technische universität
dortmund

Klassifikation von Brustkrebspatientinnen anhand vorausgewählter Gene mit charakteristischer Expressionsverteilung

Dissertation

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
an der Fakultät Statistik
der Technischen Universität Dortmund

vorgelegt von

Birte Hellwig

Dortmund, Oktober 2017

Gutachter:
Prof. Dr. Jörg Rahnenführer
Dr. Uwe Ligges

Tag der mündlichen Prüfung:
05.02.2018

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einleitung

Brustkrebs ist die häufigste bösartige Tumorerkrankung bei Frauen, in Deutschland gibt es jedes Jahr ca. 70 000 Neuerkrankungen. Im Laufe ihres Lebens erkrankt etwa jede achte Frau an Brustkrebs (RKI 2015). Dabei ist Brustkrebs eine sehr heterogene Erkrankung, für die verschiedene Risikofaktoren bekannt sind. Auf Basis von etablierten klinischen Charakteristika wird in der klinischen Praxis eine Therapieentscheidung getroffen. Eine Standardtherapie beim Vorliegen bestimmter Klassifikatoren ist dabei eine adjuvante Chemotherapie. Studien haben gezeigt, dass etwa zwei Drittel der Patientinnen, deren Lymphknoten noch keine Metastasierung aufweisen (nodal-negativer Brustkrebs), auch ohne eine adjuvante Chemotherapie metastasenfrei bleiben. Diese Patientinnen sicher zu identifizieren würde die Möglichkeit bieten, ihnen die Chemotherapie zu ersparen, die auch mit den heute erhältlichen Medikamenten noch mit starken Nebenwirkungen verbunden ist.

Seit den 1990er Jahren hat die genomweite Genexpressionsanalyse mittels Microarrays immer mehr an Bedeutung gewonnen. Diese Arrays erlauben es die Expression mehrerer Tausend Gene gleichzeitig zu messen. Das Identifizieren von Genen mit differentieller Expression zwischen verschiedenen Gewebe-Entitäten – etwa zwischen normalem Gewebe und Tumorgewebe – ist eine Standardmethode geworden. Darüber hinaus werden Genexpressionsdaten verwendet um prognostische Subgruppen in einem Patientenkolletiv zu identifizieren. Ziel der Forschung ist es die Patienten in Bezug auf Therapiewahl und Prognose möglichst gut zu unterscheiden. Dazu werden häufig Gensignaturen erstellt, die aus einer Kombination von mehreren Dutzend Genen bestehen. Die Interpretierbarkeit solcher Signaturen ist schwer. Im Fall vom Brustkrebs wurde bereits eine Vielzahl von Klassifikatoren entwickelt, die zum Teil bereits erfolgreich in der klinischen Praxis eingesetzt werden. Dabei gibt es im Wesentlichen zwei Ansätze. Zum einen werden mit Gensignaturen molekulare Subtypen des Brustkrebses definiert, zum anderen werden Risikoscores berechnet, mit deren Hilfe die Patientinnen in verschiedene Risikogruppen eingeteilt werden, die eine unterschiedliche Behandlung erhalten. Bekannte Beispiele sind der 70-Gen-Klassifikator (van't Veer et al. 2002), der 76-Gen-Klassifikator (Wang et al.

2005), Oncotype DX (Paik et al. 2004), der Genomic Grade Index (Sotiriou et al. 2006) und die PAM50-Signatur (Parker et al. 2009).

Übliche Methoden der Diskriminanzanalyse liefern keine scharfen Entscheidungsgrenzen. Für die klinische Praxis sind jedoch Klassifikatoren wichtig, die gleichzeitig eine hohe Klassifikationsgüte besitzen und leicht interpretierbar in Bezug auf die enthaltenen Gene sind. Gene mit Expressionsverteilungen, die klar zwischen einer Gruppe mit niedriger und einer Gruppe mit hoher Expression unterscheiden, wären ideale Kandidaten um sie in einem Klassifikationsverfahren zu verwenden. Diese Klassifikatoren würden eine gute Grundlage für personalisierte Medizin bieten.

Ziel der Arbeit ist es Gene mit charakteristischer Expressionsverteilung zu identifizieren und diese mittels Klassifikationsbäumen und Random Forests zu kombinieren, um Patientinnen in eine Gruppe mit Metastase innerhalb der ersten 5 Jahre und eine Gruppe ohne Metastase zu klassifizieren. Die Beobachtungszeit der Patientinnen ohne Metastase muss dabei mindestens 5 Jahre betragen. Dafür wird eine Trainingskohorte von 200 nodal-negativen unbehandelten Brustkrebspatientinnen verwendet, von denen 164 für die Klassifikation verwendet werden können (Schmidt et al. 2008). Bei der Konstruktion der Klassifikatoren stellt sich das Problem, dass die Anzahl der Patientinnen mit und ohne Metastase im Trainingsdatensatz unterschiedlich sind. Nur 28 von 164 Patientinnen in der Trainingskohorte bekommen eine frühe Metastase. Klassifikationsverfahren versuchen üblicherweise eine Klasseneinteilung zu finden, bei der möglichst wenige Patientinnen falsch klassifiziert werden. In unserem Fall erhält man durch Zuordnung aller Patientinnen in die Gruppe ohne Metastase bereits eine Fehlerrate von ca. 17 %. Ohne Adjustierung für die unterschiedlichen Gruppengrößen entscheidet sich der Klassifikator für diese (triviale) Lösung. Mögliche Methoden für den Umgang mit unbalancierten Daten sind das Einführen von unterschiedlichen Fehlklassifikationskosten oder das Verwenden von Up- oder Down-Sampling.

Bezogen auf die Problemstellung ist es sinnvoll die Klassifikationsgüte der entwickelten Klassifikatoren nicht mit der Fehlklassifikationsrate, sondern mit dem negativ prädiktiven Wert (NPV) und der Spezifität zu bewerten. Ein guter Klassifikator im Sinne der Anwendung hat einen hohen NPV-Wert und eine hohe Spezifität. Ein hoher NPV-Wert bedeutet, dass der Anteil der Patientinnen, die tatsächlich metastasenfrei bleiben, unter

den Patientinnen, die der Klassifikator in die Gruppe ohne Metastase klassifiziert, groß ist. Das ist von großer Bedeutung, weil es wichtig ist, dass Patientinnen nur dann keine Chemotherapie erhalten, wenn es sehr sicher ist, dass sie keine Metastase bekommen. Die Spezifität misst den Anteil der metastasenfreien Patientinnen unter den Patientinnen, die vom Klassifikator der metastasenfreien Gruppe zugeordnet wird. Für die klinische Relevanz des Klassifikators ist ein hoher Spezifitätswert wichtig.

In Kapitel 2 werden der medizinische und der biologische Hintergrund dieser Arbeit beschrieben. Dazu wird ein kurzer Überblick über die Brustkrebs-Erkrankung und wichtige klinische Kovariablen gegeben. Anschließend wird der biologische Hintergrund der Genexpressionsmessung erklärt, die Affymetrix-Technologie zur Messung der Genexpression und Methoden zur Vorverarbeitung dieser Daten vorgestellt, sowie das für die Analysen verwendete Datenmaterial. Im letzten Abschnitt dieses Kapitels werden die bekannten Gensignaturen, die als Referenz zur Beurteilung der in dieser Arbeit entwickelten Klassifikatoren dienen, ausführlich beschrieben. Den Abschluss dieses Abschnitts bildet ein Überblick über weitere Gensignaturen für Brustkrebs.

In Kapitel 3 werden die in dieser Arbeit verwendeten statistischen Methoden beschrieben. Dazu werden im ersten Abschnitt Methoden zur Identifizierung von Genen mit charakteristischer Expressionsverteilung vorgestellt. Die in der Arbeit verwendeten Bimodalitätsmaße werden ausführlich beschrieben. Der zweite Abschnitt beschäftigt sich mit dem Testen der prognostischen Relevanz der identifizierten Gene. Die verwendeten Methoden der Überlebenszeitanalyse werden erläutert und es wird auf das Problem des multiplen Testens eingegangen. Anschließend werden Methoden zur Enrichment-Analyse vorgestellt. Im letzten Teil dieses Kapitels werden kurz Maßzahlen zur Beurteilung der Klassifikationsgüte beschrieben und anschließend die Klassifikationsbäume und Random Forests als Klassifikationsmethoden vorgestellt.

In Kapitel 4 werden die Ergebnisse der Analysen mit den unterschiedlichen Bimodalitäts-Scores beschrieben. Dazu wird zunächst die Fragestellung untersucht, ob die verwendeten Maße auf globaler Ebene die gleiche Art von Expressionsverteilung identifizieren. Anschließend wird betrachtet, welche Form von Expressionsverteilung die Top-Gene der verschiedenen Scores aufweisen. Im folgenden Abschnitt wird die prognostische Relevanz der identifizierten Gene lokal und global untersucht. Anschließend wird überprüft, ob sich

die Ergebnisse zur Bimodalität der Gene auf zwei unabhängige Datensätze übertragen lässt. Die Analysen aus Kapitel 4 wurden bereits in Hellwig et al. (2010) beschrieben und durchgeführt, wobei sich das verwendete Datenmaterial in dieser Arbeit in einigen Punkten von dem in Hellwig et al. (2010) unterscheidet. Darum wird abschließend untersucht, ob die Ergebnisse in Bezug auf die Werte der Bimodalitätsmaße vergleichbar sind.

In Kapitel 5 wird auf die Klassifikation der Brustkrebspatientinnen eingegangen. Als Referenz dienen die Ergebnisse der etablierten Klassifikatoren auf dem verwendeten Datenmaterial. Aus den Expressionsdaten werden Klassifikatoren gebildet, wobei als erstes Klassifikationsbäume mit vorausgewählten Genen aufgestellt werden. Im nächsten Schritt werden statt Klassifikationsbäumen Random Forests mit vorausgewählten Genen zur Modellbildung verwendet. Abschließend wird auf die Vorauswahl der Gene verzichtet und es werden mit Hilfe des R-Paketes `mlr` (Bischl et al. 2016) Random Forests mit optimierten Parametern aufgestellt. Die Modellbildung geschieht jeweils auf der Trainingskohorte und die besten Modelle werden auf zwei unabhängigen Kohorten validiert. Den Abschluss bildet ein Vergleich der Ergebnisse der unterschiedlichen Ansätze.

In Kapitel 6 werden die Ergebnisse der Arbeit zusammengefasst und es wird ein Ausblick auf offene Fragestellungen gegeben.

Diese Arbeit ist im Rahmen des DFG-Projektes RA 870/5-1 „Verbesserte prognostische Signaturen aus Microarray-Studien durch Auswahl von Genen mit charakteristischen Verteilungen“ entstanden.

2 Medizinischer und biologischer Hintergrund

In diesem Kapitel werden der medizinische und der biologische Hintergrund dieser Arbeit beschrieben. Dazu wird in Kapitel 2.1 zunächst ein Überblick über die Krankheit Brustkrebs und bekannte Risikofaktoren gegeben. In Kapitel 2.2 wird der biologische Hintergrund der Genexpressionsmessung beschrieben und anschließend die Affymetrix-Technologie zum Messen der Genexpression vorgestellt. Am Ende dieses Abschnittes werden Methoden zur Vorverarbeitung der Genexpressionsdaten beschrieben. In Kapitel 2.3 wird das in dieser Arbeit verwendete Datenmaterial vorgestellt. Dabei handelt es sich um drei Brustkrebskohorten, von denen eine als Trainingsdatensatz verwendet wird und zwei als Validierungsdatensätze. In Kapitel 2.4 wird ein Überblick über bekannte Gensignaturen für Brustkrebs gegeben.

2.1 Brustkrebs

Jedes Jahr erkranken in Deutschland etwa 70 000 Frauen an Brustkrebs, damit ist Brustkrebs mit 30.8 % Anteil die häufigste bösartige Tumorerkrankung bei Frauen und die Erkrankung, die zu den häufigsten krebsbedingten Todesfällen führt. Das mittlere Lebenszeitrisiko einer Frau in Deutschland an Brustkrebs zu erkranken liegt bei 12.8 %. Die Anzahl der Neuerkrankungen ist in den letzten Jahren gestiegen, was vor allem auf die höhere Lebenserwartung und die eingeführten Früherkennungsmaßnahmen zurückzuführen ist. Durch Fortschritt in der Therapie sind die Überlebenschancen in den letzten 10 Jahren allerdings gestiegen, sodass heute insgesamt weniger Frauen an Brustkrebs versterben (RKI 2015).

Es sind zahlreiche Risikofaktoren für eine Erkrankung an Brustkrebs bekannt. Als Faktoren, die zu einem erhöhten Risiko führen an Brustkrebs zu erkranken, gelten zum Beispiel eine frühe erste oder späte letzte Regelblutung, ein höheres Alter bei der ersten Geburt oder eine Hormonersatztherapie in und nach der Menopause. Ebenso wurde in Studien beobachtet, dass Übergewicht und Bewegungsmangel nach den Wechseljahren, Alkohol

und in geringerem Maß auch Rauchen zu einem erhöhten Risiko führen. Ein verringertes Brustkrebsrisiko kann bei mehreren bzw. frühen Geburten und Stillzeiten beobachtet werden. Es ist außerdem bekannt, dass auch genetische Faktoren das Brustkrebsrisiko erhöhen können. Eine vererbbare Mutation in den Genen *BRCA1* oder *BRAC2* ist für 5-10 % aller Brustkrebserkrankungen und die Hälfte der familiär gehäuft auftretenden Fälle von Brust- und Eierstockkrebs verantwortlich (RKI 2015).

Werden Einflussfaktoren einer Krebserkrankung betrachtet, so spricht man von *prognostischen* und *prädiktiven Faktoren*. Prognostische Faktoren erlauben eine Schätzung des wahrscheinlichen Krankheitsverlaufs einer erkrankten Person. Mit Hilfe prädiktiver Faktoren können Mediziner abschätzen, wie ein Patient auf eine bestimmte Therapie anspricht bzw. wie hoch der Therapienutzen ist. Prädiktive Faktoren sind somit eine wichtige Grundlage für *personalisierte Medizin* (Buyse et al. 2011). In den letzten Jahren wurden bereits einige wichtige prognostische und prädiktive Faktoren beim Brustkrebs identifiziert:

Nodalstatus Der Nodalstatus gibt an, ob bereits Lymphknoten in der Achselhöhle von Tumorzellen befallen sind und falls ja, wie viele. Sind keine Lymphknoten befallen, so spricht man von nodal-negativem Brustkrebs (N0). In Abhängigkeit von der Anzahl der befallenen Lymphknoten gibt es noch die Stadien N1-N3.

Metastasenstatus Der Metastasenstatus gibt an, ob zum Zeitpunkt der Diagnose bereits Metastasenbildung beobachtet wurde.

Alter Das Alter ist ein bekannter Risikofaktor für Brustkrebs. Das Risiko an Brustkrebs zu erkranken steigt mit dem Alter an, das mediane Erkrankungsalter in Deutschland liegt bei 64 Jahren (RKI 2015). Bei jüngeren Frauen verläuft die Erkrankung in der Regel aggressiver.

pT-Stage Das pT-Stadium wird anhand der Tumorgröße bestimmt. Stadium 1 Tumore sind kleiner als 2 cm, Stadium 2 Tumore 2-5 cm und Stadium 3 Tumore größer als 5 cm. Sind Tumore mit der Brustwand oder der Haut verwachsen, werden sie mit Stadium 4 klassifiziert. Je höher das Stadium, desto ungünstiger ist in der Regel die Prognose.

Histologisches Grading Der histologische Grad nach Elston und Ellis (auch Nottingham-Score, vgl. Elston und Ellis (1991)) gibt das Maß der Bösartigkeit des Tumors an. Das Tumormaterial wird von einem Pathologen mikroskopisch untersucht und dieser beurteilt wie stark die Struktur des Tumorgewebes gegenüber der Struktur des ursprünglichen Gewebes verändert ist. Es gibt 4 Stufen, von gut differenziert bis undifferenziert, die mit Grad 1-4 bzw. G1-G4 bezeichnet werden.

ER-Status und PR-Status Das Tumorwachstum kann durch die Hormone Östrogen und Progesteron beeinflusst werden. Mittels immunhistochemischer Untersuchung kann im Labor festgestellt werden, ob die Tumorzellen Hormon-Rezeptoren besitzen. Die Tumore werden dann als Östrogen-Rezeptor-positiv (ER-positiv oder ER+) bzw. Progesteron-Rezeptor-positiv (PR-positiv oder PR+) bezeichnet. ER-positive Tumore können mit einer Hormontherapie (z. B. Tamoxifen oder Aromatasehemmer) behandelt werden. ER-positive Tumore haben eine günstigere Prognose.

HER2-Status HER2-Rezeptoren (Human epidermal growth factor receptor 2) sind Bindungsstellen für Wachstumsfaktoren, die die Krebszellen zum Teilen anregen. In 15-30 % der Brustkrebstumore ist HER2 überexprimiert, die Tumore besitzen dann besonders viele Rezeptoren auf der Zelloberfläche. Diese Tumore werden als HER2-positiv (HER2+) bezeichnet und gelten als besonders aggressiv. HER2-positive Tumore können mit einer Antikörper-Therapie (z. B.Trastuzumab) behandelt werden. Die Antikörper blockieren die HER2-Rezeptoren und verlangsamen so das Wachstum.

In der Vergangenheit wurden einige Computerprogramme etabliert, die den behandelnden Ärzten bei der Therapieentscheidung helfen sollen. Das bekannteste und verbreitetste Programm ist Adjuvant! (`https://www.adjuvantonline.com/` (Ravdin et al. 2001)). Es gibt auf Basis von einigen klinischen Variablen (Alter, Tumorgröße, Tumorgrad, Anzahl befallener Lymphknoten und ER-Status) das individuelle Risiko des Patienten für das Wiederauftreten der Erkrankung (Rekurrenz) oder den Tod für den Zeitraum der nächsten 10 Jahre aus. Der Zusatznutzen einer adjuvanten Hormon- oder Chemotherapie kann ebenfalls geschätzt werden.

Der NNBC-3-Algorithmus (node-negative-breast cancer-3 (Schmidt et al. 2009)) benutzt ebenfalls die klassischen klinischen Variablen um die Patienten zu klassifizieren. Der

Algorithmus klassifiziert eine Patientin als high risk, wenn mindestens eine der folgenden Voraussetzungen erfüllt sind: Tumorgrad 3, HER2-positiv, PR-negativ, Alter < 35 Jahre oder peritumorale vaskuläre Infiltration (PVI). Als low risk werden Patientinnen mit Tumorgrad 1 und pT-Stage 1 oder 2 und Patientinnen mit Tumorgrad 3 und pT-Stage 1 klassifiziert. Alle anderen Patientinnen gehören zur Gruppe mit mittlerem Risiko (intermediate risk).

2.2 Messung der Genexpression

2.2.1 Biologischer Hintergrund

Die Messung der Genexpression (oder auch Genexpressionsanalyse) erlaubt es die Genaktivität in Zellen zu quantifizieren. In jeder Körperzelle ist die komplette benötigte Information in Form von DNA gespeichert. Die Baupläne für Proteine liegen auf der DNA als Gene vor. Die Gesamtheit aller Erbinformation eines Organismus wird als Genom bezeichnet. Beim Menschen besteht es aus 23 Chromosomenpaaren und enthält insgesamt etwa 25 000 Gene, die etwa 500 000 Proteine codieren.

Die Chromosomen bestehen aus einer DNA-Doppelhelix, die sich aus zwei komplementären Strängen zusammensetzt. Jeder DNA-Strang ist eine lange Kette von sogenannten Nukleotiden. Die Nukleotide bestehen aus einer Base ((G) Guanin, (A) Adenin, (T) Thymin und (C) Cytosin), einer Phosphatgruppe und dem Zucker Desoxyribose. Die beiden Stränge der Doppelhelix sind durch Wasserstoffbrückenbindungen miteinander verbunden. Dabei binden die Basenpaare Adenin und Thymin, sowie Guanin und Cytosin aneinander. Daraus ergibt sich, dass der komplementäre DNA-Strang, kurz cDNA, durch die Folge der Nukleotide im ersten, dem sogenannten codogenen Strang, vorgegeben ist. Neben den Genen, die für Proteine codieren, besteht die DNA auch aus nicht-codierenden Abschnitten.

$$\text{DNA} \xrightarrow{\text{Transkription}} \text{mRNA} \xrightarrow{\text{Translation}} \text{Protein}$$

Die Proteinbiosynthese basiert auf einem Informationstransfersystem und beinhaltet im Wesentlichen zwei Schritte: die *Transkription* und die *Translation*. Bei der Transkription

werden zunächst durch Transkriptionsfaktoren die beiden DNA-Stränge lokal getrennt, in dem die Wasserstoffbrückenbindungen auftrennt werden. Anschließend werden die Basensequenzen des codierenden Stranges durch Anlagerung komplementärer Basen in die sogenannte *messenger RNA* (kurz mRNA) umgeschrieben. Die mRNA unterscheidet sich von der DNA insofern, dass die Nukleotide statt der Desoxyribose den Zucker Ribose enthalten. Außerdem ist die Base Thymin durch Uracil ersetzt, die sich Adenin anlagert. Die einzelnen Nukleotide werden mit einem Enzym, der RNA-Polymerase, miteinander verbunden. Nicht codierende Abschnitte werden im Anschluss entfernt, diesen Vorgang bezeichnet man als *Splicen*. Die Translation findet in den Ribosomen der Zellen statt. Jeweils drei aufeinander folgende Basen der RNA (Basentriplet genannt) codieren für eine Aminosäure des Proteins. Mir der transfer RNA (kurz tRNA) werden die Aminosäuren zu den Ribosomen transportiert. Sie besitzt auf der einen Seite einen Arm mit der komplementären Basensequenz des Basentriplets und gegenüberliegend die zugehörige Aminosäure. Die tRNA ist also für die Aminosäure spezifisch. Die Translation startet immer an einem Startcodon (UAC) und endet an einem von drei Stoppcodons.

Die genomweite Messung der Genexpression geschieht mit Hilfe von sogenannten DNA-*Microarrays* (auch Gen-Chips), wobei hier hauptsächlich zwischen cDNA- und mRNA-Microarrays unterschieden wird. Die Microarray-Technologie entstand in den 1990er Jahren. Sie erlaubt es mit verhältnismäßig kleinen Probenmengen viele Tausend Gene gleichzeitig zu untersuchen. Mit cDNA-Microarrays (auch Spotted Microarrays) können zwei Proben gleichzeitig auf einem Chip analysiert werden, dabei handelt es sich üblicherweise um die Probe eines Probanden und eine Kontrollprobe. Die Oberfläche der Chips besteht aus chemisch behandeltem Glas, das dafür sorgt, dass die DNA darauf immobilisiert wird. Für die Microarrays werden Gensequenzen aus cDNA-Bibliotheken verwendet. Für jedes Gen wird ein Spot auf dem Chip definiert. Die cDNA von Proband und Kontrolle werden mit unterschiedlichen Fluoreszenzen markiert, in der Regel rot und grün. Beide Proben werden vermischt und auf den Array aufgetragen. Dort hybridisieren sie mit den komplementären DNA-Sequenzen. Anschließend wird nicht gebundene cDNA abgewaschen und das Fluoreszenzsignal jedes Spots mittels eines Scanners ausgelesen.

