazardratio in der Untergruppe und in der Restgruppe hat sich das `WEIGHTED` Modell in den Szenarien 1.2 und 2.2 am geeignetsten erwiesen. In den anderen Szenarien (siehe Beispiel 2) liegt der Vorhersagefehler des `WEIGHTED` Modells nicht maßgeblich über dem

des besten Modells. Es bleibt jedoch in jedem Fall zu beurteilen ob eine entsprechende Verbesserung der Vorhersage nicht nur statistisch signifikant sondern auch klinisch relevant ist.

Die Schätzung der hier erzeugten Überlebenszeitmodelle wurde auf den simulierten Daten mit der Funktion `coxph` aus dem *survival* Paket durchgeführt Zur Schätzung des gewichteten Modells wurde von der `weights=` Option Gebrauch gemacht. Die entsprechenden Gewichte aus Gleichung (4.6) wurden zuvor mit Hilfe der `glm` Funktion geschätzt (`family="binomial"`). Anschließend wurden die geschätzten Überlebensfunktionen aus den angepassten Modellen mit den Funktionen `PREDmat` und `KMmat` geschätzt (siehe Abschnitt 4.1.5 sowie Anhang B).

5 Bewertung der Überlebenszeitmodelle und ihrer Vorhersagen entlang der Untergruppen

In diesem Kapitel werden die Resultate der verschiedenen Verfahren zur Modellierung der Überlebenszeit von bestimmten Patientenuntergruppen verglichen und analysiert. Die beiden zur Verfügung stehenden Kohorten aus Köln und Uppsala (siehe Kapitel 2.3) werden dabei stets getrennt betrachtet. Die Untergruppen werden jeweils durch die erhobenen klinischen Merkmale definiert. Abschnitt 5.1 gibt zunächst einen Überblick über alle betrachteten Untergruppen und zeigt Unterschiede der beobachteten Überlebensraten zwischen diesen Gruppen auf.

Für jede Untergruppe werden die in Kapitel 4.2 gezeigten Modelle `SUBGROUP`, `WEIGHTED` und `ALL` betrachtet. Dabei wird dem Versuchsaufbau in 4.2.1 entsprechend der Vorhersagefehler dieser Modelle auf der zu untersuchenden Untergruppe ermittelt. Bevor die Vorhersagefehler ausgewertet werden, wird in Abschnitt 5.2 zunächst die Schätzung der Stichprobengewichte bewertet. Diese Gewichte bilden die Basis des `WEIGHTED` Modells und können erste Aufschlüsse über einen potentiellen Erfolg oder Misserfolg der für dieses Modell zugrundeliegenden gewichteten Modellbildung geben.

Die grundlegende Analyse der Vorhersagefehler aller Modelle zu allen Untergruppen ist in Abschnitt 5.3 enthalten. Dieser Abschnitt zeigt auf, ob das `ALL` Modell homogene Vorhersagefehler auf allen Untergruppen generiert oder ob und in wie weit bestimmte Untergruppen von einer separaten Modellbildung profitieren und somit das `SUBGROUP` oder das `WEIGHTED` Modell die Überlebenszeit der Patienten in diesen Untergruppen besser vorhersagen können.

Abschließend werden in Abschnitt 5.4 die zu den Analysen benötigten Ressourcen dargelegt. Dazu zählt insbesondere die benötigte Rechenzeit für den Prozess der Modellbildung.

5.1 Überblick über die Untergruppen

Mit der Notation aus Kapitel 4 ist eine Untergruppe g definiert als die Menge derjenigen Patienten, für die an einer Untergruppenvariable U die Ausprägung $U = g$ beobachtet wurde. Wie in der Einleitung zu diesem Kapitel erwähnt, werden in dieser Arbeit ausschließlich klinische Variablen als Untergruppenvariablen aufgefasst. Andere Untergrup-

Tabelle 5.1: Köln Kohorte: Fallzahlen N, Anzahl Ereignisse $\sum d_i$, geschätzte mediane Überlebenszeit $\tilde{t}_{0,5}$ sowie längste beobachtete Risikozeit $t_{\max}$ in der Untergruppe $U = g$

U	g	N	$\sum d_i$	$\tilde{t}_{0,5}$	$t_{\max}$
Alter	jung	67	21	7,9	13,6
	mittelalt	459	132	9,3	16,8
	alt	301	120	4,5	16,1
Geschlecht	männlich	549	201	5,7	16,8
	weiblich	284	76	14,1	14,8
Histologie	AD	378	129	6,1	14,9
	CA	49	1	–	16,1
	LC	84	30	11,2	13,6
	SCLC	29	12	2,8	3,9
	SQ	260	91	7,0	16,8
	unklassifiziert	30	12	3,7	5,1
Raucherstatus	Raucher	462	168	7,1	16,8
	Ex-Raucher	177	51	6,0	16,1
	Nichtraucher	100	30	7,9	14,8
Stadium	I	418	99	10,1	16,8
	II	166	56	7,1	13,5
	III	196	89	2,7	14,1
	IV	39	25	1,4	5,8

penvariablen wie beispielsweise der Mutationsstatus eines Gens oder Kombinationen der hier betrachteten Merkmale sind ebenfalls denkbar. Beide betrachteten Patientenkohorten enthalten dieselben klinischen Merkmale der Patienten. Die folgenden Variablen sind in beiden Kohorten vorhanden und werden jeweils als Untergruppenvariable aufgefasst: Alter (3), Geschlecht (2), Histologie (3 bzw. 5), Raucherstatus (3) und Stadium (4). Dabei gibt die Zahl in Klammern jeweils die Anzahl möglicher Ausprägungen G und damit die Zahl der zu untersuchenden Untergruppen an. Zu beachten ist, dass in der Uppsala Kohorte lediglich drei Histologietypen enthalten sind, wohingegen in der größeren Köln Kohorte fünf verschiedene Typen vertreten sind (vgl. Tabellen 5.1 und 5.2). Damit ergeben sich insgesamt 15 Untergruppen in der Uppsala und 17 Untergruppen in der Köln Kohorte. Mit der in Kapitel 2.1 beschriebenen histologischen Klassifikation und der Einteilung in Stadien sind diese Untergruppenvariablen fest definiert. Die drei Alterskategorien sind die Klasse der unter 50-Jährigen, der 50- bis unter 70-Jährigen und der mindestens 70-Jährigen.

Tabelle 5.1 fasst für die Köln Kohorte neben der Fallzahl der einzelnen Gruppen, die Anzahl der Ereignisse (Todesfälle) $\sum d_i$, die mediane Überlebenszeit $\tilde{t}_{0,5}$ sowie den Zeitpunkt der längsten beobachteten Risikozeit $t_{\max}$ in jeder Untergruppe g zusammen.

Tabelle 5.2: Uppsala Kohorte: Fallzahlen N, Anzahl Ereignisse $\sum d_i$, geschätzte mediane Überlebenszeit $\tilde{t}_{0,5}$ sowie längste beobachtete Risikozeit $t_{\max}$ in der Untergruppe $U = g$

U	g	N	$\sum d_i$	$\tilde{t}_{0,5}$	$t_{\max}$
Alter	jung	8	4	9,9	14,2
	mittelalt	114	75	4,1	15,9
	alt	70	62	3,1	15,2
Geschlecht	männlich	105	79	3,4	15,9
	weiblich	87	62	4,0	14,2
Histologie	AD	105	76	4,1	15,9
	LC	23	15	5,6	14,2
	SQ	64	50	3,0	15,2
Raucherstatus	Raucher	94	70	3,6	14,8
	Ex-Raucher	83	61	3,5	15,9
	Nichtraucher	15	10	4,7	11,3
Stadium	I	126	87	4,3	15,9
	II	35	27	3,0	15,2
	III	27	24	1,1	14,8
	IV	4	3	3,5	8,9

Tabelle 5.2 bereitet diese Kennzahlen analog für die Uppsala Kohorte auf.

In der Köln Kohorte sind einige der Merkmale nicht an allen Patienten erhoben worden bzw. nicht bekannt. Am häufigsten ist dies beim Raucherstatus zu beobachten. Dieser ist von 94 der 833 Patienten nicht angegeben, sodass sich die Fallzahlen dieser Variable lediglich zu 739 addieren. Weniger schwer ins Gewicht fällt die Variable Stadium, die an 14 Patienten nicht beobachtet wurde. Das Geschlecht hingegen ist von jedem Patienten verzeichnet. Die Uppsala Kohorte enthält ausnahmslos vollständige Datensätze. Dies spricht neben der in Kapitel 2.3 festgestellten niedrigen Zensierungsrate erneut für die Qualität dieser Daten.

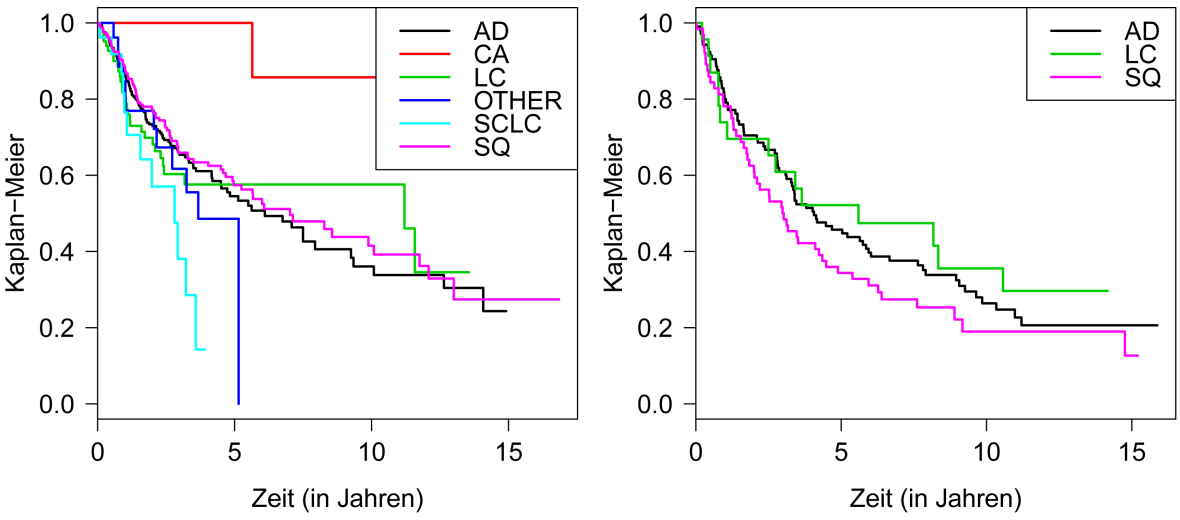


Abbildung 5.1: Kaplan-Meier-Schätzer getrennt nach Histologietypen in der Köln Kohorte (links) und in der Uppsala Kohorte (rechts). `OTHER` bezeichnet den Schätzer für unklassifizierte Patienten der Köln Kohorte.

Wie in Kapitel 2.3 angegeben beträgt die mediane Überlebenszeit $\tilde{t}_{0,5} = 7$ Jahre für Patienten der Daten aus Köln sowie $\tilde{t}_{0,5} = 3{,}5$ Jahre für die Patienten aus Uppsala. So unterscheiden sich auch die medianen Überlebenszeiten innerhalb der einzelnen Gruppen zwischen den beiden Kohorten. Entscheidend ist jedoch, dass sich die Überlebenszeiten auch innerhalb einer Kohorte zwischen den Untergruppen unterscheiden können. So liegt die geschätzte mediane Überlebenszeit der 29 SCLC Patienten aus der Köln Kohorte (Tabelle 5.1) mit $\tilde{t}_{0,5} = 2{,}8$ Jahren deutlich unter dem Median der gesamten Kohorte. Dazu passend liegt die maximal beobachtete Risikozeit lediglich bei 3,9 Jahren. Diese Beobachtung ist mit den Resultaten von Rosti u. a. (2006) annähernd konform. Hier wird sogar von einer 50 % Überlebenswahrscheinlichkeit nach nur 12 bis 20 Monaten gesprochen. Abbildung 5.1 veranschaulicht die unterschiedlichen Überlebensfunktionen gegeben der Histologie jeweils für die Patienten aus der Köln Kohorte sowie aus der Uppsala Kohorte und unterstreicht die Heterogenität der Daten im Hinblick auf die Überlebenszeiten.

Die längste beobachtete Risikozeit $t_{\max}$ in einer Untergruppe muss in jedem Fall bei der Betrachtung des Brier Scores berücksichtigt werden (siehe Abschnitt 3.3.1). Für Zeitpunkte über $t_{\max}$ hinaus sollte der Brier Score nicht betrachtet werden.

Abschließend sei auf Korrelationen der Untergruppen hingewiesen, die sich ebenfalls in den beiden vorliegenden Kohorten bestätigen. So beschreibt Edlund (2012) einen Zusammenhang zwischen Plattenepithelkarzinomen (SQ) und Rauchern. In der Köln Kohorte gehören von 218 SQ Patienten, von denen auch der Raucherstatus bekannt ist, lediglich vier zu den insgesamt 100 Nichtrauchern (p-Wert des Fisher Tests ist kleiner als 9×10^{-12}). In der Uppsala Kohorte ergibt sich ein ähnliches Bild, jedoch kann hier aufgrund der kleineren Fallzahl nicht von einem signifikanten Zusammenhang gesprochen werden. Edlund (2012) berichtet darüber hinaus von weiteren Korrelationen. So leiden beispielsweise junge männliche Patienten eher unter einem Adenokarzinom (AD).

5.2 Trennschärfe der Stichprobengewichte

Das `WEIGHTED` Modell (siehe Kapitel 4) nutzt alle Patienten einer Kohorte, um jeweils ein Überlebenszeitmodell für eine bestimmte Untergruppe $U = g$ aufzustellen. Dabei ist entscheidend, mit welcher Gewichtung ein Patient einen Beitrag zu der Modellbildung leistet. Passt der Patient gut zu der Untergruppe g, so bekommt er ein hohes Gewicht und umgekehrt. Kapitel 4.1.4 beschreibt die Schätzung der optimalen Gewichte

$$\omega_u(y, z) = \frac{p(u|y, z)}{p(u)}$$

aus Gleichung (4.6) für alle möglichen Ausprägungen (Untergruppen) $U = u$. Dabei wird die Wahrscheinlichkeit $p(u|y, z)$ über eine logistische Regression geschätzt. Im Folgenden wird die Schätzung dieser Gewichte für jede Gruppe $U = g$ bewertet. Dabei wird die Performance der logistischen Regression auf unabhängigen Testdaten ausgewertet. Die Testdaten setzen sich aus den 400-mal zufällig gezogenen Teilmengen der Daten zusammen (siehe Kapitel 4.2).

Abschnitt 5.2.1 misst zunächst die Performance der Modellierung von $p(u|y, z)$ in Abhängigkeit der Kovariablen z und der Methode der Regularisierung (Lasso oder Ridge). Abschnitt 5.2.2 geht detailliert auf bestimmte Untergruppen $U = g$ ein. Dabei wird der Einfluss der Kovariablen auf die Schätzung der Stichprobengewichte untersucht und interpretiert.

5.2.1 Performance der logistischen Regression in Abhängigkeit der Kovariablenmengen und der Modellparameter

Ein erster wichtiger Gesichtspunkt ist die Wahl geeigneter Kovariablen y und z. Es gilt zu beachten, dass für die Schätzung der Stichprobengewichte die Information über die Überlebenszeit (Kovariable y) als erklärende Variable angesehen wird. Wie in Kapitel 4.2 erläutert werden insgesamt drei Varianten in Betracht gezogen. Wie auch die Überle-

benszeitmodelle, so werden ebenfalls die Stichprobengewichte zunächst ausschließlich auf klinischen, anschließend nur auf genetischen sowie abschließend unter Verwendung beider Arten von Merkmalen gebildet. Da sich gezeigt hat, dass die Variante mit klinischen und genetischen Variablen keine Verbesserung der Performancemaßzahlen im Vergleich zu den anderen beiden Varianten nach sich zieht, wird diese Variante der Übersichtlichkeit wegen im Folgenden außer Acht gelassen. Die jeweiligen Merkmalsmengen werden durch die Kovariablen z bezeichnet. Unabhängig von z ist stets die beobachtete Risikozeit $\tilde{t}$ und der vorliegende Statusindikator δ Bestandteil der Kovariablen y. Diese Zusammensetzung entspricht dem Grundprinzip der in 4.1.3 beschriebenen Idee der Stichprobengewichte.

Als weiterer Gesichtspunkt ist die Methode der Regularisierung der Regressionskoeffizienten bei der Maximierung der Likelihood (4.8) zu nennen. Dazu werden das Lasso Verfahren L1 ($\alpha = 1$) und die Ridge Regression L2 ($\alpha = 2$) miteinander verglichen. Insgesamt ergeben sich somit für jede Untergruppe 4 Konstellationen aus Kovariablenmenge und Modellbildungsverfahren.

Fehlklassifikationsrate (ACC)

Für einen ersten Überblick über die Performance der logistischen Regression dient die Fehlklassifikationsrate, die sich aus der Differenz zwischen Eins und der in Abschnitt 4.1.4 erläuterten `ACC` (accuracy), siehe Gleichung (4.9) ergibt. Die Tabellen 5.3 und 5.4 geben die mittleren Fehlklassifikationsraten jeweils für eine Konstellation von Kovariablen und Regularisierung an.

In Tabelle 5.3 fällt zunächst auf, dass kein relevanter Unterschied zwischen den Resultaten des Lasso Verfahrens und der Ridge Regression zu existieren scheint. Die Fehlklassifikationsrate des naiven Bayes-Schätzers wird in der Köln Kohorte durch Verwendung der klinischen Merkmale für jede Untergruppenvariable U mindestens um ca. 5 % unterboten, im Fall der Untergruppenvariablen *Geschlecht* und *Histologie* sogar um leicht über 10 %. Dies kann in der Uppsala Kohorte – vermutlich aufgrund der niedrigeren Stichprobengrößen – nicht beobachtet werden.

Bei der Betrachtung von Modellen ausschließlich auf der Basis genetischer Variablen zeigt sich eine relativ gute Klassifikationsleistung für die Untergruppenvariable *Histologie*. Die Fehlerrate des naiven Bayes-Schätzers wird um 28 % (Ridge) bzw. 33 % (Lasso) unterboten (Köln Kohorte). In der Uppsala Kohorte (Tabelle 5.4) ist eine deutlichere Verbesserung erkennbar. Die Fehlklassifikationsrate ist um knapp 50 % verringert. Diese Steigerung kann durch eine kleinere Bayesrate von 0,45 (Köln 0,52) und einen leicht niedrigeren Anteil schlecht klassifizierbarer Gruppen zu erklären sein.

Letzteres wird durch die `ACC` nicht berücksichtigt. Diese Kennzahl fasst global die Anzahl richtig zugeordneter Beobachtungen (Patienten) zusammen. Können Patienten aus einer bestimmten Gruppe $U = g$ gut und aus einer anderen Gruppe schlecht klassifiziert werden, so kann dies nicht über die `ACC` ausgedrückt werden.

Tabelle 5.3: Köln Kohorte: Mittlere Fehlklassifikationsrate ($1 - \texttt{ACC}$) der regularisierten logistischen Regression auf unabhängigen Testdaten. Der Strafterm (siehe Gleichung (4.8)) gibt an, ob Lasso (L1) oder Ridge (L2) verwendet wurde. Die Modelle wurden jeweils nur mit den klinischen bzw. nur mit den genetischen Kovariablen trainiert. In Klammern ist die prozentuale Verbesserung im Vergleich zum naiven Bayes-Schätzer angegeben.

U	Strafterm	Kovariablen klinisch		Kovariablen genetisch		Bayesrate
Alter	L1	0,42	(−5)	0,44	(±0)	0,44
	L2	0,42	(−5)	0,44	(±0)	
Geschlecht	L1	0,30	(−13)	0,31	(−8)	0,34
	L2	0,30	(−11)	0,29	(−17)	
Histologie	L1	0,47	(−11)	0,35	(−33)	0,52
	L2	0,47	(−11)	0,37	(−28)	
Raucherstatus	L1	0,35	(−5)	0,37	(±0)	0,37
	L2	0,36	(−4)	0,37	(+1)	
Stadium	L1	0,46	(−4)	0,49	(+1)	0,48
	L2	0,46	(−5)	0,49	(+1)	

Tabelle 5.4: Uppsala Kohorte: Mittlere Fehlklassifikationsrate ($1 - \texttt{ACC}$) der regularisierten logistischen Regression auf unabhängigen Testdaten. Der Strafterm (siehe Gleichung (4.8)) gibt an, ob Lasso (L1) oder Ridge (L2) verwendet wurde. Die Modelle wurden jeweils nur mit den klinischen bzw. nur mit den genetischen Kovariablen trainiert. In Klammern ist die prozentuale Verbesserung im Vergleich zum naiven Bayes-Schätzer angegeben.

U	Strafterm	Kovariablen klinisch		Kovariablen genetisch		Bayesrate
Alter	L1	0,40	(−2)	0,42	(+2)	0,41
	L2	0,40	(−3)	0,41	(+1)	
Geschlecht	L1	0,43	(−4)	0,47	(+4)	0,45
	L2	0,43	(−4)	0,46	(+1)	
Histologie	L1	0,46	(+1)	0,24	(−47)	0,45
	L2	0,46	(+1)	0,23	(−49)	
Raucherstatus	L1	0,49	(−6)	0,52	(+2)	0,51
	L2	0,50	(−4)	0,52	(+2)	
Stadium	L1	0,34	(±0)	0,35	(+2)	0,34
	L2	0,34	(±0)	0,34	(±0)	

Area under the ROC Curve (AUC)

Um einen genaueren Einblick in die Performance der logistischen Modelle zu erhalten, wird die AUC (Area under the ROC curve) herangezogen (siehe Kapitel 4.1.4). Damit kann für jede Untergruppe einzeln beurteilt werden, wie gut sich diese vorhersagen lässt.

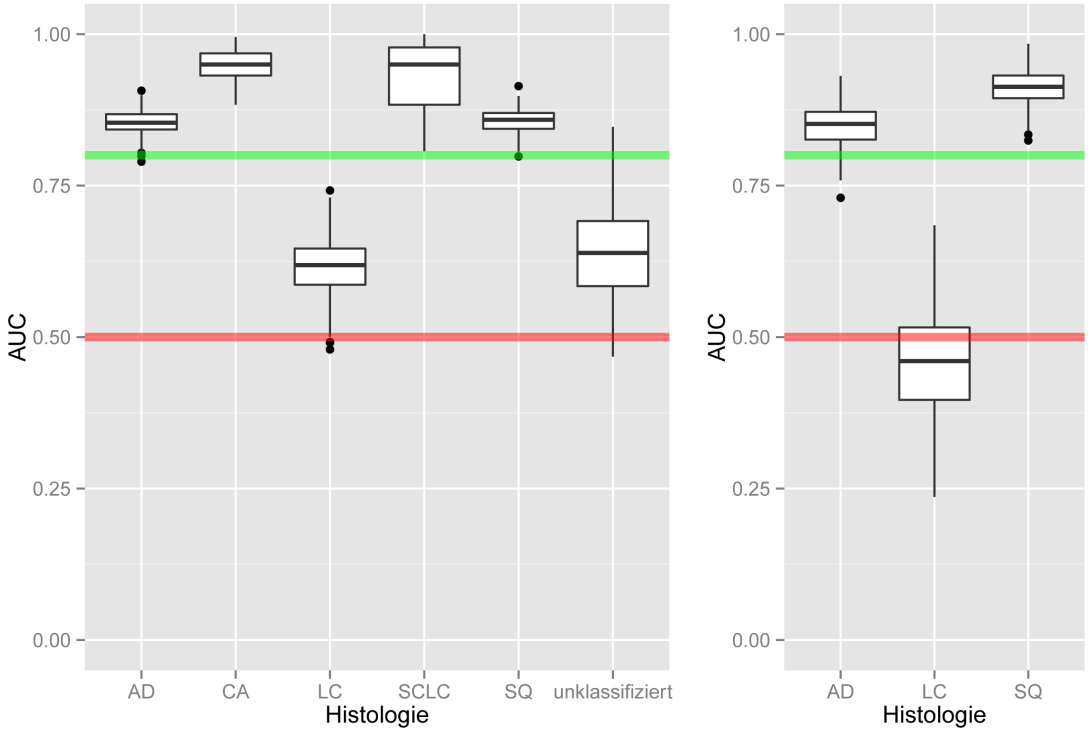


Abbildung 5.2: Area under ROC curve (AUC) für $\hat{p}(U = g|y, z)$. Die Gruppen werden durch die Variable *Histologie* definiert. Die Modelle enthalten nur genetische Variablen und wurden über eine Ridge Regression angepasst. Links sind die Werte für die Köln Kohorte und rechts für die Uppsala Kohorte abgebildet. Gute Vorhersagen werden für AUC $>$ 0,80 erzielt (grüne Linie). Zufällige Vorhersagen ergeben eine AUC von 0,50 (rote Linie).

Abbildung 5.2 zeigt exemplarisch die AUC in beiden Kohorten jeweils für alle Ausprägungen der Histologie. Großzellige Karzinome (LC) können offenbar weniger gut aus den genetischen Merkmalen vorhergesagt werden. Eine schlechtere Performance für Großzeller (mittlere AUC von 0,62 (Köln) bzw. 0,46 (Uppsala)) zieht den Wert der (global wirkenden) ACC nach unten. Die übrigen Histologietypen können über die genetischen Merkmale gut prognostiziert werden. Alle mittleren AUC-Werte liegen für beide Kohorten über 0,85. Modelle, die rein auf den klinischen Variablen basieren, erzielen eine niedrigere AUC, wobei die Klassifikationsleistungen auch hier oberhalb der Zufälligkeitsgrenze von 50 % liegen (siehe Tabellen A.1 und A.2 im Anhang). Dies zeigt am Beispiel der Histologie, dass eine mäßige globale Performance dennoch mit einer sehr guten Klassifikationsleistung für einzelne Untergruppen einhergehen kann.

5.2.2 Interpretation der Stichprobengewichte am Beispiel der Untergruppen 'Jung' und 'Nichtraucher'

In Abschnitt 5.2.1 zeigte sich zunächst, dass die Wahrscheinlichkeit $p(U = g|y, z)$ zu einem bestimmten Histologietyp g zu gehören gut mit genetischen Kovariablen z modelliert werden kann. Auch über rein klinische Faktoren lassen sich gewisse Rückschlüsse auf die Histologie ziehen (siehe Tabellen A.1 und A.2). In beiden Fällen bestimmen die modellierten Wahrscheinlichkeiten direkt die Stichprobengewichte, so dass sich auf Basis der Regressionskoeffizienten der logistischen Regression angeben lässt, welche Faktoren zu höheren Stichprobengewichten führen und damit den jeweiligen Beobachtungen (Patienten) einen stärkeren Einfluss auf das resultierende `WEIGHTED` Modell einräumen.

Im Folgenden werden die Untergruppe der unter 50-Jährigen sowie die Untergruppe der Nichtraucher näher analysiert. Auch für diese Gruppen ergibt sich im Sinne der AUC eine akzeptable Performance der logistischen Regression. Es zeigt sich darüber hinaus, dass der Einfluss bestimmter Kovariablen im Einklang mit der Literatur steht. Um die Regressionskoeffizienten näher interpretieren zu können, werden für die beiden Gruppen die Modelle auf Basis der klinischen Variablen betrachtet. Abbildung 5.3 zeigt die Boxplots der geschätzten Regressionskoeffizienten $\hat{\beta}_{gk}$ mit den vier im Mittel (basierend auf dem Median) größten absoluten Werten, wobei g die Gruppe der jungen Patienten (linke Seite) bzw. die Gruppe der Nichtraucher (rechte Seite) ist.

Die Regressionskoeffizienten geben implizit den Einfluss der Kovariablen auf die im `WEIGHTED` Modell zu verwendenden Stichprobengewichte an. Wird das `WEIGHTED` Modell zur Schätzung der Überlebenszeit von jungen bzw. nicht rauchenden Patienten erstellt, so bekommen Patienten mit einem Karzinoidtumor (CA) im Mittel ein höheres Stichprobengewicht. In beiden Fällen ist der mittlere absolute Regressionskoeffizient für den CA-Typ am größten. In Abbildung 5.3 liegt die als Indikator für ein Plattenepithel stehende Variable (SQ) an zweiter Position. Patienten dieses Histologietyps erhalten ein kleineres Stichprobengewicht. Ferner deuten Karzinome von jungen Patienten eher auf ein fortgeschritteneres Stadium hin.

Die Interpretation der jungen Patienten sowie der Nichtraucher deckt sich mit Aussagen in diversen Artikeln. So berichten etwa Brambilla u. a. (2001) von einem höheren Anteil an Adenokarzinomen bei Patienten unter 50 Jahren. Dies kann indirekt auch aus den hier erhaltenen Ergebnissen abgeleitet werden. Ebenso schreiben Subramanian und Govindan (2007), dass sich vermehrt Frauen mit Adenokarzinomen unter Nichtrauchern finden. Die gute Performance der oben betrachteten Modelle kann unter anderem durch die sehr gute Klassifikation der Histologie begründet sein, die entsprechend mit jungen Patienten bzw. mit Nichtrauchern korreliert. Sun u. a. (2007) dokumentieren bei Nichtrauchern markante Unterschiede sowohl auf klinischer als auch auf molekularer Ebene. Dies erklärt die hohen `AUC` Werte von klinischen und genetischen Modellen (siehe Tabelle A.1).

Zumindest die genannten Effekte der klinischen Variablen lassen sich in der Uppsala Kohorte bestätigen. Die gute Performance der Modelle für junge Patienten ist mit einer mittleren `AUC` von 0,70 ebenso deutlich erkennbar, wie die für Nichtraucherpatienten (`AUC` = 0,69) (Tabelle A.2). Die genetischen Variablen haben in der Uppsala Kohorte

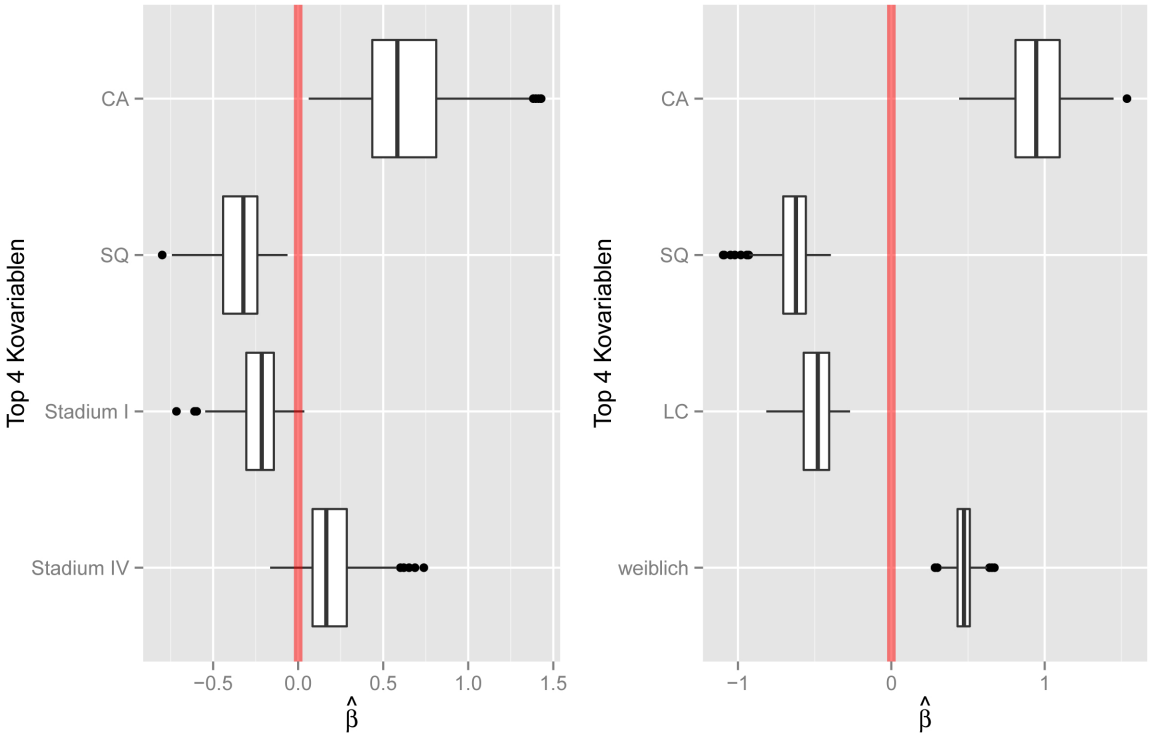


Abbildung 5.3: Regressionskoeffizienten der Top 4 Kovariablen (größter medianer absoluter Regressionskoeffizient). Die Modelle basieren auf den klinischen Variablen und modellieren die Wahrscheinlichkeit zur Gruppe der jungen Patienten (links) sowie zur Gruppe der Nichtraucher (rechts) zu gehören. Die Modelle wurden auf der Köln Kohorte mittels einer Ridge Regression angepasst.

nicht die entsprechende Relevanz im Vergleich zur Köln Kohorte. Die Fallzahlen sind mit acht Patienten unter 50 Jahren und 15 Nichtrauchern auch relativ klein. Einzig die Histologie kann hier durch die Genetik charakterisiert werden.

Abschließend lässt sich aus den bisherigen Erkenntnissen eine entscheidende Schlussfolgerung ziehen. Das `WEIGHTED` Modell ermöglicht zumindest theoretisch über die Stichprobengewichte die Heterogenität der Daten für die Modellierung der Untergruppe sinnvoll zu nutzen. So können bestimmte Korrelationen innerhalb und zwischen klinischen sowie genetischen Merkmalen berücksichtigt werden. Wie hier am Beispiel der jungen Patienten bzw. der Nichtraucher deutlich wurde, werden Patienten aus der jeweiligen Restgruppe, die der Untergruppe ähnlich im Sinne der Kovariablenausprägungen sind, für die Modellierung der Überlebenszeiten der Patienten aus der Untergruppe mitbenutzt. Ob und in wie weit die Stichprobengewichte tatsächlich einen gewinnbringenden Nutzen zur Schätzung der Überlebenszeit bestimmter Untergruppen mit sich bringen, zeigt der nachfolgende Abschnitt 5.3.

5.3 Vergleich der Vorhersagefehler

Die in Abschnitt 4.2.1 vorgestellten Modelle `SUBGROUP`, `WEIGHTED` und `ALL` werden für alle in Kapitel 5.1 betrachteten Untergruppen entsprechend trainiert und getestet. Die Bewertung der Modelle auf den Testdaten erfolgt dabei mit Hilfe des in Kapitel 3.3 vorgestellten Brier Scores. Wie in Abschnitt 3.3.2 beschrieben, werden jeweils 400 Trainings- und Testmengen bestimmt und anschließend der mittlere Vorhersagefehler auf den Testmengen ermittelt.

Abschnitt 5.3.1 gibt zunächst mittels des integrierten Brier Scores einen Überblick über die Performance aller Modelle. Für jede Konstellation von Modell, Modellbildungsalgorithmus und verwendeter Kovariablen wird der Vorhersagefehler betrachtet. Zudem wird ein Vergleich der beiden Kohorten aus Köln und Uppsala gezogen. Es gilt zu prüfen, ob sich die auf der Köln Kohorte ergebenen Resultate ebenfalls auf der kleineren Uppsala Kohorte wiederfinden. Aufgrund der niedrigeren Fallzahlen in der Uppsala Kohorte ist eine aussagekräftige Validierung der Ergebnisse für besonders kleine Untergruppen jedoch nicht zu bewerkstelligen. Dies wurde bereits bei der Analyse der Stichprobengewichte in Abschnitt 5.2 deutlich.

Anschließend wird in Abschnitt 5.3.2 der Brier Score für bestimmte Untergruppen über die Zeit hinweg analysiert. Diese Untergruppen weisen auffällige Unterschiede zwischen den drei betrachteten Modellen aus Abschnitt 4.2.1 auf. Der Vergleich der drei Modelle konzentriert sich dabei auf mögliche genetische Marker, die einen Einfluss auf die Überlebenszeit haben. Abschließend wird kurz auf die Modellbildung bei konstanter Gewichtung der Restgruppe eingegangen, wie sie in Kapitel 4.1.2 dargestellt ist.

5.3.1 Qualität der Vorhersagen getrennt nach Untergruppen, Modellen und verwendeten Kovariablen

In diesem Abschnitt wird mit Hilfe des integrierten Brier Scores $\texttt{IBS}(t^*)$, siehe Gleichung (3.14), ein Überblick über die Qualität der Vorhersagen für sämtliche betrachteten Modelle gegeben. Dabei ist zunächst zu entscheiden, bis zu welchem Zeitpunkt t^* der Brier Score integriert wird. Diese Entscheidung ist abhängig von der betrachteten Untergruppe. Wie in den Tabellen 5.1 und 5.2 zu sehen, weicht der Zeitpunkt des spätesten beobachteten Ereignisses $t_{\max}$ in einigen Untergruppen stark vom Durchschnitt ab. Dies betrifft vor allem die Untergruppen der SCLC und Stadium IV Patienten. Da der Brier Score für spätere Zeitpunkte als $t_{\max}$ nicht sinnvoll bestimmt werden kann, wird der integrierte Brier Score stets bis $t_{\max}$ ermittelt, höchstens jedoch über 10 Jahre. Damit wird die erhöhte Varianz berücksichtigt, die sich für extrem späte Zeitpunkte einstellt (siehe Abschnitt 5.3.2). Daraus ergibt sich $t^* = \min(10, t_{\max})$.

Übersicht über alle untersuchten Konstellationen

Für alle Untergruppen g der Köln sowie der Uppsala Kohorte, siehe Abschnitt 5.1, werden jeweils die Modelle `SUBGROUP`, `WEIGHTED` und `ALL` in den folgenden Varianten aufgestellt und gemäß Abschnitt 3.3.2 evaluiert. Wie in Abschnitt 4.2.1 beschrieben, werden für alle Modelle drei verschiedene Kovariablenmengen genutzt. Die klinischen und genetischen Kovariablen werden zunächst separat betrachtet. In der dritten Variante werden diese Kovariablen gemeinsam zur Modellbildung genutzt. Ferner werden jeweils drei verschiedene Algorithmen zur Modellbildung verwendet (vgl. 3.2). Dazu zählt das in Abschnitt 3.2.1 vorgestellte CoxBoost Verfahren sowie Lasso und die Ridge Regression (siehe Abschnitt 3.2.2). Zusammenfassend ergeben sich die folgenden Anzahlen:

- 18 bzw. 15 Untergruppen aus den jeweils 5 Untergruppenvariablen der Köln und Uppsala Kohorte
- 3 Kovariablenmengen (klinisch, genetisch, klinisch und genetisch)
- 3 Algorithmen zur Bildung der Überlebenszeitmodelle (CoxBoost, Lasso und Ridge Regression)
- 2 Modellbildungsalgorithmen zur Bestimmung der Stichprobengewichte (Lasso und Ridge)

Zudem wird das Verfahren zur Bestimmung der Stichprobengewichte aus Abschnitt 4.1.2 (konstantes Gewicht) mit drei verschiedene Gewichten (5 %, 10 % und 50 %) ausgewertet. Dabei werden die Modelle allerdings ausschließlich mit der Lasso Regression gebildet, um die Anzahl aller möglichen Modellvergleiche klein zu halten. Die Kaplan-Meier-Schätzer auf Basis aller Patienten sowie auf Basis der jeweiligen Untergruppe werden für alle Modelle als Referenzschätzer herangezogen.

Tabelle A.3 im Anhang fasst zunächst für die größere Köln Kohorte jeweils den integrierten Brier Score für jede Konstellation aus Modell, Modellbildungsalgorithmus und

verwendeter Kovariablenmenge sowie für die Kaplan-Meier-(Referenz)-Schätzer zusammen. Des Weiteren wird jeweils die entsprechende 2σ-Umgebung angegeben, um einen Überblick über die Streuung des `IBS` zu erhalten. Die Bestimmung der Stichprobengewichte für das `WEIGHTED` Modell erfolgte auf Basis des Lasso Verfahrens. In Abschnitt 5.2 zeigte sich bereits, dass die Wahl des Modellbildungsverfahrens sich kaum auf die resultierenden Stichprobengewichte auswirkt. Aus diesem Grund wird fortan nur die Variante mit der Lasso Regression betrachtet.

Für Karzinoidtumoren (CA) sind aufgrund nicht vorhandener Ereignisse keine Überlebenszeitmodelle sinnvoll zu schätzen. Dieser Histologietyp ist in Tabelle A.3 deshalb nicht abgedruckt. Die in diesem Abschnitt gezeigte Tabelle 5.5 zeigt ausschnittsweise die auffälligsten Resultate aus Tabelle A.3. Statt der 2σ-Umgebung ist in dieser Tabelle jeweils die relative Verbesserung eines Modells gegenüber dem besten Kaplan-Meier-(Referenz)-Schätzer der entsprechenden Untergruppe enthalten. Der Kaplan-Meier-Schätzer wird sowohl auf Basis der Patienten aus der Untergruppe als auch auf Basis aller Patienten bestimmt. Der Schätzer mit dem niedrigsten integrierten Brier Score entspricht dann dem Referenz-Schätzer für die jeweilige Untergruppe.

Performance des `ALL` Modells auf der Köln Kohorte

Zunächst wird das `ALL` Modell näher betrachtet. Dieses Modell wird unabhängig von der Untergruppe auf allen Patienten der Trainingsmenge angepasst. Das resultierende Modell ist somit für jede Untergruppe identisch. Auffällig ist die zum Teil stark unterschiedliche Qualität der Vorhersagen, die entlang der Untergruppen beobachtet werden kann. Beispielsweise beträgt der mittlere integrierte Brier Score für die Gruppe der Nichtraucher 0,136, wenn das Modell mittels Lasso aus den klinischen Variablen erstellt wird. Für diese Gruppe generiert das `ALL` Modell die besten Vorhersagen. Für die Gruppe der Ex-Raucher beträgt dieser Wert hingegen 0,227. Zum Vergleich liefert der Kaplan-Meier-Schätzer für diese Gruppen zwischen 0,204 und 0,221. Die Überlebenszeiten der Gruppe der Raucher kann mit einem mittleren Wert von 0,188 gut vorhergesagt werden. Ebenfalls auffällig gute Vorhersagen werden mit dem `ALL` Modell für junge Patienten getroffen (mittlerer integrierter Brier Score beträgt 0,147). Für ältere Patienten (über 70) wird die Überlebenswahrscheinlichkeit durch dieses Modell schlecht vorhergesagt. Das `ALL` Modell sagt die Überlebenswahrscheinlichkeit von Patienten mit einem Adeno- oder Squamouskarzinom besser vorher als der Kaplan-Meier-Schätzer. Auch für Patienten im ersten Stadium eignet sich das `ALL` Modell (mittlerer `IBS` von 0,173). Auffällig ist außerdem, dass sich die Überlebenszeiten der Frauen (0,163) im Vergleich zu Männern (0,194) im Mittel besser haben schätzen lassen. In allen erläuterten Fällen liegt der maximal über die Testmengen hinweg beobachtete mittlere integrierte Brier Score unterhalb der kritischen 0,25 Marke (siehe 2σ-Intervall in Tabelle A.3).

Die bisherigen Aussagen beziehen sich auf das `ALL` Modell unter Verwendung des Lasso Verfahrens in Verbindung mit den klinischen Merkmalen. Bei Verwendung der Ridge Regression ergeben sich annähernd identische Resultate. Die entsprechenden mittleren integrierten Brier Scores variieren nur schwach zwischen diesen beiden Modellbildungsalgorithmen. Deutlich schlechter schneidet hingegen das CoxBoost Verfahren ab. Hier

Tabelle 5.5: Mittlerer integrierter Brier Score (`IBS`) bis zum Zeitpunkt $\min(10, t_{\max})$ getrennt nach Untergruppen der Köln Kohorte. Diese Tabelle gibt ausschnittsweise die acht auffälligsten Resultate aus Tabelle A.3 wieder, wobei hier ausschließlich die Modellierung mit Lasso berücksichtigt wurde. Die Untergruppen wurden auf der Basis des kleinsten Vorhersagefehlers ausgewählt, den ein Modell in dieser Untergruppe erzielt hat. Der Wert in Klammern gibt jeweils die relative Verbesserung bzw. Verschlechterung des jeweiligen Modells gegenüber dem besten Kaplan-Meier-Schätzer (entweder auf Basis aller Patienten oder nur auf Patienten der Untergruppe geschätzt) der entsprechenden Untergruppe an. Die Modellbezeichnungen wurden wie folgt abgekürzt: `SUBGROUP` (`S`), `WEIGHTED` (`W`) und `ALL` (`A`).

Untergruppe	Modell	Kovariablen klinisch	Kovariablen genetisch	Kovariablen gemeinsam	Kaplan-Meier
	S	17.1 (−0,12)	21.8 (+0,12)	23.0 (+0,19)	21.3
< 50 Jahre	W	14.7 (−0,24)	19.3 (−0,01)	13.2 (−0,32)	
	A	14.7 (−0,24)	19.3 (−0,01)	14.5 (−0,25)	19.4
	S	18,2 (−0,13)	20,8 (±0,00)	18,1 (−0,13)	20,8
$50 - 70$ Jahre	W	17,8 (−0,10)	21,2 (+0,02)	18,3 (−0,12)	
	A	17,6 (−0,15)	21,2 (+0,02)	17,6 (−0,15)	21,1
	S	16.8 (−0,18)	20.8 (+0,01)	17.5 (−0,15)	20.6
Frauen	W	15.8 (−0,23)	20.5 (−0,00)	15.7 (−0,24)	
	A	16.3 (−0,21)	20.7 (±0,00)	16.3 (−0,21)	20.7
	S	18.8 (−0,13)	21.9 (+0,01)	18.8 (−0,13)	21.9
AD	W	18.2 (−0,16)	21.8 (+0,00)	18.4 (−0,15)	
	A	18.4 (−0,15)	21.8 (+0,00)	18.6 (−0,14)	21.7
	S	21.8 (−0,02)	23.4 (+0,05)	21.0 (−0,05)	22.9
LC	W	19.4 (−0,13)	22.0 (−0,01)	16.9 (−0,24)	
	A	19.0 (−0,14)	22.3 (+0,00)	19.0 (−0,14)	22.2
	S	19.4 (−0,12)	22.1 (+0,00)	19.2 (−0,13)	22.0
Raucher	W	18.9 (−0,14)	22.1 (+0,00)	19.0 (−0,14)	
	A	18.8 (−0,15)	22.1 (+0,00)	18.8 (−0,15)	22.0
	S	14.4 (−0,29)	21.2 (+0,04)	20.6 (+0,01)	20.9
Nichtraucher	W	12.8 (−0,37)	20.2 (−0,01)	12.3 (−0,40)	
	A	13.6 (−0,33)	20.3 (−0,00)	13.5 (−0,34)	20.4
	S	17,3 (−0,08)	18,8 (−0,01)	17,2 (−0,09)	18,9
Stadium I	W	17,1 (−0,10)	19,9 (+0,05)	18,2 (−0,04)	
	A	17,3 (−0,08)	20,0 (+0,06)	17,1 (−0,10)	20,0

liegen die beobachteten Werte fast ausschließlich oberhalb derer der L1 und L2 Regression und zwar unabhängig von der genutzten Kovariablenmenge.