2.2.2 Affymetrix-Technologie

Viele in Datenbanken wie dem Gene Expression Omnibus (GEO, `http://www.ncbi.nlm.nih.gov/geo/`, Edgar et al. (2002)) frei verfügbare Datensätze wurden auf RNA-Microarrays (auch Oligonukleotid-Microarrays genannt) der Firma Affymetrix gemessen. Die Firma vertreibt die Microarrays unter dem Markennamen GeneChip®. Anders als bei den cDNA-Arrays wird einzelsträngige DNA mit Hilfe von Photolithographie direkt auf dem Chip synthetisiert (in situ) (Schulze und Downward 2001). Dabei wird jedes zu messende Gen auf dem Array durch mindestens ein sogenanntes *Probe Set* repräsentiert, das aus 11-20 Oligonukleotid-Paaren (den *Probe Pairs*) besteht. Oligonukleotide sind kurze Nukleotid-Sequenzen, auf dem Affymetrix-Array bestehen sie aus 25 Basen. Jedes Probe Pair besteht aus zwei *Probe Cells*, dem *Perfect Match* (PM) und dem *Mismatch* (MM). Das Perfect Match enthält die Basensequenz, die exakt komplementär zu einer für das jeweilige Gen spezifischen Sequenz ist. Bei der Sequenz des Mismatches ist die mittlere Base im Vergleich zur Originalsequenz durch die komplementäre Base ausgetauscht. Damit ist das Mismatch also nicht spezifisch für das entsprechende Gen. Es soll zur Kontrolle einer nicht-spezifischen Hybridisierung verwendet werden. Eine schematische Darstellung des Aufbaus des Chips ist in Abbildung 2.1 zu sehen.

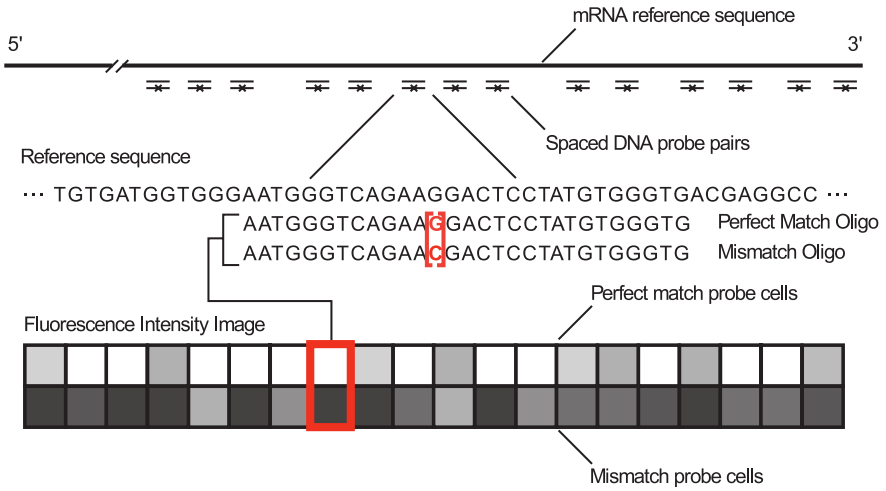

Abbildung 2.1: Affymetrix Array Design (aus Lipshutz et al. (1999)).

Alle in dieser Arbeit analysierten Genexpressionsdaten wurden auf Microarrays des Typs HG-U133A gemessen. Dieser Chip ist 1.28 cm × 1.28 cm groß und enthält 22 283 Probe Sets. Die Probe Pairs eines Probe Sets sind bei diesem Array über den Chip verteilt, um räumliche Effekte auszuschließen. Um die Genexpression zu messen wird RNA aus den zu untersuchenden Zellen extrahiert, gelabelt, auf den Microarray aufgetragen und hybridisiert. Anschließend werden die nicht hybridisierten Anteile abgewaschen und der Chip mit fluoreszierendem Farbstoff eingefärbt, der an der gelabelten RNA bindet. Danach wird jedes Probe Cell eines Chips mit einer Auflösung von 64 Pixel gescannt. Da ein 16-bit Bild verwendet wird, kann jeder Pixel einen Wert zwischen 0 und 2^{-16} haben. Die Intensitätswerte aller Pixel werden in der DAT-File gespeichert. Über das gescannte Bild kann dann mit einer Affymetrix-Software ein Gitter gelegt werden um die Position jedes Probe Cells zu finden.

Aus jedem DAT-File wird ein sogenanntes CEL-File erstellt, das für jedes Probe Cell nur noch einen einzigen Intensitätswert enthält. Dieser Wert berechnet sich als 75 %-Quantil aller Intensitätswerte des Probe Cells, wobei die Pixel am Rand weggelassen werden. Sämtliche Informationen zu den Affymetrix Microarrays sind in Boes (2007) zu finden. Für die Analysen der Genexpressionsdaten bildet das CEL-File den Ausgangspunkt. Um für jedes Probe Set einen einzelnen Expressionswert zu erhalten, müssen die Rohdaten zunächst vorverarbeitet werden.

2.2.3 Vorverarbeitung der Daten

Um aus den CEL-File pro Probe Set einen Expressionswert zu erhalten, sind verschiedene Vorverarbeitungsschritte notwendig. In der Vergangenheit wurde zunächst der Algorithmus der *Microarray Suite* (MAS) 5.0 von Affymetrix (Affymetrix 2002) verwendet, im Laufe der Jahre aber eine Vielzahl von anderen Methoden vorgeschlagen. Bekannte Methoden sind zum Beispiel RMA (Irizarry et al. 2003), gcRMA (Wu et al. 2004), MBEI (Li und Hung Wong 2001) und PLIER (Affymetrix 2005). Eine Übersicht der Methoden ist in Göhlmann und Talloen (2009) zu finden. Die meisten Methoden beinhalten die Schritte Hintergrundkorrektur, Normalisierung und Zusammenfassen.

Bei der Hintergrundkorrektur soll ein Signal entfernt werden, das beim Scannen der Chips im Hintergrund auftritt. Gründe für dieses Signal können zum Beispiel Eigenfluoreszenz

der verwendeten Reagenzien und Streulicht sein. Die Stärke dieses Signals soll geschätzt und dann von den gemessenen Werten entfernt werden. Die Varianz zwischen den verschiedenen Messungen ist oft sehr groß. Eine Normalisierung soll gewährleisten, dass die Messwerte der einzelnen Arrays einer Gruppe von I Arrays vergleichbar sind. Es existieren verschiedene Normalisierungs-Methoden. Sie lassen sich im Wesentlichen in zwei Gruppen einteilen: in Verfahren mit einem Baseline-Chip und Verfahren, die die Informationen aller Arrays benutzen. Nach der Normalisierung werden die normalisierten Werte zu einem Expressionswert pro Probe Set zusammengefasst.

Bei MAS 5.0 werden die Signale global und lokal hintergrundkorrigiert. Dafür wird der Array zunächst in rechteckige Zonen eingeteilt (meist 16). Für jede Zone wird dann ein Hintergrundwert und ein Rauschwert bestimmt. Hintergrundwert und Rauschwert für eine einzelne Zelle werden als gewichtetes Mittel der Werte aller Zonen berechnet, wobei die Größe des Gewichtes vom Abstand zum jeweiligen Zonen-Mittelpunkt abhängt.

Bei der lokalen Hintergrundkorrektur werden die Werte des Mismatches verwendet. Wie zuvor beschrieben sollen sie ein Maß für die unspezifische Hybridisierung darstellen. Daher möchte man die Signale korrigieren, indem man die MM-Werte von den PM-Werten subtrahiert. Da es vorkommen kann, dass ein MM-Wert größer als der zugehörige PM-Wert ist, was zu einem negativen Signal führen würde, wurde das Ideal Mismatch (IM) eingeführt. Dieses ersetzt das Mismatch, wenn dessen Intensität größer als die des Perfect Matches ist. Die Subtraktion der MM-Werte ist jedoch nicht immer der geeignete Weg, um eine Korrektur für nicht-spezifische Bindung vorzunehmen. Nach Irizarry et al. (2003) zeigen empirische Ergebnisse, dass die mathematische Subtraktion nicht die biologische Subtraktion übersetzt. Daher schlagen Bolstad et al. (2003) vor, die MM-Werte komplett zu ignorieren und nur die PM-Werte zu nutzen.

Um die Intensitäten aller Probes eines Probe Sets zu einem Wert zusammenzufassen wird Tukey's Biweight auf den logarithmierten Intensitätswerten verwendet. Zur Normalisierung werden die Expressionswerte bei MAS 5.0 mit einem Faktor multipliziert, der so gewählt ist, dass nach der Normalisierung der mittlere Intensitätswert eines Arrays über alle Gene für jeden Array gleich ist (Standardwert: 500 oder 600). Der MAS 5.0-Algorithmus beinhaltet auch die sogenannten *Detection Calls*, diese sollen Auskunft darüber geben, ob ein Signal, das von einem Probe Set gemessen wurde, tatsächlich von

dem hybridisierten Transkript stammt. Dafür wird mit dem Wilcoxon-Test getestet, ob es einen signifikanten Unterschied zwischen den PM- und den MM-Werten gibt. Basierend auf dem p-Wert werden die Probe Sets dann als „present“, „marginal“ oder „absent“ markiert (s. Affymetrix (2002) für Details).

Eine anderer Ansatz zur Hintergrundkorrektur ist Teil der Methode RMA (Robust Multi-array Average) von Irizarry et al. (2003). Die Methode basiert auf der Idee, dass die Intensität eines Perfect Matches sich als eine Mischung aus einem normalverteilten Hintergrundsignal und einem exponentialverteilten Signal beschreiben lässt. Der hintergrundkorrigierte PM-Wert ist dann der Erwartungswert des Signals gegeben dem ursprünglichen PM-Wert. Die Hintergrundkorrektur wird für jeden Array getrennt durchgeführt. Um die Verteilungen der Probe-Intensitäten einer Gruppe von I Arrays anzugleichen verwenden Irizarry et al. (2003) im Anschluss die Quantil-Normalisierung. Dabei werden die Verteilungen der Probe-Intensitäten so aneinander angepasst, dass die Quantile für alle Arrays gleich sind. Eine genaue Erklärung der Vorgehensweise ist bei Bolstad et al. (2003) zu finden.

Beim Zusammenfassen der normalisierten Probe-Werte berücksichtigen Irizarry et al. (2003) die Beobachtung von Li und Hung Wong (2001), dass die Variabilität zwischen den einzelnen Probes eines Probe Sets typischerweise größer ist als die Variabilität eines bestimmten Probes zwischen verschiedenen Arrays. Sie gehen davon aus, dass sich der Intensitätswert Y_{ijn} des j-ten Probes ($j = 1, \ldots, J_n$) des n-ten Probe Sets ($n = 1, \ldots, N$) auf dem i-ten Array ($i = 1, \ldots, I$) darstellen lässt als

$$Y_{ijn} = \theta_{in} + \phi_{jn} + \epsilon_{ijn}, \tag{2.1}$$

wobei θ_{in} die Expression des n-ten Probe Sets auf dem i-ten Chip ist, ϕ_{jn} der Effekt des j-ten Probes des n-ten Probe Sets und ϵ_{ijn} der Messfehler. Dabei wird vorausgesetzt, dass sich die Probe-Effekte eines Probe Sets zu Null aufsummieren. Die Schätzung in diesem Modell geschieht dann mittels Median-Polish.

Nachteil dieser Standardmethode ist, dass sie nicht geeignet ist, wenn einzelne Arrays analysiert werden sollen, was in der Praxis häufig der Fall ist. Außerdem sind Datensätze, die nicht gemeinsam vorverarbeitet wurden, nicht vergleichbar (Batch-Effekt). Um diese Nachteile zu adressieren, wurde von McCall et al. (2010) ein Algorithmus mit dem Namen

fRMA (frozen Robust Multi-array Average) vorgeschlagen. Dazu wird das Modell aus Gleichung (2.1) wie folgt erweitert:

$$Y_{ijkn} = \theta_{in} + \phi_{jn} + \gamma_{jkn} + \epsilon_{ijkn}.$$

Dabei wird der Index $k = 1, \ldots, K$ für einen Batch-Effekt γ eingeführt, der die Variabilität der Probe-Effekte über die Batches erklären soll. Für Batch k soll $\phi_{jn} + \gamma_{jkn}$ der Batch-spezifische Probe-Effekt für Probe j in Probe Set n sein. Die Varianz von γ ist in diesem Modell Probe-spezifisch $\mathrm{Var}(\gamma_{jkn}) = \tau_{jn}^2$. Außerdem wird zugelassen, dass die Varianz eines Probes innerhalb eines Batches ebenfalls Probe-abhängig ist, das heißt $\mathrm{Var}(\epsilon_{ijkn}) = \sigma_{jn}^2$.

Um einzelne Arrays oder kleine Batches von Arrays vorverarbeiten zu können wurde zunächst eine Referenzgruppe aus 850 Proben von unterschiedlichen Gewebetypen aus öffentlichen Datenbanken erstellt. Die Idee ist es, diese Gruppe als Referenzverteilung für die Quantil-Normalisierung zu verwenden, sowie für die Schätzung der Parameter ϕ, τ und σ. Neue Arrays werden wie bei RMA hintergrundkorrigiert. Anschließend wird eine Quantil-Normalisierung mit der Referenzverteilung durchgeführt. Bei der Zusammenfassung der Werte eines Probe Sets wird zunächst von jedem Intensitätswert der globale Batch-Effekt abgezogen. Anschließend werden die Expressionswerte mit einer robusten Methode geschätzt. Dabei werden unterschiedliche Methoden verwendet, abhängig davon ob einzelne Arrays oder Batches vorverarbeitet werden sollen (s. McCall et al. (2010) für Details).

Die in dieser Arbeit verwendeten Expressionsdaten wurden mit fRMA vorverarbeitet. Dafür wurde die Funktion *frma* aus dem Bioconductor-Paket `frma` (McCall et al. 2010) mit den Standardeinstellungen verwendet. Dabei werden die Expressionswerte als gewichtetes Mittel der Probes eines Probe Sets berechnet. Die Gewichte stammen dabei aus einer M-Schätzer-Methode und werden durch die Summe der vorher berechneten Schätzer für die Varianz zwischen den Batches (τ) und innerhalb der Batches (σ) geteilt. Die Zuordnung der Gensymbole zu den Probe Sets wurde mit Hilfe des Paketes `hgu133a.db` (Carlson 2016) durchgeführt. Wenn in der Arbeit von Genen gesprochen wird, sind eigentlich die Probe Sets gemeint.

2.3 Brustkrebskohorten

Mainz-Kohorte

Die Mainz-Kohorte (Schmidt et al. 2008) besteht aus 200 nodal-negativen Brustkrebs-Patientinnen, die zwischen 1988 und 1998 in der Klinik für Geburtshilfe und Frauenheilkunde der Johann-Gutenberg-Universität Mainz behandelt wurden. Alle Patientinnen wurden operiert, wobei bei 75 Patientinnen eine modifizierte radikale Mastektomie durchgeführt wurde und 125 Patientinnen eine brusterhaltende Operation mit anschließender Strahlentherapie erhielten. Keine der Patientinnen erhielt eine adjuvante systemische Therapie. Eine Übersicht der Verteilung der klinischen Variablen dieser Kohorte ist in Tabelle A.2 zu finden. Zum Zeitpunkt der Operation gab es bei keiner Patientin einen Hinweis auf Metastasierung.

Ein Zielereignis der Studie war das Auftreten einer Fernmetastase, die Ereigniszeit bis zum Eintreten dieses Ereignisses wird als *Metastasis Free Survival* (MFS) bezeichnet. Patientinnen, die verstorben sind, wurden zum Todeszeitpunkt zensiert. Insgesamt trat bei 47 Patientinnen eine Fernmetastase auf, davon bei 28 Patientinnen innerhalb der ersten 5 Jahre nach der Operation. 136 Patientinnen wurden mindestens 5 Jahre beobachtet und bekamen keine Metastase. 17 Patientinnen, die metastasenfrei blieben, wurden kürzer als 5 Jahre beobachtet. Die Genexpression aus Tumormaterial wurde mit dem HG-U133A Array der Firma Affymetrix gemessen. Der Datensatz kann unter der Nummer GSE11121 aus der GEO-Datenbank (Edgar et al. 2002) heruntergeladen werden.

Rotterdam-Kohorte

Bei der Rotterdam-Kohorte handelt es sich um die Studienkohorte, die zur Entwicklung des 76-Gen-Klassifikators von Wang et al. (2005) verwendet wurde (vgl. Kapitel 2.4.2). Aus der Tumorbank des Erasmus Medical Centers in Rotterdam, Niederlande, wurden gefrorene Tumorproben von nodal-negativen Brustkrebspatientinnen entnommen, die zwischen 1980 und 1985 behandelt wurden und dabei keine neoadjuvante oder adjuvante systemische Therapie erhielten. Insgesamt wurden zunächst 436 Proben ausgewählt, jedoch wurden später 150 Patientinnen ausgeschlossen. Die Gründe für einen Ausschluss

waren schlechte RNA-Qualität (77), zu geringer Tumorgehalt (53) und unzureichende Chip-Qualität (20). Insgesamt umfasst die Rotterdam-Kohorte damit 286 Patientinnen. 219 (76.6 %) der Patientinnen erhielten eine brusterhaltende Operation, die übrigen 67 (23.4 %) eine modifizierte radikale Mastektomie.

Das Zielereignis der Studie von Wang et al. (2005) war wie bei der Mainz-Studie das Entwickeln einer Fernmetastase. 107 der 280 Patientinnen entwickelten eine Fernmetastase, davon 93 innerhalb der ersten 5 Jahre nach der Operation. 168 wurden mindestens 5 Jahre beobachtet und blieben metastasenfrei (vgl. Tabelle A.3). Die Genexpressionsmessungen wurden ebenfalls mit Chips des Typs HG-U133A durchgeführt. Die Daten der Rotterdam-Kohorte können aus der GEO-Datenbank (GSE2034) heruntergeladen werden. Bis auf den ER-Status liegen zu dieser Kohorte keine individuellen Angaben zu den klinischen Charakteristika der Patientinnen vor. Die Veröffentlichung von Wang et al. (2005) enthält lediglich eine tabellarische Zusammenfassung für die Gesamtkohorte. Diese wurde in Tabelle A.3 übernommen.

Transbig-Kohorte

Der zweite Datensatz, der zur Validierung der Ergebnisse herangezogen werden soll, wird in dieser Arbeit als Transbig-Kohorte bezeichnet. Die Daten stammen vom internationalen TRANSBIG-Konsortium (Translating molecular knowledge into early breast cancer management: building on the Breast International Group network for improved treatment tailoring). Im Gegensatz zu den anderen beiden vorgestellten Studien handelt es sich hierbei um eine multizentrische Studie. Die Tumorproben und klinischen Daten stammen aus fünf europäischen Zentren: dem Institute Gustave Roussy, Villejuif, Frankreich; dem Karolinska Institutet, Stockholm, Schweden; dem Centre René Huguenin, Saint-Cloud, Frankreich; dem Guy's Hospital, London, England und dem John Radcliffe Hospital, Oxford, England.

Bei allen Patientinnen dieser Kohorte wurde der Brustkrebs zwischen 1980 und 1998 diagnostiziert. Die Kohorte umfasst ausschließlich nodal-negative Brustkrebspatientinnen, die keine adjuvante systemische Therapie erhielten. Patientinnen, die bereits vorher eine bösartige Tumorerkrankung hatten oder unter beidseitigen Mammakarzinomen litten,

wurden ausgeschlossen. Die RNA wurde aus gefrorenen Tumorproben extrahiert und die Genexpression mit HG-U133A Chips der Firma Affymetrix gemessen. Die für diese Arbeit analysierten Genexpressionsdaten wurden aus der GEO-Datenbank heruntergeladen. Dabei setzt sich die Kohorte aus nodal-negativen unbehandelten Patientinnen aus zwei Datensätzen (GSE6532 (n=198) (Loi et al. 2007) und GSE7390 (n=137) (Desmedt et al. 2007)) zusammen, die ursprünglich zur Validierung des 70-Gen-Klassifikators (vgl. Kapitel 2.4.1) verwendet wurden. Die Datensätze überlappen sich teilweise, das heißt es kann vorkommen, dass eine Patientin in beiden Datensätzen enthalten ist. Die Messung der Genexpression wurde für diese Patientinnen (n=41) zweimal durchgeführt. Für unsere Analysen wurden die Duplikate entfernt, verwendet wurden dann nur die Messungen aus dem Datensatz GSE7390. Für insgesamt 14 Patientinnen der kombinierten Kohorte sind keine Ereigniszeiten vorhanden, daher wurden sie bei den weiteren Analysen nicht berücksichtigt. Insgesamt umfasst die Transbig-Kohorte damit 280 Patientinnen. Eine Übersicht der GSM-Nummern der verwendeten Arrays ist in Tabelle A.4 zu finden. Eine Probe ist durch die GSM-Nummer in der GEO-Datenbank eindeutig identifizierbar.

Zielereignisse der Studie waren das Entwickeln einer Fernmetastasierung oder der Tod der Patientinnen. Insgesamt erlebten 82 der Patientinnen das Zielereignis, davon 52 innerhalb der ersten 5 Jahre nach der Operation. 180 Patientinnen wurden mindestens 5 Jahre beobachtet und erlebten kein Ereignis (vgl. Tabelle A.5).

Für die Transbig-Kohorte ist keine Information über den immunhistochemisch bestimmten HER2-Status der Patientinnen verfügbar. Um die auf der Mainz-Kohorte entwickelten Klassifikatoren mit klinischen Kovariablen auf der Transbig-Kohorte validieren zu können, wurde der HER2-Status für die Transbig-Kohorte aus den Expressionswerten des entsprechenden Gens (*ERBB2*, Probe Set 216836_s_at) bestimmt. Dazu wurde das R-Paket `mclust` verwendet um ein Normalverteilungsmischungsmodell mit zwei Komponenten an die Expressionswerte anzupassen. Das modellbasierte Clustern wird in Kapitel 3.1.1 beschrieben.

2.4 Bekannte Gensignaturen

Personalisierte Medizin spielt in den letzten Jahren eine immer größer werdende Rolle. Adjuvante Chemotherapie hat dazu geführt, dass sich die Prognose von Brustkrebspatientinnen verbessert hat. Allerdings würden ungefähr zwei Drittel der Patientinnen mit nodal-negativem Brustkrebs auch ohne adjuvante Chemotherapie die ersten 10 Jahren nach der Diagnose überleben. Diesen Patientinnen könnte die Chemotherapie, die mit starken Nebenwirkungen verbunden ist, erspart werden. In den vergangenen Jahren wurden bereits einige Gensignaturen etabliert, die bei der Therapieentscheidung von Brustkrebs helfen sollen, und teilweise bereits in der klinischen Praxis eingesetzt werden. Die bekanntesten Gensignaturen werden in diesem Kapitel vorgestellt.

2.4.1 70-Gen-Klassifikator

Der 70-Gen-Klassifikator wurde von van't Veer et al. (2002) vorgestellt. Ziel der Arbeit war es einen Klassifikator zu konstruieren, der vorhersagt, welche Patientinnen keine adjuvante Therapie benötigen. Dafür wurden Proben von insgesamt 78 jungen nodal-negativen Patientinnen ausgewählt. 34 der Patientinnen hatten eine Fernmetastase innerhalb der ersten 5 Jahre nach der Operation, 44 Patientinnen waren mindestens 5 Jahre mestastasenfrei.

Aus dem gefrorenen Tumormaterial wurde RNA isoliert, die zur Herstellung von cRNA genutzt wurde, dabei wurde zweimal hybridisiert um eine Referenz zu erhalten. Dafür wurde eine gepoolte Probe erstellt, die für jeden Tumor die gleiche Menge cRNA erhält. Die Genexpression wurde auf einem cRNA-Microarray gemessen, der ungefähr 25 000 menschliche Gene enthält. Die Fluoreszenz-Signale wurden quantifiziert, normalisiert und ein Intensitätsverhältnis in Bezug auf das Signal des Referenz-Pools gebildet. Dann wurden ca. 5000 Gene ausgewählt, die in mindestens 3 der 78 Tumoren in Bezug auf die Referenz signifikant reguliert sind (van't Veer et al. 2002).