Werden anstelle der klinischen Variablen lediglich die genetischen Merkmale als Kovariablen in die Modelle aufgenommen, so zeigt sich in keiner Konstellation eine adäquate Vorhersageleistung der Modelle. Die mittleren integrierten Brier Scores sind ausnahmslos auf dem Niveau des jeweils besten Kaplan-Meier-Schätzers. Ein anderes Bild ergibt sich, sobald beide Merkmalstypen (klinisch und genetisch) gemeinsam als Kovariablen genutzt werden. Wie in Abschnitt 4.2.1 beschrieben, werden die klinischen Variablen in dieser Variante nicht bestraft, im Sinne der regularisierten Modellbildung. Das `ALL` Modell generiert damit vergleichbare Vorhersagen, wie bei ausschließlicher Verwendung der klinischen Variablen. Abbildung 5.4 zeigt jeweils die ersten acht Boxplots der geschätzten Koeffizienten geordnet nach absteigendem medianem Wert für beide Modelle. Die Modelle ohne genetische Variablen weisen eine deutlich geringere Streuung der geschätzten Koeffizienten auf. Die für die Vorhersage ausschlaggebenden Variablen sind in beiden Modellen annähernd identisch. Die Hinzunahme der genetischen Variablen bringt in diesem Fall also keine Verbesserung mit sich. Dies zeigt sich zudem an den Häufigkeiten, mit denen im Lasso Verfahren genetische Kovariablen selektiert werden (siehe Tabellen A.7 und A.8).

Vergleich mit den Modellen `SUBGROUP` und `WEIGHTED` (Köln Kohorte)

Ein Vergleich der Modelle `SUBGROUP` und `WEIGHTED` mit dem `ALL` Modell zeigt für einige Untergruppen merkliche Verbesserungen in Bezug auf die Schätzung der Überlebenszeiten. Zur besseren Übersicht werden im Folgenden die Ergebnisse der Lasso Methode vorgestellt und diskutiert (siehe Tabelle 5.5). Einen ausführlichen Blick auf sämtliche Ergebnisse gibt Tabelle A.3. Den bedeutendsten Unterschied zeigt die Untergruppe der großzelligen Karzinome (LC). Hier weist das `WEIGHTED` Modell unter Verwendung aller Kovariablen einen um 11 % verringerten mittleren integrierten Brier Score (0,169) auf als das beste konkurrierende Modell. Dieser Wert liegt außerdem um 24 Prozentpunkte unter dem des Kaplan-Meier-Schätzers, wohingegen der des `ALL` Modells lediglich um 14 % darunter liegt (vgl. Tabelle 5.5). Das `ALL` Modell liefert danach mit einem mittleren `IBS` von 0,190 die besten Vorhersagen (siehe dazu auch Abbildung 5.7). Für das `ALL` Modell spielt es entgegen dem `WEIGHTED` Modell keine Rolle, ob nur die klinischen oder alle Kovariablen genutzt werden. Die Vorhersage verbessert oder verschlechtert sich dadurch nicht. Eine detailliertere Analyse dieser Untergruppe ist im folgenden Abschnitt 5.3.2 enthalten. Ebenfalls die besten Vorhersagen liefert das `WEIGTHED` Modell für die Untergruppe der jungen Patienten, für Nichtraucher sowie für die Untergruppe weiblicher Patienten. In der Gruppe junger Patienten verringert sich der gemittelte `IBS` um 10 Prozentpunkte, wenn die genetischen Variablen hinzugenommen werden. Die Gewichte können über die genetischen Variablen besser geschätzt werden. Das `ALL` Modell generiert hingegen anders als für die Untergruppe der Großzeller ebenfalls gute Vorhersagen (mittlerer `IBS` von 0,145). Allerdings zeigt sich hier eine hohe Varianz der ermittelten Werte (siehe Tabelle A.3). In der Gruppe der Nichtraucher zeigen sich ähnliche Zusammenhänge.

Das `SUBGROUP` Modell scheint wie in den in Abschnitt 4.2.2 skizzierten Szenarien 1.1

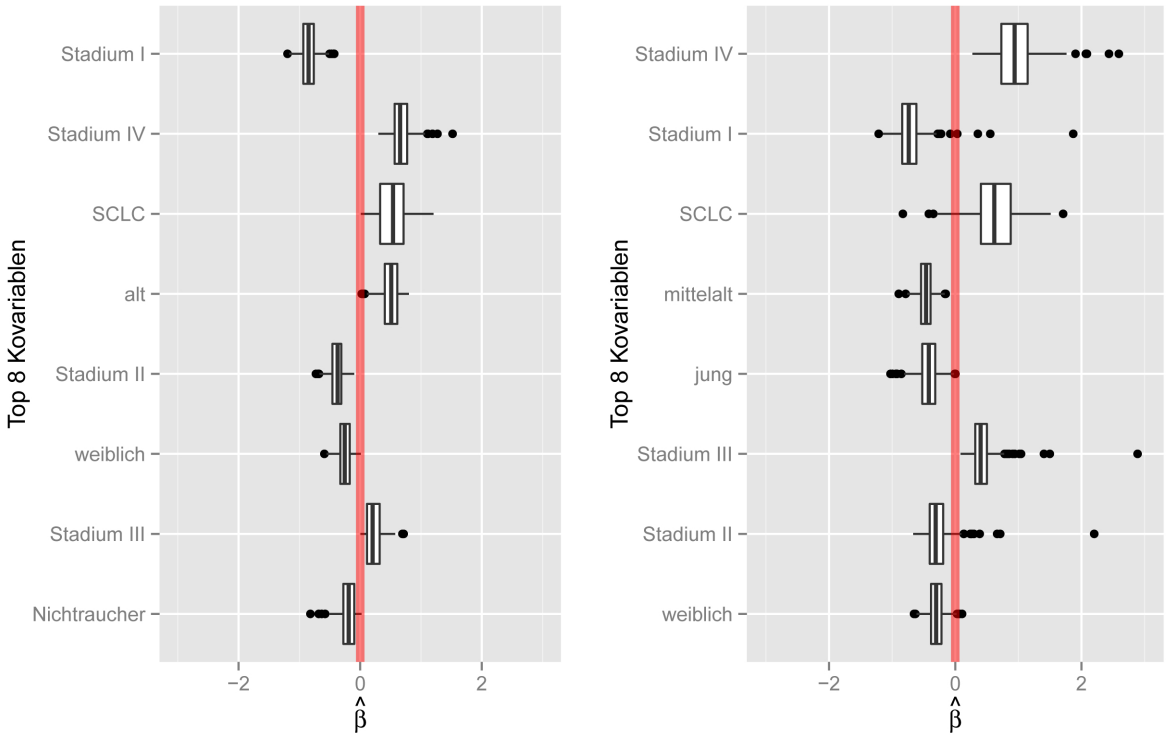


Abbildung 5.4: Verteilung der acht wichtigsten Kovariablen des `ALL` Modells ohne (links) und mit (rechts) Berücksichtigung der genetischen Variablen, absteigend geordnet nach dem größten medianen absoluten Regressionskoeffizienten aus den 400 Trainingsmengen. Die Modelle wurden mit Lasso trainiert. Die klinischen Variablen wurden nicht bestraft ($\lambda_k = 0$), wenn diese gemeinsam mit den genetischen Variablen verwendet wurden.

und 2.1 ausreichend für die Untergruppen der Stadium I Patienten und der Adenokarzinome zu sein (Tabelle A.3). Die zusätzlich gewonnene Fallzahl durch das `ALL` Modell bzw. die gewichtete Modellbildung durch das `WEIGHTED` Modell erbringen keine Verbesserung der Vorhersagen für diese Gruppen. Bei den Plattenepithelkarzinomen (SQ) verringert sich hingegen der Vorhersagefehler. Dieses Bild zeigt sich ebenfalls bei den Rauchern. Dies kann dadurch begründet werden, dass bei Rauchern häufig ebendieser Histologietyp diagnostiziert wird (vgl. Kapitel 2.1 sowie Abschnitt 5.2.2).

Uppsala Kohorte

In der deutlich kleineren Uppsala Kohorte zeigen sich weitaus weniger Unterschiede zwischen den einzelnen Modellen für die jeweiligen Untergruppen. Tabelle A.4 zeigt analog zu Tabelle A.3 die jeweiligen mittleren integrierten Brier Scores und deren Streuung. Das `ALL` Modell erzeugt unabhängig von der Untergruppe gleichermaßen schlechte Vorhersagen der Überlebenszeit. Der mittlere integrierte Brier Score liegt ausnahmslos auf dem Niveau der jeweiligen Kaplan-Meier-Schätzer. Das `SUBGROUP` Modell liefert ebenso keine guten Vorhersagen. Generell lässt sich feststellen, dass die Varianz der Schätzer im Vergleich zur Köln Kohorte größer ist, was aufgrund der Stichprobengrößen nicht verwunderlich scheint. Einzig auffällige Untergruppe in der Uppsala Kohorte ist die Gruppe der Nichtraucher. Der mittlere `IBS` für das `WEIGHTED` Modell liegt hier mit 0,185 Punkten knapp 15 % unter dem des Kaplan-Meier-Schätzers (`IBS` = 0,217). Das entsprechende 2σ-Intervall liegt zudem unter der kritischen 0,25 Grenze. Dies gilt sowohl bei der Lasso als auch bei der Ridge Regression, wenn neben den klinischen Kovariablen außerdem die genetischen Variablen zur Modellbildung genutzt werden. Die genetischen Variablen deuten auf eine zusätzliche Information bezüglich der Überlebenszeiten hin. Dies wird im nächsten Abschnitt ausführlich beleuchtet.

Konstante Stichprobengewichte

Abschließend zeigen sich für viele der hier betrachteten Untergruppen Parallelen zwischen den Resultaten der individuellen Gewichtung der Patienten im `WEIGHTED` Modell und der einheitlichen (konstanten) Gewichtung, wie sie in Abschnitt 4.1.2 beschrieben wurde. Die Tabellen A.5 und A.6 fassen die entsprechenden mittleren integrierten Brier Scores für dieses Verfahren zusammen. Einzig für die Gruppe der Großzeller (LC) ist das `WEIGHTED` Modell mit der individuellen Gewichtung auch diesen Modellen überlegen.

Zusammenfassend lässt sich sagen, dass sich bei den Vorhersagemodellen für die Überlebenszeiten eine klare Heterogenität zwischen den Untergruppen abzeichnet. In einigen Untergruppen generieren die Modelle offenbar geeignetere Vorhersagen als in anderen Untergruppen. Die gewichtete Modellierung über das `WEIGHTED` Modell liefert auffällig häufig den kleinsten Vorhersagefehler. Ein Beweis dazu liefern die Ergebnisse aufgrund fehlender statistischer Tests jedoch nicht. Zudem ist die Streuung der einzelnen Werte extrem hoch.

5.3.2 Analyse auffälliger Untergruppen hinsichtlich des Einflusses genetischer Faktoren auf die Überlebenszeit

Im vorangegangenen Abschnitt 5.3.1 zeigten sich vor allem die Untergruppen der jungen Patienten unter 50 Jahren, Frauen, Nichtraucher sowie die Untergruppe der Patienten mit großzelligem (LC) Karzinom besonders auffällig in Bezug auf die Vorhersage der Überlebenszeit. Das Überleben ließ sich bei Patienten aus diesen Untergruppen am besten vorhersagen. Dieser Unterschied in der Qualität der Vorhersage in bestimmten Untergruppen betraf gleichermaßen das `ALL` Modell wie auch die Modelle `SUBGROUP` und `WEIGHTED`. Dieser Abschnitt untersucht im Besonderen die Qualität der Vorhersagen des `WEIGHTED` Modells im Vergleich mit den beiden anderen Modellen. Die Konzentration liegt dabei auf möglichen genetischen Faktoren, die einen Einfluss auf die Überlebenszeit der Patienten besitzen. Hierzu werden die Resultate des Lasso Verfahrens herangezogen.

Die Häufigkeit, mit der ein genetischer Faktor (Kovariable) in das Modell selektiert wurde, kann Aufschluss über dessen Relevanz bezüglich des Einflusses auf die Überlebenszeit geben. Dazu sei zunächst ein Überblick über die Häufigkeiten gegeben, mit der genetische Kovariablen in den Modellen mit gemeinsamer Kovariablenmenge (klinische und genetische Variablen) ausgewählt werden. Des Weiteren werden für die genannten Untergruppen die Brier Scores der Modelle über die Zeit hinweg betrachtet. Damit kann die Qualität der Vorhersagen zu jedem Zeitpunkt bewertet werden. Die Modelle können detaillierter miteinander verglichen werden.

Anzahl selektierter genetischer Variablen

Tabelle A.7 gibt für das `ALL` Modell sowie je Untergruppe für die Modelle `SUBGROUP` und `WEIGHTED` die relative Häufigkeit an, mit der in den 400 Trainingsmengen mindestens eine der genetischen Variablen in das jeweilige geschätzte Modell aufgenommen wurde. Dabei zeigt sich, dass in beiden Kohorten in lediglich 1/4 der Trainingsmengen das `ALL` Modell, welches für alle Untergruppen identisch ist, mindestens auch eine genetische Variable enthält.

Tabelle A.8 gibt analog zu Tabelle A.7 die relative Häufigkeit der am häufigsten selektierten genetischen Variable an. Das `ALL` Modell selektiert über die Trainingsmengen hinweg maximal nur in 5 % (Köln Kohorte) bzw. 9 % (Uppsala Kohorte) der Fälle dieselbe genetische Variable. Somit ist kein genetischer Faktor stabil im `ALL` Modell enthalten. Die genetischen Variablen liefern in diesem Modell keinen relevanten Beitrag zur Schätzung der Überlebenszeit.

Das `WEIGHTED` Modell selektiert in der Köln Kohorte häufiger dieselbe genetische Variable als die beiden anderen Modelle. Für die vier zuvor genannten Untergruppen liefert dieses Modell erhöhte Häufigkeiten im Vergleich zum `SUBGROUP` Modell. Eine der auffälligsten Untergruppen in der Köln Kohorte ist die Gruppe der großzelligen Karzinome. Das `WEIGHTED` Modell wählt die häufigste genetische Region in 40 % der Trainingsmengen aus. In knapp 70 % der Stichproben wird mindestens eine genetische Variable ausgewählt. Das `SUBGROUP` Modell liefert hier lediglich in 10 % der Trainingsmengen dieselbe Region. Außerdem enthalten hier mit einer Häufigkeit von 20 % weniger Modelle

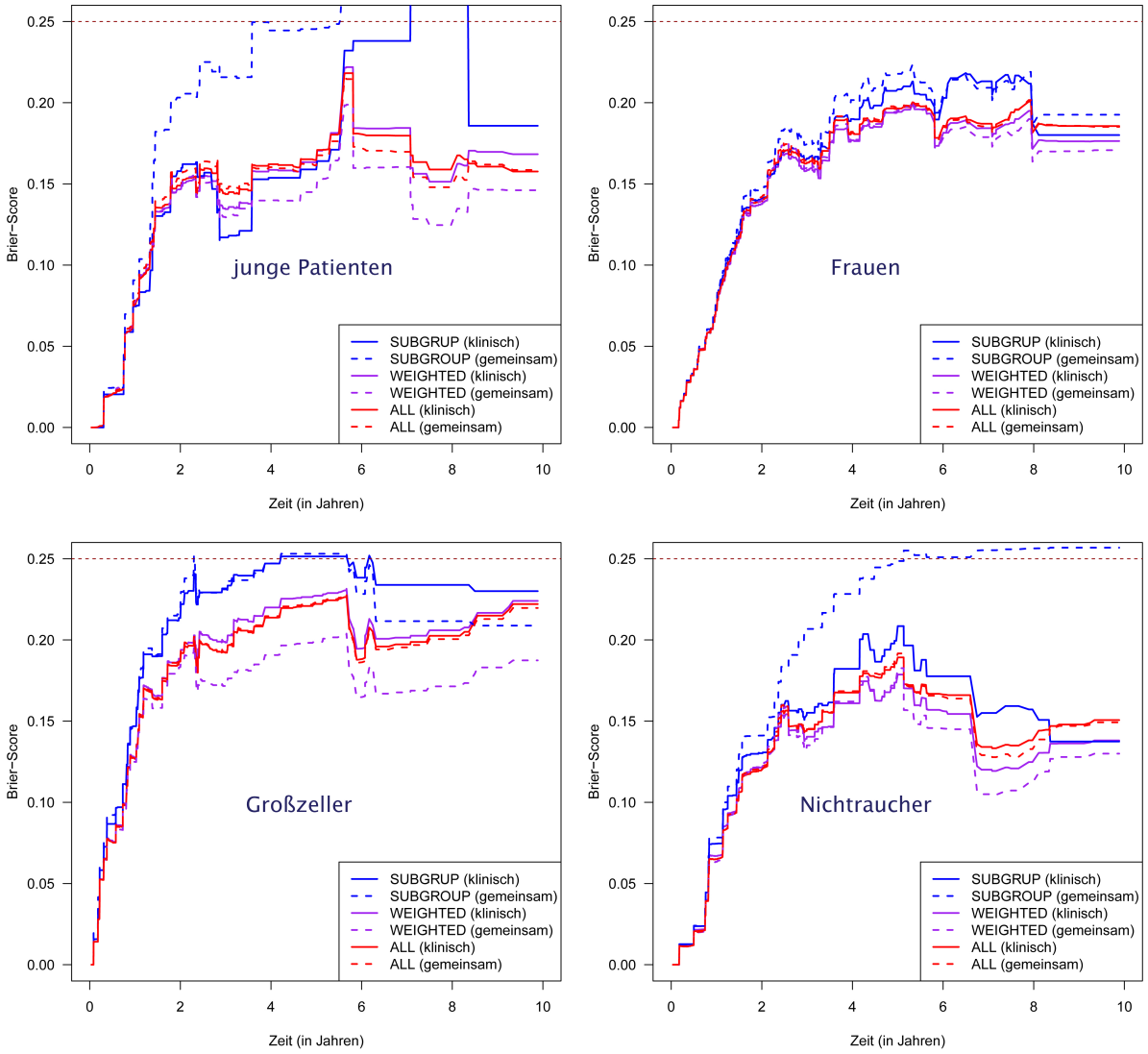


Abbildung 5.5: Köln Kohorte: Mittlerer Brier Score über die Zeit für die Überlebenszeitmodelle auf Basis der klinischen Variablen (durchgezogene Linien) sowie auf Basis klinischer und genetischer Kovariablen (gestrichelte Linien) getrennt für die Modelle `SUBGROUP` (blau), `WEIGHTED` (violett) und `ALL` (rot) zur Schätzung der Überlebenszeit der Gruppe der jungen Patienten unter 50 Jahren (links oben), der Frauen (rechts oben), Patienten mit großzelligem (LC) Karzinom (unten links) sowie der Nichtraucher (unten rechts).

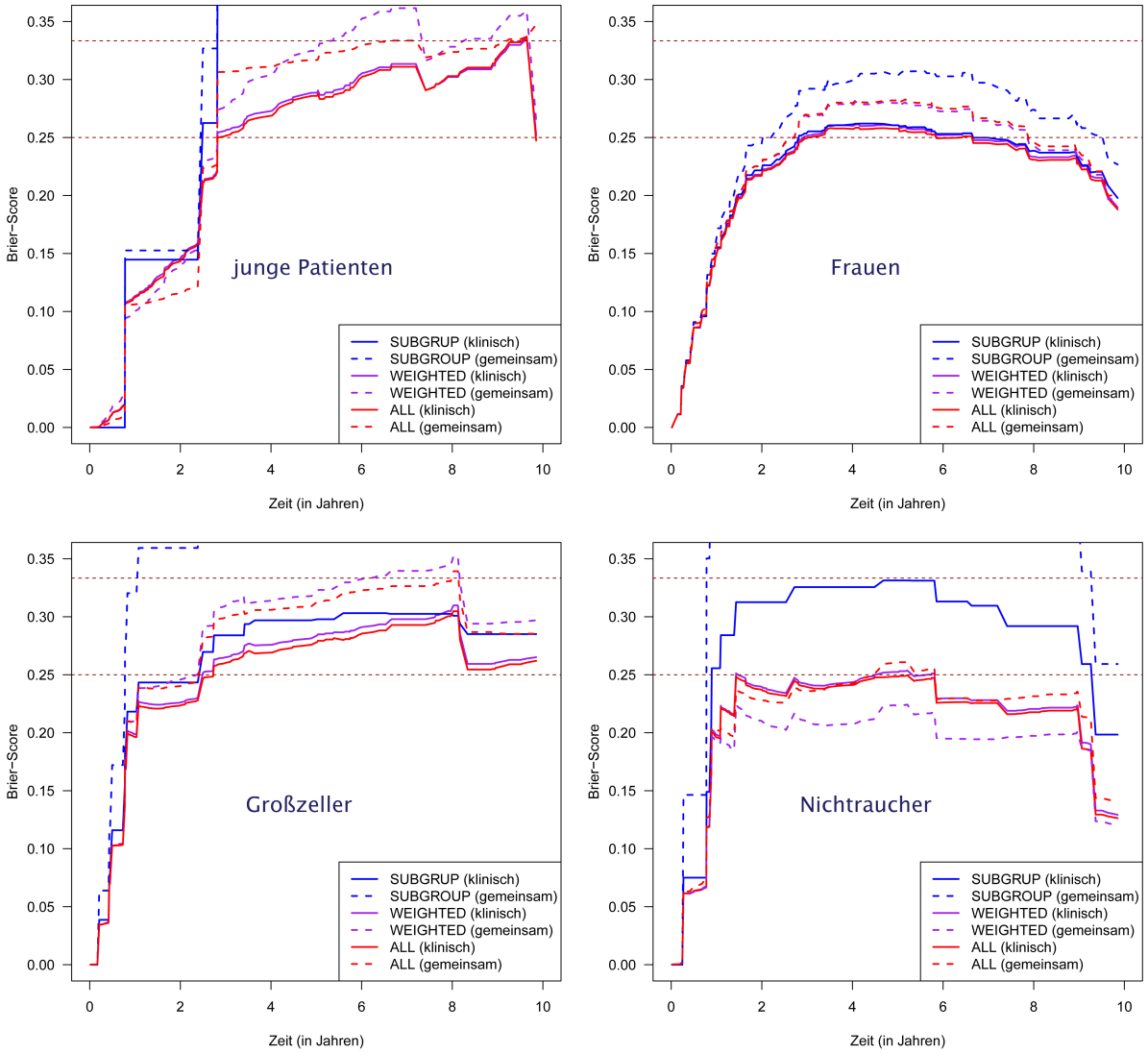


Abbildung 5.6: Uppsala Kohorte: Mittlerer Brier Score über die Zeit für die Überlebenszeitmodelle auf Basis der klinischen Variablen (durchgezogene Linien) sowie auf Basis klinischer und genetischer Kovariablen (gestrichelte Linien) getrennt für die Modelle `SUBGROUP` (blau), `WEIGHTED` (violett) und `ALL` (rot) zur Schätzung der Überlebenszeit der Gruppe der jungen Patienten unter 50 Jahren (links oben), der Frauen (rechts oben), Patienten mit großzelligem (LC) Karzinom (unten links) sowie der Nichtraucher (unten rechts).

eine genetische Region als das ALL Modell. Für die Untergruppen der kleinzelligen Karzinome (SCLC) sowie für die Gruppe der Frauen können ebenfalls extreme Unterschiede beobachtet werden. Hier ist die häufigste Region in 41 bzw. 42 % der Trainingsmengen enthalten und in 82 bzw. 64 % der Fälle ist stets auch eine genetische Variable in dem Modell enthalten.