Für diese Gene wurde die Korrelation mit dem klinischen Outcome (Metastase vs. keine Metastase) der Patientinnen berechnet. Für 231 Gene konnte ein Korrelationskoeffizient $\rho < -0.3$ oder $\rho > 0.3$ beobachtet werden. Die Signifikanz der Korrelation wurde mit

einem Permutationstest überprüft. Im nächsten Schritt wurden die 231 Gene auf Basis des Absolutwertes des Korrelationskoeffizienten geordnet. In einem schrittweisen Verfahren wurde dann der Klassifikator optimiert. Dabei wurden schrittweise jeweils 5 Gene aus der geordneten Genliste dem Klassifikator hinzugefügt. Zur Beurteilung der Prognosegüte wurde eine Leave-one-out-Kreuzvalidierung durchgeführt. Für jede der Patientengruppen wurde dafür das arithmetische Mittel der Expressionswerte der jeweiligen Gene über die Proben in der Gruppe berechnet. Für die Probe, die klassifiziert werden sollte, wurde nun die Korrelation der Expressionswerte mit den Expressionsprofilen der beiden Gruppen berechnet und die Probe der Gruppe mit der größeren Korrelation zugeordnet. Anschließend wurden die Fehler gezählt. Die Anzahl der Fehlklassifikationen sank in diesem Verfahren zunächst mit dem Hinzufügen von Genen, der minimale Wert der Fehlklassifikationen wurde bei 70 Genen erreicht. Danach stieg der Wert wieder an, was mit dem Einführen von Rauschen durch die Hinzunahme von Genen begründet werden kann. Der finale Klassifikator wurde somit mit 70 Genen gebildet. Die Zuordnung einer Patientin zur Gruppe mit guter oder schlechter Prognose geschieht auf Basis der Korrelation mit dem durchschnittlichen Expressionsprofil der Gruppe mit guter Prognose. Für die Korrelation wird dafür ein Threshold benötigt, ab dem eine Patientin der Gruppe mit guter Prognose zugeordnet wird. Dieser wurde zunächst als der Wert gewählt, bei dem die Anzahl der Fehlklassifikationen am kleinsten ist.

Der Klassifikator sagte für 65 der 78 Patienten das Outcome richtig voraus, dabei wurden 5 Patientinnen mit Metastasenbildung der Gruppe mit guter Prognose zugeordnet und 8 Patientinnen ohne Metastase falsch klassifiziert. Da eine falsche Klassifikation der Patientinnen mit Metastase als schwerwiegender angesehen wird, wurde der Threshold für den Korrelationskoeffizienten verschoben, sodass nur 10 % (3 von 34) dieser Patientinnen fehlklassifiziert werden, der resultierende Klassifikator hatte also eine größere Sensitivität. Dafür werden 12 der Patientinnen ohne Metastase falsch zugeordnet.

Die Autoren validierten ihren Klassifikator auf einer Kohorte von 19 nodal-negativen jungen Patientinnen, 12 mit einer Fernmetastase innerhalb der ersten 5 Jahre nach der Operation und 7 Patientinnen, die für mindestens 5 Jahre metastasenfrei blieben. Mit beiden Thresholds wurde das gleiche Klassifikationsergebnis erreicht, jeweils eine Patientin beider Gruppen wird falsch klassifiziert. Im Vergleich mit den Konsensus-Richtlinien des NIH und St. Gallen, die auf klinischen Charakteristika beruhen, erkennt der Klassifikator

ähnlich viele Patientinnen mit Metastasenbildung, ordnet aber weniger Patientinnen ohne Metastase der Gruppe mit schlechter Prognose zu (ähnliche Sensitivität, höherer NPV).

In ihrer folgenden Arbeit (van de Vijver et al. 2002) validierten die Autoren den 70-Gen-Klassifikator auf einer Kohorte von 295 jungen (Alter < 55 Jahre), behandelten Patientinnen mit Stage 1 oder Stage 2 Brustkrebs, dabei waren 151 nodal-negativ und 144 nodal-positiv. Die Patientinnen wurden mittels des Klassifikators in zwei Gruppen eingeteilt und die Survivalkurven in den beiden Gruppen mit dem Log-Rank-Test verglichen. Dabei konnte gezeigt werden, dass sich die Survivalkurven signifikant unterscheiden, wobei die Patientinnen, die vom Klassifikator der Gruppe mit schlechter Prognose zugeordnet werden, tatsächlich eine schlechtere Prognose hatten. Auch in einer weiteren Studie konnte der Klassifikator auf einer unabhängigen Kohorte validiert werden (Buyse et al. 2006).

Aus dem Klassifikator wurde ein kommerzieller Test für Formalin-fixiertes Paraffin-eingebettetes Gewebe (FFPE-Gewebe) mit dem Namen MammaPrint® (Hersteller: Agendia) entwickelt (Glas et al. 2006). Dieser wurde von der U.S. Food and Drug Administration (FDA) 2007 für nodal-negative Patientinnen zugelassen. Die klinische Validierung soll in verschiedenen prospektiven Studien, etwa der MINDACT-Studie (Microarray In Node negative Disease may Avoid ChemoTherapy (Cardoso et al. 2007)) des TRANSBIG-Konsortiums (vgl. Kapitel 2.3) oder der MINT-Studie (Multi-Institutional Neo-adjuvant Therapy MammaPrint Project) erfolgen. Erste Ergebnisse der MINDACT-Studie zeigen, dass die 70-Gen-Signatur einen zusätzlichen Nutzen zu den etablierten klinischen Variablen bei der Identifizierung von Patientinnen hat, die adjuvante Chemotherapie benötigen (Cardoso et al. 2016).

2.4.2 76-Gen-Klassifikator

Der 76-Gen-Klassifikator wurde von Wang et al. (2005) vorgestellt. Entwickelt wurde er auf der Rotterdam-Kohorte (vgl. Kapitel 2.3). Die Expressionswerte der Gene, die mit dem Affymetrix HG-U133A-Array gemessen wurden, wurden mit MAS 5.0 vorverarbeitet. Die Signale wurden auf einen Intensitätswert von 600 skaliert. 17 819 wurden in zwei oder mehr Arrays als „present“ gekennzeichnet.

Für die weiteren Analysen wurden die Expressionswerte zunächst standardisiert, indem die Expressionswerte eines Gens durch den Median der Expressionswerte dieses Gens über alle Proben geteilt wurden. Die Autoren verwenden zwei Ansätze um einen Klassifikator zu erstellen, der Patientinnen unterscheidet, die eine Fernmetastase bekommen oder mindestens 5 Jahre metastasenfrei bleiben. Im ersten Ansatz werden alle Proben zufällig in eine Trainings- (n=80) und eine Testmenge (n=206) eingeteilt. Beim zweiten Ansatz wird eine Analyse stratifiziert nach dem ER-Status durchgeführt. Auch hier wurden die Proben beider Gruppen wieder in Trainings- und Testmenge eingeteilt (ER+: 80/129, ER-: 35/42). Zur Bestimmung der optimalen Größe der Trainingsmenge wurde eine Resampling-Methode verwendet.

Zur Genauswahl wurden auf den Trainingsmengen jeweils univariate Cox-Modelle aufgestellt, um Gene zu finden, die mit metastasenfreiem Überleben assoziiert sind. Um den Effekt des multiplen Testens zu reduzieren und die Robustheit der ausgewählten Gene zu testen wurde ein Bootstrap-Verfahren verwendet. Aus der Trainingsmenge wurden jeweils 400 Bootstrap-Stichproben der gleichen Stichprobengröße gezogen. Auf jeder Bootstrap-Stichprobe wurde für jedes Gen ein Bootstrap-Score berechnet, indem die größten und kleinsten 5 % der p-Werte entfernt und anschließend über die Inversen der verbleibenden p-Werte gemittelt wurde. Die Scores wurden verwendet um eine geordnete Genliste zu erstellen. Um die optimale Kombination von Genen zu bestimmen, wurden schrittweise Gene dieser Liste hinzugefügt und eine ROC-Analyse durchgeführt, bis die Lösung mit dem maximalen AUC gefunden wurde.

Für den nicht-stratifizierten Ansatz wurde ein Klassifikator mit 35 Genen konstruiert, der jedoch nur „moderate“ Klassifikationsgüte hat. Für die Gruppe der ER-positiven Proben wurden 60 Gene, für die Gruppen der ER-negativen Proben 16 Gene ausgewählt. Aus diesen Genen und dem ER-Status konstruierten Wang et al. (2005) ein Cox-Modell allen 286 Proben. Für jeden Patienten kann ein Relapse-Score berechnet werden, der als Linearkombination von gewichteten Expressionswerten mit standardisierten Cox-Regressionskoeffizienten als Gewichte konstruiert wird:

$$\text{Relapse-Score} = A \cdot I + \sum_{i=1}^{60} I \cdot w_i x_i + B \cdot (1 - I) + \sum_{j=1}^{16} (1 - I) \cdot w_j x_j.$$

Dabei ist I die Indikatorvariable für den ER-Status, das heißt $I = 1$, wenn die Probe

ER-positiv ist, und $I = 0$, wenn die Probe ER-negativ ist. w_i und w_j sind standardisierte Cox-Regressionskoeffizienten für die Marker für ER-positive bzw. ER-negative Proben und x_i und x_j die Expressionswerte der entsprechenden Marker. Der Cutoff für den Score wurde so aus der ROC-Kurve bestimmt, dass bei einer Sensitivität von 100 % die Spezifität maximal ist. Für die Konstanten wurden die Werte $A = 313.5$ und $B = 280$ gewählt, um den Cutoff des Scores für ER-positive und ER-negative Proben auf den Wert 0 zu zentrieren.

Zur Validierung wurde der Klassifikator auf die Testmenge von 171 Proben angewendet. Die resultierende ROC-Kurve hatte einen AUC-Wert von 0.694, eine Sensitivität von 93 % (52/56) und eine Spezifität von 48 % (55/115). Als Kontrolle wurde ein Klassifikator aus 76 zufällig ausgewählten Genen erstellt, dieser hatte auf der Testmenge einen AUC-Wert von 0.515, Sensitivität von 91 % und Spezifität von 12 %. Die Autoren zeigten außerdem, dass der Relapse-Score im univariaten und multivariaten Cox-Modell (adjustiert auf klinische Parameter) einen signifikanten Einfluss auf das metastasenfreie Überleben hat.

In einer späteren Arbeit (Foekens et al. 2006) zeigten die Autoren, dass der 76-Gen-Klassifikator in einer Kohorte von 180 nodal-negativen Brustkrebspatientinnen Patientinnen, die in den ersten 5 Jahren nach der Operation eine Metastase bekamen, mit 90 % Sensitivität (27/30) und 47 % Spezifität (75/150) vorhersagt. Auch in der Cox-Analyse hatte der Relapse-Score wieder einen signifikanten Einfluss auf das metastasenfreie Überleben, sowohl in der univariaten als auch in der multivariaten Analyse. In einer weiteren Analyse (Zhang et al. 2009) konnten die Autoren zeigen, dass der 76-Gen-Klassifikator dabei hilft unter Tamoxifen-behandelten Patientinnen eine Subgruppe mit schlechter Prognose zu identifizieren, die von der Tamoxifen-Behandlung profitiert.

2.4.3 Oncotype DX

Oncotype DX® (Hersteller: Genomic Health) ist ein kommerzieller Assay für FFPE-Gewebe, der auf dem 21-Gen-Score von Paik et al. (2004) basiert. Für die Entwicklung des Recurrence-Scores wurde FFPE-Gewebe von Tumoren von 447 nodal-negativen, ER-positiven Brustkrebspatientinnen verwendet. Es wurde eine Methode der Reversen Transkriptase-Polymerase-Kettenreaktion (RT-PCR) entwickelt, um die Genexpression

in dem FFPE-Gewebe zu quantifizieren. Die Autoren wählten 250 Kandidatengene aus der Literatur aus und überprüften den Zusammenhang der Expression mit der Rekurrenz in drei Datensätzen mit insgesamt 447 Patientinnen (n=233, n=78, n=136). Der große Datensatz enthielt dabei nur nodal-negative, ER-positive Patientinnen, die mit Tamoxifen behandelt wurden. Die anderen beiden Datensätze waren dagegen heterogen und enthielten teilweise nodal-positive und ER-negative Patientinnen, auch wurden die Patientinnen zum Teil mit Chemotherapie behandelt.

Für neun der 250 Gene konnte eine Korrelation mit der Rekurrenz in allen drei Studien beobachtet werden (unadjustierter p-Wert kleiner 0.05). Für weitere fünf Gene konnte in allen drei Kohorten ein p-Wert kleiner 0.10 beobachtet werden und 9 Gene hatten in zwei Studien einen p-Wert kleiner 0.05. Insgesamt wurden demnach 21 Gene ausgewählt. Die Gene wurden nachfolgend auf Basis von Korrelation der Expressionswerte und/oder biologischer Funktion in verschiedene Gruppen einteilt: Proliferations- (5 Gene), GRB7- (2 Gene), Östrogen- (4 Gene) und Invasions-Gruppe (2 Gene) und eine Referenzgruppe (5 Gene). Drei Gene (*GSTM1*, *CD68*, *BAG1*) wurden keiner Gruppe zugeordnet.

Der Rekurrenz-Score wurde auf Basis der 21 Gene in den drei Kohorten mittels Regression konstruiert. Die Berechnung für neue Proben erfolgt in einem schrittweisen Verfahren. Zunächst werden die Expressionswerte der 16 anderen Gene relativ zu den Werten der fünf Referenzgene normalisiert. Die normalisierten Werte liegen dann im Bereich zwischen 0 und 15. Für Proliferations-, GRB7-, Östrogen- und Invasions-Gruppe werden Scores berechnet, die den (zum Teil gewichteten) Mittelwerten der Expressionswerte der Gene in der jeweiligen Gruppe entsprechen (die genaue Berechnung kann Tabelle F.1 entnommen werden). Mit diesen Scores wird dann zunächst ein unskalierter Rekurrenz-Score $\mathrm{RS_U}$ berechnet:

$$\begin{aligned}\mathrm{RS_U} &= 0.47 \cdot GRB7\text{-Score} - 0.34 \cdot \text{Östrogen-Score} + 1.04 \cdot \text{Proliferations-Score} \\ &\quad + 0.10 \cdot \text{Invasions-Score} + 0.05 \cdot CD68 - 0.08 \cdot GSTM1 - 0.07 \cdot BAG1.\end{aligned}$$

Dabei stammen die Koeffizienten aus der Regression in den drei Trainingskohorten. Der skalierte Rekurrenz-Scores RS wird dann (gemäß Paik et al. (2004)) wie folgt bestimmt:

$$\mathrm{RS} = \begin{cases} 0 & \text{für } \mathrm{RS_U} < 0 \\ 20 \cdot (\mathrm{RS_U} - 6.7) & \text{für } 0 \leq \mathrm{RS_U} \leq 100 \\ 100 & \text{für } \mathrm{RS_U} > 100 \end{cases} .$$

Werte von RS kleiner 0 werden dann wieder auf 0 gesetzt und Werte größer 100 auf 100, sodass der Wertebereich des Rekurrenz-Scores zwischen 0 und 100 liegt.

Auf Basis des Rekurrenz-Scores werden die Patientinnen in drei Gruppen eingeteilt: niedriges Risiko (RS $<$ 18), mittleres Risiko (18 $\leq$ RS $<$ 31), hohes Risiko (RS $>$ 31). Die Cutoffs für die Einteilung wurden auf Basis der Ergebnisse der homogenen Kohorte mit den 233 Patientinnen bestimmt. Paik et al. (2004) validierten den Rekurrenz-Score auf einer Kohorte von 668 Patientinnen, die mit Tamoxifen behandelten wurden. Es konnte gezeigt werden, dass die Zeit bis zur Rekurrenz in der Gruppe mit hohem RS signifikant kürzer ist als in der Gruppe mit niedrigem RS ($p < 0.001$).

In großen retrospektiven Studien wurde nachgewiesen, dass der Rekurrenz-Score sowohl prognostisch als auch prädiktiv ist (Habel et al. 2006; Paik et al. 2006; Albain et al. 2010; Dowsett et al. 2010). In einer großen prospektiven Studie mit HER2-negativen, nodal-negativen Patientinnen (Sparano et al. 2015) konnte gezeigt werden, dass Patientinnen mit einem Rekurrenz-Score zwischen 0 und 10 sicher mit Hormontherapie als einzige systemischer Therapie behandelt werden konnten. Auf eine zusätzliche Chemotherapie konnte verzichtet werden.

2.4.4 Genomic Grade Index

Der *Genomic Grade Index* (GGI, in der ersten Veröffentlichung *Gene Expression Grade Index*) wurde von Sotiriou et al. (2006) als Gensignatur vorgestellt, die entwickelt wurde um die Bestimmung des histologischen Grades zu verbessern. Der histologische Grad eines Tumors ist in der klinischen Praxis ein wichtiger Prognosefaktor und Basis für eine Therapieentscheidung. Die Tumore werden in Grad 1-3 eingeteilt, wobei Tumore des Grades 2 ein mittleres Rekurrenzrisiko aufweisen. Für diese Gruppe ist es schwer eine Therapieentscheidung zu treffen. Der GGI soll helfen innerhalb dieser Gruppe zwischen Patientinnen mit hohem und niedrigem Risiko zu unterscheiden.

Für die Konstruktion des GGI verwendeten die Autoren eine Kohorte von 64 ER-positiven Patientinnen, deren Tumore Grad 1 (n=33) oder Grad 3 (n=31) aufwiesen, und die mit Tamoxifen behandelt wurden. Die Genexpression wurde auf einem Affymetrix HG-U133A-Array gemessen und die Rohdaten mit RMA vorverarbeitet. Die Proben wurden verwendet um Gene zu finden, die mit dem histologischen Grad assoziiert sind. Es wurden dabei nur ER-positive Proben verwendet, da es eine Korrelation zwischen ER-Status und histologischem Grad gibt – ER-negative Tumore haben überwiegend mittleren oder hohen Grad – und die Autoren vermeiden wollten Gene zu finden, die nur über den ER-Status mit dem histologischen Grad verbunden sind.

Da die Trainingsmenge aus Proben von zwei verschiedenen Laboren stammt, benutzten die Autoren zur Berechnung der differentiellen Expression der Gene zwischen G1- und G3-Tumoren die Standardised Mean Difference von Hedges und Olkin (1985), einen Score aus einem meta-analytischen Verfahren. Um für das multiple Testproblem zu adjustieren wurde die maxT-Statistik von Westfall und Young (1993) mit einer Erweiterung von Korn et al. (2004) verwendet, die die False Discovery Rate kontrolliert. Dabei werden Abhängigkeiten zwischen den Genen berücksichtigt.

Mit dieser Methode wurden 128 Probe Sets identifiziert, die zu 97 Genen gehören. Der Großteil dieser Gene war in G3-Tumoren überexprimiert. der Genomic Grade Index wird definiert als

$$\mathrm{GGI} = \mathit{scale}\left(\sum_{j\in \mathrm{G}_3} x_j - \sum_{j\in \mathrm{G}_1} x_j - \mathit{offset}\right),$$

dabei sind *scale* und *offset* Datensatz-spezifische Transformationsparameter, sodass die mittlere Expression von G1-Tumoren - 1 und von G3-Tumoren + 1 ist. x_j ist der Expressionswert des j-ten Gens und G_1 und G_3 sind die Gengruppen mit erhöhten Expressionswerten in G1- bzw. G3-Tumoren. Patientinnen mit negativem GGI-Wert werden der Gruppe mit niedrigem Grad (GG1) zugeordnet, Patientinnen mit einem GGI-Wert von 0 oder höher der Gruppe mit hohem Grad (GG3).

In einem Testdatensatz von 125 Patientinnen, die keine systemische Therapie erhalten hatten, konnten ähnliche Expressionsmuster der Gene in den G1- und G3-Tumoren beobachtet werden. Auch in drei weiteren Datensätzen aus öffentlichen Datenbanken mit insgesamt 479 Patientinnen, die zum Teil behandelt wurden, konnten diese Muster erkannt

werden. Zur Validierung des GGI wurden die Proben der vier Validierungskohorten gepoolt und der Zusammenhang mit dem Relapse-Free-Survival (RFS) mit Hilfe von Kaplan-Meier-Kurven und dem Log-Rank-Test untersucht. In der Untergruppe der Patientinnen mit histologischem Grad 2 konnte beobachtet werden, dass Patientinnen mit GG3 ein signifikant kürzeres RFS aufwiesen (Hazard Ratio: 3.61, $p < 0.001$). Auch für die Gesamtkohorte konnte ein signifikanter Unterschied im RFS zwischen GG3 und GG1 gezeigt werden (Hazard Ratio: 2.83, $p < 0.001$). In einem multivariaten Cox-Modell mit klinischen Variablen konnte ein signifikanter Zusammenhang zwischen dem GGI und dem RFS nachgewiesen werden.

In einer weiteren Studie (Loi et al. 2007) wurde auf einer Kohorte von 666 ER-positiven Patientinnen gezeigt, dass der GGI sowohl auf der Subkohorte ohne systemische Behandlung (n=417) als auch auf der mit Tamoxifen behandelten Gruppe signifikant mit Prognose assoziiert war. Liedtke et al. (2009) konnten zeigen, dass mit dem GGI das Ansprechen auf eine Chemotherapie vorhergesagt werden kann. Der GGI wird in Europa als MapQuant Dx® Test vermarktet, der auf der Affymetrix-Technologie basiert. Darüber hinaus wurde ein 8-Gen-Klassifikator für FFPE-Gewebe basierend auf qRT-PCR entwickelt (genannt PCR-GGI Toussaint et al. 2009), der auf 4 Genen des herkömmlichen RNA-GGI und 4 Referenzgenen basiert.

2.4.5 Weitere Gensignaturen

Neben den in dieser Arbeit betrachteten Gensignaturen für Brustkrebs existiert noch eine Vielzahl von weiteren Klassifikatoren, die auf Genexpressionsdaten basieren. In diesem Abschnitt wird ein Überblick über diese Methoden gegeben. Im Wesentlichen gibt es zwei Ansätze. Zum einen werden mit Gensignaturen molekulare Subtypen des Brustkrebses definiert, zum anderen werden Risikoscores berechnet, mit deren Hilfe die Patientinnen in verschiedene Risikogruppen eingeteilt werden. Die im Folgenden dargestellten Methoden wurden in dieser Arbeit nicht als Referenzmethoden verwendet, da wir unabhängig von der molekularen Subgruppe Metastasenbildung vorhersagen möchten.

Perou et al. (2000) haben untersucht, ob sich die Beobachtung, dass Brustkrebs eine sehr heterogene Erkrankung ist, auch auf Genexpressionsebene machen lässt. Mit Hilfe

von hierarchischem Clustern von Genen, die gut zwischen verschiedenen Tumorarten differenzierten, wurden die Tumore in vier molekulare Subgruppen geclustert, die als *intrinsische Subtypen* bezeichnet werden und sich im Wesentlichen im ER- und HER2-Status unterscheiden: Luminal (ER+/HER2-), HER2-enriched (HER2+), Basal-like (ER-/HER2-) und normal-like. In einer späteren Veröffentlichung (Sørlie et al. 2001) wurde das gleiche Verfahren auf einer größeren Kohorte angewendet und gezeigt, dass sich der Luminal-Subtyp weiter aufteilen lässt, wobei schnell proliferierende Tumore (Luminal B) von langsam proliferierenden Tumoren (Luminal A) unterschieden werden. Außerdem konnte gezeigt werden, dass sich die molekularen Subgruppen in Bezug auf Prognose unterscheiden.

Eine Weiterentwicklung dieser Studien ist die PAM50-Signatur (Parker et al. 2009). Die Signatur basiert auf 50 ausgewählten Genen, die am besten zwischen den Subtypen differenzieren (10 pro Subtyp). Zum Kombinieren der Gene wurde PAM (Prediction Analysis for Microarrays) verwendet, was auf der Nearest Shrunken Centroids Methode von Tibshirani et al. (2002) basiert. Außerdem wurde ein Modell zur Vorhersage des Rückfall-Risikos entwickelt (ROR), das auf einer Kombination von PAM50-Subtyp und Tumorgröße und Tumorgrad basiert. Die prognostische Güte von PAM50 wurde in klinischen Studien bestätigt (Nielsen et al. 2010; Filipits et al. 2014). Es wurde ein auf qRT-PCR basierender Assay entwickelt, der heute kommerziell unter dem Namen Prosigna® vertrieben wird.

Eine weitere Gensignatur zum Identifizieren von molekularen Subtypen und Berechnung individueller Risikoscores für die verschiedenen Subtypen ist GENIUS (Gene Expression progNostic Index Using Subtypes) von Haibe-Kains et al. (2012). Die Methode besteht aus drei Schritten. Im ersten Schritt werden mit Hilfe einer Methode des fuzzy-Clustering Subtypen identifiziert, indem für jede Patientin die Wahrscheinlichkeit berechnet wird zu einer der drei molekularen Subtypen (ER-/HER2-, HER2+ und ER+/HER2-) zu gehören. Im zweiten Schritt wird für jeden Subtyp eine individuelle prognostische Signatur entwickelt bzw. werden existierende Signaturen verwendet. Abschließend wird ein Risikoscore durch Kombination der Ergebnisse aus den ersten beiden Schritten bestimmt.

Der Breast Cancer Index (BCI) von Ma et al. (2008) ist ein Assay für FFPE-Gewebe, der auf der Kombination des 5-Gen-Klassifikators MGI (Molecular Grade Index) und

dem Verhältnis der Gene *HOXB13* und *IL17BR*, die mit dem ER-Status assoziiert sind, basiert. Die fünf Gene des MGI wurden auf Basis der Korrelation mit dem Tumorgrad und ihrer funktionalen Annotation ausgewählt und mittels Hauptkomponentenanalyse zu einem Klassifikator kombiniert. *HOXB13* und *IL17BR* wurden in einer Analyse von Tumoren von Frauen mit ER-positivem Brustkrebs im frühen Stadium und Tamoxifen-Monotherapie als differentiell exprimiert zwischen Tumoren mit Rekurrenz und Tumoren ohne Rekurrenz identifiziert. Dabei war *HOXB13* in Tumoren mit Rekurrenz und *IL17BR* in Tumoren ohne Rekurrenz überexprimiert.

EndoPredict (Filipits et al. 2011) ist ebenfalls ein qRT-PCR basierender Assay, der die Expression von acht mit Krebs assoziierten Genen und drei Referenzgenen misst. Er wurde auf einer Kohorte von ER-positiven, HER2-negativen Patienten entwickelt. Der EP-Score ist eine Linearkombination der standardisierten Expressionswerte der acht Krebs-Gene. Die Koeffizienten wurden dabei aus der multivariaten Cox-Regression mit metastasenfreier Zeit als Zielvariable entnommen. Als weiteren Risikoscore schlagen die Autoren den EPclin-Score vor, der als Linearkombination von EP-Score mit der Tumorgröße und der Anzahl der befallenen Lymphknoten berechnet wird. Anhand der Scores werden die Patienten in eine Niedrig- und eine Hoch-Risiko-Gruppe eingeteilt, die entsprechenden Cutoffs wurden auf einer Trainingskohorte ermittelt. Die Scores wurden in zwei großen Studien evaluiert.

3 Statistische Methoden

In diesem Kapitel werden die statistischen Methoden beschrieben, die zur Analyse der Daten in dieser Arbeit verwendet werden. In Kapitel 3.1 werden zunächst die Methoden zur Identifizierung von Genen mit charakteristischer Expressionsverteilung vorgestellt. Grundlage dafür bilden Clusterverfahren, die in Kapitel 3.1.1 beschrieben werden. In Kapitel 3.1.2 erfolgt eine Beschreibung der Bimodalitätsmaße, die in dieser Arbeit verwendet werden. Abschließend erfolgt ein kurzer Überblick über weitere Bimodalitätsmaße. In Kapitel 3.2 werden die Methoden zusammengefasst, die verwendet werden um die prognostische Relevanz der bimodalen Gene zu untersuchen. Dazu erfolgt zunächst eine Einführung in die Methoden der Überlebenszeitanalyse (Kapitel 3.2.1) und anschließend wird auf das Problem des multiplen Testens eingegangen (Kapitel 3.2.5). Methoden zur Analyse von Genlisten in Bezug auf Enrichment mit prognostischen Genen werden in Kapitel 3.3 beschrieben.

Der letzte Abschnitt des Methodenteils behandelt das Kombinieren von einzelnen Genen zu Klassifikatoren. Ein kurzer Überblick über Maßzahlen zur Bewertung der Klassifikationsgüte wird in Kapitel 3.4.1 gegeben. Als Klassifikationsverfahren werden Klassifikationsbäume (Kapitel 3.4.2) und Random Forests (Kapitel 3.4.3) beschrieben.

Alle Analysen werden mit der Statistik-Software R (R Core Team 2016) und Bioconductor (Huber et al. 2015) durchgeführt.

3.1 Finden von Genen mit charakteristischer Expressionsverteilung

Ziel ist es Patientinnen anhand der Expression eines bestimmten Gens in zwei Gruppen einzuteilen. Dabei gibt zwei Hauptansätze:

1. Identifikation von Genen mit klarer Clusterstruktur in der Expressionsverteilung.
2. Identifikation von Expressionsverteilungen mit einer Hauptverteilung und zusätzlichen Ausreißern mit extrem hohen oder niedrigen Expressionswerten in einer Untergruppe der Samples.

Beim ersten Ansatz erhält man die Patienteneinteilung direkt aus dem Ergebnis des Clusterverfahrens. Beim zweiten Ansatz lässt sich eine Hauptgruppe und eine Ausreißergruppe definieren. Im Folgenden sollen zunächst zwei Clusterverfahren vorgestellt werden. Danach werden Scores für die Bimodalität von univariaten Verteilungen vorgestellt. Die Bimodalitätsmaße wurden bereits in Hellwig et al. (2010) beschrieben und verglichen.

3.1.1 Clusteranalyse

k-means-Algorithmus

Ein einfacher Ansatz Patienten anhand der Expressionswerte eines bestimmten Gens zwei Gruppen zuzuordnen ist das Verwenden eines Cluster-Algorithmus. Der *k-means-Algorithmus* ist ein bekanntes und häufig verwendetes Cluster-Verfahren.

Seien im univariaten Fall die Beobachtungen $x_1, \ldots, x_n$ und $k < n$ verschiedene Startwerte für die Clusterzentren gegeben. Der k-means-Algorithmus teilt die Beobachtungen in k Cluster auf, sodass die *Within Cluster Sum of Squares*

$$\mathrm{WSS} = \sum_{j=1}^{k} \sum_{x \in C_j} (x - \bar{x}_j)^2$$

minimiert wird. Dabei ist C_j, $j = 1, \ldots, k$, der j-te Cluster, n_j die Anzahl der Elemente im j-ten Cluster und $\bar{x}_j = \frac{1}{n_j} \sum_{x \in C_j} x$ das zugehörige Clusterzentrum. Der Algorithmus ist ein iterativer Prozess, der zwischen zwei Schritten alterniert:

1. Zuordnen der Beobachtungen zu einem Cluster basierend auf ihrem Abstand zu den Clusterzentren.
2. Aktualisieren der Clusterzentren auf Basis der neuen Elemente in einem Cluster.

Das Ergebnis des k-means-Algorithmus hängt von der Wahl der Startwerte für die Clusterzentren ab (MacQueen 1967). Da hier für ein bestimmtes Gen eine Patientengruppe mit niedrigen und eine Patientengruppe mit hohen Expressionswerten unterschieden werden sollen, werden als Startwerte für die Clusterzentren das Minimum und das Maximum aller Expressionswerte des Gens gewählt.

Modellbasiertes Clustern

Die Annahme bei diesem Clusterverfahren ist es, dass sich die Verteilung Y eines Gens mit bimodaler Expressionsverteilung als Mischung zweier Normalverteilungen Y_1 und Y_2 darstellen lässt. Seien μ_1 und μ_2 die Erwartungswerte dieser Normalverteilungen und σ_1^2 und σ_2^2 die zugehörigen Varianzen, also $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ und $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Dann lässt sich die Verteilungsfunktion von Y schreiben als:

$$Y = \Delta \cdot Y_1 + (1 - \Delta) \cdot Y_2,$$

wobei $\Delta \in \{0, 1\}$ mit $P(\Delta = 1) = p$ (Hastie et al. 2009). p ist die Wahrscheinlichkeit zu Y_1 zu gehören. Sei ϕ_θ die Dichte der Normalverteilung mit $\theta = (\mu, \sigma^2)$. Dann ist die Dichte von Y gegeben durch

$$g_Y(y) = p\phi_{\theta_1}(y) + (1 - p)\phi_{\theta_2}(y).$$

Die Parameter des Mischungsmodells können mit dem Expectation-Maximization (EM) Algorithmus bestimmt werden. Die Parameter der Verteilung sind

$$\theta = (p, \theta_1, \theta_2) = (p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2).$$

Die log-Likelihood für eine Stichprobe der Größe n ist gegeben durch

$$\ell(\theta; Z) = \sum_{i=1}^{n} \log[p \cdot \phi_{\theta_1}(y_i) + (1 - p) \cdot \phi_{\theta_2}(y_i)].$$

Zum Anpassen des Modells an die Daten muss die log-Likelihood-Funktion $\ell(\theta; Z)$ maximiert werden. Dies ist numerisch schwer. Deswegen wird ein einfacherer Ansatz gewählt. Dazu wird eine nicht-beobachtbare latente Variable Δ_i eingeführt, die die Werte 0 oder 1 annehmen kann:

$$\Delta_i = \begin{cases} 1, & \text{wenn } Y_i \text{ aus Modell 1 ist} \\ 0, & \text{wenn } Y_i \text{ aus Modell 2 ist} \end{cases}.$$

Unter der Annahme, dass alle Δ_i bekannt sind, lässt sich die log-Likelihood-Funktion wie folgt schreiben:

$$\begin{aligned} \ell_0(\theta; Z, \Delta) &= \sum_{i=1}^{n} [\Delta_i \cdot \log[p \cdot \phi_{\theta_1}(y_i)] + (1 - \Delta_i) \cdot \log[(1-p) \cdot \phi_{\theta_2}(y_i)]] \\ &= \sum_{i=1}^{n} [\Delta_i \cdot \log[p \cdot \phi_{\theta_1}(y_i)]] + \sum_{i=1}^{n} [(1 - \Delta_i) \cdot \log[(1-p) \cdot \phi_{\theta_2}(y_i)]] \\ &= \sum_{i=1}^{n} [\Delta_i \cdot \log[\phi_{\theta_1}(y_i)] + (1 - \Delta_i) \cdot \log[\phi_{\theta_2}(y_i)]] \\ &\quad + \sum_{i=1}^{n} [\Delta_i \cdot \log[p] + (1 - \Delta_i) \cdot \log[(1-p)]]. \end{aligned}$$

Die Maximum-Likelihood-Schätzer für μ_1 und σ_1^2 wären hier das arithmetische Mittel und die Varianz der Beobachtungen, für die $\Delta_i = 1$ ist. Analog können μ_2 und σ_2^2 durch arithmetisches Mittel und Varianz der Beobachtungen geschätzt werden, für die $\Delta_i = 0$ ist. Der Parameter p lässt sich durch den Anteil von Beobachtungen für die gilt $\Delta_i = 1$ schätzen.

In Wahrheit sind die Werte für Δ_i jedoch nicht bekannt, sie werden daher durch ihren Erwartungswert, gegeben die übrigen Parameter und die Beobachtungen, ersetzt. Diese Erwartungswerte nennt man auch *Reponsibilities*, sie sind gegeben durch

$$\gamma_i(\theta) = E(\Delta_i | \theta, Z) = P(\Delta_i = 1 | \theta, Z).$$

Der EM-Algorithmus wird zur Schätzung der Parameter verwendet. Im Expectation-Schritt werden die Reponsibilities für jede der Beobachtungen berechnet, basierend auf der Dichte mit den aktuellen Parametern. Im Maximization-Schritt werden die

Reponsibilities verwendet um die Parameterschätzer upzudaten, dazu wird gewichtete Maximum-Likelihood-Schätzung verwendet (Algorithmus 1).

Algorithmus 1 EM-Algorithmus für das modellbasierte Clustern (vgl. Hastie et al. (2009), S. 275).

1. Wähle Startwerte für die Parameter $\hat{\mu}_1$, $\hat{\sigma}_1^2$, $\hat{\mu}_2$, $\hat{\sigma}_2^2$ und $\hat{p}$.
2. **Expectation-Schritt:** Berechne die Reponsibilities:

$$\hat{\gamma}_i = \frac{\hat{p} \cdot \phi_{\hat{\theta}_1}(y_i)}{\hat{p} \cdot \phi_{\hat{\theta}_1}(y_i) + (1-\hat{p}) \cdot \phi_{\hat{\theta}_2}(y_i)}.$$

3. **Maximization-Schritt:** Berechne die gewichteten Mittelwerte und Varianzen:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^n \hat{\gamma}_i y_i}{\sum_{i=1}^n \hat{\gamma}_i}, \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^n \hat{\gamma}_i (y_i - \hat{\mu}_1)^2}{\sum_{i=1}^n \hat{\gamma}_i},$$
$$\hat{\mu}_2 = \frac{\sum_{i=1}^n (1-\hat{\gamma}_i) y_i}{\sum_{i=1}^n (1-\hat{\gamma}_i)}, \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^n (1-\hat{\gamma}_i)(y_i - \hat{\mu}_2)^2}{\sum_{i=1}^n (1-\hat{\gamma}_i)}$$

und $\hat{p} = \sum_{i=1}^n \hat{\gamma}_i / n$.

4. Wiederhole Schritt 2 und 3 so lange, bis der Algorithmus konvergiert.

Zum Anpassen der Modelle an unsere Daten benutzen wir das R-Paket `mclust` (Fraley und Raftery 2002; Fraley und Raftery 2006). Dabei wird durch die Einstellung G=2 immer eine Lösung mit 2 Komponenten erzwungen. Um die Startwerte des EM-Algorithmus zu bestimmen, verwendet die Funktion *Mclust* hierarchisches Clustern. Wird keine Voraussetzung zur Gleichheit der Varianzen gemacht, so entscheidet die Funktion auf Basis des Bayesschen Informationskriteriums (engl. Bayesian Information Criterion (BIC) (Schwarz 1978)), ob gleiche oder ungleiche Varianzen verwendet werden. Das BIC ist definiert als

$$\text{BIC} = -2 \cdot \ell(\theta; Z) + (\log n) \cdot d,$$

wobei n die Stichprobengröße und d die Anzahl der Parameter im Modell ist. Es wird das Modell mit dem minimalen BIC ausgewählt. Die Definition des Bimodality Index in Kapitel 3.1.2 setzt voraus, dass die beiden Komponenten die gleiche Varianz haben. Bei den anderen Maßen, die auf dem modellbasierten Clustern beruhen, gibt es keine Voraussetzung für die Varianzen.

Wir möchten die Patienten anhand der Expressionswerte eines Gens in eine Gruppe mit niedrigen und eine Gruppe mit hohen Werten einteilen. Beim modellbasierten Clustern kann es vorkommen, dass Beobachtungen, die am Rand der Verteilungen liegen, der anderen Komponente zugeordnet werden (zwei Schnittpunkte der Dichtefunktionen). Dies macht in unserer Anwendung keinen Sinn. Daher werden die Clusterergebnisse gemäß Abbildung 3.1 korrigiert. Dafür werden die Schnittpunkte der Dichtefunktionen der beiden Komponenten verwendet. Die Berechnung der Schnittpunkte ist im Anhang (Kapitel B) zu finden. Beispiele für die Korrektur sind in Abbildung 3.2 dargestellt.

In Beispiel (a) liegt ein Schnittpunkt der Dichtefunktionen zwischen den Mittelwerten, der zweite Schnittpunkt liegt am rechten Rand des Wertebereiches der Expressionswerte, sodass zwei Beobachtungen mit hohen Expressionswerten dem Cluster mit niedrigen Expressionswerten zugeordnet werden. Wir wählen zur Korrektur den Schnittpunkt zwischen den beiden Mittelwerten als Cutoff für die Gruppeneinteilung. In Beispiel (b) liegen die beiden Schnittpunkte der Dichtefunktionen links und rechts von den Mittelwerten. In diesem Fall zählen wir die Beobachtungen, die links und rechts der Schnittpunkte liegen (hier links 39 und rechts 14 Beobachtungen). Als Cutoff für die Gruppeneinteilung wird nun der Schnittpunkt gewählt, bei dem die Anzahl der Beobachtungen im Randbereich größer ist. In diesem Beispiel ist das der linke Schnittpunkt.

Wie aus Abbildung 3.1 hervorgeht, ist der Schnittpunkt nur definiert, wenn gilt:

$$(\hat{\mu}_1 - \hat{\mu}_2)^2 \geq 2(\hat{\sigma}_1^2 - \hat{\sigma}_2^2) \log\left(\frac{\hat{p} \cdot \hat{\sigma}_2}{(1-\hat{p}) \cdot \hat{\sigma}_1}\right).$$

Wenn diese Bedingung nicht erfüllt ist, werden die Patientinnen nur einer Gruppe zugeordnet. Wenn man die Gleichheit der Varianzen der beiden Komponenten voraussetzt, so haben die beiden Dichtefunktionen genau einen Schnittpunkt. Diesen wählen wir als Cutoff für die Einteilung der Patientinnen. Es kann vorkommen, dass der Schnittpunkt außerhalb des Wertebereichs der Expressionswerte liegt. In diesem Fall gibt es ebenfalls nur eine Gruppe. Bei der Analyse der Daten der Mainz-Kohorte ist dies bei 129 Genen der Fall.

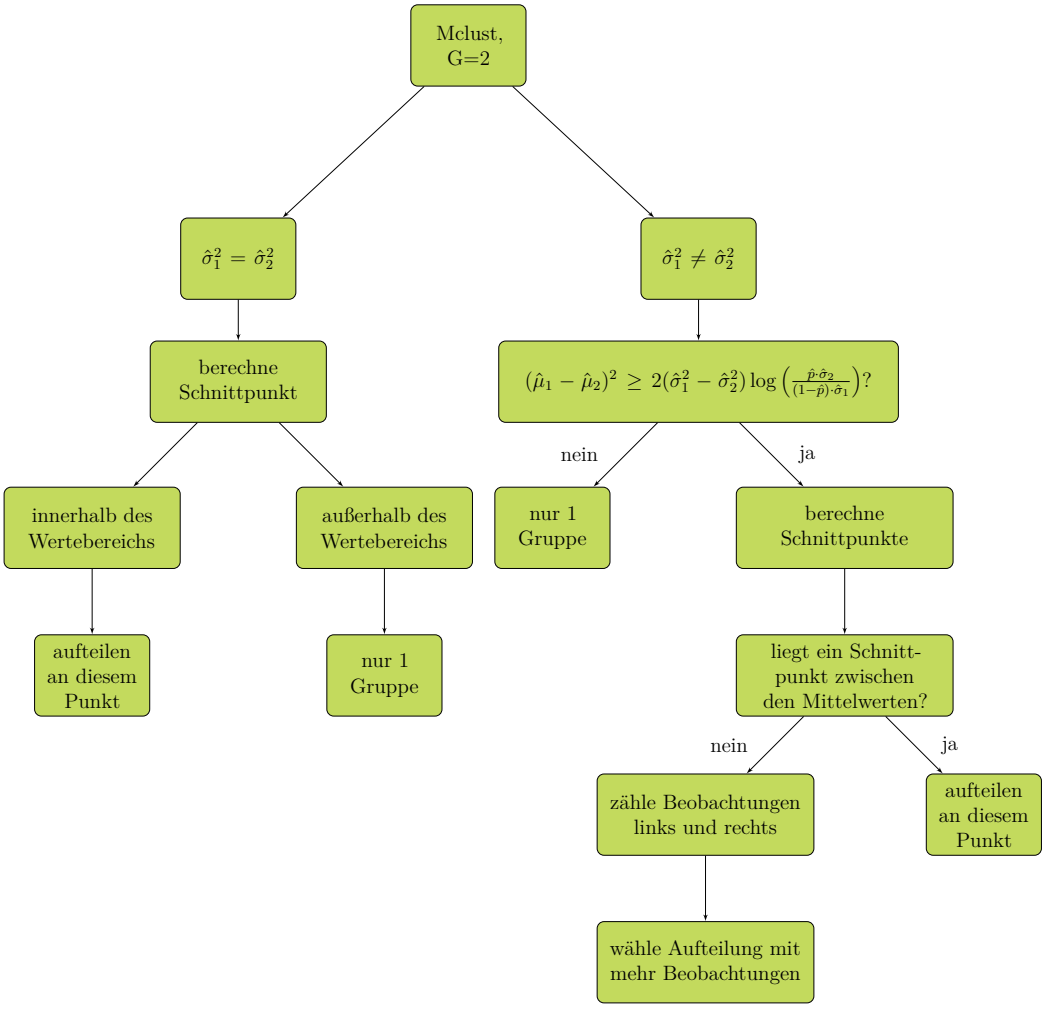

Abbildung 3.1: Korrektur der Clusterergebnisse, schematische Darstellung.

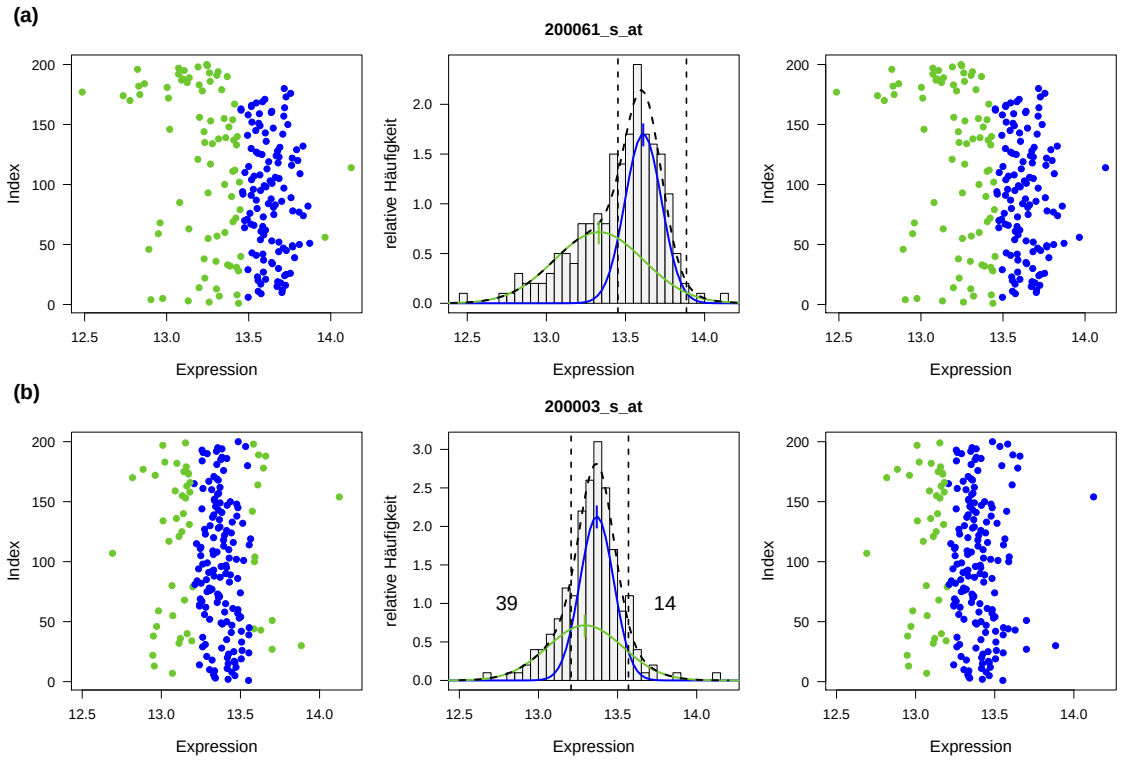

Abbildung 3.2: Beispiele für die Korrektur der Ergebnisse des modellbasierten Clusterns.