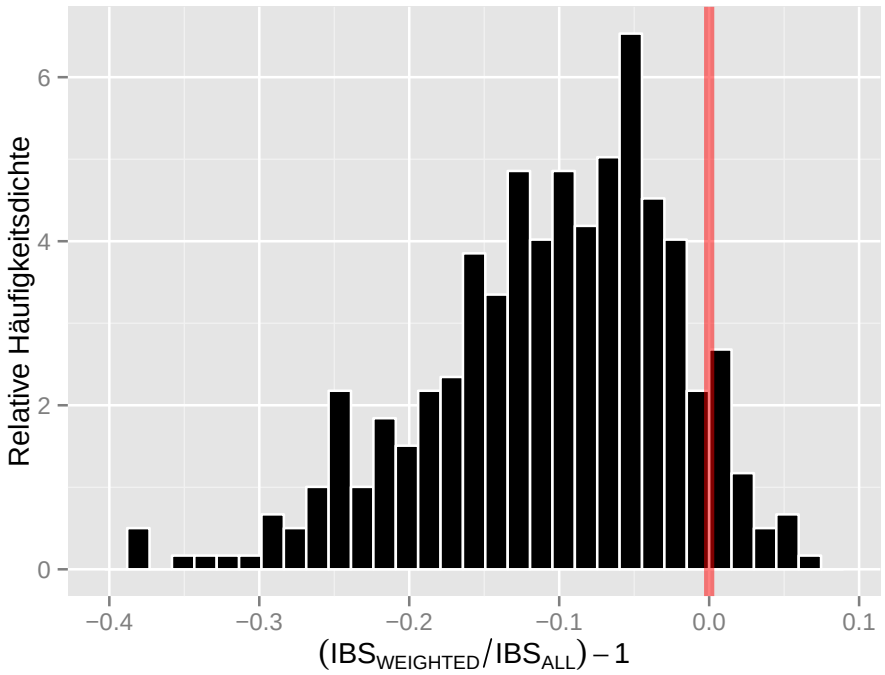


Abbildung 5.7: Relative Unterschiede der integrierten Brier Scores des WEIGHTED und ALL Modells für die Untergruppe der großzelligen Karzinome (LC). Die Modelle wurden auf der Köln Kohorte mit dem Lasso Verfahren auf Basis aller Kovariablen (klinisch und genetisch) gebildet (vgl. Abbildung 5.5 links unten). Die integrierten Brier Scores wurden jeweils auf denselben Testdaten miteinander verglichen.

Brier Score über die Zeit für Modelle mit und ohne genetische Variablen

Wie bereits in Abschnitt 5.3.1 angedeutet, liegt der mittlere Brier Score des WEIGHTED Modells für die vier zuvor genannten Untergruppen (junge Patienten unter 50 Jahren, Frauen, Nichtraucher, großzellige Karzinome) in der Köln Kohorte unterhalb der Brier Scores aller anderen Modelle. Abbildung 5.5 vergleicht für die Köln Kohorte die mittleren Brier Scores der Modelle über die Zeit hinweg. Dabei werden sowohl die Modelle mit klinischen und genetischen Variablen betrachtet als auch Modelle rein auf Basis der klinischen Variablen. Das WEIGHTED Modell generiert für Zeitpunkte ab 2 Jahren konstant bessere Vorhersagen als alle anderen Modelle (siehe Abbildung 5.5 links unten).

Abbildung 5.6 zeigt die mittleren Brier Scores analog für diese Gruppen aus der Uppsala Kohorte.

Abbildung 5.7 zeigt für diese Untergruppe den relativen Unterschied des `WEIGHTED` Modells im Verglich zum `ALL` Modell jeweils auf Basis aller Kovariablen (klinisch und genetisch). Der Unterschied entspricht dem prozentualen Anteil, um den der integrierte Brier Score des `WEIGHTED` Modells dem des `ALL` Modells überlegen ist. Der Vergleich eines `IBS` Paares ($\texttt{IBS}_{\texttt{WEIGHTED}}$ gegen $\texttt{IBS}_{\texttt{ALL}}$) basiert dabei auf derselben Testmenge. Diese Abbildung zeigt, dass das `WEIGHTED` Modell dem `ALL` Modell in fast allen 400 Trainings/Test-Aufteilungen überlegen ist.

Für die anderen Untergruppen liefert das `WEIGHTED` Modell keine besseren Ergebnisse als die anderen beiden Modelle. Zwar können auffällige Unterschiede in den Häufigkeiten selektierter genetischer Variablen beispielsweise beim `WEIGHTED` Modell der SCLC Patienten beobachtet werden (Tabelle A.8). Dennoch generiert das Modell keine guten Vorhersagen, weshalb der Brier Score für diese Untergruppe in Abbildung 5.5 nicht dargestellt ist. In diesem Fall ist die Stichprobengröße zur Validierung des Modells vermutlich nicht ausreichend groß. Die Testmenge der SCLC Patienten enthält lediglich 11 Patienten mit im Mittel vier Ereignissen (vgl. Tabelle 5.1). Für die Untergruppe der Frauen werden zwar vermehrt Gene in das `WEIGHTED` Modell mit aufgenommen, der Brier Score dieses Modells zeigt auf den Testdaten jedoch keine relevante Verbesserung der Schätzung der Überlebenszeit (Abbildungen 5.5 und 5.6 jeweils rechts oben). In der Köln Kohorte verbessert sich die Vorhersage unter dem `WEIGHTED` Modell mit genetischen und klinischen Variablen für die Untergruppen der jungen Patienten (Abbildung 5.5 links oben) sowie der Nichtraucher (Abbildung 5.5 rechts unten) marginal. Diese Auffälligkeiten lassen sich in der Uppsala Kohorte jedoch nicht wiederfinden (siehe Abbildung 5.6). Aufgrund der instabilen Kurven und der hohen Varianz der Werte (siehe Tabellen A.3 und A.4) kann in keiner Kohorte von einem relevanten Unterschied zwischen einem Untergruppenmodell und dem jeweiligen `ALL` Modell gesprochen werden.

Insgesamt kann nicht von mehr als einem schwachen zusätzlichen Nutzen der genetischen Variablen gesprochen werden. In keiner untersuchten Untergruppe wird dieselbe Region in über der Hälfte der Trainingsmengen selektiert. In der Uppsala Kohorte wird in keinem Modell eine genetische Variable in über einem Drittel der Trainingsmengen ausgewählt. Die beispielsweise für die Großzeller der Köln Kohorte beobachtete Selektions-Häufigkeit kann in der Uppsala Kohorte nicht annähernd bestätigt werden. Ferner finden sich keine genetischen Regionen unter den Top Acht einflussreichsten Kovariablen in den Modellen `WEIGHTED` und `SUBGROUP` in allen betrachteten Untergruppen. Dies gilt gleichermaßen für die Köln und Uppsala Kohorte. Es kann dadurch im besten Fall von schwachen genetischen Effekten auf die Überlebenszeit ausgegangen werden. Ein schwacher Effekt würde zudem erklären, weshalb auf der kleineren Uppsala Kohorte die erhöhte Auswahl der genetischen Regionen, wie sie beispielsweise bei den Großzellern der Köln Kohorte beobachtet wurde, nicht bestätigt werden konnte.

5.4 Ressourcenauswertung

Die Berechnungen zur Schätzung und zur Validierung aller in dieser Arbeit vorgestellten Modelle und Verfahren sowie die Erstellung sämtlicher Grafiken erfolgte mit der Statistiksoftware R (R Core Team, 2012) in Version 2.15.2. Dabei kamen jeweils die in den entsprechenden Abschnitten erläuterten Zusatzpakete beziehungsweise eigenhändig implementierte Funktionen zum Einsatz. Das Trainieren aller Überlebenszeitmodelle wurde, wie in Abschnitt 3.3.2 erläutert, auf dem Rechencluster der Technischen Universität Dortmund LiDOng (**Li**nux cluster **DO**rtmund **n**ext **g**eneration) durchgeführt. Hierzu wurden die Pakete *BatchJobs* (Bischl u. a., 2012b) und *BatchExperiments* (Bischl u. a., 2012a) zur Hilfe genommen. Die Evaluierung der geschätzten Modelle auf den entsprechenden Testdaten wurde auf lokalen Computern getätigt. Das Trainieren der Modelle wurde sowohl aufgrund der hohen Anzahl einzelner Jobs als auch wegen der hohen Rechenzeit auf das Batch-System LiDOng ausgelagert. Dieser Abschnitt fasst hierzu die benötigte Rechenzeit zusammen und beleuchtet Unterschiede im Hinblick auf einzelne zur Modellbildung verwendeter Verfahren.

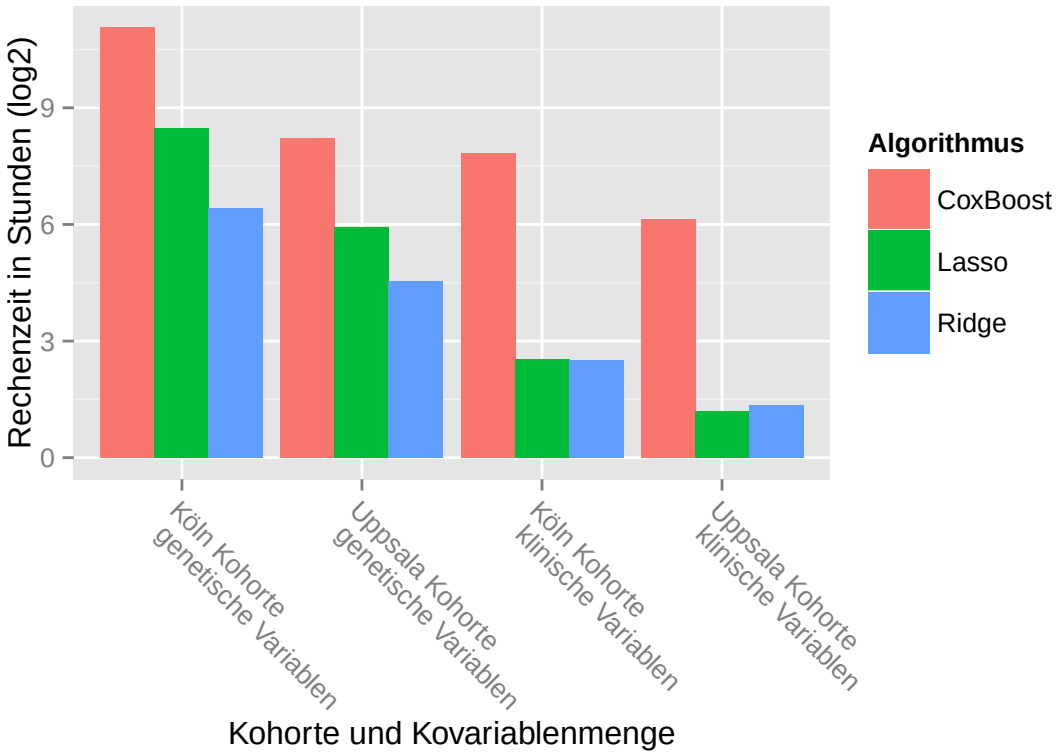


Abbildung 5.8: Logarithmierte Summe der Laufzeiten von je 4 000 Jobs getrennt nach Kohorte, Kovariablenmenge und Modellbildungsalgorithmus. Die 4 000 Jobs setzen sich aus je 400 Bootstrapstichproben pro Untergruppenvariable und Verfahren zur Schätzung der (individuellen) Stichprobengewichte zusammen.

Aufsummiert betrug die Zeit für alle gerechneten Jobs 5 686 Stunden (237 Tage). In einem einzelnen Job wurden für eine der 400 Aufteilungen in Trainings- und Testmenge und für eine der jeweils 5 Untergruppenvariablen die betrachteten Modelle `ALL`, `SUBGROUP` und `WEIGHTED` (inklusive der Berechnung der Stichprobengewichte) trainiert. Insgesamt ergaben sich für jede Kohorte 54 000 Jobs. Darunter fällt die Unterscheidung der Methoden zur Berechnung der Stichprobengewichte aus den Abschnitten 4.1.4 und 4.1.2 sowie die verschiedenen Verfahren zur Bildung der Überlebenszeitmodelle (CoxBoost, Lasso und Ridge). Ebenso wurden alle Modelle mit jeweils unterschiedlichen Kovariablenmengen gebildet (klinisch und/oder genetisch).

In Abbildung 5.8 sind die Laufzeiten der einzelnen Jobs getrennt nach der Anzahl Beobachtungen (Köln oder Uppsala Kohorte) sowie der Anzahl Kovariablen (klinisch oder genetisch) und den Algorithmen zur Modellbildung gegenübergestellt. Zunächst fällt auf, dass der CoxBoost Algorithmus (aus dem gleichnamigem R-Paket) in allen betrachteten Fällen deutlich langsamer ist als das Lasso Verfahren und die Ridge Regression (beide aus dem *glmnet* Paket). Die größte Differenz zeigt sich bei den kleinsten Anzahlen für Kovariablen und Beobachtungen (Uppsala Kohorte in Verbindung mit den klinischen Variablen); hier benötigt das CoxBoost Verfahren in etwa das 30-fache der Zeit. Leicht geringer ist dieser Unterschied für die größere Köln Kohorte. Die Differenz der Rechenzeit nimmt ab, je mehr Kovariablen für die Modellbildung verwendet werden. Hier zeigt sich zudem ein Unterschied zwischen dem Lasso Verfahren und der Ridge Regression. Im äußersten Fall (Köln Kohorte mit genetischen Variablen) benötigt Lasso etwa das 4-fache der Zeit im Vergleich zu Ridge. Der CoxBoost Algorithmus nimmt in Bezug zu Lasso in etwa das 6-fache der Zeit in Anspruch.

6 Zusammenfassung und Diskussion

Die Heterogenität klinischer und vor allem genetischer Merkmale innerhalb vieler Krebspatientenkohorten steht im Widerspruch zur Annahme eines einheitlichen Überlebenszeitmodells für alle Patienten. Es ist somit in Frage zu stellen, ob und inwieweit ein bestimmter Zusammenhang zwischen einem bzw. mehreren Merkmalen und der Zielvariablen innerhalb der gesamten Patientenkohorte Gültigkeit besitzt. Vielmehr liegt die Vermutung nahe, dass gewisse Effekte lediglich in einer Teilmenge (Untergruppe) der Kohorte vorzufinden sind.

In dieser Arbeit wurden Untersuchungen durchgeführt, für bestimmte Untergruppen von Lungenkrebspatienten ein optimales Vorhersagemodell für die Überlebenswahrscheinlichkeit zu identifizieren. Dieses Vorgehen unterscheidet sich von der üblichen Verfahrensweise ein Modell für die gesamte Kohorte zu finden. Erstmals wurde hier die Vorhersagegenauigkeit von Überlebenszeitmodellen mit klinischen und genetischen Merkmalen kreuzvalidiert für diverse Untergruppen mit dem Modell für alle Patienten verglichen. Die im Zuge der Kooperation mit dem *Clinical Lung Cancer Genome Project* zur Verfügung stehende weltweit größte Lungenkrebskohorte (Köln Kohorte) hat diesen Ansatz möglich gemacht. Zur weiteren Validierung der Ergebnisse stand eine zweite unabhängige Kohorte aus der Kooperation mit der Universität Uppsala (Schweden) bereit.

Zunächst wurden die klinischen und genetischen Daten der beiden Kohorten aus Köln und Uppsala aufbereitet. Für beide Kohorten standen neben fünf klinischen Variablen jeweils Daten aus einer Array-CGH Chip Verarbeitung der Firma Affymetrix zur Verfügung. Eine entsprechende einheitliche Vorverarbeitung dieser genetischen Daten setzte den Grundstein zur Validierung der Ergebnisse zwischen den beiden Kohorten.

Ein erster Schwerpunkt dieser Arbeit und wichtiger Eckstein für die Analyse von Untergruppen mit hochdimensionalen Daten war die Einführung eines neuen Modells, welches speziell die Überlebenswahrscheinlichkeit von Patienten einer Untergruppe modelliert und mit Hilfe aller Patienten geschätzt werden kann. Dies ermöglicht auch für kleinere Untergruppen ein geeignetes Modell zu finden. Dieses Modell nutzt ähnlich wie lokale Regressionsverfahren eine gewichtete Likelihood. Es unterscheidet sich von solchen Verfahren durch seine theoretische Herleitung und einfache Interpretation. Die zugrundeliegende Idee der dabei genutzten gewichteten Likelihood stammt ursprünglich von Bickel (2009) und wurde hier in den Kontext der Überlebenszeitanalyse übertragen und das Verfahren wurde entsprechend adaptiert. Jede Beobachtung (Patient) trägt dabei mit einem für sie individuellen Gewicht zur Modellbildung bei. Dieses Gewicht ist umso höher, je wahrscheinlicher die entsprechende Beobachtung denjenigen aus der Untergruppe entspricht. Diese individuellen Stichprobengewichte wurden mit der in Abschnitt 4.1.4 beschriebenen logistischen Regression geschätzt.

Ein weiterer entscheidender Schwerpunkt war der Aufbau sowie die Durchführung ei-

nes statistischen Designs auf Basis einer erschöpfenden Suche zur Erkennung auffälliger Untergruppen. Dafür wurde jede Ausprägung aller zur Verfügung stehenden klinischen Variablen jeweils als Untergruppe betrachtet. Es sollten zum einen relevante Untergruppen identifiziert werden und zum anderen Unterschiede im Hinblick auf die verwendeten Modelle aufgedeckt werden. Insbesondere der zusätzliche Nutzen der genetischen Merkmale gegenüber den klinischen Merkmalen in den Modellen sollte herausgestellt werden. Dazu setzten sich die Kovariablen der Modelle aus unterschiedlichen Mengen zusammen. Hierbei wurden sowohl ausschließlich klinische bzw. genetische Merkmale als auch beide Arten gemeinsam berücksichtigt. Zudem wurde der Einfluss der Modellbildungsalgorithmen untersucht. In dem verwendeten Design wurden der CoxBoost Algorithmus, die Lasso und die Ridge Regression herangezogen. Sämtliche Modelle wurden dabei jeweils über ein Bootstrapverfahren mit 400 zufällig ausgewählten Trainingsdaten geschätzt und anschließend auf den entsprechenden Testdaten evaluiert (siehe Kapitel 4.2).

Neben dem Modell auf Basis einer gewichteten Likelihood aus Kapitel 4.1 wurden weitere Modelle untersucht und mit diesem verglichen. Es wurde ein auf allen Patienten der Trainingsmenge gelerntes Überlebenszeitmodell betrachtet. Für dieses Modell wurde somit die Untergruppenzugehörigkeit der Patienten außer Acht gelassen. Die Vorhersagegenauigkeit wurde jedoch stets auf Testdaten mit Patienten einer bestimmten Untergruppe gemessen und mit dem gewichteten Modell verglichen. Weiterhin wurde ein Modell ausschließlich unter Verwendung der Patienten der Untergruppe trainiert und ebenfalls den anderen Modellen unter Berücksichtigung der Vorhersagequalität gegenübergestellt. Schließlich wurden Modelle für die Untergruppe erstellt, indem die übrigen, nicht zu der Untergruppe gehörenden Patienten mit einem konstanten Gewicht in die Modellbildung einfließen (siehe Abschnitt 4.1.2).

Als Vergleichskriterium der Modelle diente der Brier Score bzw. der über die Zeit integrierte Brier Score. Der Brier Score vergleicht zu jedem Zeitpunkt die durch ein Modell geschätzte Überlebensfunktion mit der tatsächlich bei einem Patienten beobachteten Überlebensfunktion. Die Brier Scores der Modelle wurden sowohl untereinander als auch mit dem Brier Score des Kaplan-Meier-Schätzers verglichen. Der Brier Score des Kaplan-Meier-Schätzers wurde für jede Untergruppe ermittelt und stellte den Referenzwert für die betrachteten Modelle dar.

Die Analyse der Untergruppen in den beiden Patientenkohorten zeigte eine hohe Variabilität der Vorhersagefehler zwischen den einzelnen Aufteilungen in Trainings- und Testmengen. Die Verwendung der genetischen Variablen allein erzeugte generell keine sinnvollen Modelle. Bei den klinischen Variablen zeigten sich beispielsweise für das Alter oder das Tumorstadium die zu erwartenden Effekte. Die genetischen Faktoren konnten gegebenenfalls in Verbindung mit den klinischen Variablen die Vorhersage der Überlebenszeiten verbessern. In der deutlich kleineren Uppsala Kohorte konnte für keine Untergruppe ein für die jeweilige Gruppe charakteristisches Überlebenszeitmodell angepasst werden. Alle Modelle, wie auch das Modell auf der gesamten Kohorte, zeigten keine wesentlichen Verbesserungen gegenüber dem Kaplan-Meier-Schätzer. In der weitaus größeren Köln Kohorte konnten hingegen leichte Unterschiede in den Überlebenszeitmodellen ausgemacht werden. Auffälligste Gruppe ist die der großzelligen Tumoren. Für diese Untergruppe generierte das Verfahren über die individuellen Stichprobengewichte

die mit Abstand besten Vorhersagen. Hierbei gingen klinische und genetische Variablen gemeinsam in die Modellbildung ein. Die übrigen Modelle, darunter das rein auf der Untergruppe trainierte Modell sowie die Modelle unter Verwendung einer konstanten Gewichtung, lieferten keine brauchbaren Vorhersagen. Allerdings kann nur von einem schwachen Effekt der genetischen Variablen auf die Überlebenszeit ausgegangen werden. Die Stabilitätsanalyse des Modells zeigte, dass in weniger als der Hälfte der erzeugten Trainingsmengen dieselbe genetische Variable in das Modell aufgenommen wurden. Allerdings ist diese Art der Stabilitätsanalyse problematisch, sollten viele der Variablen untereinander korreliert sein und so jeweils abwechselnd in den Modellen enthalten sein.