3.1.2 Bimodalitäts-Scores

Im Folgenden werden verschiedene Scores zur Beurteilung der Bimodalität von Expressionsverteilungen vorgestellt. Die ersten beiden Maße sollen quantifizieren, wie gut die Cluster-Einteilung des k-means-Algorithmus ist. Es stellt sich für jedes Gen die Frage, ob eine Einteilung der Expressionswerte in zwei Gruppen die Verteilung besser darstellt als eine einzige Gruppe. Dazu konstruieren wir zwei Maße, die auf der Zerlegung der *Total Sum of Squares* $\mathrm{TSS} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ beruhen. Man kann zeigen, dass gilt

$$\begin{aligned} \mathrm{TSS} &= \mathrm{BSS} + \mathrm{WSS} \\ &= \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2 + \sum_{j=1}^{k}\sum_{x \in C_j}(x - \bar{x}_j)^2, \end{aligned}$$

wobei BSS die *Between Cluster Sum of Squares* und WSS die *Within Cluster Sum of Squares* ist, also die Variabilität zwischen den Clustern bzw. innerhalb der Cluster.

Variance Reduction Score

Der *Variance Reduction Score* (VRS) wird definiert als Quotient von WSS und TSS:

$$\mathrm{VRS} := \frac{\mathrm{WSS}}{\mathrm{TSS}}.$$

VRS misst den Grad der Varianzreduktion beim Aufteilen der Daten in zwei Gruppen. Der Wert des Scores liegt im Intervall $[0; 1]$. Ein niedriger Score spricht für eine sinnvolle Einteilung in zwei Gruppen, da ein großer Anteil der Variabilität innerhalb der Daten durch die Variabilität zwischen den Clustern erklärt werden kann, die Variabilität innerhalb der Cluster ist klein.

Weighted Variance Reduction Score

Der *Weighted Variance Reduction Score* (WVRS) ist eine gewichtete Version des VRS. Er misst die Reduktion der Varianz unabhängig von der Clustergröße, also der Anzahl der Elemente innerhalb eines Clusters. Im Zähler des Scores steht das arithmetische

Mittel der Varianzen innerhalb der Cluster. Im Nenner steht die Stichprobenvarianz. Wir definieren

$$\text{WVRS} := \frac{\frac{1}{2}\left(\frac{1}{n_1}\sum_{x\in C_1}(x-\bar{x}_1)^2 + \frac{1}{n_2}\sum_{x\in C_2}(x-\bar{x}_2)^2\right)}{\frac{1}{n}\cdot \text{TSS}}.$$

Wir wählen den nicht-korrigierten Schätzer für die Varianzen, da es vorkommen kann, dass es Aufteilungen der Patientinnen gibt, bei denen in einer Gruppe nur eine Beobachtung ist.

Dieser Score kann auch Gruppeneinteilungen mit extrem unterschiedlichen Gruppengrößen entdecken, da die Varianz kleiner Cluster hier denselben Einfluss hat wie die Varianz großer Cluster. Kleine Werte für WVRS sprechen wieder für eine sinnvolle Gruppeneinteilung. Anders als VRS kann WVRS auch Werte größer 1 annehmen.

Dip-Test auf Unimodalität

Der *dip-Test* auf Unimodalität wurde von Hartigan und Hartigan (1985) vorgeschlagen. Die *dip-Statistik* ist definiert als die maximale Differenz einer empirischen Verteilungsfunktion und derjenigen unimodalen Verteilungsfunktion, die die maximale Differenz minimiert. Damit misst sie den Abstand einer Verteilungsfunktion von der Unimodalität.

Definiere für zwei beschränkte Funktionen F und G $\rho(F,G) := \sup_x |F(x) - G(x)|$. Definiere $\rho(F,\mathscr{A}) := \inf_{G\in\mathscr{A}} \rho(F,G)$ für eine Klasse $\mathscr{A}$ von beschränkten Funktionen. $\mathscr{U}$ sei die Klasse der unimodalen Verteilungsfunktionen. Der *dip* einer Verteilungsfunktion ist dann definiert als $D(F) := \rho(F,\mathscr{U})$. Dabei gilt $D(F_1) \leq D(F_2) + \rho(F_1,F_2)$ und $D(F) = 0$ für $F \in \mathscr{U}$ und $D(F) > 0$ für $F \notin \mathscr{U}$. Die größte konvexe Minorante von F auf $(-\infty;a]$ ist $\sup G(x)$ für $x \leq a$, wobei das Supremum über alle Funktionen G genommen wird, die auf $(-\infty;a]$ konvex und an keiner Stelle größer als F sind. Analog ist die kleinste konkave Majorante von F auf $[a;\infty)$ $\inf H$ für alle $x \geq a$, wobei das Infimum über alle Funktionen H bestimmt wird, die auf dem Intervall $[a;\infty)$ konkav und an keiner Stelle kleiner als F sind.

Die Definition der dip-Statistik wird ausgeweitet auf beschränkte Funktionen F, die auf $(-\infty;0]$ und $[1;\infty)$ konstant sind. Die dip-Statistik kann dann auch definiert werden als $D(F) = \rho(F,\mathscr{V})$, wobei $\mathscr{V}$ die Klasse von Funktionen ist, die konstant auf $(-\infty;0]$ und $[1;\infty)$ und für einige m, $0 \leq m \leq 1$, auf $[0,m]$ konvex und auf $[m;1]$ konkav sind.

Hartigan und Hartigan (1985) zeigen, dass diese Definition konsistent zu der vorherigen Definition ist und beide Definitionen für Verteilungsfunktion auf $[0;1]$ gelten.

Die Nullverteilung der dip-Statistik ist die Gleichverteilung. Sei F_n eine empirische Verteilungsfunktion, Hartigan und Hartigan (1985) zeigen, dass $\sqrt{n}D(F_n)$ für die Gleichverteilung asymptotisch positiv und für unimodale Verteilungen, deren Dichten vom Modus exponentiell abfallen, asymptotisch 0 ist. Es kann gezeigt werden, dass für die Gleichverteilung auf $(0;1)$ gilt $\sqrt{n}D(F_n) \to D(B)$ für $n \to \infty$, wobei B eine Brownsche Brücke ist.

Algorithmus 2 Bestimmung der größten konvexen Minorante und der kleinsten konkaven Majorante von F_n (vgl. Hartigan (1985)).

1: Starte mit $x_L = x_1$, $x_U = x_n$ und $D = 0$.

2: Berechne die größte konvexe Minorante G und die kleinste konkave Majorante H von F_n auf $[x_L; x_U]$, bezeichne die Punkte, an denen die Kurven F_n berühren, mit $g_1, g_2, \ldots, g_k$ und $h_1, h_2, \ldots, h_m$.

3: Setze voraus, dass $d = \max|G(g_i) - H(g_i)| > \max|G(h_j) - H(h_j)|$, und dass das Maximum an der Stelle $h_j \leq g_i \leq h_{j+1}$ liegt. Definiere $x_L^0 = g_i$ und $x_U^0 = h_{j+1}$.

4: Setze voraus, dass $d = \max|G(h_j) - H(h_j)| \geq \max|G(g_i) - H(g_i)|$, und dass das Maximum an der Stelle $g_i \leq h_j \leq g_{i+1}$ liegt. Definiere $x_L^0 = g_i$ und $x_U^0 = h_j$.

5: Wenn $d \leq D$, stoppe und setze dip $= D/2$.

6: Wenn $d > D$ setze

$$D = \max\left(\sup_{x_L \leq x \leq x_L^0} |G(x) - F(X)|, \sup_{x_U \leq x \leq x_U^0} |H(x) - F(X)|\right).$$

7: Setze $x_L = x_L^0$, $x_U = x_U^0$ und kehre zu Schritt 2 zurück.

Die dip-Statistik wird numerisch bestimmt. Seien dazu $x_1, x_2, \ldots, x_n$ die geordneten Beobachtungen. Dann gibt es $n(n-1)/2$ mögliche Intervalle $(x_L; x_U)$, in denen der Modus liegen kann. Berechne für jedes Intervall $(x_i; x_j)$ die größte konvexe Minorante von F_n auf $(-\infty; x_i)$ und die kleinste konkave Majorante von F_n auf $(x_j; \infty)$ gemäß Algorithmus 2.

Zur Berechnung der dip-Statistik verwenden wir das R-Paket `diptest` (Mächler 2009). Dieses stellt auch eine Tabelle mit Quantilen für $D(F_n)$ für verschiedene Stichprobenumfänge basierend auf 1 000 000 Simulationen bereit. p-Werte können durch Interpolation zwischen den Quantilen berechnet werden.

Kurtosis

Teschendorff et al. (2006) haben ein Verfahren mit dem Namen PACK (Profile Analysis using Clustering and Kurtosis) vorgeschlagen um Gene mit bimodaler Expressionsverteilung zu finden, das auf modellbasiertem Clustern und *Kurtosis* basiert. Das Verfahren besteht aus zwei Schritten. Zunächst wird ein Clusteralgorithmus verwendet um für jedes Gen die optimale Anzahl an Clustern zu bestimmen. Für die Gene, für die eine bimodale Verteilung angenommen werden kann, wird dann die Kurtosis berechnet. Diese kann dazu verwendet werden die Gene zu ordnen.

Die Kurtosis einer Zufallsvariable X ist verwandt mit dem 4. zentralen Moment und kann definiert werden als

$$K(X) = \frac{E[(X - \bar{X})^4]}{E[(X - \bar{X})^2]^2} - 3.$$

Für eine Normalverteilung gilt $E[(X - \bar{X})^4] = 3 \cdot E[(X - \bar{X})^2]^2$ und damit $K(X) = 0$. Mit dieser Definition der Kurtosis lässt sich also sagen, dass die Kurtosis den Abstand der Wölbung einer Verteilungsfunktion von der Wölbung der Normalverteilung misst. Betrachtet man eine Mischung von zwei Normalverteilungen mit gleichgroßer Masse, so ist die Kurtosis negativ, da die gemeinsame Verteilung eine flachere Wölbung als die Normalverteilung besitzt. Bei einer Mischung von zwei Normalverteilungen mit stark unterschiedlich großer Masse ist die Kurtosis positiv, da die gemeinsame Verteilung steiler ist und einseitig einen schwereren Rand hat als eine Normalverteilung mit gleicher Varianz (vgl. Abbildung 3.3).

Seien $x = (x_1, \ldots, x_n)$ die Expressionswerte eines Gens, dann ist ein unverzerrter Schätzer für die Kurtosis gegeben durch:

$$K(x) = \frac{n(n+1)\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)(n-2)(n-3)\hat{\sigma}^4} - \frac{3(n-1)^2}{(n-2)(n-3)} \quad \text{mit} \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

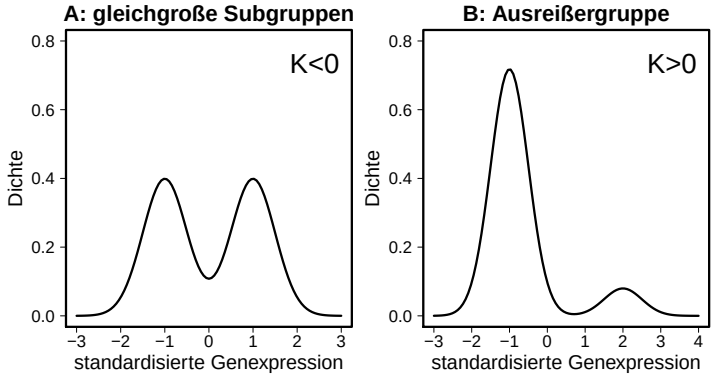

Abbildung 3.3: Beispiele für die Kurtosis von Mischungsmodellen (reproduziert nach Teschendorff et al. (2006)).

Wir verwenden die Kurtosis wie die anderen Maßzahlen zum Ordnen der Gene, ohne vorher zu filtern. Dabei bekommen einmal die Gene mit positiver Kurtosis kleine Ränge und einmal die Gene mit negativer Kurtosis. Zu beachten ist, dass eine Mischung von Normalverteilungen mit dem Verhältnis 80/20 ebenfalls eine Kurtosis nahe 0 haben kann (Wang et al. 2009). Das bedeutet, dass bei diesem Maß unter Umständen interessante Kandidatengene übersehen werden können.

Likelihood Ratio

Ertel und Tozeren (2008) haben vorgeschlagen zur Identifikation von Genen mit bimodaler Expressionsverteilung das Likelihood Ratio einer Normalverteilung und eines Normalverteilungs-Mischungsmodells zu verwenden.

Um einen Score zu berechnen verwenden wir `mclust` zur Anpassung eines Normalverteilungsmodells mit einer Komponente und eines Mischungsmodells mit zwei Komponenten an die Expressionswerte eines Gens (vgl. Kapitel 3.1.1). Seien L_1 und L_2 die zugehörigen Likelihoods, dann berechnet sich das Likelihood Ratio gemäß

$$\mathrm{LR} = \frac{L_2}{L_1}.$$

Kleine Ratios sind ein Hinweis dafür, dass eine Funktion unimodal ist. Ein großes Ratio spricht dafür, dass die Expressionswerte eine bimodale Verteilungsfunktion besitzen.

Bimodality Index

Der Bimodality Index (BI) wurde von Wang et al. (2009) als Kriterium zum Identifizieren und Ordnen von bimodalen Genexpressionsverteilungen vorgestellt. Die zentrale Annahme dieser Methode ist es, dass sich die Verteilungsfunktion eines Gens mit bimodaler Expressionsverteilung als Mischung zweier Normalverteilungen mit Mittelwerten μ_1 und μ_2 und gemeinsamer Standardabweichung σ darstellen lässt. Der standardisierte Abstand δ zwischen den beiden Gruppen wird definiert als

$$\delta = \frac{|\mu_1 - \mu_2|}{\sigma}.$$

Möchte man Gene mit bimodaler Expressionsverteilung finden, so ist die Nullhypothese $\delta = 0$ und die Alternative $\delta > 0$. Zur Definition des Bimodality Index wird zunächst die Fallzahlplanung für ein Experiment mit 2 normalverteilten Gruppen mit Mittelwerten μ_1 und μ_2 und gemeinsamer Standardabweichung σ betrachtet. Will man einen bestimmten Abstand $\delta > 0$ in einem Experiment mit gleichgroßen Gruppen nachweisen, so benötigt man eine Stichprobe der Größe

$$N = \frac{4\left(Z_{\alpha/2} + Z_\beta\right)^2}{\delta^2},$$

wobei $Z_{\alpha/2}$ und Z_β die Quantile der Standardnormalverteilung sind, die zum gewünschten Signifikanzniveau α und der Güte gehören. Um die gleiche Güte bei unterschiedlich großen Gruppen der Größe Mp und $M(1-p)$ zu erreichen, so sollte M so gewählt werden, dass die Varianz des geschätzten standardisierten Abstandes bei ungleich großen Gruppen dem bei gleich großen Gruppen entspricht:

$$M = \frac{\left(Z_{\alpha/2} + Z_\beta\right)^2}{p(1-p)\delta^2}.$$

Der Bimodality Index wird definiert als

$$\mathrm{BI} = [p(1-p)]^{1/2} \cdot \delta,$$

wobei p die Wahrscheinlichkeit für die 1. Komponente ist. Der Bimodality Index kann

aus den Daten geschätzt werden, indem p durch den Anteil der Beobachtungen in der 1. Komponente ersetzt wird und μ_1, μ_2 und σ durch die entsprechenden Schätzer aus dem modellbasierten Clustern.

Outlier-Sum-Statistik

Eine weitere Methode Proben anhand ihrer Genexpressionswerte in zwei Gruppen einzuteilen basiert auf einer Methode von Tibshirani und Hastie (2007). Diese Methode erlaubt es Gene zu finden, deren Expressionswert in einer Subgruppe von Proben ungewöhnlich groß oder klein sind. In ihrer Arbeit gehen Tibshirani und Hastie (2007) davon aus, dass es Proben von erkrankten Personen (z. B. Tumormaterial) und Proben einer Referenzgruppe (z. B. gesunde Zellen) gibt. In unserem Fall haben wir nur Tumorproben, deshalb wenden wir die Methode modifiziert an.

Bezeichne x_{ij} die Expressionswerte des i-ten Gens. Sei med_i der Median des i-ten Gens und mad_i die zugehörige mittlere absolute Abweichung (englisch: *median absolute deviation*), d. h. $\text{mad}_i = 1.4826 \cdot \text{med}[|x_{ij} - \text{med}_i|]$. Die Konstante 1.4826 entspricht dabei ungefähr $1/\Phi^{-1}(0.75)$, wobei Φ die Verteilungsfunktion der Standardnormalverteilung ist. Die Standardisierung mit diesem Faktor sorgt dafür, dass der Schätzer für Normalverteilungen ein konsistenter Schätzer der Standardabweichung ist. Die Expressionswerte des i-ten Gens werden zunächst standardisiert durch

$$x'_{ij} = \frac{x_{ij} - med_i}{mad_i}.$$

Bezeichne $q_r(i)$ das r-te Perzentil der empirischen Verteilungsfunktion der x_{ij}-Werte und $\text{IQR}(i) = q_{75}(i) - q_{25}(i)$ den Interquartilsabstand. Mit Hilfe dieser Größen werden dann alle Werte, die kleiner als $\text{IQR}(i) - q_{25}(i)$ oder größer als $\text{IQR}(i) + q_{75}(i)$ sind, als Ausreißer definiert. Die Outlier-Sum-Statistik für positive Ausreißer wird definiert als

$$W_i = \sum_j x'_{ij} \cdot I[x'_{ij} > q_{75}(i) + \text{IQR}(i)],$$

und analog für negative Ausreißer

$$W'_i = \sum_j x'_{ij} \cdot I[x'_{ij} < q_{25}(i) - \text{IQR}(i)].$$

W_i bzw. W_i' ist groß, wenn es viele Ausreißer gibt oder einige wenige Ausreißer mit extrem großen Werten. Gibt es keine Ausreißer, so gilt $W_i = 0$ und $W_i' = 0$. Die Outlier-Sum-Statistik ist definiert als das Maximum von $|W_i|$ und $|W_i'|$.

Für die weiteren Analysen nehmen wir an, dass die Ausreißer, die zur Outlier-Sum-Statistik gehören, eine Gruppe bilden und alle anderen Proben die andere.

3.1.3 Weitere Bimodalitätsmaße

Zusätzlich zu den in dieser Arbeit untersuchten Maße für Bimodalität wurden in der Literatur weitere Maße vorgeschlagen. Das Maß von Bessarabova et al. (2010) basiert auf Minimierung der Within Sum of Squares. Es wird angenommen, dass die Expressionswerte eines Gens eine Mischung aus zwei Normalverteilungen darstellt. Für die Berechnung des Maßes werden die Expressionswerte eine Gens geordnet und für jede mögliche Partition der Proben in zwei Gruppen die WSS berechnet. Gewählt wird die Einteilung, die die WSS minimiert. Als Maß wird dann eine Art t-Test-Statistik $\tau = \frac{u-l}{\sqrt{\gamma/M}}$ verwendet. Dabei sind u und l die arithmetischen Mittel der Expressionswerte in den beiden Gruppen, γ die WSS und M die Gesamtzahl der Proben. Große Werte von τ sprechen für eine bimodale Verteilung. Der Score verwendet also ähnlich wie der Bimodality Index einen standardisierten Abstand zwischen den Gruppenmitteln. Ein Unterschied der beiden Methoden ist allerdings, dass dieser Abstand beim Bimodality Index mit dem Faktor $[p(1-p)]^{1/2}$ multipliziert wird (wobei p dem Anteil der Proben in der ersten Gruppe entspricht), sodass Einteilungen mit gleichgroßen Gruppen höhere Werte bekommen. Ein Nachteil der Methode von Bessarabova et al. (2010) ist es, dass sie sehr anfällig gegenüber Ausreißern ist, da diese durch die Methode einer Gruppe zugeordnet werden und alle anderen Werte der anderen Gruppe. Die Autoren schlagen vor, alle Gruppen, die weniger als 5 % der Proben enthalten, als Ausreißer zu betrachten. Es existieren auch noch weitere Maße für Bimodalität, die auf dem standardisierten Abstand zwischen den Gruppenmitteln basieren, wie etwa der *Bimodality Seperation Score* von Zhang et al. (2003).

Der *Bimodality Coefficient* aus der CLUSTER Prozedur der statistischen Software SAS (SAS Institute Inc 2008) setzt die Schiefe der Verteilung ins Verhältnis zur Kurtosis.

Große Werte des Maßes sprechen für eine bimodale Verteilung. Verteilungen mit schweren Rändern haben kleine Werte, unabhängig von der Anzahl der Modalwerte. Es existieren noch weitere Maße für Bimodalität, die auf Kurtosis und Schiefe basieren. Eine gute Übersicht ist in Knapp (2007) zu finden.

3.2 Testen der prognostischen Relevanz

Auf Basis der in Kapitel 3.1.1 vorgestellten Clusterverfahren werden die Patientinnen in Abhängigkeit von den Expressionswerten eines bestimmten Gens in zwei Gruppen eingeteilt. Auch die Outlier-Sum-Methode liefert zwei Patientengruppen. Mit Hilfe von Methoden der Überlebenszeitanalyse können die Überlebensfunktionen in den resultierenden Gruppen grafisch dargestellt werden und miteinander verglichen werden. Diese Analyse erlaubt es uns Aussagen über die prognostische Relevanz eines Gens zu treffen.

3.2.1 Grundlagen der Überlebenszeitanalyse

In der Überlebenszeitanalyse, auch Ereigniszeitanalyse genannt, betrachtet man den Zeitraum vom Startzeitpunkt bis zum Eintritt eines bestimmten Ereignisses. Im Fall von Krebs-Daten kann das der Zeitraum vom Tag der Diagnose bis zum Tod sein. Häufig verwendet man aber auch andere Ziel-Ereignisse, etwa das Auftreten eines Rezidives nach der Operation oder das Auftreten von Fernmetastasen. Für jedes Individuum definiert seine Überlebenszeit den Zeitraum zwischen dem festgelegten Start-Zeitpunkt und dem Eintritt des Ziel-Ereignisses. Nicht immer wird der Zeitpunkt des Eintretens des Ereignisses für alle Personen beobachtet. Man spricht in diesem Fall von zensierten Ereignissen. Ist das Ereignis eingetreten, bevor die Beobachtung beginnt, so nennt man die Ereigniszeit links-zensiert. Bei Krebs-Daten kommt es häufig vor, dass nicht bei allen Patienten im Beobachtungszeitraum ein Ereignis eintritt. Die Ereigniszeiten dieser Patienten werden dann als rechts-zensiert bezeichnet.

Eine zentrale Annahme der Überlebenszeitanalyse ist, dass die Zensierungen unabhängig von den Ereigniszeiten sind. Bei den meisten Studien kann davon ausgegangen werden,

dass diese Annahme erfüllt ist. Verletzt wäre die Annahme zum Beispiel, wenn das vorzeitige Ausscheiden eines Patienten aus der Studie auf einen besonders guten oder schlechten Gesundheitszustand zurückzuführen ist. Dies kann zu einer systematischen Verzerrung der Ergebnisse führen. Die Survivalfunktion S ist definiert als Wahrscheinlichkeit, dass ein Ereignis erst nach einem Zeitpunkt t eintritt, das heißt es gilt

$$S(t) = P(T > t),$$

wobei T eine positive Zufallsvariable sei, die die Überlebenszeit beschreibt. Die Beziehung der Survivalfunktion zur Verteilungsfunktion F von T ist leicht zu erkennen, es gilt $S(t) = 1 - F(t) \forall t$. S ist eine monoton fallende Funktion und es gilt $\lim_{t\to 0} S(t) = 1$ und $\lim_{t\to\infty} S(t) = 0$. Ist T eine stetige Zufallsvariable gilt, ebenfalls

$$S(t) = P(T > t) = \int_t^\infty f(t)dt,$$

das heißt die Survivalfunktion ist das Integral der Dichtefunktion $f(t)$. In Kapitel 3.2.2 wird der Kaplan-Meier-Schätzer als Schätzer der Survivalfunktion vorgestellt.

Ein weiterer wichtiger Begriff der Überlebenszeitanalyse ist der der *Hazardrate*. Die Hazardrate ist die unmittelbare Ausfallrate zum Zeitpunkt t, bedingt darauf, dass ein Individuum den Zeitpunkt t erreicht hat. Sie wird definiert als

$$\lambda(t) = \lim_{h\to 0} \frac{P(t \leq T < t + h | T \geq t)}{h}.$$

Es gilt $\lambda(t) \geq 0$ für alle $t \geq 0$. Ist T eine stetige Zufallsvariable, so gilt

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t} \ln[S(t)].$$

Eine verwandte Größe ist die sogenannte kumulative Hazardrate $\Lambda(t)$, diese ist definiert als

$$\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u = \int_0^t \frac{f(u)}{S(u)}\mathrm{d}u = -\int_0^t \frac{\frac{\mathrm{d}}{\mathrm{d}u}S(u)}{S(u)}\mathrm{d}u = -\ln(S(t)).$$

Und damit ist

$$S(t) = \exp(-\Lambda(t)) = \exp\left[-\int_0^t \lambda(u)du\right].$$

In Kapitel 3.2.3 wird der Nelson-Aalen-Schätzer als Schätzer für die kumulative Hazardrate vorgestellt. Auf diesem Schätzer basiert der Log-Rank-Test, der in Kapitel 3.2.4 beschrieben wird.

3.2.2 Kaplan-Meier-Schätzer

Bezeichne $t_i,\ i = 1, \ldots, T$ die geordneten Zeitpunkte, an denen ein Ereignis eintritt. Sei $Y_i,\ i = 1, \ldots, T$ die Anzahl der Personen unter Risiko zum Zeitpunkt t_i, das heißt die Anzahl der Personen, die bis unmittelbar vor t_i kein Ereignis erlebt haben, und $d_i,\ i = 1, \ldots, T$ die Anzahl der Ereignisse zum Zeitpunkt t_i. Dann ist der Kaplan-Meier-Schätzer definiert als

$$\hat{S}(t) = \begin{cases} 1 & \text{, für } t < t_1 \\ \prod\limits_{t_i \leq t} \left(1 - \frac{d_i}{Y_i}\right) & \text{, für } t \geq t_1 \end{cases}.$$

Man berechnet den Schätzer demnach iterativ als Produkt der geschätzten bedingten Wahrscheinlichkeiten, dass ein Ereignis im Intervall $[t_i; t_{i+1})$ nicht eintritt, unter der Bedingung, dass es sich vor t_i noch nicht ereignet hat. Man nennt den Schätzer daher auch Produkt-Limit-Schätzer. Der Kaplan-Meier-Schätzer wurde von Kaplan und Meier (1958) vorgeschlagen und ist der am häufigsten verwendete Schätzer für die Survivalfunktion.

$\hat{S}(t)$ ist eine monoton fallende Treppenfunktion mit Sprungstellen an den Ereigniszeitpunkten t_i. Zu beachten ist, dass die Höhe des Sprunges zum Zeitpunkt t_i nicht nur von der Anzahl der Ereignisse d_i zu diesem Zeitpunkt abhängt, sondern auch von den zensierten Daten bis zum Zeitpunkt t_i. Diese gehen über die Menge der Personen unter Risiko in die Berechnung ein. Der Kaplan-Meier-Schätzer ist ein effizienter Schätzer für die Survivalfunktion für rechts-zensierte Daten. Wenn keine Zensierungen vorliegen, gilt $\hat{S}(t) = 1 - \hat{F}(t)$, wobei $\hat{F}$ die empirische Verteilungsfunktion der Überlebenszeiten ist.

Für Zeiten größer als der letzte beobachtete Zeitpunkt ist der Kaplan-Meier-Schätzer nicht wohldefiniert. Wenn die Beobachtung, die zum größten Beobachtungszeitpunkt t_{max} gehört, zensiert ist, dann ist der Wert von $S(t)$ für alle größeren Zeitpunkte unbestimmt, da nicht bekannt ist, ob die letzte Person das Ereignis erlebt hätte, wenn die Ereigniszeit nicht zensiert wäre (vgl. Klein und Moeschberger (2005), Kapitel 4.2). Es wurden verschiedene Methoden vorgestellt $\hat{S}(t)$ für größere Zeitpunkte zu schätzen. Zum Beispiel schlägt Efron (1967) vor $\hat{S}(t)$ für alle $t > t_{max}$ durch 0 zu schätzen, was der Annahme entspricht, dass die letzte Person unmittelbar nach t_{max} das Ereignis erlebt. Gill (1980) schlägt vor $\hat{S}(t)$ für alle $t > t_{max}$ durch $\hat{S}(t_{max})$ zu schätzen. Dies entspricht der Annahme, dass die letzte Person das Ereignis zum Zeitpunkt ∞ erlebt. Der Schätzer von Efron ist negativ verzerrt, der Schätzer von Gill positiv verzerrt (vgl. Klein und Moeschberger (2005), Kapitel 4.2). Asymptotisch sind die beiden Schätzer äquivalent, aber für kleine bis mittlere Stichprobenumfänge hat der Schätzer von Gill einen kleineren Bias und MSE (Klein 1991). Er wird deswegen bevorzugt verwendet. Die Funktion *survfit* der R-Paketes `survival` verwendet den Schätzer von Gill.

Der Kaplan-Meier-Schätzer kann wegen der Beziehung der Survivalfunktion zur kumulativen Hazardrate auch zum Schätzen von $\Lambda(t)$ verwendet werden: $\hat{\Lambda}(t) = -\ln(\hat{S}(t))$. Ein weiterer Schätzer für $\Lambda(t)$ ist der Nelson-Aalen-Schätzer, der im folgenden Abschnitt vorgestellt wird.

3.2.3 Nelson-Aalen-Schätzer

Der Nelson-Aalen Schätzer wurde erstmals von Nelson (1972) vorgestellt und unabhängig davon von Aalen (1978) hergeleitet. Dieser nichtparametrische Schätzer ist für alle $t \leq t_{max}$ definiert als

$$\tilde{\Lambda}(t) = \begin{cases} 0 & \text{, für } t < t_1 \\ \sum_{t_i \leq t} \frac{d_i}{Y_i} & \text{, für } t \geq t_1 \end{cases} .$$

Dabei sei wieder Y_i die Anzahl der Personen unter Risiko zum Zeitpunkt t_i und d_i die Anzahl der Ereignisse zum Zeitpunkt t_i.

Für kleinere Stichprobenumfänge ist der Nelson-Aalen-Schätzer ein besserer Schätzer für die kumulative Hazardrate als der Schätzer, der auf dem Kaplan-Meier-Schätzer beruht (Klein und Moeschberger 2005).

Die Varianz des Schätzers kann geschätzt werden durch

$$\hat{\sigma}^2_\Lambda(t) = \sum_{t_i \leq t} \frac{d_i}{Y_i^2}.$$

Aus dem Nelson-Aalen-Schätzer kann gemäß $\tilde{S}(t) = \exp(-\tilde{\Lambda}(t))$ ein alternativer Schätzer für die Survivalfunktion hergeleitet werden.

3.2.4 Log-Rank-Test

Auf Basis der in Kapitel 3.1.1 vorgestellten Clusterverfahren werden die Patientinnen in Abhängigkeit von den Expressionswerten eines bestimmten Gens in zwei Gruppen eingeteilt. Auch die Outlier-Sum-Methode liefert zwei Patientengruppen. Die Hazardraten der beiden Gruppen können mit Hilfe des Log-Rank-Tests verglichen werden. Es wird die Nullhypothese getestet, dass sich die Hazardraten in den beiden Gruppen nicht unterscheiden gegen die Alternativhypothese, dass es mindestens einen Zeitpunkt $t \leq \tau$ gibt, an dem sich die Hazardraten unterscheiden. Formal haben wir also folgendes Testproblem:

$$H_0\text{: } \lambda_1(t) = \lambda_2(t) \text{ für alle } t \leq \tau \qquad \textit{vs.}$$

$$H_1\text{: es gibt mindestens ein } t \leq \tau \text{ für das gilt } \lambda_1(t) \neq \lambda_2(t).$$

Dabei ist τ der größte Zeitpunkt, an dem sich in beiden Gruppen noch mindestens eine Person unter Risiko befindet. Zum Überprüfen dieser Hypothesen verwenden wir den Log-Rank-Test, der erstmals von Mantel (1966) vorgeschlagen wurde. Dieser Test basiert auf dem Nelson-Aalen-Schätzer für die kumulative Hazardrate und wurde für $k \geq 2$ entwickelt. Wir beschränken uns hier jedoch auf den Fall $k = 2$, da wir ausschließlich die Hazardraten zweier Gruppen vergleichen möchten. Es handelt sich um einen nichtparametrischen Test für rechts-zensierte Daten.

Seien dafür $t_1 < t_2 < \ldots < t_D$ die verschiedenen Ereigniszeitpunkte in der gepoolten Stichprobe. Dann bezeichne d_{ij} die Anzahl der Ereignisse in der j-ten Stichprobe von Y_{ij} Personen unter Risiko zum Zeitpunkt t_i, $j = 1, 2$, $i = 1, \ldots, D$. Sei $d_i = d_{i1} + d_{i2}$ die Anzahl der Ereignisse zum Zeitpunkt t_i in der gepoolten Stichprobe und $Y_i = Y_{i1} + Y_{i2}$ die zugehörige Anzahl von Personen unter Risiko. Dann ist die Teststatistik des Log-Rank-Tests gegeben durch

$$Z := \frac{\sum_{i=1}^{D} \left[d_{i1} - Y_{i1} \left(\frac{d_i}{Y_i} \right) \right]}{\sqrt{\sum_{i=1}^{D} \frac{Y_{i1}}{Y_i} \left(1 - \frac{Y_{i1}}{Y_i} \right) \left(\frac{Y_i - d_i}{Y_i - 1} \right) d_i}}.$$

Wenn die Nullhypothese wahr ist und die Stichprobengrößen in den beiden Gruppen ähnlich sind, dann ist Z asymptotisch standardnormalverteilt. H_0 wird dann abgelehnt, wenn $|Z| > Z_{1-\alpha/2}$, wobei $Z_{1-\alpha/2}$ das $1 - \alpha/2$-Quantil der Standardnormalverteilung ist. Der Log-Rank-Test hat die größte Power, wenn die Hazardraten der zwei Gruppen proportional zueinander sind (Klein und Moeschberger 2005).

Unterscheiden sich die Stichprobengrößen sehr stark, so ist die Annahme der asymptotischen Standardnormalverteilung nicht mehr erfüllt (Latta 1981; Kellerer und Chmelevsky 1983). Dies gilt insbesondere wenn eine Gruppe weniger als 5 Personen enthält. Da viele Gene eine Aufteilung mit extrem unterschiedlich großen Gruppen liefern, wird in dieser Arbeit ein Permutationstest verwendet um die p-Werte zu bestimmen. Dazu wurden 100 000 zufällige Permutationen der Patienten erzeugt und für jede mögliche Kombination von Gruppengrößen die Teststatistik des Log-Rank-Tests berechnet. Bei vorgegebener Anzahl von Patienten in beiden Gruppen ist der p-Wert für eine Teststatistik dann der Anteil der Teststatistiken, die größer sind als der beobachtete Wert.

Bei Verwendung des k-means-Algorithmus wie in Kapitel 3.1.1 beschrieben erhalten wir immer zwei Patientengruppen. Beim modellbasierten Clustern kann es sein, dass alle Patientinnen in eine Gruppe eingeteilt werden. Ebenso kann es vorkommen, dass bei der Outlier-Sum-Methode keine Ausreißergruppe gefunden wird. In diesen Fällen kann der Log-Rank-Test nicht angewendet werden. Für die weiteren Analysen wird den entsprechenden Genen der p-Wert 1 zugewiesen.

3.2.5 Multiples Testen

Wir verwenden den in Kapitel 3.2.4 beschriebenen Log-Rank-Test um für jedes Gen die Hazardraten in den durch die verschiedenen Methoden bestimmten Gruppen zu vergleichen und so die prognostische Relevanz des jeweiligen Gens zu beurteilen. Pro Methode bedeutet das 22 283 Tests, es handelt sich also um ein multiples Testproblem.

Bei einem statistischen Test unterscheidet man zwischen zwei Fehlerarten. Als Fehler 1. Art bezeichnet man den Fehler eine wahre Nullhypothese abzulehnen. Als Fehler 2. Art wird es bezeichnet eine falsche Nullhypothese fälschlicherweise nicht abzulehnen. Das lokale Signifikanzniveau α eines Hypothesentests kontrolliert den Fehler 1. Art. Die Wahrscheinlichkeit eine falsche Nullhypothese korrekterweise abzulehnen ist also $1 - \alpha$. Werden m unabhängige wahre Nullhypothesen gleichzeitig getestet, so würde man erwarten, dass durchschnittlich $m \cdot \alpha$ p-Werte p kleiner α sind. Aus diesem Grund reicht es nicht aus das lokale Signifikanzniveau zu kontrollieren, die p-Werte müssen für das multiple Testen adjustiert werden.

Es gibt verschiedene Ansätze um das multiple Testniveau α einzuhalten (Foulkes 2009). Eine Art von Methoden kontrolliert die sogenannte *Familywise Error Rate* (FWER), womit die Wahrscheinlichkeit bezeichnet wird mindestens eine wahre Nullhypothese abzulehnen. Die bekannteste und einfachste Methode ist die von Bonferroni, bei der sich der adjustierte p-Wert als Produkt des alten p-Wertes und der Anzahl der durchgeführten Tests ergibt. Wenn das Produkt größer als 1 ist, wird der p-Wert auf 1 gesetzt. Diese Methode ist sehr konservativ und für hochdimensionale Daten wenig geeignet. Die ebenfalls sehr bekannte Methode von Bonferroni-Holm (Holm 1979) ist weniger konservativ, für hochdimensionale Daten jedoch auch wenig geeignet. Für dieses Verfahren werden die m p-Werte der Größe nach geordnet. Seien $p_{(1)}, \ldots, p_{(m)}$ die geordneten p-Werte, es gelte also $p_{(i)} \leq p_{(i+1)}$, für $i = 1, \ldots, m-1$. Der adjustierte p-Wert berechnet sich als:

$$\bar{p}_{(i)} = \min\left(p_{(i)} \cdot \frac{m}{m-i}, 1\right).$$

Ein anderer Ansatz ist die Kontrolle der sogenannten *False Discovery Rate* (FDR). Als diese wird der erwartete Anteil von irrtümlich abgelehnten Nullhypothesen unter der Gesamtheit von abgelehnten Nullhypothesen definiert. Bei diesem Ansatz wird akzeptiert,

dass bei Vergrößerung von m öfter wahre Nullhypothesen abgelehnt werden. Mit den Bezeichnungen aus Tabelle 3.1 kann die FDR formal definiert werden als FDR $= E\left(\frac{V}{R}\right)$ für $R > 0$ und FDR $= 0$ für $R = 0$, wobei V eine Zufallsvariable ist, die die fälschlicherweise abgelehnten Nullhypothesen beschreibt, und R eine Zufallsvariable, die die insgesamt abgelehnten Nullhypothesen beschreibt. Es soll gelten: FDR $\leq \alpha$, das heißt, der Anteil falsch abgelehnter Nullhypothesen unter allen abgelehnten Nullhypothesen soll kleiner als das globale Testniveau sein. Hierbei wird also eine gewisse Anzahl von fälschlicherweise abgelehnten Hypothesen von vornherein zugelassen. Die Anzahl hängt davon ab, wie viele Nullhypothesen insgesamt abgelehnt werden.

Tabelle 3.1: Anzahl der Fehler beim Testen von m Nullhypothesen (reproduziert nach Benjamini und Hochberg (1995))

.

	nicht abgelehnt	abgelehnt	Total
wahre Nullhypothesen	U	V	m_0
wahre Alternativen	T	S	$m - m_0$
Total	m - R	R	m

Die häufig verwendete Methode von Benjamini und Hochberg (1995) kontrolliert die FDR. Seien $p_{(1)}, \ldots, p_{(m)}$ wieder die der Größe nach geordneten p-Werte von m Hypothesentests. Dann können die adjustierten p-Werte $\tilde{p}_{(i)}$ iterativ bestimmt werden wie folgt:

$$\tilde{p}_{(i)} = \min\left(p_{(i)} \cdot \frac{m}{i}, \tilde{p}_{(i+1)}\right), \; i = m-1, \ldots, 1,$$

wobei gilt $\tilde{p}_{(m)} = p_{(m)}$. Das heißt, die Adjustierung beginnt mit dem größten p-Wert und läuft dann schrittweise über die geordneten p-Werte. $\tilde{p}$ wird in der Literatur auch *q-Wert* genannt. Abgelehnt werden nur die Hypothesen, deren zugehörige q-Werte kleiner als das globale Testniveau sind. Wenn bereits $\tilde{p}_{(m)} \leq \alpha$ gilt, dann werden alle Nullhypothesen abgelehnt (Benjamini und Hochberg 1995). Diese Adjustierungs-Methode ist weniger konservativ als die Bonferroni-Holm-Methode und wird daher seit eineinhalb Dekaden als Standardmethode für hochdimensionale Genexpressionsdaten verwendet (Tsai et al. 2003). Die Benjamini-Hochberg-Methode wird auch in dieser Arbeit verwendet um die p-Werte der Log-Rank-Tests zu adjustieren.

3.3 Analysieren von Genlisten in Bezug auf Enrichment mit prognostischen Genen

Es soll untersucht werden, ob sich unter den Genen mit auffälligen Bimodalitäts-Scores besonders viele Gene befinden, bei denen die resultierenden Gruppen sich in Bezug auf Prognose unterscheiden. Dazu verwenden wir zwei unterschiedliche Methoden. Für beide Ansätze müssen die Gene in Bezug auf die Bimodalitätsmaße geordnet werden. Beim Fisher-Test werden für den jeweiligen Bimodalitäts-Score und den zugehörigen p-Wert des Log-Rank-Tests Cutoffs festgelegt und damit eine Kontingenztafel aufgestellt. Wir testen dann auf Unabhängigkeit in dieser Vierfeldertafel. Mit Hilfe des Kolmogorov-Smirnov-Tests wird überprüft, ob die Ränge der Gene mit kleinen p-Werten des Log-Rank-Tests gleichverteilt sind.

Fisher-Test

Wir legen für jeden Bimodalitäts-Score und den zugehörigen p-Wert des Log-Rank-Tests einen Cutoff fest. Die M Gene mit den kleinsten p-Werten werden als prognostisch bezeichnet. Sei N die Anzahl der Gene auf dem Array und K die Anzahl der Top-Gene des jeweiligen Scores, dann erhalten wir die Kontingenztafel aus Tabelle 3.2.

Tabelle 3.2: Kontingenztafel für den Fisher-Test.

	unter Top-Genen	nicht unter Top-Genen	$\sum$
prognostisch	X	$M - X$	M
nicht prognostisch	$K - X$	$N - M - K + X$	$N - M$
$\sum$	K	$N - K$	N

Dann ist X hypergeometrisch verteilt mit $X \sim \mathcal{H}(K, N, N - M)$. Für eine Realisierung x dieser Teststatistik lässt sich der p-Werte des Fisher-Tests berechnen gemäß

$$p = 1 - \sum_{i=1}^{x} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}}.$$

Kolmogorov-Smirnov-Test

Anders als beim Fisher-Test wird beim Test vom Kolmogorov-Smirnov-Typ kein Cutoff für den Bimodalitäts-Score festgelegt. Die Gene auf Basis des jeweiligen Scores geordnet und die Gene mit den M kleinsten p-Werten des Log-Rank-Tests ausgewählt. Dass die Gene mit den auffälligsten Bimodalitäts-Scores auch prognostisch sind, ist dann gleichbedeutend damit, dass die M prognostischen Gene in der geordneten Genliste kleine Ränge haben. Wenn es keinen Zusammenhang zwischen Bimodalitäts-Score und prognostischer Bedeutung der Gene gibt, so kann man annehmen, dass die Ränge der prognostischen Gene gleichverteilt sind.

Der Kolmogorov-Smirnov-Test ist ein nichtparametrischer Test auf Gleichheit einer Verteilung mit einer vorgegebenen Referenzverteilung. Die Teststatistik misst den Abstand zwischen der empirischen Verteilungsfunktion der Stichprobe und der kumulativen Verteilungsfunktion der Referenzverteilung. Wir verwenden den einseitigen Kolmogorov-Smirnov-Test um die Nullhypothese zu testen, dass die Verteilung der Ränge der prognostischen Gene nicht größer ist als die Verteilungsfunktion einer Gleichverteilung.

Sei F_n die empirischen Verteilungsfunktion und F die Verteilungsfunktion der Referenzverteilung, dann ist die Teststatistik des einseitigen Kolmogorov-Smirnov-Tests gegeben durch

$$D_n^+ = \sup_x [F(x) - F_n(x)].$$

Die Nullverteilung der Statistik wird unter der Nullhypothese bestimmt, dass die Stichprobe aus der Referenzverteilung stammt. Wenn F stetig ist, hängt die Verteilung von D_n^+ nicht von F ab, sondern nur von der Stichprobengröße n. Für $n > 40$ können approximative p-Werte über die asymptotische Verteilung von $\sqrt{n}D_n^+$ berechnet werden (Conover 1999).

Zur Visualisierung des Enrichments kann die sogenannte *Running-Sum-Statistik* geplottet werden. Sei dazu N die Gesamtzahl der Gene und M die Anzahl der prognostischen Gene. Die Statistik berechnet sich dann wie folgt:

- Wenn ein Gen zu den prognostischen Genen gehört, addiere $N - M$.
- Wenn es nicht zu den prognostischen Genen gehört, subtrahiere M.

Die Gesamtsumme ist immer 0. Die maximale Abweichung von 0 könnte alternativ als Teststatistik verwendet werden. Die p-Werte können dann bestimmt werden, indem man die Gene zufällig permutiert und die zugehörige Teststatistik berechnet.

3.4 Kombinieren von Genen zu Klassifikatoren

3.4.1 Maßzahlen für die Vorhersagegüte

Die Vorhersagegüte eines Klassifikationsverfahrens kann mit Hilfe verschiedener Maßzahlen beurteilt werden. Betrachte dafür zunächst die Kontingenztafel für das Klassifikationsergebnis (Tabelle 3.3).

Tabelle 3.3: Kontingenztafel für die Klassifikationsergebnisse.

		wahre Gruppenzugehörigkeit		
		D+	D-	$\sum$
Klassifikationsergebnis	T+	n_{11}	n_{12}	$n_{1\cdot}$
	T-	n_{21}	n_{22}	$n_{2\cdot}$
	$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	n

Die Anzahl der Klassifikationsfehler berechnet sich mit der Notation dieser Kontingenztafel gemäß $n_{12} + n_{21}$. Die *Sensitivität*, auch Richtig-Positiv-Rate (englisch: true positive rate (TPR)) genannt, ist die bedingte Wahrscheinlichkeit, dass das Klassifikationsverfahren eine Person als krank klassifiziert (T+) gegeben der Voraussetzung, dass die Person tatsächlich krank ist (D+), formell TPR = P(T+|D+). Sie kann aus der Kontingenztafel geschätzt werden als Anteil der als krank klassifizierten Personen unter allen Erkrankten ($n_{11}/n_{\cdot 1}$).