Neben der Untergruppenanalyse auf den beiden Patientenkohorten wurden die Modelle zudem auf simulierten Daten evaluiert. Dabei wurden unterschiedlich starke Effekte der Kovariablen auf die Überlebenszeit innerhalb einer Untergruppe sowie der dazugehörigen Restgruppe erzeugt. Unter Variation der Gesamtstichprobengröße und der Untergruppengröße sowie der Stärke der Effekte wurden verschiedenste Szenarios erstellt (siehe Abschnitt 4.2.2). Die Vor- und Nachteile der verschiedenen Modelle konnten durch diese Simulation aufgezeigt werden, indem die Überlegenheit einzelner Modelle in bestimmten Szenarien veranschaulicht wurde. Es zeigte sich, dass selbst unter einfachen Rahmenbedingungen diverse Szenarien denkbar sind, so dass eines der Modelle den jeweils anderen vorzuziehen wäre. Insbesondere wurde ein Szenario skizziert, in dem ausschließlich das über die individuellen Gewichte erstellte Überlebenszeitmodell geeignete Vorhersagen liefert, wohingegen die sonstigen Modelle nicht besser abschneiden als das Referenzmodell (Nullmodell) in Form des Kaplan-Meier-Schätzers.

Die in dieser Arbeit aufgeführte Verfahrensweise zur Entdeckung und Beurteilung unterschiedlicher Überlebenszeitmodelle für bestimmte Patientenuntergruppen bietet an vielen Stellen die Möglichkeit zur Optimierung oder Ergänzung bzw. Erweiterung. Im Folgenden werden diesbezüglich weitere Aussichten diskutiert.

Andere Untergruppen können untersucht werden. Die Analyse muss sich dabei keineswegs auf klinische Variablen beschränken. Untergruppen können beispielsweise ebenso durch genetische Variablen charakterisiert sein. Als Untergruppe könnten daraufhin Patienten mit einer bestimmten Mutation angesehen werden. Auch die Kombination mehrerer Untergruppenvariablen ist denkbar. Aus einer Dreierkombination heraus könnten etwa junge rauchende Männer separat untersucht werden. Daneben müssen sich die Gruppen nicht zwangsläufig aus beobachtbaren Variablen zusammensetzen. Zum Beispiel können über eine Clusteranalyse neue Untergruppen gefunden werden. Diese ließen sich über ein entsprechendes Cluster Profiling ebenfalls interpretieren. Hierdurch könnten relevante Untergruppen berücksichtigt werden, die sich aus mehreren Faktoren zusammensetzen. Binder u. a. (2012) haben ein vergleichbares Verfahren für eine Fall-Kontroll-Studie bei Harnblasenkrebspatienten mit SNP Daten angewandt. Sie stellen ein Clusterverfahren namens CLR (cluster-localized regression) vor, welches die zu untersuchenden Patientenuntergruppen definiert. Die Vorhersage konnte durch ihr Modell verbessert werden. Allerdings betrachten sie die Vorhersagegenauigkeit stets auf der gesamten Kohorte. Dies stellt einen wesentlichen Unterschied zu der hier gemachten

Vorgehensweise dar.

Die Analyse immer weiterer Untergruppen zieht viele multiple Vergleiche nach sich, wodurch eine entsprechende Adjustierung unumgänglich wird. Alternativ könnten im Vorfeld bestimmte potentiell relevante Untergruppen für die weitere Analyse ausgewählt werden.

Die für das in Kapitel 4.1 vorgestellte Verfahren benötigten individuellen Stichprobengewichte könnten zusätzlich in einer gesonderten Kreuzvalidierungsschleife optimiert werden. Die Gewichte würden somit nicht in demselben Trainingsschritt wie die darauf aufbauenden Überlebenszeitmodelle ermittelt und einer Überanpassung an die Trainingsdaten würde entgegengewirkt. Ein derartiges Vorgehen würde allerdings die Stichprobengröße für jede einzelne Phase der Modellbildung und Evaluierung weiter reduzieren, was hinsichtlich der von vornherein kleinen Untergruppengrößen nachteilig ist. Bei der jetzigen Strategie, die Gewichte und das Überlebenszeitmodell in einem Trainingsschritt zu schätzen wird dafür ein Optimismus bei der Modellbildung auf den Trainingsdaten in Kauf genommen. Der Vorhersagefehler auf den Testdaten kann hierdurch gegebenenfalls größer werden. Somit ist es möglich, dass die hier vorgestellten Verfahren in Wahrheit besser sind als in Kapitel 5 zu erkennen ist. Neben dem individuellen Gewicht kann ebenso das in Abschnitt 4.1.2 beschriebene konstante Gewicht separat optimiert werden. Es ließe sich ferner erweitern und ein spezielles Gewicht für jede Ausprägung der Untergruppenvariablen bestimmen. Über diese Überlegungen hinaus können die Stichprobengewichte über alternative Verfahren ermittelt werden. Beispielsweise ließen sich bei der zuvor skizzierten Anwendung einer Clusteranalyse die Gewichte aus einem entsprechenden Abstandsmaß ableiten. Ebenfalls ließen sich nach biologischen Kriterien ähnliche Gruppen identifizieren.

Die Wahl geeigneter Modelle bzw. Modellbildungsverfahren kann sowohl für die Schätzung der Stichprobengewichte als auch für die Vorhersage der Überlebenswahrscheinlichkeit von entscheidender Bedeutung sein, wodurch es sich lohnt über eventuelle Verbesserungen oder Alternativen nachzudenken. Ein Ansatz, auf dem jedes hier aufgestellte Modell beruht, ist die Regularisierung der Regressionskoeffizienten (vgl. Kapitel 3.2). Statt über einen Strafterm in der Likelihood die Anzahl der in dem Modell enthaltenen Kovariablen zu kontrollieren könnten a-priori bestimmte Variablen ausgewählt werden. Oft werden diejenigen Variablen selektiert, welche eine hohe Varianz der beobachteten Werte aufweisen oder Variablen, die nach biologischem Vorwissen relevant sind. Darüber hinaus können weitere Modellbildungsverfahren verwendet und die Ergebnisse miteinander verglichen werden. Beispielsweise könnte ein Vergleich mit Elastischen Netzen (Hastie u. a., 2009), Survival Trees (Zhang und Singer, 2010), Logic Regression (Ruczinski u. a., 2003) oder Supervised PCR (Bair u. a., 2006) durchgeführt werden. Besonders im Hinblick auf kategorielle Variablen, wie sie bei klinischen Variablen häufig vorzufinden sind, könnte sich die Verwendung der Grouped Lasso Methode (siehe Bühlmann und van de Geer, 2011, Kapitel 4) bezahlt machen. Hierbei werden die einzelnen Dummy-Variablen ausschließlich gebündelt in das Modell aufgenommen bzw. ausgelassen.

Neben der Modellbildung spielt ebenso die Bewertung der Vorhersagen eine wichtige Rolle. Die Entwicklung eines statistischen Tests auf Unterschiede zwischen den Brier Scores in zwei oder mehr Stichproben könnte ein interessanter zukünftiger Forschungs-

schwerpunkt sein. Hierbei sind mehrere Schwierigkeiten zu umgehen. Über das in dieser Arbeit angewandte stratifizierte Subsampling wird die tatsächliche Variabilität nicht exakt widergespiegelt. Da die Testmengen nicht disjunkt voneinander sind, wird die Varianz unterschätzt. Außerdem müsste der Unterschied bezüglich der Fallzahl zwischen einem Modell auf Basis aller Patienten und einem Modell auf Basis einer Teilmenge (Untergruppe) der Daten mit berücksichtigt werden. Alternativ zum Gebrauch des Brier Scores kann beispielsweise der Concordance Index (Harrell u. a., 1996) betrachtet werden. Dieser ermöglicht hingegen nicht, wie beim Brier Score die Vorhersagen zu jedem Zeitpunkt zu bewerten. Andererseits sollte die Darstellungsform des Brier Scores überdacht werden. Unterschiede zu relativ frühen Zeitpunkten (Brier Score nahe bei Null) können per bloßem Hinschauen weniger gut erkannt werden als Unterschiede auf der Höhe der medianen Überlebenszeit. Hierzu ließe sich der Brier Score auf eine horizontale Gerade projizieren.

Ein weiterer Gesichtspunkt im Hinblick auf die Entdeckung relevanter Untergruppen ist die Frage nach der benötigten Stichprobengröße, um eine relativ kleine bedeutsame Untergruppe erkennen zu können. Dazu bedarf es geeigneter Simulationsstudien. Schlussendlich ist die Analyse diverser Untergruppen keineswegs auf den Kontext der Überlebenszeit der Patienten beschränkt. Andere Zielvariablen wie etwa die Resistenz gegenüber potentiellen Wirkstoffen sind ebenfalls denkbar. Auch sollte die Verwendung alternativer genetischer Daten wie zum Beispiel Expressionsdaten in Erwägung gezogen werden.

Literaturverzeichnis

[Affymetrix 2006] Affymetrix: *Affymetrix GeneChip Human Mapping 500K Array Set Data Sheet.* 2006

[Affymetrix 2009] Affymetrix: *Affymetrix Genome-Wide Human SNP Array 6.0 Data Sheet.* 2009

[Altman 1992] Altman, Naomi S.: An introduction to kernel and nearest-neighbor nonparametric regression. In: *The American Statistician* 46 (1992), Nr. 3, S. 175–185

[Atkeson u. a. 1997] Atkeson, Christopher G. ; Moore, Andrew W. ; Schaal, Stefan: Locally Weighted Learning. In: *Artificial Intelligence Review* 11 (1997), S. 11–73

[Bair u. a. 2006] Bair, Eric ; Hastie, Trevor ; Paul, Debashis ; Tibshirani, Robert: Prediction by Supervised Principal Components. In: *Journal of the American Statistical Association* 101 (2006), März, Nr. 473, S. 119–137

[Beroukhim u. a. 2007] Beroukhim, Rameen ; Getz, Gad ; Nghiemphu, Leia ; Barretina, Jordi ; Hsueh, Teli ; Linhart, David ; Vivanco, Igor ; Lee, Jeffrey C. ; Huang, Julie H. ; Alexander, Sethu ; Du, Jinyan ; Kau, Tweeny ; Thomas, Roman K. ; Shah, Kinjal ; Soto, Horacio ; Perner, Sven ; Prensner, John ; Debiasi, Ralph M. ; Demichelis, Francesca ; Hatton, Charlie ; Rubin, Mark a. ; Garraway, Levi a. ; Nelson, Stan F. ; Liau, Linda ; Mischel, Paul S. ; Cloughesy, Tim F. ; Meyerson, Matthew ; Golub, Todd a. ; Lander, Eric S. ; Mellinghoff, Ingo K. ; Sellers, William R.: Assessing the Significance of Chromosomal Aberrations in Cancer: Methodology and Application to Glioma. In: *Proceedings of the National Academy of Sciences of the United States of America* 104 (2007), Dezember, Nr. 50, S. 20007–12

[Bhatia u. a. 2012] Bhatia, Sangeeta ; Frangioni, John V. ; Hoffman, Robert M. ; Iafrate, A J. ; Polyak, Kornelia: The Challenges Posed by Cancer Heterogeneity. In: *Nature Biotechnology* 30 (2012), Juli, Nr. 7, S. 604–10

[Bianchi u. a. 2007] Bianchi, Fabrizio ; Nuciforo, Paolo ; Vecchi, Manuela ; Bernard, Loris ; Tizzoni, Laura ; Marchetti, Antonio ; Buttitta, Fiamma ; Felicioni, Lara ; Nicassio, Francesco ; Di Fiore, Pier P.: Survival Prediction of Stage I Lung Adenocarcinomas by Expression of 10 Genes. In: *Journal of Clinical Investigation* 117 (2007), November, Nr. 11, S. 3436–44

[Bickel 2009] Bickel, Steffen: *Learning under Differing Training and Test Distributions*, Universität Potsdam, Dissertation, 2009

[Bickel u. a. 2008] Bickel, Steffen ; Bogojeska, Jasmina ; Lengauer, Thomas ; Scheffer, Tobias: Multi-Task Learning for HIV Therapy Screening. In: *Proceedings of the 25th International Conference on Machine Learning - ICML '08*. New York, NY, USA : ACM Press, 2008, S. 56–63. – ISBN 9781605582054

[Binder 2011] Binder, Harald: *CoxBoost: Cox Models by Likelihood Based Boosting for a Single Survival Endpoint or Competing Risks*, 2011. – URL `http://cran.r-project.org/package=CoxBoost`

[Binder u. a. 2012] Binder, Harald ; Müller, Tina ; Schwender, Holger ; Golka, Klaus ; Steffens, Michael ; Hengstler, Jan G. ; Ickstadt, Katja ; Schumacher, Martin: Cluster-Localized Sparse Logistic Regression for SNP Data. In: *Statistical Applications in Genetics and Molecular Biology* 11 (2012), Januar, Nr. 4

[Binder u. a. 2011] Binder, Harald ; Porzelius, Christine ; Schumacher, Martin: An Overview of Techniques for Linking High-Dimensional Molecular Data to Time-to-Event Endpoints by Risk Prediction Models. In: *Biometrical Journal* 53 (2011), Nr. 2, S. 170–189

[Binder und Schumacher 2008] Binder, Harald ; Schumacher, Martin: Allowing for Mandatory Covariates in Boosting Estimation of Sparse High-Dimensional Survival Models. In: *BMC Bioinformatics* 9 (2008), Januar, S. 14

[Bischl u. a. 2012a] Bischl, Bernd ; Lang, Michel ; Mersmann, Olaf: *BatchExperiments: Statistical Experiments on Batch Computing Clusters*, 2012. – URL `http://cran.r-project.org/package=BatchExperiments`

[Bischl u. a. 2012b] Bischl, Bernd ; Lang, Michel ; Mersmann, Olaf: *BatchJobs: Batch Computing with R*, 2012. – URL `http://cran.r-project.org/package=BatchJobs`

[Bischl u. a. 2012c] Bischl, Bernd ; Lang, Michel ; Mersmann, Olaf ; Rahnenführer, Jörg ; Weihs, Claus: Computing on High Performance Clusters with R: Packages BatchJobs and BatchExperiments / SFB 876, TU Dortmund University. 2012. – Forschungsbericht

[Bogojeska u. a. 2010] Bogojeska, Jasmina ; Bickel, Steffen ; Altmann, André ; Lengauer, Thomas: Dealing with Sparse Data in Predicting Outcomes of HIV Combination Therapies. In: *Bioinformatics* 26 (2010), September, Nr. 17, S. 2085–92

[Bogojeska und Lengauer 2012] Bogojeska, Jasmina ; Lengauer, Thomas: Hierarchical Bayes Model for Predicting Effectiveness of HIV Combination Therapies. In: *Statistical Applications in Genetics and Molecular Biology* 11 (2012), Januar, Nr. 3, S. Article 11

[Bøvelstad u. a. 2007] Bøvelstad, Hege M. ; Nygård, Ståge ; Størvold, H L. ; Aldrin, Magne ; Borgan, Ørnulf ; Frigessi, Arnoldo ; Lingjaerde, Ole C.: Predicting Survival from Microarray Data–A Comparative Study. In: *Bioinformatics* 23 (2007), August, Nr. 16, S. 2080–7

[Brambilla u. a. 2001] Brambilla, E ; Travis, William D. ; Colby, T V. ; Corrin, B ; Shimosato, Y: The New World Health Organization Classification of Lung Tumours. In: *European Respiratory Journal* 18 (2001), Dezember, Nr. 6, S. 1059–1068

[Breslow 1974] Breslow, Norman E.: Covariance Analysis of Censored Survival Data. In: *Biometrics* 30 (1974), März, Nr. 1, S. 89–99

[Brier 1950] Brier, Glenn W.: Verification of Forecasts Expressed in Terms of Probablity. In: *Monthly Weather Review* 78 (1950), Januar, Nr. 1, S. 1–3

[Bühlmann und van de Geer 2011] Bühlmann, Peter ; van de Geer, Sara: *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin, Heidelberg : Springer, 2011 (Springer Series in Statistics). – 556 S. – ISBN 978-3-642-20191-2

[Cetin u. a. 2011] Cetin, Karynsa ; Ettinger, David S. ; Hei, Yong-Jiang ; O'Malley, Cynthia D.: Survival by Histologic Subtype in Stage IV Non-Small Cell Lung Cancer Based on Data from the Surveillance, Epidemiology and End Results Program. In: *Clinical Epidemiology* 3 (2011), Januar, S. 139–48

[Chiaretti u. a. 2004] Chiaretti, Sabina ; Li, Xiaochun ; Gentleman, Robert ; Vitale, Antonella ; Vignetti, Marco ; Mandelli, Franco ; Ritz, Jerome ; Foa, Robin: Gene Expression Profile of Adult T-cell Acute Lymphocytic Leukemia Identifies Distinct Subsets of Patients with Different Response to Therapy and Survival. In: *Blood* 103 (2004), April, Nr. 7, S. 2771–8

[Cleveland 1979] Cleveland, William S.: Robust Locally and Smoothing Weighted Regression Scatterplots. In: *Journal of the American Statistical Association* 74 (1979), Nr. 368, S. 829–836

[Cleveland und Devlin 1988] Cleveland, William S. ; Devlin, Susan J.: Locally Weighted Regression : An Approach to Regression Analysis by Local Fifing. In: *Journal of the American Statistical Association* 83 (1988), Nr. 403, S. 596–610

[Cox 1972] Cox, David R.: Regression Models and Life Tables. In: *Journal of the Royal Statistical Society. Series B* 34 (1972), Nr. 2, S. 187–220

[Dijkman u. a. 2009] Dijkman, Bernadette ; Kooistra, Bauke ; Bhandari, Mohit: How to Work with a Subgroup Analysis. In: *Canadian Journal of Surgery* 52 (2009), Dezember, Nr. 6, S. 515–22

[Edlund 2012] Edlund, Karolina: *Molecular Characterisation and Prognostic Biomarker Discovery in Human Non-Small Cell Lung Cancer*, Uppsala Universitet, Dissertation, 2012

[Efron 1977] Efron, Bradley: The Efficiency of Cox's Likelihood Function for Censored Data. In: *Journal of the American Statistical Association* 72 (1977), September, Nr. 359, S. 557–565

[Fawcett 2006] Fawcett, Tom: An Introduction to ROC Analysis. In: *Pattern Recognition Letters* 27 (2006), Juni, Nr. 8, S. 861–874

[Friedman u. a. 2010] Friedman, Jerome ; Hastie, Trevor ; Tibshirani, Robert: Regularization Paths for Generalized Linear Models via Coordinate Descent. In: *Journal of Statistical Software* 33 (2010), Nr. 1, S. 1–22

[Garber u. a. 2001] Garber, Mitchell E. ; Troyanskaya, Olga G. ; Schluens, Karsten ; Petersen, Simone ; Thaesler, Zsuzsanna ; Pacyna-Gengelbach, Manuela ; van de Rijn, Matt ; Rosen, Glenn D. ; Perou, Charles M. ; Whyte, Richard I. ; Altman, Russ B. ; Brown, Patrick O. ; Botstein, David ; Petersen, Iver: Diversity of Gene Expression in Adenocarcinoma of the Lung. In: *Proceedings of the National Academy of Sciences of the United States of America* 98 (2001), November, Nr. 24, S. 13784–9

[Gelman u. a. 2003] Gelman, Andrew ; Carlin, John B. ; Stern, Hal S. ; Rubin, Donald B.: *Bayesian Data Analysis*. 2. Boca Raton, FL, USA : Chapman and Hall/CRC, 2003. – 696 S. – ISBN 1-58488-388-X

[George und McCulloch 1993] George, Edward I. ; McCulloch, Robert E.: Variable Selection via Gibbs Sampling. In: *Journal of the American Statistical Association* 88 (1993), September, Nr. 423, S. 881

[Gerds und Schumacher 2006] Gerds, Thomas A. ; Schumacher, Martin: Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. In: *Biometrical Journal* 48 (2006), Nr. 6, S. 1029–1040

[Govindan u. a. 2012] Govindan, Ramaswamy ; Ding, Li ; Griffith, Malachi ; Subramanian, Janakiraman ; Dees, Nathan D. ; Kanchi, Krishna L. ; Maher, Christopher A. ; Fulton, Robert ; Fulton, Lucinda ; Wallis, John ; Chen, Ken ; Walker, Jason ; McDonald, Sandra ; Bose, Ron ; Ornitz, David ; Xiong, Donghai ; You, Ming ; Dooling, David J. ; Watson, Mark ; Mardis, Elaine R. ; Wilson, Richard K.: Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers. In: *Cell* 150 (2012), September, Nr. 6, S. 1121–34

[Graf u. a. 1999] Graf, Erika ; Schmoor, Claudia ; Sauerbrei, Willi ; Schumacher, Martin: Assessment and Comparison of Prognostic Classification Schemes for Survival Data. In: *Statistics in Medicine* 18 (1999), Nr. 17-18, S. 2529–45

[Hammerschmidt und Wirtz 2009] Hammerschmidt, Stefan ; Wirtz, Hubert: Lung Cancer: Current Diagnosis and Treatment. In: *Deutsches Ärzteblatt international* 106 (2009), Dezember, Nr. 49, S. 809–18; quiz 819–20

[Harrell u. a. 1996] Harrell, Frank E. ; Lee, Kerry L. ; Mark, Daniel B.: Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. In: *Statistics in Medicine* 15 (1996), Nr. 4, S. 361–387

[Hastie u. a. 2009] Hastie, Trevor ; Tibshirani, Robert ; Friedman, Jerome: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. New York, NY, USA : Springer, 2009 (Springer Series in Statistics). – 746 S. – ISBN 978-0-387-84857-0

[Hoerl und Kennard 1970] Hoerl, Arthur E. ; Kennard, Robert W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems. In: *Technometrics* 12 (1970), Februar, Nr. 1, S. 55

[Ibrahim u. a. 1999] Ibrahim, Joseph G. ; Chen, Ming-Hui ; Maceachern, Steven N.: Bayesian Variable Selection for Proportional Hazards Models. In: *Canadian Journal of Statistics* 27 (1999), Nr. 4, S. 701–717

[Ibrahim u. a. 2001] Ibrahim, Joseph G. ; Chen, Ming-Hui ; Sinha, Debajyoti: *Bayesian Survival Analysis*. New York, NY, USA : Springer, 2001 (Springer Series in Statistics). – 481 S. – ISBN 0-387-95277-2

[Jung u. a. 2003] Jung, Yunjae ; Park, Haesun ; Du, Ding-Zhu ; Drake, Barry L.: A Decision Criterion for the Optimal Number of Clusters in Hierarchical Clustering. In: *Journal of Global Optimization* (2003), S. 1–22

[Kalbfleisch und Prentice 2002] Kalbfleisch, John D. ; Prentice, Ross L.: *The Statistical Analysis of Failure Time Data*. 2. Hoboken, NJ, USA : John Wiley & Sons, 2002. – 462 S. – ISBN 978-0-471-36357-6

[Kammers u. a. 2011] Kammers, Kai ; Lang, Michel ; Hengstler, Jan G. ; Schmidt, Marcus ; Rahnenführer, Jörg: Survival Models with Preclustered Gene Groups as Covariates. In: *BMC Bioinformatics* 12 (2011), Januar, S. 478

[Kaplan und Meier 1958] Kaplan, Edward L. ; Meier, Paul: Nonparametric Estimation from Incomplete Observations. In: *Journal of the American Statistical Association* 53 (1958), Juni, Nr. 282, S. 457–481

[Klein und Moeschberger 2003] Klein, John P. ; Moeschberger, Melvin L.: *Survival Analysis: Techniques for Censored and Truncated Data*. 2. New York, NY, USA : Springer, 2003 (Statistics for Biology and Health). – 536 S. – ISBN 978-0-387-95399-1

[Loader 1999] Loader, Clive: *Local Regression and Likelihood*. New York : Springer, 1999. – 290 S

[Mavaddat u. a. 2010] Mavaddat, Nasim ; Antoniou, Antonis C. ; Easton, Douglas F. ; Garcia-Closas, Montserrat: Genetic Susceptibility to Breast Cancer. In: *Molecular Oncology* 4 (2010), Juni, Nr. 3, S. 174–91

[Micke u. a. 2011] Micke, Patrick ; Edlund, Karolina ; Holmberg, Lars ; Kultima, Hanna G. ; Mansouri, Larry ; Ekman, Simon ; Bergqvist, Michael ; Scheibenflug, Lena ; Lamberg, Kristina ; Myrdal, Gunnar ; Berglund, Anders ; Andersson, Annsofie ; Lambe, Mats ; Nyberg, Fredrik ; Thomas, Andrew ; Isaksson, Anders ; Botling, Johan: Gene Copy Number Aberrations Are Associated with Survival in Histologic Subgroups of Non-Small Cell Lung Cancer. In: *Journal of Thoracic Oncology* 6 (2011), November, Nr. 11, S. 1833–40

[Mogensen u. a. 2012] Mogensen, Ulla B. ; Ishwaran, Hemant ; Gerds, Thomas A.: Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. In: *Journal of Statistical Software* 50 (2012), Nr. 11, S. 1–23

[Netzer und Rahnenführer 2012] Netzer, Christian ; Rahnenführer, Jörg: Sample Size Estimation for Cancer Progression Models. In: *International Journal of Computational Bioscience* 1 (2012), Nr. 1

[Olshen u. a. 2004] Olshen, Adam B. ; Venkatraman, E S. ; Lucito, Robert ; Wigler, Michael: Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data. In: *Biostatistics* 5 (2004), Oktober, Nr. 4, S. 557–72

[Prat u. a. 2013] Prat, Aleix ; Adamo, Barbara ; Cheang, Maggie C U. ; Anders, Carey K. ; Carey, Lisa A. ; Perou, Charles M.: Molecular Characterization of Basal-like and Non-Basal-Like Triple-Negative Breast Cancer. In: *Oncologist* 18 (2013), Januar, Nr. 2, S. 123–33

[Pschyrembel 2007] Pschyrembel, Willibald: *Pschyrembel Klinisches Wörterbuch.* 261. Berlin : Walter de Gruyter, 2007

[R Core Team 2012] R Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing (Veranst.), 2012. – URL `http://www.r-project.org/`

[Robert Koch-Institut 2012] Robert Koch-Institut: *Krebs in Deutschland 2007/2008*. Bd. 8. 8. Berlin : Robert Koch-Institut und die Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V., 2012. – ISBN 978-3-89606-214-7

[Rosti u. a. 2006] Rosti, G ; Bevilacqua, G ; Bidoli, P ; Portalone, L ; Santo, A ; Genestreti, G: Small Cell Lung Cancer. In: *Annals of Oncology* 17 Suppl 2 (2006), März, S. ii5–10

[Ruczinski u. a. 2003] Ruczinski, Ingo ; Kooperberg, Charles ; LeBlanc, Michael: Logic Regression. In: *Journal of Computational and Graphical Statistics* 12 (2003), September, Nr. 3, S. 475–511

[Ruppert und Wand 1994] Ruppert, D ; Wand, M P.: Multivariate Locally Weighted Least Squares Regression. In: *The Annals of Statistics* 22 (1994), Nr. 3, S. 1346–1370

[Russell u. a. 2011] Russell, Prudence A. ; Wainer, Zoe ; Wright, Gavin M. ; Daniels, Marissa ; Conron, Matthew ; Williams, Richard A.: Does Lung Adenocarcinoma Subtype Predict Patient Survival?: A Clinicopathologic Study Based on the New International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society International Multidisciplinary Lung Adenocarcinoma Classification. In: *Journal of Thoracic Oncology* 6 (2011), September, Nr. 9, S. 1496–504

[Silverman und Jones 1989] Silverman, B. W. ; Jones, M. C.: E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). In: *International Statistical Review / Revue Internationale de Statistique* 57 (1989), Nr. 3, S. pp. 233–238

[Simon u. a. 2011] Simon, Noah ; Friedman, Jerome ; Hastie, Trevor ; Tibshirani, Robert: Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. In: *Journal of Statistical Software* 39 (2011), Nr. 5, S. 1–13

[Sing u. a. 2005] Sing, Tobias ; Sander, Oliver ; Beerenwinkel, Niko ; Lengauer, Thomas: ROCR: Visualizing Classifier Performance in R. In: *Bioinformatics* 21 (2005), Oktober, Nr. 20, S. 3940–1

[Sing u. a. 2012] Sing, Tobias ; Sander, Oliver ; Beerenwinkel, Niko ; Lengauer, Thomas: *ROCR: Visualizing the Performance of Scoring Classifiers*, 2012. – URL `http://cran.r-project.org/package=ROCR`

[Sinha u. a. 2003] Sinha, Debajyoti ; Ibrahim, Joseph G. ; Chen, Ming-Hui: A Bayesian Justification of Cox's Partial Likelihood. In: *Biometrika* 90 (2003), September, Nr. 3, S. 629–641

[Sobin und Compton 2010] Sobin, Leslie H. ; Compton, Carolyn C.: TNM Seventh Edition: What's New, What's Changed: Communication from the International Union Against Cancer and the American Joint Committee on Cancer. In: *Cancer* 116 (2010), November, Nr. 22, S. 5336–9

[Solinas-Toldo u. a. 1997] Solinas-Toldo, S ; Lampel, S ; Stilgenbauer, S ; Nickolenko, J ; Benner, A ; Döhner, H ; Cremer, T ; Lichter, P: Matrix-Based Comparative Genomic Hybridization: Biochips to Screen for Genomic Imbalances. In: *Genes, Chromosomes & Cancer* 20 (1997), Dezember, Nr. 4, S. 399–407

[Subramanian und Govindan 2007] Subramanian, Janakiraman ; Govindan, Ramaswamy: Lung Cancer in Never Smokers: A Review. In: *Journal of Clinical Oncology* 25 (2007), Februar, Nr. 5, S. 561–70

[Sun u. a. 2007] Sun, Sophie ; Schiller, Joan H. ; Gazdar, Adi F.: Lung Cancer in Never Smokers–A Different Disease. In: *Nature Reviews: Cancer* 7 (2007), Oktober, Nr. 10, S. 778–90

[Therneau 2012] Therneau, Terry M.: *A Package for Survival Analysis in S*, 2012. – URL `http://cran.r-project.org/package=survival`

[Therneau und Grambsch 2000] Therneau, Terry M. ; Grambsch, Patricia M.: *Modeling Survival Data: Extending the Cox Model.* New York, NY, USA : Springer, 2000. – ISBN 0-387-98784-3

[Tibshirani 1996] Tibshirani, Robert: Regression Shrinkage and Selection via the Lasso. In: *Journal of the Royal Statistical Society. Series B* 58 (1996), Nr. 1, S. 267–288

[Tibshirani 1997] Tibshirani, Robert: The Lasso Method for Variable Selection in the Cox Model. In: *Statistics in Medicine* 16 (1997), Februar, Nr. 4, S. 385–95

[Tibshirani u. a. 2004] Tibshirani, Robert ; Johnstone, Iain ; Hastie, Trevor ; Efron, Bradley: Least Angle Regression. In: *Annals of Statistics* 32 (2004), April, Nr. 2, S. 407–499

[Toloşi 2012] Toloşi, Laura: *Finding Regions of Aberrant DNA Copy Number Associated with Tumor Phenotype*, Universität des Saarlandes, Dissertation, 2012

[Turner und Reis-Filho 2012] Turner, Nicholas C. ; Reis-Filho, Jorge S.: Genetic Heterogeneity and Cancer Drug Resistance. In: *Lancet Oncology* 13 (2012), April, Nr. 4, S. e178–85

[Tutz und Binder 2004] Tutz, Gerhard ; Binder, Harald: Localized Regression / Ludwig-Maximilians-Universität München. 2004. – Forschungsbericht

[Tutz und Binder 2007] Tutz, Gerhard ; Binder, Harald: Boosting Ridge Regression. In: *Computational Statistics & Data Analysis* 51 (2007), August, Nr. 12, S. 6044–6059

[Venkatraman und Olshen 2007] Venkatraman, E S. ; Olshen, Adam B.: A Faster Circular Binary Segmentation Algorithm for the Analysis of Array CGH Data. In: *Bioinformatics* 23 (2007), März, Nr. 6, S. 657–63

[Verweij und Van Houwelingen 1993] Verweij, Pierre J M. ; Van Houwelingen, Hans C.: Cross-Validation in Survival Analysis. In: *Statistics in Medicine* 12 (1993), Dezember, Nr. 24, S. 2305–2314

[Verweij und Van Houwelingen 1994] Verweij, Pierre J M. ; Van Houwelingen, Hans C.: Penalized Likelihood in Cox Regression. In: *Statistics in Medicine* 13 (1994), Dezember, Nr. 23-24, S. 2427–2436

[Yap u. a. 2012] Yap, Timothy A. ; Gerlinger, Marco ; Futreal, P A. ; Pusztai, Lajos ; Swanton, Charles: Intratumor Heterogeneity: Seeing the Wood for the Trees. In: *Science Translational Medicine* 4 (2012), März, Nr. 127, S. 127ps10

[Zhang und Singer 2010] Zhang, Heping ; Singer, Burton H.: *Recursive Partitioning and Applications*. 2. New York, NY, USA : Springer, 2010 (Springer Series in Statistics). – ISBN 978-1-4419-6823-4

[Zhu und Hastie 2004] Zhu, Ji ; Hastie, Trevor: Classification of Gene Microarrays by Penalized Logistic Regression. In: *Biostatistics* 5 (2004), Juli, Nr. 3, S. 427–443

[Zou und Hastie 2005] Zou, Hui ; Hastie, Trevor: Regularization and Variable Selection via the Elastic Net. In: *Journal of the Royal Statistical Society: Series B* 67 (2005), April, Nr. 2, S. 301–320

A Ergänzende Tabellen

Tabelle A.1: Köln Kohorte: Mittlere `AUC` in Prozent

U	g	Modell			
		klinisch		genetisch	
		L1	L2	L1	L2
Alter	jung	66	66	58	62
	mittelalt	62	62	52	51
	alt	64	65	54	54
Geschlecht	männlich	69	70	69	74
	weiblich	69	70	69	74
Histologie	AD	64	64	85	85
	CA	88	88	92	95
	LC	57	56	62	62
	SCLC	62	56	90	93
	SQ	72	72	86	86
	unklassifiziert	58	58	58	64
Raucherstatus	Raucher	66	66	62	60
	Ex-Raucher	58	58	55	54
	Nichtraucher	84	83	81	82
Stadium	I	64	64	54	55
	II	62	63	52	60
	III	67	67	54	56
	IV	71	72	55	56

Tabelle A.2: Uppsala Kohorte: Mittlere `AUC` in Prozent

U	g	Modell			
		klinisch		genetisch	
		L1	L2	L1	L2
Alter	jung	55	70	46	47
	mittelalt	63	63	49	48
	alt	65	65	49	49
Geschlecht	männlich	58	59	52	55
	weiblich	58	59	52	55
Histologie	AD	56	59	85	85
	LC	48	51	53	46
	SQ	57	60	88	91
Raucherstatus	Raucher	55	54	49	48
	Ex-Raucher	54	54	49	48
	Nichtraucher	60	69	52	53
Stadium	I	52	53	50	53
	II	47	48	48	49
	III	54	55	55	62
	IV	48	47	46	51

Tabelle A.3: Köln Kohorte: Mittlerer integrierter Brier Score (IBS) bis zum Zeitpunkt $\min(10, t_{\max})$ für die Modelle SUBGROUP (S), WEIGHTED (W) und ALL (A) sowie für den jeweiligen Kaplan-Meier-Schätzer getrennt nach Untergruppen, Modellbildungsverfahren und verwendeter Kovariablen. In eckigen Klammern ist die jeweilige 2σ-Umgebung auf den 400 Trainings/Test-Aufteilungen angegeben.

U	g	Modell	Methode / Kovariablen: CoxBoost klinisch	CoxBoost genetisch	CoxBoost gemeinsam	Lasso klinisch	Lasso genetisch	Lasso gemeinsam	Ridge klinisch	Ridge genetisch	Ridge gemeinsam	Kaplan-Meier
Alter	jung	S	17.1 [6.5, 27.8]	22.0 [12.9, 31.0]	21.3 [13.9, 28.6]	17.1 [4.8, 29.3]	21.8 [13.2, 30.4]	23.0 [8.4, 37.6]	16.7 [6.0, 27.5]	20.9 [12.3, 29.6]	23.0 [8.2, 37.7]	21.3 [13.9, 28.6]
		W	18.4 [11.0, 25.8]	19.3 [14.8, 23.8]	19.4 [15.4, 23.4]	14.7 [6.2, 23.1]	19.3 [14.9, 23.7]	13.2 [5.3, 21.1]	14.8 [6.9, 22.8]	19.0 [14.4, 23.6]	13.1 [5.4, 20.9]	
		A	18.0 [12.4, 23.7]	19.3 [15.0, 23.6]	19.4 [15.4, 23.4]	14.7 [6.8, 22.5]	19.3 [15.1, 23.5]	14.5 [5.6, 23.4]	15.0 [7.7, 22.2]	19.2 [14.9, 23.4]	14.4 [5.7, 23.2]	19.4 [15.4, 23.4]
	mittelalt	S	18.8 [16.5, 21.0]	20.8 [18.8, 22.9]	20.8 [19.3, 22.4]	18.2 [15.2, 21.2]	20.8 [18.8, 22.9]	18.1 [14.5, 21.7]	18.0 [15.3, 20.7]	20.9 [19.1, 22.6]	18.0 [14.5, 21.6]	20.8 [19.3, 22.4]
		W	18.9 [17.1, 20.6]	21.3 [20.0, 22.5]	21.1 [20.3, 22.0]	17.8 [15.2, 20.4]	21.2 [20.0, 22.5]	18.3 [15.4, 21.2]	17.8 [15.3, 20.4]	21.1 [20.1, 22.1]	18.2 [15.4, 21.1]	
		A	18.9 [17.2, 20.6]	21.3 [20.1, 22.4]	21.1 [20.3, 22.0]	17.6 [15.1, 20.2]	21.2 [20.1, 22.4]	17.6 [14.7, 20.5]	17.7 [15.1, 20.2]	21.1 [20.2, 22.1]	17.6 [14.8, 20.4]	21.1 [20.3, 22.0]
	alt	S	21.2 [18.5, 23.9]	21.7 [19.3, 24.1]	21.5 [19.4, 23.7]	21.4 [17.8, 24.9]	21.6 [19.3, 24.0]	21.6 [16.7, 26.5]	21.6 [18.1, 25.1]	21.6 [19.4, 23.7]	21.6 [16.7, 26.4]	21.5 [19.4, 23.7]
		W	21.0 [18.2, 23.9]	22.7 [20.8, 24.5]	22.6 [20.9, 24.2]	20.8 [16.8, 24.8]	22.6 [20.8, 24.5]	22.3 [18.3, 26.3]	21.0 [17.1, 24.9]	22.5 [20.9, 24.2]	22.2 [18.2, 26.2]	
		A	21.5 [18.8, 24.2]	22.7 [20.9, 24.6]	22.6 [20.9, 24.2]	20.9 [16.8, 25.0]	22.7 [20.8, 24.6]	21.0 [16.6, 25.5]	21.1 [17.1, 25.1]	22.7 [21.0, 24.4]	21.0 [16.5, 25.5]	22.6 [20.9, 24.2]
Geschlecht	männlich	S	21.0 [19.5, 22.4]	22.2 [21.0, 23.5]	22.1 [21.3, 22.9]	19.7 [17.2, 22.1]	22.2 [21.0, 23.5]	19.6 [16.9, 22.4]	19.8 [17.6, 22.1]	22.2 [21.1, 23.2]	19.5 [16.8, 22.3]	22.1 [21.3, 22.9]
		W	20.8 [19.1, 22.6]	22.2 [21.0, 23.3]	22.1 [21.1, 23.2]	19.5 [17.0, 21.9]	22.2 [21.1, 23.3]	19.3 [16.6, 22.1]	19.5 [17.3, 21.8]	22.1 [21.2, 23.1]	19.3 [16.6, 22.0]	
		A	20.7 [18.9, 22.6]	22.3 [21.0, 23.5]	22.1 [21.1, 23.2]	19.4 [16.8, 21.9]	22.2 [21.1, 23.4]	19.3 [16.6, 22.1]	19.5 [17.0, 21.9]	22.2 [21.1, 23.2]	19.3 [16.5, 22.0]	22.1 [21.1, 23.2]
	weiblich	S	16.9 [13.3, 20.5]	20.8 [18.0, 23.7]	20.6 [18.3, 22.9]	16.8 [12.5, 21.2]	20.8 [18.0, 23.6]	17.5 [11.3, 23.7]	16.6 [12.3, 20.8]	20.7 [18.2, 23.2]	17.5 [11.3, 23.7]	20.6 [18.3, 22.9]
		W	17.1 [14.2, 20.0]	20.7 [19.0, 22.3]	20.7 [19.3, 22.0]	15.8 [12.1, 19.6]	20.5 [18.8, 22.3]	15.7 [11.6, 19.8]	15.9 [12.2, 19.5]	20.5 [18.9, 22.1]	15.7 [11.6, 19.7]	
		A	17.6 [15.2, 20.1]	20.7 [19.1, 22.2]	20.7 [19.3, 22.0]	16.3 [12.6, 19.9]	20.7 [19.2, 22.2]	16.3 [12.5, 20.2]	16.3 [12.8, 19.8]	20.7 [19.2, 22.1]	16.3 [12.4, 20.1]	20.7 [19.3, 22.0]
Histologie	AD	S	18.8 [16.4, 21.3]	21.9 [20.5, 23.4]	21.9 [20.7, 23.1]	18.8 [15.7, 22.0]	21.9 [20.4, 23.4]	18.8 [14.8, 22.8]	18.7 [15.7, 21.7]	21.8 [20.5, 23.0]	18.8 [14.8, 22.7]	21.9 [20.7, 23.1]
		W	19.2 [16.9, 21.6]	21.8 [20.5, 23.1]	21.7 [20.6, 22.8]	18.2 [14.8, 21.7]	21.8 [20.4, 23.1]	18.4 [14.7, 22.0]	18.3 [15.0, 21.5]	21.6 [20.4, 22.8]	18.3 [14.7, 21.9]	
		A	19.0 [17.0, 21.1]	21.8 [20.5, 23.1]	21.7 [20.6, 22.8]	18.4 [15.2, 21.6]	21.8 [20.5, 23.1]	18.6 [15.2, 22.1]	18.4 [15.3, 21.6]	21.7 [20.5, 22.9]	18.6 [15.1, 22.0]	21.7 [20.6, 22.8]
	LC	S	22.2 [16.3, 28.1]	23.4 [17.6, 29.2]	22.9 [18.2, 27.5]	21.8 [14.1, 29.4]	23.4 [17.8, 28.9]	21.0 [10.0, 32.0]	20.7 [13.3, 28.2]	22.2 [16.7, 27.8]	20.9 [10.1, 31.7]	22.9 [18.2, 27.5]
		W	21.7 [14.7, 28.7]	22.5 [18.5, 26.5]	22.2 [18.4, 26.0]	19.4 [11.4, 27.4]	22.0 [17.6, 26.5]	16.9 [9.0, 24.8]	19.5 [11.7, 27.2]	21.7 [17.4, 25.9]	17.3 [8.5, 26.1]	
		A	20.9 [14.3, 27.5]	22.2 [18.2, 26.3]	22.2 [18.4, 26.0]	19.0 [10.4, 27.7]	22.3 [18.2, 26.3]	19.0 [9.9, 28.0]	19.2 [10.8, 27.5]	22.2 [18.1, 26.3]	19.0 [10.0, 28.0]	22.2 [18.4, 26.0]
	SCLC	S	24.0 [10.0, 37.9]	22.5 [8.6, 36.3]	21.8 [10.1, 33.6]	24.6 [8.2, 41.0]	23.0 [9.2, 36.9]	36.1 [8.9, 63.3]	24.0 [8.3, 39.6]	22.2 [9.5, 34.9]	36.5 [9.7, 63.3]	21.8 [10.1, 33.6]
		W	20.5 [11.0, 30.0]	21.8 [9.9, 33.6]	21.9 [9.6, 34.1]	20.9 [9.4, 32.4]	20.9 [10.4, 31.3]	21.5 [7.8, 35.2]	19.8 [9.6, 30.1]	19.8 [10.6, 29.0]	22.8 [6.5, 39.0]	
		A	21.7 [9.3, 34.2]	21.8 [9.6, 34.0]	21.9 [9.6, 34.1]	24.0 [8.2, 39.9]	21.8 [9.5, 34.1]	23.4 [5.4, 41.5]	22.1 [8.5, 35.7]	21.6 [9.9, 33.2]	23.4 [5.4, 41.4]	21.9 [9.6, 34.1]
	SQ	S	20.9 [18.5, 23.2]	22.1 [19.9, 24.2]	21.8 [20.2, 23.4]	21.0 [18.2, 23.7]	22.0 [19.8, 24.2]	21.4 [17.6, 25.2]	21.0 [18.8, 23.2]	21.8 [20.2, 23.5]	21.3 [17.7, 24.9]	21.8 [20.2, 23.4]
		W	20.6 [18.3, 22.9]	21.5 [20.1, 22.9]	21.5 [20.2, 22.8]	19.8 [17.1, 22.4]	21.6 [20.0, 23.3]	19.9 [17.1, 22.8]	19.8 [17.2, 22.4]	21.7 [20.2, 23.1]	19.9 [17.1, 22.6]	
		A	20.1 [17.7, 22.4]	21.7 [20.1, 23.3]	21.5 [20.2, 22.8]	19.6 [16.9, 22.3]	21.7 [20.0, 23.3]	19.7 [16.8, 22.6]	19.7 [17.0, 22.3]	21.7 [20.2, 23.3]	19.7 [16.8, 22.5]	21.5 [20.2, 22.8]
	unklassifiziert	S	21.0 [12.0, 30.0]	21.4 [11.3, 31.5]	20.4 [12.2, 28.6]	21.6 [11.4, 31.8]	21.5 [11.2, 31.9]	24.0 [5.5, 42.5]	21.1 [11.9, 30.4]	21.7 [11.7, 31.8]	23.9 [6.0, 41.8]	20.4 [12.2, 28.6]
		W	19.0 [13.3, 24.7]	19.8 [11.8, 27.8]	19.8 [11.8, 27.7]	18.5 [11.3, 25.7]	19.9 [11.8, 28.0]	18.4 [9.9, 26.9]	18.7 [11.7, 25.7]	20.2 [12.0, 28.5]	18.8 [10.1, 27.4]	
		A	19.0 [12.2, 25.7]	19.9 [11.8, 27.9]	19.8 [11.8, 27.7]	18.6 [10.1, 27.2]	19.8 [11.8, 27.8]	19.4 [9.6, 29.1]	19.1 [10.6, 27.5]	19.9 [11.8, 28.0]	19.4 [9.7, 29.1]	19.8 [11.8, 27.7]
Raucherstatus	Raucher	S	19.9 [18.2, 21.6]	22.1 [20.9, 23.4]	22.0 [21.1, 23.0]	19.4 [17.2, 21.5]	22.1 [20.9, 23.4]	19.2 [16.7, 21.7]	19.3 [17.4, 21.2]	22.1 [21.1, 23.2]	19.2 [16.7, 21.6]	22.0 [21.1, 23.0]
		W	19.6 [17.9, 21.2]	22.1 [20.9, 23.3]	22.0 [21.0, 23.0]	18.9 [16.7, 21.0]	22.1 [20.9, 23.3]	19.0 [16.7, 21.2]	18.9 [16.8, 20.9]	22.1 [21.0, 23.2]	18.9 [16.7, 21.2]	
		A	19.8 [18.1, 21.6]	22.2 [20.9, 23.4]	22.0 [21.0, 23.0]	18.8 [16.6, 20.9]	22.1 [20.9, 23.3]	18.8 [16.5, 21.0]	18.7 [16.7, 20.8]	22.1 [21.0, 23.3]	18.7 [16.5, 21.0]	22.0 [21.0, 23.0]
	Ex-Raucher	S	22.9 [18.6, 27.2]	22.3 [18.7, 26.0]	22.1 [18.9, 25.3]	23.6 [17.3, 30.0]	22.3 [18.6, 26.1]	25.7 [14.9, 36.6]	23.9 [17.2, 30.7]	22.2 [18.8, 25.6]	25.6 [15.0, 36.1]	22.1 [18.9, 25.3]
		W	22.7 [17.9, 27.4]	21.4 [19.2, 23.7]	21.3 [19.5, 23.1]	22.3 [14.7, 30.0]	21.4 [19.3, 23.5]	22.5 [14.0, 31.0]	22.6 [15.1, 30.1]	21.4 [19.4, 23.3]	22.4 [13.9, 30.8]	
		A	22.7 [17.8, 27.6]	21.5 [19.3, 23.6]	21.3 [19.5, 23.1]	22.7 [14.9, 30.6]	21.4 [19.3, 23.5]	22.8 [14.3, 31.3]	23.0 [15.4, 30.6]	21.3 [19.4, 23.2]	22.8 [14.3, 31.3]	21.3 [19.5, 23.1]
	Nichtraucher	S	15.0 [9.5, 20.5]	21.2 [16.0, 26.4]	20.9 [16.4, 25.3]	14.4 [7.9, 20.9]	21.2 [15.9, 26.5]	20.6 [14.9, 26.3]	15.0 [9.1, 20.8]	20.9 [16.2, 25.7]	20.5 [14.8, 26.3]	20.9 [16.4, 25.3]
		W	20.6 [15.4, 25.8]	20.0 [17.2, 22.8]	20.4 [18.7, 22.2]	12.8 [7.8, 17.9]	20.2 [17.3, 23.2]	12.3 [7.4, 17.2]	13.3 [8.4, 18.2]	19.7 [16.9, 22.4]	12.3 [7.3, 17.2]	
		A	16.6 [13.4, 19.8]	20.3 [18.2, 22.3]	20.4 [18.7, 22.2]	13.6 [9.4, 17.8]	20.3 [18.3, 22.3]	13.5 [8.8, 18.2]	14.1 [10.0, 18.2]	20.2 [18.3, 22.1]	13.5 [8.8, 18.2]	20.4 [18.7, 22.2]