Die *Spezifität* wird auch als Richtig-Negativ-Rate (englisch: true negative rate (TNR)) bezeichnet. Sie ist die bedingte Wahrscheinlichkeit, dass das Verfahren eine nicht erkrankte Person (D-) als nicht-erkrankt (T-) klassifiziert (TNR = P(T-|D-)) und wird geschätzt durch $n_{22}/n_{\cdot 2}$. Die bedingte Wahrscheinlichkeit, dass eine Person erkrankt ist, wenn das Verfahren ihn als erkrankt klassifiziert (PPV = P(D+|T+)), wird als *positiv*

prädiktiver Wert (englisch: positive predictive value (PPV)) bezeichnet. Analog ist der *negativ prädiktive Wert* (englisch: negative predictive value (NPV)) die bedingte Wahrscheinlichkeit, dass eine Person nicht erkrankt ist, wenn sie als nicht-erkrankt klassifiziert wird (NPV = P(D-|T-)). Die Schätzer für NPV und PPV sind nur dann gut, wenn die Prävalenz der Erkrankung derjenigen in der Zielpopulation entspricht. Eine Übersicht über die Maßzahlen und ihrer Schätzer aus der Kontingenztafel ist in Tabelle 3.4 zu finden.

Wenn alle Personen als erkrankt klassifiziert werden, kann NPV nicht durch den oben angegebenen Schätzer geschätzt werden da $n_{2\cdot} = 0$ ist. Wir setzen in diesem Fall den NPV-Wert auf 0.

Tabelle 3.4: Übersicht über die Maßzahlen für die Vorhersagegüte und ihre Schätzer.

Maßzahl	Definition	Schätzer
Klassifikationsfehler		$n_{12} + n_{21}$
Sensitivität (TPR)	P(T+|D+)	$n_{11}/n_{\cdot 1}$
Spezifität (TNR)	P(T-|D-)	$n_{22}/n_{\cdot 2}$
positiv prädiktiver Wert (PPV)	P(D+|T+)	$n_{11}/n_{1\cdot}$
negativ prädiktiver Wert (NPV)	P(D-|T-)	$n_{22}/n_{2\cdot}$

3.4.2 Klassifikationsbäume

Entscheidungsbäume stellen eine Klassifikationsmethode dar, die sich dadurch auszeichnet, dass die Klassifikationsregeln leicht verständlich sind. Ein Baum besteht aus Knoten, die sich anhand einer Splittingvariablen in zwei Unterknoten aufteilen. Am Anfang des Baumes steht der Wurzelknoten, der alle Beobachtungen enthält. Die Terminal- oder Endknoten des Baumes werden auch als Blätter bezeichnet. Die Daten werden rekursiv partioniert. Ein bekannter Algorithmus für Klassifikationsbäume ist der CART-Algorithmus (Classification and Regression Trees), der von Breiman et al. (1984) entwickelt wurde. Die Methode ist in dem R-Paket `rpart` (Recursive Partitioning and Regression

Trees (Therneau et al. 2015)) implementiert, das in dieser Arbeit für die Konstruktion von Klassifikationsbäumen verwendet wird.

Es sollen n Beobachtungen anhand von m Variablen in C Klassen eingeteilt werden. Dabei werden nur binäre Splits gemacht, das bedeutet, ein Knoten A wird in genau zwei Unterknoten A_L und A_R aufgeteilt. An jedem Knoten steht jede Variable als mögliche Splittingvariable zur Verfügung. Es wird der Split gewählt, der die Reduktion der Unreinheit des Knotens maximiert. Das Verfahren wird so lange durchgeführt, bis ein Stopp-Kriterium erfüllt wird (z.B. Mindestanzahl von Objekten in einem Knoten) oder keine Verbesserung mehr erzielt werden kann. Ein Endknoten wird dann der Klasse zugeordnet, die am häufigsten vertreten ist.

Ein mögliches Unreinheitsmaß ist der Gini-Index definiert als $\text{Gini} = 1 - \sum_{i=1}^{C} p_i^2$, wobei p_i für $i = 1, \ldots, C$ die Wahrscheinlichkeit angibt, zur i-ten Klasse zu gehören. Der Gini-Index ist maximal, wenn alle p_i gleich sind, das heißt wenn jede Klasse an diesem Knoten gleich wahrscheinlich ist. Der Gini-Index ist auch interpretierbar als Wahrscheinlichkeit einer Fehlklassifikation, wenn ein Objekt zufällig mit den Wahrscheinlichkeiten $(p_1, p_2, \ldots, p_C)$ aus C Klassen gezogen und dann mit den gleichen Wahrscheinlichkeiten einer der C Klassen zugeordnet wird:

$$\sum_i \sum_{j \neq i} p_i p_j = \sum_i \sum_j p_i p_j - \sum_i p_i^2.$$

Diese Definition des Gini-Index erlaubt es auch einen generalisierten Gini-Index zu definieren:

$$G(p) = \sum_i \sum_j L(i,j) p_i p_j,$$

wobei $L(i,j)$ den Verlust angibt, der auftritt, wenn man ein Objekt der Klasse i fälschlicherweise der Klasse j zuordnet. Hiermit können unterschiedliche Kosten für die Fehlklassifikation von Objekten verschiedener Klassen berücksichtigt werden. Endknoten werden der Klasse mit den kleinsten erwarteten Fehlklassifikationskosten zugeordnet. Im Allgemeinen wird gelten: $L(i,i) = 0$, das heißt die Klassifikation eines Objektes in die richtige Klasse verursacht keine Kosten. Im Zwei-Klassen-Fall hat diese Gewichtung keinen Einfluss auf den Gini-Index, da gilt $p_1 = 1 - p_2$, und damit $G(p) = (L(1,2) + L(2,1)) p_1 p_2$. Damit spielt das Verhältnis zwischen den Fehlklassifikationskosten keine Rolle bei der

Wahl des Splits. Der generalisierte Gini-Index ist nicht in `rpart` implementiert.

Eine Alternative zum generalisierten Gini-Index, die auch im Zwei-Klassen-Fall anwendbar ist, ist das Verwenden von sogenannten *altered priors*, also veränderten Klassenwahrscheinlichkeiten. Die Klassenwahrscheinlichkeiten werden bei der Konstruktion des Klassifikationsbaumes bei der Wahl des besten Splits verwendet. In die Berechnung des Risikos eines Knotens A geht für jede Klasse das Produkt der Klassenwahrscheinlichkeiten π_i und der Fehlklassifikationskosten ein. Die Idee dieses Ansatzes ist es, die Klassenwahrscheinlichkeiten zu verändern, indem eine Verlust-Matrix der folgenden Form verwendet wird:

$$L(i,j) = \begin{cases} L_i & \text{für } i \neq j \\ 0 & \text{für } i = j \end{cases}, \text{ für } i = 1, \ldots, C, j = 1, \ldots, C.$$

Mit dieser Verlust-Matrix können die veränderten Klassenwahrscheinlichkeiten dann berechnet werden gemäß:

$$\tilde{\pi}_i = \frac{\pi_i L_i}{\sum_j \pi_j L_j}.$$

Für den Zwei-Klassen-Fall sind diese Klassenwahrscheinlichkeiten eindeutig. Für $C > 2$ und beliebige Verlust-Matrizen verwendet `rpart` die obige Formel mit $L_i = \sum_j L(i,j)$. Anschaulich bedeutet die Verwendung der veränderten Klassenwahrscheinlichkeiten im Fall von zwei Klassen, dass sich bei Erhöhung der Fehlklassifikationskosten für die erste Klasse die Klassenwahrscheinlichkeit für die erste Klasse erhöht. Im Fall von gleichen Fehlklassifikationskosten entspricht $\tilde{\pi}$ den normalen Klassenwahrscheinlichkeiten. Die veränderten Klassenwahrscheinlichkeiten werden zur Berechnung des Unreinheitsmaßes bei der Wahl des besten Splits verwendet. Dazu werden die normalen Klassenwahrscheinlichkeiten durch die veränderten Klassenwahrscheinlichkeiten ersetzt. Zur Berechnung des tatsächlichen Risikos des Splits verwendet `rpart` dann die normalen Klassenwahrscheinlichkeiten und Verluste. Die veränderten Klassengewichte sollen dem Unreinheitsmaß helfen den besten Split in Bezug auf die Fehlklassifikationskosten zu wählen (s. Therneau et al. (2015) für Details).

Die angepassten Bäume sind in der Regel zu komplex und zu stark an die Daten angepasst (*Overfitting*). Dem kann man mit dem Stutzen (*Pruning*) des Baumes entgegenwirken.

Betrachte dafür die Zerlegung der Kosten eines Baumes T

$$R_\alpha(T) = R(T) + \alpha|T|,$$

dabei ist $R(T)$ das Risiko von T, $|T|$ die Anzahl der Endknoten von T und α ein Strafterm (Komplexitätsparameter), der die Anzahl der Endknoten des Baumes bestraft. Dann kann man zeigen, dass es für jedes α genau einen kleinsten Unterbaum T_α von T gibt, der die Kosten $R_\alpha(T)$ minimiert. Der Komplexitätsparameter α kann mittels Kreuzvalidierung bestimmt werden. Bei `rpart` werden während der Kreuzvalidierung das Risiko und der zugehörige Standardfehler berechnet. Als optimaler Baum wird der Baum gewählt, der innerhalb von einer Standardabweichung des Baumes mit den kleinsten kreuzvalidierten Kosten liegt. Eine weitere Möglichkeit Overfitting zu vermeiden, ist es durch das Festlegen bestimmter Parameter schon beim Konstruieren des Baumes zu verhindern, dass der Baum zu weit verzweigt. Dafür kann man zum Beispiel den Komplexitätsparameter vorgeben oder eine Mindestanzahl von Objekten für einen Knoten fordern, damit er für einen Split berücksichtigt wird. Bei `rpart` geschieht das durch die Parameter *cp* und *minsplit*.

Ein großer Nachteil der Klassifikationsbäume ist ihre Instabilität (Hastie et al. 2009). Schon kleine Änderungen der Stichprobe können dazu führen, dass der Baum sich stark verändert. Der Grund dafür ist das schrittweise Vorgehen im Algorithmus. Wird an der Wurzel des Baumes ein anderer Split gemacht, setzt sich das für die weiteren Splits fort. Ein Ansatz um die Varianz zu verringern ist das sogenannte *Bagging*, bei dem mehrere Baummodelle aufgestellt werden, über die dann gemittelt wird. Dieser Ansatz wird bei den Random Forests verwendet, die im nächsten Abschnitt beschrieben werden.

3.4.3 Random Forests

Random Forests (Breiman 2001) sind ein Klassifikationsverfahren, bei dem eine große Anzahl von dekorrelierten Entscheidungsbäumen gebildet wird, über die dann gemittelt wird. Die Grundlage dafür bildet das Bagging. Ebenso wie die Baummodelle aus Kapitel 3.4.2 können sie auch als Regressionsverfahren verwendet werden.

Algorithmus 3 Random Forests (vgl. Hastie et al. (2009), S. 588).

1. Für $b = 1, \ldots, B$:
   a Ziehe aus den Trainingsdaten eine Bootstrap-Stichprobe Z^* der Größe N.
   b Bilde für die Bootstrap-Stichprobe einen Baum T_b, indem rekursiv für jeden Knoten die folgenden Schritte wiederholt werden, bis eine minimale Knotengröße $n_{\min}$ erreicht wird:
      1. Wähle zufällig $m = \sqrt{p}$ Variablen aus den p Variablen aus.
      2. Wähle aus den m Variablen die beste Variable mit dem besten Split im Sinne der Minimierung der Unreinheit aus.
      3. Verwende die Variable aus 2. um den Knoten in zwei Unterknoten aufzuteilen.
2. Output: Menge von Bäumen $\{T_b\}_1^B$.

Ordne eine neue Beobachtung der Klasse zu, für die sich die Mehrzahl der Bäume entscheidet. Bei Gleichstand wähle eine zufällige Zuordnung.

Die Idee des Baggings ist es über viele verrauschte, aber approximativ unverzerrte Modelle zu mitteln und so die Varianz zu verringern. Die Bäume, die beim Bagging generiert werden, sind identisch verteilt. Das heißt der Erwartungswert der gemittelten Bäume ist gleich dem Erwartungswert jedes einzelnen Baumes. Das bedeutet, dass sich der Bias durch das Bagging nicht reduzieren lässt (Hastie et al. 2009). Ein Mittel aus B identisch verteilten Zufallsvariablen mit Varianz σ^2 und paarweiser Korrelation ρ hat die Varianz $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$. Das heißt wenn B größer wird, wird der 2. Term kleiner. Der 1. Term hängt aber nicht von B ab, sondern nur von ρ. Die Idee der Random Forests ist es daher die Varianz zu verringern, indem die Korrelation zwischen den Bäumen reduziert wird. Das soll dadurch erreicht werden, dass beim Bilden des Baumes an jedem Knoten nur eine Untermenge m von Variablen für den Split zugelassen wird. Man wählt häufig $m = \sqrt{p}$, wobei p die Anzahl der Variablen ist. Der Algorithmus zur Konstruktion eines Random Forest ist in Algorithmus 3 dargestellt.

Die Fehlerrate der Random Forests kann mit der sogenannten *Out-of-Bag* (OOB)-Stichprobe geschätzt werden. In jeder Bootstrap-Runde kann für die Beobachtungen, die nicht in der Bootstrap-Stichprobe sind, mit dem konstruierten Baum eine Vorhersage

gemacht werden. Die Vorhersagen aller Läufe können zusammengefasst werden um einen OOB-Schätzer für die Fehlerrate zu erhalten. Das Verfahren ähnelt der Kreuzvalidierung, bedeutet aber kaum zusätzliche Rechenzeit. Die Variablenwichtigkeit kann ebenfalls mit den OOB-Beobachtungen geschätzt werden. Dafür werden beim b-ten Baum die Werte der j-ten Variable für die OOB-Stichprobe permutiert, während die anderen Variablen gleich bleiben. Die Fehlerrate wird berechnet und mit der originalen Fehlerrate verglichen. Die Verringerung der Accuracy (definiert als Anteil der richtig klassifizierten Beobachtungen unter allen Beobachtungen) wird über alle Bäume gemittelt und dann als Maß für die Variablenwichtigkeit verwendet.

Stark unbalancierte Gruppengrößen stellen für die Klassifikation mit Random Forests im Allgemeinen ein Problem dar. Da der Algorithmus die Gesamtfehlerrate minimiert, ist er ungeeignet für Klassifikationsprobleme, bei denen die kleinere Gruppe möglichst korrekt vorhergesagt werden soll (Chen et al. 2004). Eine Idee dieses Problem zu lösen, wäre es die Bootstrap-Stichprobe stratifiziert zu ziehen. Allerdings wird der Effekt der ungleich großen Gruppen dadurch nicht vollständig behoben (Chen et al. 2004). Es sind jedoch andere Ansätze zur Problemlösung möglich. Zum einen können Klassengewichte eingeführt werden, die Fehlklassifikationen der verschiedenen Klassen unterschiedlich gewichten. Zum anderen kann man mit *Up-* oder *Down-Sampling* die Stichprobengrößen der beiden Gruppen aneinander angleichen. Up-Sampling bedeutet, dass man die Stichprobengröße in der kleineren Gruppe künstlich erhöht, indem man die Beobachtungen aus dieser Gruppe mehrfach verwendet. Beim Down-Sampling wird ein Teil der Beobachtungen der größeren Gruppe nicht für die Konstruktion des Klassifikators verwendet. Beide Verfahren haben Nachteile. Up-Sampling der kleineren Klasse kann dabei zu Überanpassung (Overfitting) führen, wohingegen Down-Sampling zu Informationsverlust führt, da ein Großteil der Stichprobe der größeren Gruppe nicht verwendet wird.

Zur Konstruktion der Random Forests in Kapitel 5.3 verwenden wir das R-Paket `randomForest` (Liaw und Wiener 2002). Mit dem Parameter *mtry* lässt sich die Anzahl von Variablen festlegen, die an jedem Knoten zur Auswahl stehen, Default-Wert ist $\sqrt{p}$. Die Mindestanzahl von Beobachtungen in den Endknoten kann mit dem Parameter *nodesize* festgelegt werden. Bei unbalancierten Daten können mit dem Parameter *classwt* unterschiedliche Gewichte für die verschiedenen Klassen festgelegt werden. Die Gewichte werden ähnlich wie die Verlust-Werte bei den Klassifikationsbäumen (vgl. Kapitel 3.4.2)

für die Berechnung des Gini-Index beim Splitten verwendet. Down-Sampling ist über den Parameter *sampsize* implementiert.

Das R-Paket `mlr` (Bischl et al. 2016), das in Kapitel 5.4 verwendet wird, benutzt eine andere Implementierung der Random Forests, nämlich die des Paketes `ranger` (Wright und Ziegler 2017). Statt des Parameters *classwt* gibt es bei `ranger` den Parameter *case.weights*. Damit können Gewichte für die Beobachtungen festgelegt werden. Beobachtungen mit höheren Gewichten werden mit höherer Wahrscheinlichkeit in die Bootstrap-Stichprobe aufgenommen. Im Fall von zwei Gruppen kann ein einzelner Wert angegeben werden. Dieser entspricht dem Gewicht für eine Beobachtung aus der positiven Klasse.

4 Identifizieren von Genen mit charakteristischer Expressionsverteilung

Im Folgenden werden die Ergebnisse der Analysen der verschiedenen Bimodalitätsmaße auf den Genexpressionsdaten von 200 nodal-negativen unbehandelten Brustkrebspatientinnen (Mainz-Kohorte) beschrieben. In Kapitel 4.1 wird zunächst der Zusammenhang zwischen den verschiedenen Bimodalitätsmaßen global untersucht, um die Frage zu beantworten, ob die Scores die gleichen Gene als bimodal identifizieren. In Kapitel 4.2 werden die charakteristischen Expressionsverteilungen gezeigt, die die unterschiedlichen Scores identifizieren. In Kapitel 4.3 wird die prognostische Relevanz der Top-Gene der Bimodalitätsmaße mit Hilfe von Kaplan-Meier-Plots und dem Log-Rank-Test untersucht. Zur globalen Untersuchung der prognostischen Relevanz wird eine Enrichment-Analyse mit Fisher- und Kolmogorov-Smirnov-Tests durchgeführt. In Kapitel 4.4 wird untersucht, ob sich die charakteristische Form der Expressionsverteilung in zwei unabhängigen Datensätzen (Rotterdam- und Transbig-Kohorte) bestätigen lässt. Dazu werden die Bimodalitätsmaße für diese Datensätze berechnet und mit denen der Mainz-Kohorte verglichen. Diese Analysen wurden zum Großteil bereits in Hellwig et al. (2010) veröffentlicht. Allerdings unterscheidet sich das Datenmaterial in dieser Arbeit von dem in der früheren Arbeit verwendeten. In der Arbeit von 2010 bestand die Mainz-Kohorte nur aus 194 Patientinnen und die Expressionsdaten wurden nicht mit fRMA mit RMA vorverarbeitet. Außerdem wurde keine Korrektur der Clusterergebnisse durchgeführt und eine andere Ereigniszeit verwendet. In Kapitel 4.5 werden die Werte der Bimodalitäts-Scores aus den beiden Analysen miteinander verglichen.

4.1 Finden die Scores die gleichen Gene?

Wir stellen uns zunächst die Frage, ob die Bimodalitäts-Scores dieselben Gene als bimodal bzw. nicht-bimodal identifizieren. Dafür betrachten wir die paarweise Korrelation der berechneten Werte für die sieben Scores (Tabelle 4.1). Da die Werte vom Likelihood Ratio stark nach oben streuen, wurden die Werte logarithmiert. Bei der Pearson-Korrelation beobachten wir den größten Wert für das logarithmierte Likelihood Ratio (log(LR)) und die Outlier-Sum-Statistik (0.853). Den zweitgrößten Wert haben log(LR) und Kurtosis (0.655), den drittgrößten der Weighted Variance Reduction Score (WVRS) und Kurtosis (0.518). Bimodality Index (BI) und Variance Reduction Score (VRS) sind negativ miteinander korreliert (-0.490). Auffällig ist, dass die dip-Statistik die betragsmäßig kleinsten Korrelationskoeffizienten zu allen anderen Scores aufweist.

Tabelle 4.1: Pearson- und Spearman-Korrelation der Bimodalitäts-Scores.

Pearson-Korrelation							
	VRS	WVRS	Dip	OS	Kurtosis	log(LR)	BI
VRS	1.000	0.318	-0.102	-0.109	0.125	-0.180	-0.490
WVRS	0.318	1.000	-0.061	0.367	0.518	0.521	-0.111
Dip	-0.102	-0.061	1.000	-0.107	-0.044	-0.037	0.060
OS	-0.109	0.367	-0.107	1.000	0.376	0.853	0.261
Kurtosis	0.125	0.518	-0.044	0.376	1.000	0.655	-0.213
log(LR)	-0.180	0.521	-0.037	0.853	0.655	1.000	0.199
BI	-0.490	-0.111	0.060	0.261	-0.213	0.199	1.000
Spearman-Korrelation							
	VRS	WVRS	Dip	OS	Kurtosis	log(LR)	BI
VRS	1.000	0.701	-0.090	0.392	0.627	0.161	-0.493
WVRS	0.701	1.000	-0.131	0.760	0.862	0.605	-0.084
Dip	-0.090	-0.131	1.000	-0.143	-0.084	-0.052	0.016
OS	0.392	0.760	-0.143	1.000	0.721	0.747	0.232
Kurtosis	0.627	0.862	-0.084	0.721	1.000	0.674	-0.167
log(LR)	0.161	0.605	-0.052	0.747	0.674	1.000	0.454
BI	-0.493	-0.084	0.016	0.232	-0.167	0.454	1.000

Bei der Spearman-Korrelation wird der größte Wert für WVRS und Kurtosis beobachtet (0.862). Paare mit hoher Korrelation sind außerdem: Outlier Sum und WVRS (0.760),

Outlier Sum und log(LR) (0.747), Outlier Sum und Kurtosis (0.721), WVRS und VRS (0.701) und log(LR) und Kurtosis (0.674). Auch hier haben der Bimodality Index und VRS einen negativen Korrelationskoeffizienten und die kleinsten absoluten Werte werden für die dip-Statistik beobachtet.