Fortsetzung auf der nächsten Seite

Tabelle A.3: Köln Kohorte: Mittlerer integrierter Brier Score (`IBS`) bis zum Zeitpunkt $\min(10, t_{\max})$ für die Modelle `SUBGROUP` (`S`), `WEIGHTED` (`W`) und `ALL` (`A`) sowie für den jeweiligen Kaplan-Meier-Schätzer getrennt nach Untergruppen, Modellbildungsverfahren und verwendeter Kovariablen. In eckigen Klammern ist die jeweilige 2σ-Umgebung auf den 400 Trainings/Test-Aufteilungen angegeben.

U	g	Modell	Methode / Kovariablen									
			CoxBoost			Lasso			Ridge			Kaplan-Meier
			klinisch	genetisch	gemeinsam	klinisch	genetisch	gemeinsam	klinisch	genetisch	gemeinsam	
Stadium	I	S	17.7 [15.2, 20.2]	18.8 [16.1, 21.5]	18.9 [16.7, 21.0]	17.3 [14.0, 20.5]	18.8 [16.2, 21.3]	17.2 [13.6, 20.9]	17.3 [14.2, 20.4]	18.8 [16.4, 21.1]	17.3 [13.6, 20.9]	18.9 [16.7, 21.0]
		W	18.2 [16.3, 20.1]	19.9 [18.7, 21.1]	20.0 [19.1, 20.9]	17.1 [14.9, 19.4]	19.9 [18.8, 20.9]	18.2 [16.0, 20.4]	17.2 [14.9, 19.5]	19.7 [18.6, 20.8]	18.2 [16.0, 20.3]	
		A	18.2 [16.5, 19.9]	20.0 [18.9, 21.2]	20.0 [19.1, 20.9]	17.3 [14.9, 19.6]	20.0 [18.9, 21.1]	17.1 [14.6, 19.7]	17.3 [15.0, 19.6]	20.0 [18.9, 21.1]	17.1 [14.5, 19.7]	20.0 [19.1, 20.9]
	II	S	22.2 [19.1, 25.3]	22.7 [18.7, 26.7]	22.2 [19.7, 24.7]	22.5 [18.6, 26.4]	22.7 [18.2, 27.1]	22.7 [16.3, 29.1]	22.3 [18.6, 25.9]	22.4 [19.4, 25.5]	22.6 [16.6, 28.6]	22.2 [19.7, 24.7]
		W	21.7 [19.2, 24.3]	21.6 [19.5, 23.8]	21.5 [19.6, 23.3]	21.9 [17.3, 26.5]	21.6 [19.5, 23.7]	22.0 [17.5, 26.5]	22.0 [17.6, 26.5]	21.7 [19.7, 23.8]	22.0 [17.6, 26.3]	
		A	21.9 [18.4, 25.4]	21.7 [19.5, 24.0]	21.5 [19.6, 23.3]	22.0 [17.4, 26.6]	21.7 [19.4, 23.9]	22.4 [17.4, 27.3]	22.0 [17.7, 26.4]	21.8 [19.7, 23.9]	22.3 [17.4, 27.1]	21.5 [19.6, 23.3]
	III	S	21.7 [17.3, 26.0]	21.6 [17.1, 26.1]	21.5 [17.3, 25.7]	21.9 [16.8, 27.0]	21.5 [17.1, 26.0]	21.4 [14.1, 28.6]	21.5 [17.2, 25.9]	21.4 [17.0, 25.8]	21.3 [14.0, 28.6]	21.5 [17.3, 25.7]
		W	21.6 [16.8, 26.4]	23.7 [21.2, 26.3]	24.0 [21.5, 26.6]	20.3 [14.2, 26.3]	23.5 [20.7, 26.3]	22.1 [17.4, 26.8]	20.8 [15.0, 26.6]	23.1 [20.5, 25.7]	22.1 [17.4, 26.9]	
		A	21.6 [18.0, 25.3]	24.0 [21.3, 26.7]	24.0 [21.5, 26.6]	20.7 [13.8, 27.5]	24.0 [21.4, 26.6]	20.8 [13.2, 28.3]	20.8 [14.1, 27.4]	23.9 [21.2, 26.5]	20.8 [13.3, 28.3]	24.0 [21.5, 26.6]
	IV	S	19.8 [11.2, 28.4]	20.1 [11.2, 29.0]	19.8 [11.0, 28.5]	20.4 [10.6, 30.1]	20.1 [11.0, 29.3]	24.3 [10.1, 38.5]	20.0 [10.6, 29.3]	20.0 [11.6, 28.4]	24.1 [10.0, 38.1]	19.8 [11.0, 28.5]
		W	20.5 [13.6, 27.3]	28.7 [21.6, 35.9]	28.8 [21.6, 36.1]	20.3 [14.3, 26.2]	28.8 [21.6, 35.9]	27.3 [19.2, 35.3]	20.3 [14.5, 26.1]	28.5 [21.4, 35.7]	27.2 [19.2, 35.1]	
		A	24.8 [17.8, 31.8]	29.1 [21.7, 36.5]	28.8 [21.6, 36.1]	18.7 [10.6, 26.8]	29.1 [21.7, 36.4]	19.0 [9.5, 28.5]	19.0 [12.2, 25.8]	29.1 [21.8, 36.4]	18.9 [9.6, 28.2]	28.8 [21.6, 36.1]

Tabelle A.4: Uppsala Kohorte: Mittlerer integrierter Brier Score (IBS) bis zum Zeitpunkt $\min(10, t_{\max})$ für die Modelle SUBGROUP (S), WEIGHTED (W) und ALL (A) sowie für den jeweiligen Kaplan-Meier-Schätzer getrennt nach Untergruppen, Modellbildungsverfahren und verwendeter Kovariablen. In eckigen Klammern ist die jeweilige 2σ-Umgebung auf den 400 Trainings/Test-Aufteilungen angegeben.

U	g	Modell	Methode / Kovariablen: CoxBoost klinisch	CoxBoost genetisch	CoxBoost gemeinsam	Lasso klinisch	Lasso genetisch	Lasso gemeinsam	Ridge klinisch	Ridge genetisch	Ridge gemeinsam	Kaplan-Meier
Alter	jung	S	29.2 [4.8, 53.7]	27.2 [2.9, 51.5]	26.3 [4.6, 47.9]	31.5 [4.8, 58.2]	26.9 [3.7, 50.2]	35.9 [5.0, 66.9]	26.2 [4.1, 48.3]	26.4 [3.2, 49.6]	45.4 [18.4, 72.4]	26.3 [4.6, 47.9]
		W	23.8 [13.2, 34.5]	23.7 [12.7, 34.7]	23.5 [13.3, 33.7]	24.2 [12.4, 36.1]	23.8 [12.7, 34.9]	26.6 [7.9, 45.2]	24.2 [13.6, 34.8]	23.5 [13.4, 33.5]	26.4 [8.0, 44.8]	
		A	23.9 [12.2, 35.6]	23.8 [12.7, 34.9]	23.5 [13.3, 33.7]	24.1 [11.8, 36.4]	23.8 [12.8, 34.8]	26.1 [5.4, 46.8]	24.1 [12.5, 35.8]	23.5 [13.4, 33.6]	26.0 [5.3, 46.8]	23.5 [13.3, 33.7]
	mittelalt	S	22.0 [19.8, 24.1]	22.7 [20.0, 25.4]	22.3 [20.5, 24.1]	22.1 [19.6, 24.6]	22.6 [20.2, 25.0]	23.7 [19.6, 27.9]	22.4 [20.1, 24.8]	22.4 [20.5, 24.4]	23.7 [19.5, 27.9]	22.3 [20.5, 24.1]
		W	22.3 [20.3, 24.3]	22.5 [20.1, 25.0]	22.3 [20.2, 24.4]	22.2 [20.1, 24.4]	22.5 [20.1, 24.9]	23.0 [19.7, 26.3]	22.3 [20.4, 24.3]	22.3 [20.2, 24.5]	23.0 [19.7, 26.2]	
		A	22.1 [20.1, 24.2]	22.5 [20.1, 24.9]	22.3 [20.2, 24.4]	22.2 [20.1, 24.2]	22.5 [20.1, 24.9]	22.8 [19.6, 26.0]	22.2 [20.1, 24.3]	22.3 [20.2, 24.5]	22.7 [19.6, 25.9]	22.3 [20.2, 24.4]
	alt	S	20.1 [16.3, 24.0]	20.1 [16.1, 24.1]	20.0 [16.5, 23.4]	20.2 [16.2, 24.1]	20.1 [16.4, 23.7]	23.6 [16.8, 30.3]	20.2 [16.3, 24.0]	20.0 [16.4, 23.7]	23.5 [16.8, 30.2]	20.0 [16.5, 23.4]
		W	20.2 [17.4, 23.1]	20.4 [18.1, 22.6]	20.3 [18.3, 22.2]	20.3 [17.6, 23.0]	20.3 [18.2, 22.4]	21.8 [18.0, 25.5]	19.9 [16.8, 23.1]	20.3 [18.3, 22.3]	21.7 [18.1, 25.4]	
		A	20.2 [17.6, 22.9]	20.4 [18.1, 22.6]	20.3 [18.3, 22.2]	20.4 [17.4, 23.3]	20.4 [18.2, 22.6]	21.5 [16.4, 26.5]	20.2 [17.2, 23.2]	20.3 [18.3, 22.3]	21.5 [16.4, 26.5]	20.3 [18.3, 22.2]
Geschlecht	männlich	S	21.6 [19.1, 24.0]	22.0 [19.3, 24.6]	21.7 [19.6, 23.7]	21.9 [19.2, 24.6]	21.9 [19.4, 24.4]	23.5 [18.5, 28.5]	21.8 [19.2, 24.4]	21.8 [19.5, 24.0]	23.4 [18.4, 28.4]	21.7 [19.6, 23.7]
		W	21.5 [19.3, 23.6]	21.7 [19.6, 23.9]	21.5 [19.7, 23.2]	21.5 [19.2, 23.8]	21.7 [19.6, 23.8]	21.8 [18.4, 25.3]	21.4 [19.2, 23.7]	21.6 [19.7, 23.5]	21.8 [18.4, 25.1]	
		A	21.3 [19.2, 23.4]	21.7 [19.5, 24.0]	21.5 [19.7, 23.2]	21.4 [19.1, 23.6]	21.7 [19.5, 24.0]	21.9 [18.3, 25.4]	21.4 [19.2, 23.5]	21.6 [19.7, 23.4]	21.8 [18.3, 25.3]	21.5 [19.7, 23.2]
	weiblich	S	22.2 [19.5, 24.9]	22.3 [19.3, 25.2]	21.9 [19.6, 24.2]	22.3 [19.2, 25.4]	22.2 [19.3, 25.2]	25.6 [19.9, 31.3]	22.2 [19.3, 25.1]	22.1 [19.6, 24.6]	25.5 [20.0, 31.0]	21.9 [19.6, 24.2]
		W	21.8 [19.2, 24.3]	21.7 [19.2, 24.2]	21.6 [19.3, 23.9]	22.1 [19.2, 25.0]	21.7 [19.3, 24.1]	23.3 [19.0, 27.6]	21.9 [19.2, 24.7]	21.7 [19.4, 24.0]	23.3 [19.0, 27.6]	
		A	21.8 [19.3, 24.3]	21.7 [19.2, 24.2]	21.6 [19.3, 23.9]	21.9 [19.2, 24.6]	21.7 [19.2, 24.2]	23.5 [19.1, 27.9]	21.8 [19.2, 24.4]	21.7 [19.4, 23.9]	23.5 [19.1, 27.8]	21.6 [19.3, 23.9]
Histologie	AD	S	22.1 [19.8, 24.3]	22.2 [19.4, 25.0]	21.8 [20.0, 23.7]	22.2 [19.7, 24.7]	22.1 [19.6, 24.6]	23.3 [18.7, 28.0]	21.7 [19.3, 24.1]	21.9 [19.9, 24.0]	23.3 [18.6, 27.9]	21.8 [20.0, 23.7]
		W	21.5 [19.5, 23.6]	21.7 [19.8, 23.5]	21.7 [19.7, 23.6]	21.7 [19.4, 24.0]	22.1 [19.3, 24.9]	22.4 [18.7, 26.2]	21.6 [19.4, 23.8]	21.8 [19.9, 23.8]	22.4 [18.6, 26.2]	
		A	21.6 [19.5, 23.7]	21.9 [19.5, 24.4]	21.7 [19.7, 23.6]	21.6 [19.4, 23.8]	21.9 [19.5, 24.3]	22.4 [18.7, 26.1]	21.5 [19.4, 23.7]	21.8 [19.7, 23.8]	22.4 [18.7, 26.1]	21.7 [19.7, 23.6]
	LC	S	25.5 [18.5, 32.5]	24.8 [17.2, 32.3]	24.6 [19.7, 29.6]	26.7 [16.1, 37.4]	25.2 [16.6, 33.9]	37.6 [17.7, 57.4]	26.1 [16.9, 35.3]	25.2 [17.7, 32.8]	37.8 [18.1, 57.5]	24.6 [19.7, 29.6]
		W	24.9 [19.5, 30.2]	23.6 [19.0, 28.3]	23.3 [19.2, 27.4]	25.3 [19.5, 31.1]	23.8 [18.5, 29.2]	28.5 [20.0, 37.0]	25.2 [20.0, 30.4]	23.4 [19.1, 27.8]	28.4 [20.0, 36.8]	
		A	24.5 [19.6, 29.4]	23.6 [18.8, 28.4]	23.3 [19.2, 27.4]	24.8 [19.4, 30.3]	23.6 [19.0, 28.1]	27.7 [19.0, 36.3]	24.8 [19.7, 29.9]	23.2 [19.2, 27.3]	27.6 [19.0, 36.2]	23.3 [19.2, 27.4]
	SQ	S	21.0 [17.0, 25.0]	21.1 [16.7, 25.6]	20.7 [17.1, 24.4]	21.3 [16.8, 25.8]	21.1 [16.7, 25.5]	22.4 [15.9, 28.9]	21.2 [16.9, 25.5]	20.8 [17.2, 24.4]	22.3 [16.0, 28.5]	20.7 [17.1, 24.4]
		W	20.4 [17.6, 23.2]	20.7 [18.5, 22.8]	20.7 [18.5, 22.8]	20.4 [17.1, 23.7]	20.9 [17.8, 24.0]	21.5 [16.7, 26.2]	20.4 [17.2, 23.6]	20.8 [18.0, 23.6]	21.4 [16.6, 26.1]	
		A	20.4 [17.7, 23.1]	20.9 [18.4, 23.3]	20.7 [18.5, 22.8]	20.5 [17.5, 23.4]	20.9 [18.3, 23.5]	20.8 [16.2, 25.4]	20.5 [17.7, 23.3]	20.7 [18.5, 22.9]	20.7 [16.1, 25.3]	20.7 [18.5, 22.8]
Raucherstatus	Raucher	S	21.8 [19.1, 24.5]	21.9 [18.8, 25.0]	21.5 [19.1, 23.9]	21.9 [18.8, 24.9]	21.9 [18.7, 25.2]	25.5 [19.4, 31.7]	21.8 [19.1, 24.6]	21.7 [19.1, 24.3]	25.4 [19.5, 31.4]	21.5 [19.1, 23.9]
		W	21.5 [18.9, 24.0]	21.4 [18.9, 24.0]	21.2 [19.2, 23.2]	21.8 [18.7, 24.9]	21.4 [18.9, 24.0]	23.3 [18.8, 27.7]	21.7 [18.8, 24.6]	21.3 [19.2, 23.4]	23.2 [18.8, 27.6]	
		A	21.4 [19.2, 23.7]	21.5 [18.8, 24.1]	21.2 [19.2, 23.2]	21.6 [18.9, 24.3]	21.5 [18.9, 24.0]	23.4 [18.8, 28.0]	21.5 [18.9, 24.1]	21.3 [19.2, 23.4]	23.3 [18.8, 27.9]	21.2 [19.2, 23.2]
	Ex-Raucher	S	22.5 [19.5, 25.6]	22.7 [19.3, 26.1]	22.3 [19.5, 25.1]	22.8 [19.1, 26.6]	22.7 [19.3, 26.0]	24.0 [18.0, 30.0]	22.6 [19.2, 26.0]	22.4 [19.5, 25.2]	23.9 [17.9, 29.9]	22.3 [19.5, 25.1]
		W	21.9 [19.3, 24.6]	22.2 [19.2, 25.2]	21.9 [19.5, 24.3]	22.1 [19.2, 24.9]	22.1 [19.2, 25.0]	21.8 [17.8, 25.9]	22.0 [19.2, 24.8]	22.0 [19.5, 24.4]	21.8 [17.9, 25.7]	
		A	21.8 [19.1, 24.5]	22.1 [19.2, 24.9]	21.9 [19.5, 24.3]	21.9 [19.0, 24.7]	22.1 [19.2, 25.1]	22.2 [18.0, 26.4]	21.8 [19.1, 24.5]	21.9 [19.5, 24.4]	22.1 [18.0, 26.3]	21.9 [19.5, 24.3]
	Nichtraucher	S	26.5 [15.8, 37.2]	27.1 [15.2, 38.9]	25.5 [15.8, 35.2]	28.0 [14.2, 41.7]	27.5 [15.2, 39.8]	39.7 [16.8, 62.6]	25.6 [16.6, 34.5]	26.7 [15.5, 37.8]	39.1 [16.0, 62.2]	25.5 [15.8, 35.2]
		W	21.5 [15.9, 27.2]	21.7 [16.2, 27.1]	21.7 [16.4, 26.9]	21.0 [14.9, 27.1]	21.7 [16.0, 27.4]	18.5 [12.7, 24.3]	21.1 [14.9, 27.2]	21.6 [16.3, 26.9]	18.5 [12.9, 24.2]	
		A	20.7 [15.4, 26.1]	21.6 [16.1, 27.2]	21.7 [16.4, 26.9]	20.7 [14.6, 26.9]	21.6 [16.1, 27.1]	21.2 [10.3, 32.1]	20.9 [15.6, 26.2]	21.6 [16.4, 26.8]	21.2 [10.3, 32.1]	21.7 [16.4, 26.9]
Stadium	I	S	21.5 [19.4, 23.6]	21.9 [19.6, 24.3]	21.6 [20.0, 23.2]	21.6 [19.3, 23.8]	21.9 [19.5, 24.2]	23.1 [19.1, 27.1]	21.7 [19.5, 24.0]	21.7 [19.9, 23.5]	23.1 [19.2, 27.1]	21.6 [20.0, 23.2]
		W	21.5 [19.6, 23.3]	21.9 [19.6, 24.2]	21.7 [19.7, 23.6]	21.5 [19.4, 23.7]	21.8 [19.6, 24.1]	22.3 [19.1, 25.5]	21.6 [19.6, 23.7]	21.7 [19.8, 23.6]	22.2 [19.1, 25.4]	
		A	21.4 [19.5, 23.3]	21.9 [19.5, 24.3]	21.7 [19.7, 23.6]	21.5 [19.5, 23.5]	21.9 [19.6, 24.1]	22.1 [19.1, 25.0]	21.5 [19.5, 23.5]	21.7 [19.7, 23.8]	22.0 [19.1, 24.9]	21.7 [19.7, 23.6]
	II	S	23.3 [17.1, 29.5]	23.5 [16.6, 30.3]	22.8 [17.6, 28.0]	23.7 [15.9, 31.5]	23.5 [16.5, 30.4]	27.9 [16.0, 39.9]	23.8 [16.1, 31.5]	23.0 [17.9, 28.1]	27.8 [16.0, 39.5]	22.8 [17.6, 28.0]
		W	21.6 [17.6, 25.6]	21.7 [18.1, 25.4]	21.5 [18.4, 24.7]	21.6 [17.6, 25.7]	21.8 [18.1, 25.5]	21.8 [16.3, 27.4]	21.8 [17.5, 26.1]	21.7 [18.6, 24.7]	21.8 [16.4, 27.2]	
		A	21.8 [17.2, 26.3]	21.7 [18.1, 25.3]	21.5 [18.4, 24.7]	21.9 [17.0, 26.8]	21.7 [18.1, 25.3]	23.0 [15.9, 30.1]	21.7 [17.3, 26.2]	21.5 [18.4, 24.7]	22.9 [15.8, 30.0]	21.5 [18.4, 24.7]
	III	S	19.6 [10.2, 29.1]	18.7 [8.6, 28.8]	19.2 [10.1, 28.4]	19.8 [10.2, 29.3]	18.7 [8.5, 28.9]	28.1 [15.1, 41.1]	19.7 [11.0, 28.4]	18.6 [7.9, 29.2]	27.7 [14.2, 41.3]	19.2 [10.1, 28.4]
		W	21.8 [16.7, 26.8]	20.8 [17.4, 24.2]	20.8 [17.8, 23.8]	21.7 [17.2, 26.1]	20.6 [17.2, 24.0]	24.2 [17.8, 30.6]	21.2 [16.3, 26.1]	20.5 [17.1, 23.9]	24.2 [17.9, 30.6]	
		A	21.4 [16.5, 26.3]	20.8 [17.5, 24.2]	20.8 [17.8, 23.8]	21.6 [15.6, 27.6]	20.8 [17.6, 24.1]	22.2 [12.0, 32.5]	21.3 [15.6, 26.9]	20.9 [17.8, 24.0]	22.2 [12.0, 32.5]	20.8 [17.8, 23.8]

Tabelle A.5: Köln Kohorte: Mittlerer integrierter Brier Score (`IBS`) analog zu Tabelle A.3. Zur Modellbildung (Lasso) wurde das in Abschnitt 4.1.2 beschriebene Verfahren der konstanten Gewichtung der Beobachtungen aus der Restgruppe verwendet. Dabei wurden a-priori die Gewichte $\nu = 0{,}05, 0{,}1, 0{,}5$ benutzt.

U	g	Gewicht ν	Kovariablen klinisch	Kovariablen genetisch	Kovariablen gemeinsam
Alter	jung	0.05	13.4 [6.0, 20.7]	20.7 [14.4, 27.1]	12.6 [4.8, 20.3]
		0.1	13.0 [6.3, 19.6]	20.3 [14.9, 25.6]	12.4 [5.3, 19.5]
		0.5	13.3 [6.5, 20.2]	19.4 [15.0, 23.7]	12.9 [5.9, 19.9]
	mittelalt	0.05	18.1 [15.0, 21.2]	20.9 [19.2, 22.7]	18.0 [14.4, 21.6]
		0.1	18.1 [15.0, 21.1]	21.0 [19.3, 22.7]	17.9 [14.4, 21.4]
		0.5	18.0 [15.3, 20.7]	21.1 [19.8, 22.4]	18.0 [14.9, 21.2]
	alt	0.05	21.3 [17.9, 24.7]	22.2 [20.2, 24.2]	21.4 [17.0, 25.7]
		0.1	21.2 [17.8, 24.6]	22.3 [20.4, 24.2]	21.3 [17.2, 25.4]
		0.5	21.5 [17.9, 25.1]	22.4 [20.8, 24.1]	21.7 [17.8, 25.6]
Geschlecht	männlich	0.05	19.6 [16.8, 22.3]	22.2 [21.1, 23.3]	19.5 [16.6, 22.5]
		0.1	19.6 [16.8, 22.3]	22.2 [21.1, 23.3]	19.5 [16.6, 22.4]
		0.5	19.4 [16.8, 22.0]	22.2 [21.1, 23.3]	19.3 [16.3, 22.4]
	weiblich	0.05	16.4 [12.4, 20.5]	20.7 [18.6, 22.9]	16.5 [11.7, 21.4]
		0.1	16.2 [12.3, 20.1]	20.7 [18.7, 22.6]	16.2 [11.7, 20.7]
		0.5	15.8 [12.3, 19.4]	20.6 [19.0, 22.2]	15.8 [12.1, 19.6]
Histologie	AD	0.05	18.8 [15.7, 21.9]	21.8 [20.5, 23.1]	18.7 [14.8, 22.7]
		0.1	18.7 [15.4, 21.9]	21.8 [20.5, 23.0]	18.6 [14.8, 22.5]
		0.5	18.3 [15.1, 21.6]	21.8 [20.6, 22.9]	18.5 [14.9, 22.1]
	LC	0.05	20.9 [12.9, 29.0]	22.8 [18.1, 27.6]	19.3 [10.1, 28.5]
		0.1	19.8 [11.6, 28.0]	22.6 [18.1, 27.1]	18.7 [10.1, 27.2]
		0.5	18.8 [10.5, 27.2]	22.3 [18.7, 25.8]	18.3 [9.8, 26.9]
	SCLC	0.05	21.2 [9.8, 32.6]	21.1 [10.4, 31.8]	21.6 [8.7, 34.6]
		0.1	21.5 [10.0, 33.0]	20.5 [10.0, 31.0]	20.9 [8.3, 33.5]
		0.5	22.0 [9.5, 34.5]	21.2 [9.7, 32.7]	21.9 [8.7, 35.1]
	SQ	0.05	20.5 [17.9, 23.1]	21.8 [19.8, 23.8]	20.7 [17.2, 24.1]
		0.1	20.3 [17.6, 23.0]	21.8 [20.0, 23.6]	20.4 [17.1, 23.7]
		0.5	19.7 [16.9, 22.5]	21.6 [20.1, 23.2]	19.8 [16.7, 22.9]
	unklassifiziert	0.05	19.6 [11.9, 27.3]	19.9 [12.7, 27.1]	18.1 [8.8, 27.5]
		0.1	18.6 [10.9, 26.4]	19.7 [12.7, 26.8]	18.0 [9.2, 26.8]
		0.5	18.2 [10.4, 26.0]	19.8 [12.4, 27.1]	18.3 [9.6, 26.9]
Raucherstatus	Raucher	0.05	19.2 [17.3, 21.1]	22.1 [20.7, 23.6]	19.1 [16.8, 21.3]
		0.1	19.1 [17.2, 21.1]	22.1 [20.8, 23.4]	19.0 [16.8, 21.3]
		0.5	18.9 [16.9, 20.9]	22.1 [20.8, 23.4]	18.9 [16.8, 21.0]
	Ex-Raucher	0.05	23.9 [16.8, 31.0]	21.8 [18.8, 24.7]	24.1 [15.0, 33.3]
		0.1	23.6 [16.1, 31.2]	21.6 [19.0, 24.3]	23.6 [14.8, 32.5]
		0.5	22.7 [14.7, 30.7]	21.4 [19.3, 23.6]	22.7 [14.5, 31.0]
	Nichtraucher	0.05	13.3 [7.3, 19.3]	20.8 [16.9, 24.7]	13.6 [5.5, 21.8]
		0.1	13.0 [7.6, 18.5]	20.7 [17.5, 23.9]	13.1 [6.9, 19.3]
		0.5	13.5 [9.1, 17.9]	20.3 [18.2, 22.4]	13.4 [8.7, 18.0]
Stadium	I	0.05	17.0 [14.0, 19.9]	19.0 [16.6, 21.5]	16.9 [13.7, 20.2]
		0.1	16.9 [14.2, 19.6]	19.1 [16.9, 21.3]	16.8 [13.9, 19.8]
		0.5	17.5 [15.2, 19.7]	19.7 [18.5, 20.9]	17.4 [15.0, 19.8]
	II	0.05	21.8 [18.7, 24.9]	22.1 [19.0, 25.2]	22.0 [17.0, 27.0]
		0.1	21.6 [18.4, 24.7]	21.9 [18.9, 25.0]	21.8 [17.2, 26.3]
		0.5	21.8 [18.0, 25.6]	21.6 [19.5, 23.6]	22.1 [18.0, 26.2]
	III	0.05	22.6 [18.1, 27.1]	22.6 [18.8, 26.3]	20.8 [15.4, 26.3]
		0.1	21.9 [17.1, 26.6]	22.5 [19.1, 26.0]	20.7 [15.7, 25.8]
		0.5	21.9 [17.6, 26.1]	23.5 [20.8, 26.1]	21.7 [17.1, 26.4]
	IV	0.05	21.9 [12.6, 31.2]	23.6 [15.9, 31.3]	20.4 [12.1, 28.6]
		0.1	22.8 [14.6, 31.1]	25.3 [17.6, 33.0]	21.9 [14.4, 29.5]
		0.5	26.7 [19.0, 34.3]	28.3 [21.2, 35.5]	26.7 [19.0, 34.4]

Tabelle A.6: Uppsala Kohorte: Mittlerer integrierter Brier Score (`IBS`) analog zu Tabelle A.4. Zur Modellbildung (Lasso) wurde das in Abschnitt 4.1.2 beschriebene Verfahren der konstanten Gewichtung der Beobachtungen aus der Restgruppe verwendet. Dabei wurden a-priori die Gewichte $\nu = 0{,}05, 0{,}1, 0{,}5$ benutzt.

U	g	Gewicht ν	Kovariablen klinisch	Kovariablen genetisch	Kovariablen gemeinsam
Alter	jung	0.05	27.5 [12.3, 42.7]	23.4 [11.7, 35.1]	34.6 [15.0, 54.1]
		0.1	26.3 [12.5, 40.2]	23.4 [12.5, 34.3]	32.6 [12.9, 52.2]
		0.5	25.5 [12.2, 38.8]	23.9 [12.6, 35.2]	28.2 [9.5, 46.8]
	mittelalt	0.05	22.1 [19.5, 24.7]	22.6 [19.9, 25.3]	23.7 [19.6, 27.9]
		0.1	22.1 [19.6, 24.6]	22.7 [20.0, 25.3]	23.6 [19.7, 27.4]
		0.5	22.1 [20.1, 24.2]	22.6 [20.1, 25.0]	23.1 [19.8, 26.4]
	alt	0.05	20.3 [17.7, 22.9]	20.3 [16.8, 23.7]	22.8 [17.3, 28.3]
		0.1	20.4 [17.6, 23.1]	20.2 [17.1, 23.4]	22.4 [17.4, 27.4]
		0.5	20.3 [18.2, 22.4]	20.2 [18.0, 22.4]	21.5 [17.8, 25.3]
Geschlecht	männlich	0.05	21.7 [19.3, 24.1]	21.8 [19.5, 24.1]	23.0 [18.8, 27.2]
		0.1	21.7 [19.1, 24.3]	21.8 [19.6, 24.0]	22.7 [18.7, 26.7]
		0.5	21.4 [19.3, 23.5]	21.7 [19.6, 23.9]	22.0 [18.5, 25.4]
	weiblich	0.05	22.2 [19.2, 25.2]	22.1 [19.3, 24.9]	25.3 [20.1, 30.5]
		0.1	22.1 [19.4, 24.9]	22.0 [19.4, 24.7]	24.9 [20.1, 29.6]
		0.5	22.0 [19.3, 24.7]	21.7 [19.3, 24.1]	23.5 [19.4, 27.7]
Histologie	AD	0.05	22.2 [19.6, 24.8]	21.9 [19.6, 24.3]	23.2 [18.6, 27.8]
		0.1	22.1 [19.5, 24.7]	22.0 [19.6, 24.4]	23.0 [18.5, 27.5]
		0.5	21.8 [19.4, 24.2]	22.0 [19.6, 24.4]	22.4 [18.4, 26.4]
	LC	0.05	25.4 [15.8, 35.0]	23.8 [17.6, 29.9]	31.1 [19.1, 43.0]
		0.1	25.4 [16.3, 34.4]	23.5 [17.4, 29.7]	29.6 [19.1, 40.0]
		0.5	25.4 [19.2, 31.7]	23.7 [19.0, 28.5]	27.5 [19.1, 36.0]
	SQ	0.05	21.2 [17.6, 24.8]	21.0 [17.2, 24.9]	22.9 [17.1, 28.6]
		0.1	21.1 [17.9, 24.3]	21.0 [17.6, 24.3]	22.3 [16.9, 27.8]
		0.5	20.7 [17.4, 24.0]	20.9 [18.1, 23.7]	21.1 [16.5, 25.7]
Raucherstatus	Raucher	0.05	21.7 [18.7, 24.8]	21.5 [18.8, 24.3]	25.3 [19.7, 30.9]
		0.1	21.6 [19.1, 24.2]	21.5 [19.0, 23.9]	24.8 [19.7, 30.0]
		0.5	21.7 [19.0, 24.3]	21.4 [18.7, 24.1]	23.6 [19.2, 28.1]
	Ex-Raucher	0.05	22.6 [19.2, 26.0]	22.4 [19.1, 25.8]	23.8 [18.4, 29.1]
		0.1	22.6 [19.5, 25.7]	22.3 [19.5, 25.2]	23.4 [18.5, 28.3]
		0.5	22.1 [19.3, 24.9]	22.2 [19.3, 25.0]	22.3 [18.4, 26.2]
	Nichtraucher	0.05	22.9 [15.2, 30.7]	25.5 [15.1, 36.0]	23.9 [12.8, 35.0]
		0.1	22.2 [15.6, 28.8]	24.5 [15.4, 33.6]	22.0 [12.9, 31.1]
		0.5	20.5 [14.6, 26.4]	22.1 [16.0, 28.2]	18.9 [13.1, 24.7]
Stadium	I	0.05	21.6 [19.5, 23.8]	21.9 [19.5, 24.2]	23.1 [19.4, 26.7]
		0.1	21.6 [19.5, 23.8]	21.9 [19.3, 24.5]	23.0 [19.4, 26.5]
		0.5	21.5 [19.5, 23.6]	21.9 [19.8, 24.0]	22.4 [19.4, 25.4]
	II	0.05	22.5 [16.8, 28.3]	22.9 [16.6, 29.2]	25.0 [17.0, 33.1]
		0.1	22.5 [17.3, 27.8]	22.5 [17.2, 27.9]	24.1 [17.0, 31.2]
		0.5	21.8 [18.2, 25.4]	21.9 [17.9, 25.8]	22.0 [16.7, 27.4]
	III	0.05	21.3 [15.4, 27.1]	18.2 [12.3, 24.2]	24.4 [15.2, 33.5]
		0.1	21.0 [16.3, 25.7]	18.5 [13.3, 23.7]	23.3 [15.1, 31.6]
		0.5	21.3 [17.1, 25.5]	20.5 [17.0, 24.0]	23.9 [17.3, 30.4]

Tabelle A.7: Köln Kohorte: Relative Häufigkeit (in Prozent) der Modelle mit mindestens einer genetischen Kovariable nach Modellbildung mit Lasso. Die Modelle wurden mit allen Kovariablen (klinisch und genetisch) trainiert. Die klinischen Variablen sind unpenalisiert stets in allen Modellen enthalten.

U	g	Köln		Uppsala	
		SUBGROUP	WEIGHTED	SUBGROUP	WEIGHTED
Alter	jung	4	31	20	29
	mittelalt	32	21	11	32
	alt	21	18	14	31
Geschlecht	männlich	30	15	22	26
	weiblich	34	64	13	15
Histologie	AD	45	41	14	13
	LC	20	69	30	27
	SCLC	13	82	-	-
	SQ	22	36	16	43
	unklassifiziert	19	62	-	-
Raucherstatus	Raucher	24	21	12	23
	Ex-Raucher	20	34	13	29
	Nichtraucher	4	33	25	24
Stadium	I	33	21	11	28
	II	35	23	10	32
	III	26	30	29	41
	IV	12	31	0	40
ALL		24		25	

Tabelle A.8: Köln Kohorte: Relative Häufigkeit (in Prozent) der am häufigsten ausgewählten genetischen Kovariable (Region) nach Modellbildung mit Lasso. Die Modelle wurden mit allen Kovariablen (klinisch und genetisch) trainiert. Die klinischen Variablen sind unpenalisiert stets in allen Modellen enthalten.

U	g	Köln		Uppsala	
		SUBGROUP	WEIGHTED	SUBGROUP	WEIGHTED
Alter	jung	2	19	14	8
	mittelalt	16	8	2	8
	alt	8	3	4	8
Geschlecht	männlich	17	4	7	10
	weiblich	16	42	4	3
Histologie	AD	18	22	4	4
	LC	10	40	20	14
	SCLC	6	41	-	-
	SQ	10	15	4	23
	unklassifiziert	9	34	-	-
Raucherstatus	Raucher	5	4	4	10
	Ex-Raucher	11	19	3	12
	Nichtraucher	1	14	15	11
Stadium	I	18	6	2	6
	II	30	12	4	12
	III	16	12	22	12
	IV	7	16	-	-
ALL		5		9	

B Dokumentation der R Funktionen

```
#' Function to determine individual survival probability estimates at given time points
#'
#' @param time.train
#' @param status.train
#' @param RS.train
#' @param RS.test
#' @param time.eval
#' @param weights
#' @return matrix
PREDmat <- function(time.train, status.train, RS.train, RS.test, time.eval, weights = NULL){
  time <- time.train; status <- status.train
  t.unique <- sort(unique(time[status == 1L]))
  if (is.null(weights)) {
    #  Breslow Estimate for Baseline Hazard
    alpha <- numeric(length(t.unique))
    for (i in seq_along(t.unique)) {
      alpha[i] <- sum(time[status == 1L] == t.unique[i]) / sum(exp(RS.train[time >= t.unique[i]]))
    }
  } else {
    #  Weighted Breslow Estimation
    alpha <- numeric(length(t.unique))
    for (i in seq_along(t.unique)) {
      alpha[i] <- sum((time == t.unique[i] & status == 1L) * weights) /
        sum((weights * exp(RS.train))[time >= t.unique[i]])
    }
  }
  #  Cumulative Baseline Hazard
  hazard <- cumsum(alpha)
  #  Expand the Breslow Estimate on all times in "time.eval"
  hazard.eval <- numeric(length(time.eval))
  for (i in seq_along(time.eval)) {
    hazard.eval[i] <- max(0, hazard[t.unique <= time.eval[i]])
  }
  mat <- t(sapply(RS.test, function(rs) exp(-hazard.eval)^exp(rs)))
  colnames(mat) <- time.eval
  return(mat)
}



#' Function to construct the Prediction Matrix with KM Estimates
#'
#' @param time.train
#' @param status.train
#' @param ntest
#' @param time.eval
#' @return matrix
KMmat <- function(time.train, status.train, ntest, time.eval){
  # KM-Estimate on training set
  km.train <- survfit(Surv(time.train, status.train) ~ 1L)
  # expand KM-Estimate on all time points
  km  <- numeric(length(time.eval))
  for (i in seq_along(km)) {
    km[i] <- min(1, km.train$surv[km.train$time <= time.eval[i]])
  }
  mat <- matrix(km, nrow = ntest, ncol = length(time.eval), byrow = TRUE)
  colnames(mat) <- time.eval
  return(mat)
}
```

```
#' Function to determine observation weights
#' based on regularized multi-class logistic regression
#'
#' @param x.train [matrix]
#' @param x.test [matrix]
#' @param y.train [factor]
#' @param y.test [factor]
#' @param method [character(1)]
#'  can either be one of c("lasso", "ridge") or any numeric value in the interval (0,1)
#' @return [list]
#'
EstimateWeights <- function(x.train, x.test, y.train, y.test, penalty=NULL, method) {
  if (method == "lasso" | method == "ridge") {
    if (method == "lasso") alpha = 1L
    if (method == "ridge") alpha = 0L
    #  fit multinomial logistic regression
    require(glmnet)
    if (is.null(penalty)) {
      penalty <- rep(1, ncol(x.train))
    }
    mn.fit <- tryCatch(cv.glmnet(x = x.train,
                                 y = y.train,
                                 # type.measure     = "class",       # default: "diviance"
                                 # type.multinomial = "grouped",     # default: "ungrouped"
                                 penalty.factor = penalty,
                                 standardize = FALSE,                # default: TRUE
                                 alpha       = alpha,                # default: 1 (L1/Lasso)
                                 family      = "multinomial",
                                 maxit       = 100000L),               # default: 10^5
                       error = function(e) NULL
    )
    if (is.list(mn.fit)) {
    mn.lambda <- c(lambda.min = mn.fit$lambda.min,
                   lambda.1se = mn.fit$lambda.1se)
    mn.beta   <- coef(mn.fit$glmnet.fit, s = mn.fit$lambda.min)

    mn.beta2 <- lapply(mn.beta,
                       function(m) {b = m@x; n = m@Dimnames[[1]][m@i+1L]; names(b) = n; return(b)})

    # predicted class on test set
    pred.test.class <- predict(mn.fit,
                               newx = x.test,
                               s    = "lambda.min",
                               type = "class")
    # prediction accuracy on test set
    acc <- numeric(2); names(acc) <- c("mn", "naive")
    acc["mn"]    <- sum(pred.test.class[, 1L] == y.test) / length(y.test)
    acc["naive"] <- sum(rep(names(which.max(table(y.test))), length(y.test)) == y.test) / length(y.test)

    # predicted probabilities on test set
    pred.test.prob   <- predict(mn.fit,
                                newx = x.test,
                                s    = "lambda.min",
                                type = "response")
    # AUC per group outcome on test set
    require(ROCR)
    auc <- numeric(nlevels(y.test)); names(auc) <- levels(y.test)
    for (i in seq_along(levels(y.test))) {
      # calculate AUC values (per group outcome) on test sets
      prediction <- pred.test.prob[, i, ]
      label      <- as.numeric(y.test == levels(y.test)[i])
      pred.test  <- prediction(prediction, label)
      perf.test  <- performance(pred.test, "tpr", "fpr")
      auc.test   <- performance(pred.test, "auc", fpr.stop = 1)@y.values[[1L]]
      auc[i]     <- auc.test
    }
```

```
    ### calculate weights
    # predicted probabilities on training set
    pred.train.prob <- predict(mn.fit,
                               newx = x.train,
                               s    = "lambda.min",
                               type = "response")
    # relative frequencies of group outcomes
    frequencies <- table(y.train) / length(y.train)
    # compute likelihood ratios (weights)
    weights.mat <- t(t(pred.train.prob[, , 1L]) / as.numeric(frequencies))
    # out: matrix of dimension (# training samples) x (# groups)
    } else {
      weights.mat <- matrix(1, nrow = length(y.train), ncol = nlevels(y.train))
      mn.beta2 <- mn.lambda <- acc <- auc <- NULL
    }

    return(list(weightmat = weights.mat,
                beta      = mn.beta2,
                lambda    = mn.lambda,
                acc       = acc,
                auc       = auc))
  } else {
    nu <- as.numeric(method)
    weights.mat <- sapply(levels(y.train),
                          function(g) ifelse(y.train == g, 1L, nu))
    return(list(weightmat = weights.mat))
  }

}
```

```
#' Function to draw a stratified random sample
#' for use in training test scenarios with BatchExperiments.
#'
#' @param static [list]
#'  List of input data.
#'  Must contain 3 elements:
#'  1) genes:  matrix of features
#'  2) groups: data.frame of named factor variables each containing group information
#'  3) surv:   Surv object of survival data
#' @param ratio [numeric]
#'  specifies the ratio of training and test sample sizes
#' @param group.type [character]
#'  specifies which factor variable of the 'group' data.frame is used for stratified sampling
#' @return [list]
#'  index of training and test samples. also returns group.type and group information
#'
Subsample <- function(static, ratio, group.type) {

  group    <- static$groups[[group.type]]
  complete <- complete.cases(static$groups)
  index    <- seq_along(group)[complete]

  # draw stratified sample (for each group outcome)
  train <- unlist(lapply(split(index, group[complete]),
                         function(id, ratio) sample(id, round(length(id) * ratio)),
                         ratio = ratio),
                  use.names = FALSE
  )
  test  <- setdiff(index, train)

  return(list(train      = train,
              test       = test,
              group      = group,
              group.type = group.type)
  )
}
```