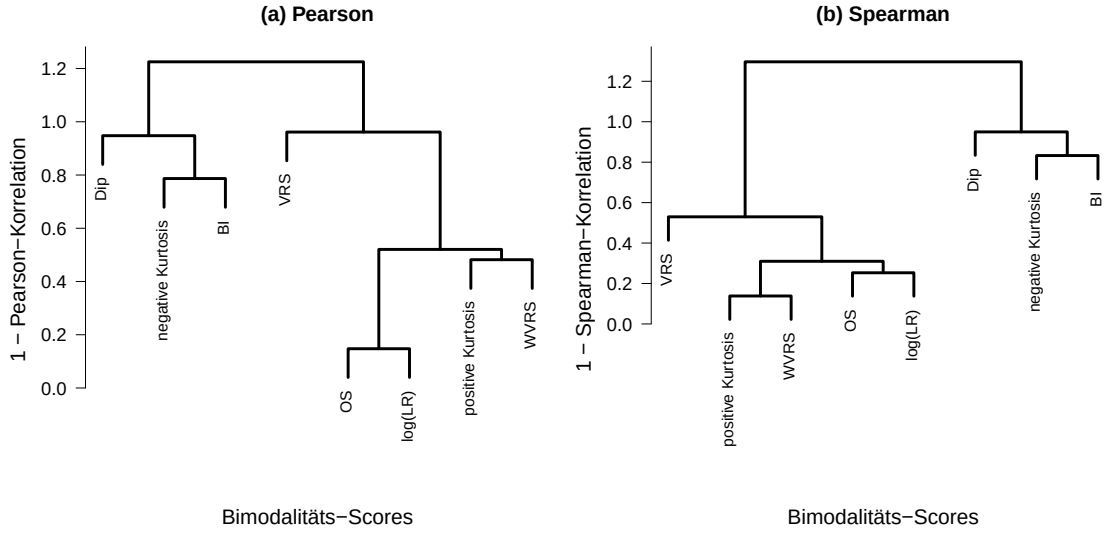

Abbildung 4.1: Dendrogramme für die Korrelation der Bimodalitäts-Scores. Als Distanzmatrix für das hierarchische Clustern mit Average Linkage wurde 1 − Korrelationsmatrix der Scores verwendet.

Zur grafischen Darstellung der Beziehung zwischen den Maßen wurden Dendrogramme gezeichnet (Abbildung 4.1). Als Distanzmatrix wurde dabei 1 − Korrelationsmatrix der Scores verwendet verwendet und das hierarchische Clustern mit Average Linkage durchgeführt. Der Score Kurtosis ist in den Dendrogrammen zweimal enthalten (negative und positive Kurtosis). Hierbei bedeutet negative Kurtosis, dass die Werte der Kurtosis so geordnet werden, dass die negativen Werte die kleinsten Ränge in der geordneten Liste haben. In beiden Grafiken sieht man zwei Gruppen. Auf der einen Seite sind die dip-Statistik, Bimodality Index und die negative Kurtosis und auf der anderen Seite die übrigen Scores. In der ersten Gruppe sind die Pearson-Korrelationen nahe bei 0. Die Scores in der anderen Gruppe liegen näher beieinander, wobei VRS eine Ausnahme bildet. Die hoch korrelierten Maße log(LR) und Outlier Sum sowie Bimodality Index und WVRS liegen jeweils nah beieinander und bilden zusammen ein Cluster. Unter Verwendung

des Spearman-Korrelationskoeffizienten sind die Abstände der Scores innerhalb dieses Clusters kleiner als bei der Pearson-Korrelation.

Die Ergebnisse zu den Korrelationen der Scores decken sich mit den Ergebnissen aus Hellwig et al. (2010). Auf globaler Ebene scheint es insgesamt Scores zu geben, die die gleiche Art von Expressionsverteilungen identifizieren. Bei der Interpretation muss allerdings beachtet werden, dass nur die Maße log(LR) und Outlier Sum sowie Bimodality Index und WVRS hoch korreliert sind, wohingegen die Bimodalitätsmaße im zweiten Cluster große Abstände zu diesen Maßen und zueinander haben. Im nächsten Abschnitt werden nun die Expressionsverteilungen der Top-Gene der verschiedenen Bimodalitätsmaße betrachtet.

4.2 Charakteristische Expressionsverteilungen

Um die Frage zu beantworten welche Art von Verteilungen wir mittels der verschiedenen Bimodalitäts-Maße identifizieren können, betrachten wir Histogramme der Expressionswerte der Top6-Gene jedes Scores (Abbildungen 4.2 und 4.3).

Bei den Top-Genen von VRS gibt es keine klare Struktur. Es sind sowohl Gene darunter, deren Expressionsverteilungen eine klare bimodale Struktur aufweisen, als auch Gene mit einer großen Hauptverteilung und einigen Ausreißern. Bei den Top-Genen von WVRS gibt es immer eine unimodale Hauptverteilung und einen Ausreißer mit einem besonders großen Wert.

Die dip-Statistik findet Gene mit einer deutlich bimodalen Expressionsverteilung, bei der sich die beiden Verteilungen gut trennen lassen. Insgesamt gibt es 30 Gene mit einem Wert der dip-Statistik, der größer als das 95 %-Quantil (0.037) für $n = 200$ aus der Tabelle des Paketes `diptest` ist. Diese besitzen also unadjustiert signifikant eine Expressionsverteilung, die nicht unimodal und damit mindestens bimodal ist. Die dip-Statistik ist das einzige der betrachteten Maße, das eine Aussage über die Signifikanz der Ergebnisse erlaubt.

Alle Top6-Gene der Outlier-Sum-Statistik besitzen eine Expressionsverteilung, die aus einer Hauptverteilung mit niedrigen Expressionswerten und einer Ausreißergruppe mit großer Varianz besteht (Abbildung 4.2, schwerer rechter Rand). Ähnliches ist für die Ex-

pressionsverteilungen der Top6-Gene des Likelihood Ratios zu beobachten (Abbildung 4.3). Die Gene mit den größten positiven Kurtosis-Werten haben eine Expressionsverteilung mit unimodaler Hauptverteilung und einigen zusätzlichen Ausreißern mit hohen Werten. Die Expressionsverteilungen der Gene mit den größten negativen Kurtosis-Werten dagegen sind dagegen bimodal mit ähnlich großen Gruppen oder sie besitzen eine flache Verteilung mit einem Peak im niedrigen Expressionsbereich (*SCGB1D2*). Beim Bimodality Index sind die Expressionsverteilungen der Top6-Gene bimodal, wobei es sowohl ähnlich starke Gruppen gibt, als auch unterschiedlich große Gruppen.

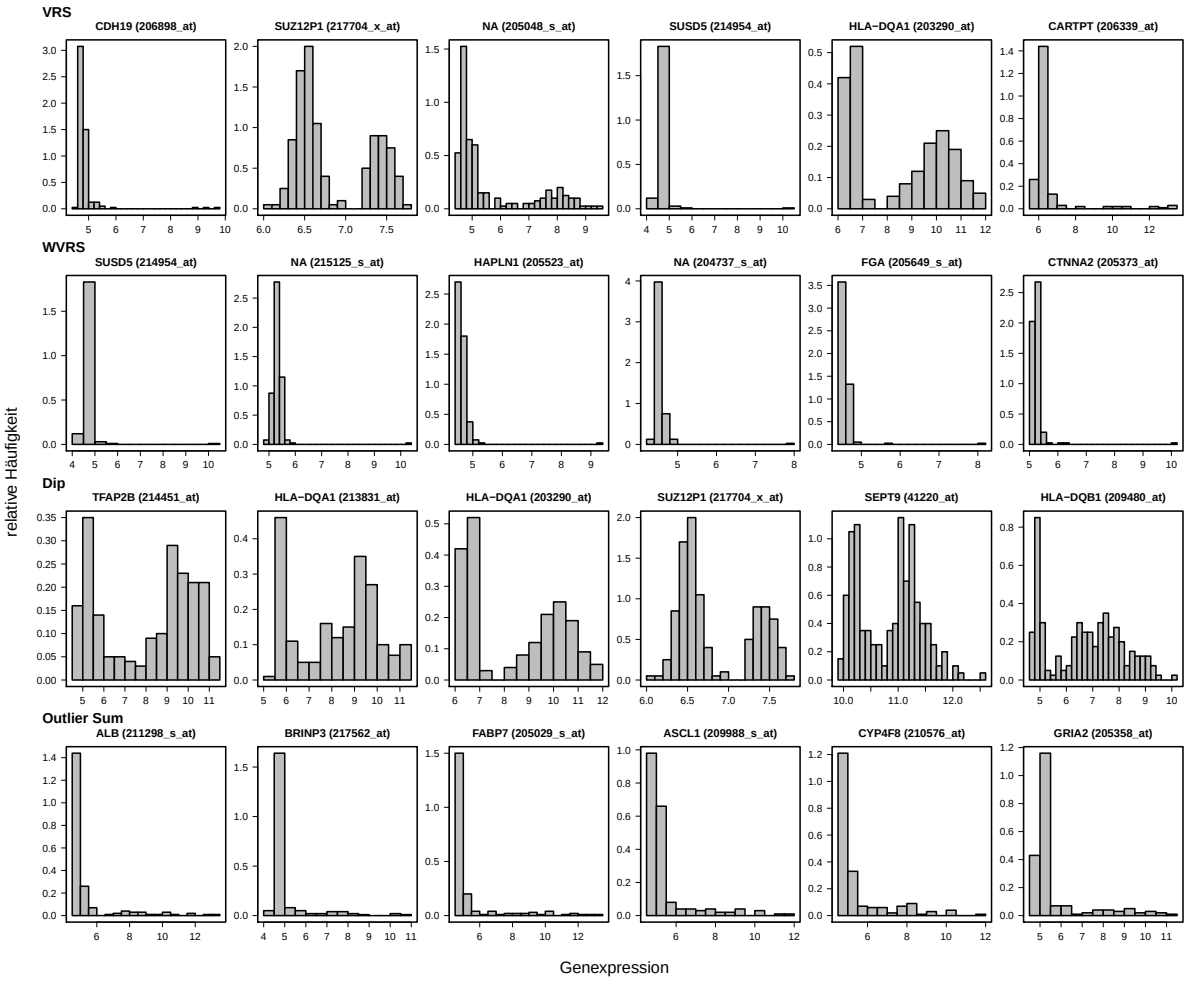

Abbildung 4.2: Histogramme der Expressionswerte der Top6-Gene der Scores VRS, WVRS, dip-Statistik und Outlier-Sum-Statistik.

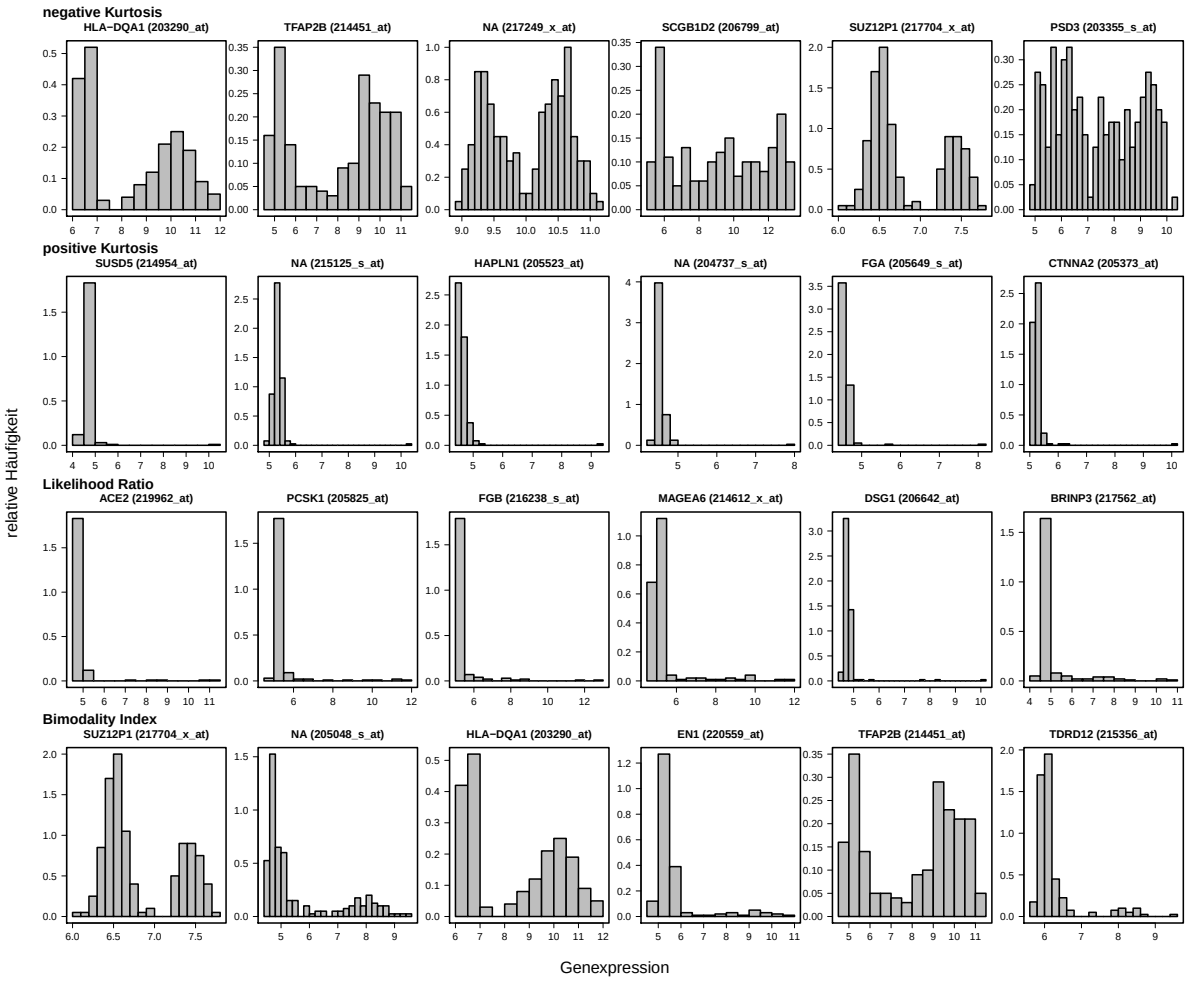

Abbildung 4.3: Histogramme der Expressionswerte der Top6-Gene der Scores negative und positive Kurtosis, Likelihood Ratio und Bimodality Index.

Insgesamt finden die verschiedenen Bimodalitätsmaße sehr unterschiedliche Expressionsverteilungen, wobei es drei Haupttypen gibt: klare bimodale Verteilungen (dip, negative Kurtosis, Bimodality Index), unimodale Verteilungen mit einem oder wenigen Ausreißern (VRS, WVRS, positive Kurtosis) und unimodale Verteilungen mit vielen Ausreißern mit großer Varianz (Outlier Sum, Likelihood Ratio). Einige Gene tauchen bei mehreren Scores unter den Top6 auf: *SUZ12P1* (VRS, dip, negative Kurtosis, BI), *SUSD5* (VRS, WVRS, positive Kurtosis) und *TFAP2B* (dip, negative Kurtosis, BI).

4.3 Prognostische Relevanz der Top-Gene

Um zu untersuchen, ob die Gene mit den auffälligsten Bimodalitäts-Scores Subgruppen mit unterschiedlicher Prognose generieren, benutzen wir den Log-Rank-Test zum Vergleich der Survivalkurven der beiden Gruppen. Dabei ist zu beachten, dass die Gruppeneinteilungen für das gleiche Gen sich für die verschiedenen Maße unterscheiden können, weil sie mit verschiedenen Verfahren vorgenommen wurden. Bei VRS, WVRS und dip-Statistik verwenden wir die Einteilungen aus dem k-means-Algorithmus, beim Bimodality Index das Ergebnis des modellbasierten Clusterns mit zwei Komponenten und gleicher Varianz, bei Kurtosis und Likelihood Ratio das Ergebnis des modellbasierten Clusterns mit zwei Komponenten und freier Wahl der Varianz und bei der Outlier-Sum-Statistik bilden jeweils die Ausreißer eine eigene Gruppe.

Wie in Kapitel 3.2.4 beschrieben, ist die Annahme der asymptotischen Normalverteilung unter der Nullhypothese nicht mehr erfüllt, wenn wir Gruppeneinteilungen mit extrem unterschiedlichen Größen haben, besonders wenn weniger als 5 Beobachtungen in der kleineren Gruppe sind. Darum verwenden wir nicht den p-Wert des Log-Rank-Tests, sondern berechnen den p-Wert mit einem Permutationstest. Bei einigen Genen werden alle Patientinnen in eine Gruppe eingeteilt, sodass kein Wert der Log-Rank-Teststatistik berechnet werden kann. In diesem Fall setzen wir den p-Wert auf 1.

Für die Top6-Gene der acht Bimodalitätsmaße sind die zugehörigen Kaplan-Meier-Kurven in Abbildungen 4.4 bis 4.7 dargestellt. Der angegebene p-Wert stammt jeweils aus dem Permutationstest und ist unadjustiert. Bei der Mehrzahl der Gene unterscheiden sich die Survivalkurven nicht. Bei den Top-Genen von WVRS gibt es jeweils nur einen Ausreißer, der eine eigene Gruppe bildet. Abhängig davon, ob die zugehörige Patientin das Zielereignis erlebt oder die Beobachtungszeit zensiert ist, bleibt die Kurve bei 100 % oder fällt auf 0.

Für die Top10-Gene jedes Bimodalitäts-Scores sind in Tabelle 4.2 die p-Werte des Log-Rank-Tests zu finden. Die p-Werte wurden mit der Methode von Benjamini und Hochberg (1995) für das multiple Testen adjustiert, wobei die Adjustierung pro Methode durchgeführt wurde. Der einzige adjustierte p-Wert, der kleiner 5 % ist, gehört zu dem Gen *DSG1*, das den 8. Rang bei VRS hat. Auf Basis der Expressionswerte dieses Gens werden

die Patientinnen mit dem k-means-Algorithmus in eine Gruppe von 197 Patientinnen mit niedrigen Expressionswerten und eine Gruppe von 3 Patientinnen mit hohen Expressionswerten eingeteilt. Diese Patientinnen fallen alle aus (maximale Ereigniszeit 1.4 Jahre), was zu einem signifikanten Unterschied zwischen den Survivalkurven der beiden Gruppen führt. *DSG1* ist auch unter den Top10-Genen vom Likelihood Ratio (Rang 5), hier ist allerdings nur der unadjustierte p-Wert kleiner 5 % (0.037), der adjustierte p-Wert dagegen 0.235. Die Unterschiede in den p-Werten für das gleiche Gen bei verschiedenen Scores lassen sich dadurch erklären, dass die Gruppeneinteilung unterschiedlich ist. Der k-means-Algorithmus teilt nur die 3 Beobachtungen mit den größten Expressionswerten in eine Gruppe ein, beim modellbasierten Clustern werden 6 Beobachtungen der Gruppe mit hohen Expressionswerten zugeordnet.

Insgesamt liefern also die meisten Gene mit den Top-Werten der Bimodalitätsmaße keine Gruppen mit unterschiedlicher Prognose. Allerdings wurden bis jetzt jeweils nur die Top10 jedes Gens betrachtet. Es kann trotzdem sein, dass viele prognostische Gene oben auf den nach den Scores geordneten Listen stehen, also niedrige Ränge bezüglich Score und p-Wert des Log-Rank-Tests haben. Um zu überprüfen, ob das für die verschiedenen Scores gilt, verwenden wir die in Kapitel 3.3 beschriebenen Tests.

Mit dem Fisher-Test können wir für fest vorgegebene Cutoffs für Score und p-Wert überprüfen, ob die Zahlen in der resultierenden Vierfeldertafel unabhängig sind. Das Ergebnis hängt hier stark von dem gewählten Cutoff ab. Der Kolmogorov-Smirnov-Test erlaubt eine globale Aussage darüber, ob sich auffällig viele der prognostischen Gene in einem bestimmten Bereich der geordneten Genliste befinden. Wenn Score und prognostische Relevanz der Gene unabhängig voneinander sind, so würde man erwarten, dass die Ränge der prognostischen Gene bezüglich des jeweiligen Scores gleichverteilt sind. Bei der Interpretation des Testergebnisses muss man allerdings beachten, dass ein kleiner p-Wert nicht zwingend bedeutet, dass alle prognostischen Gene kleine Ränge haben. Es kann ebenso sein, dass man bei den mittleren Rängen besonders viele prognostische Gene findet. Es ist daher hilfreich sich zusätzlich die zugehörigen Grafiken der Running-Sum-Statistik anzusehen (Abbildung 4.8). In diesen sind die Ränge der 200 prognostischen Gene als Striche dargestellt. Der angegebene p-Wert stammt aus dem Kolmogorov-Smirnov-Test.

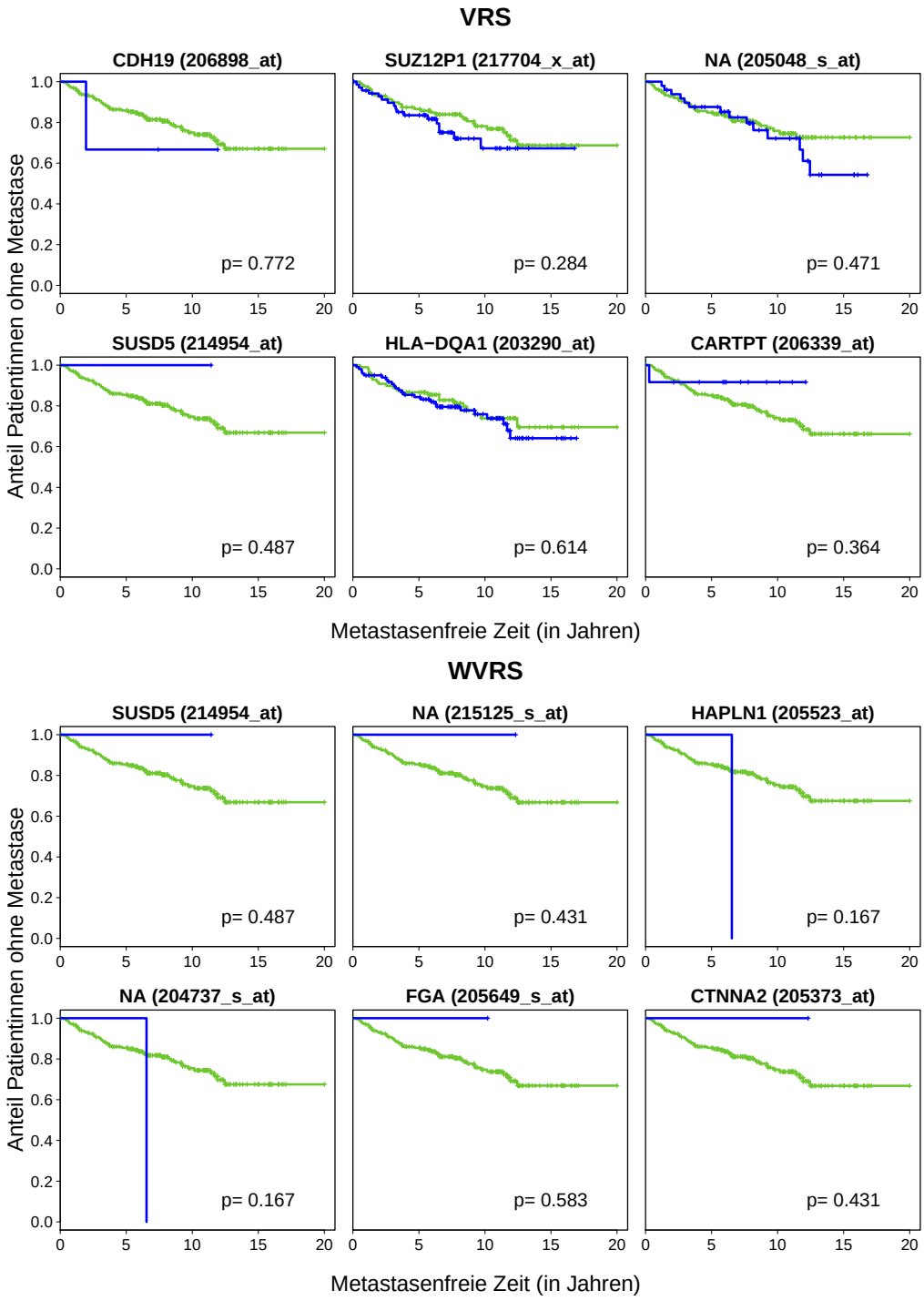

Abbildung 4.4: Kaplan-Meier-Kurven für die Top6-Gene der Scores VRS und WVRS. Die grüne Kurve gehört jeweils zur Gruppe mit den niedrigen Expressionswerten, die blaue zu der Gruppe mit den hohen Expressionswerten des Gens.

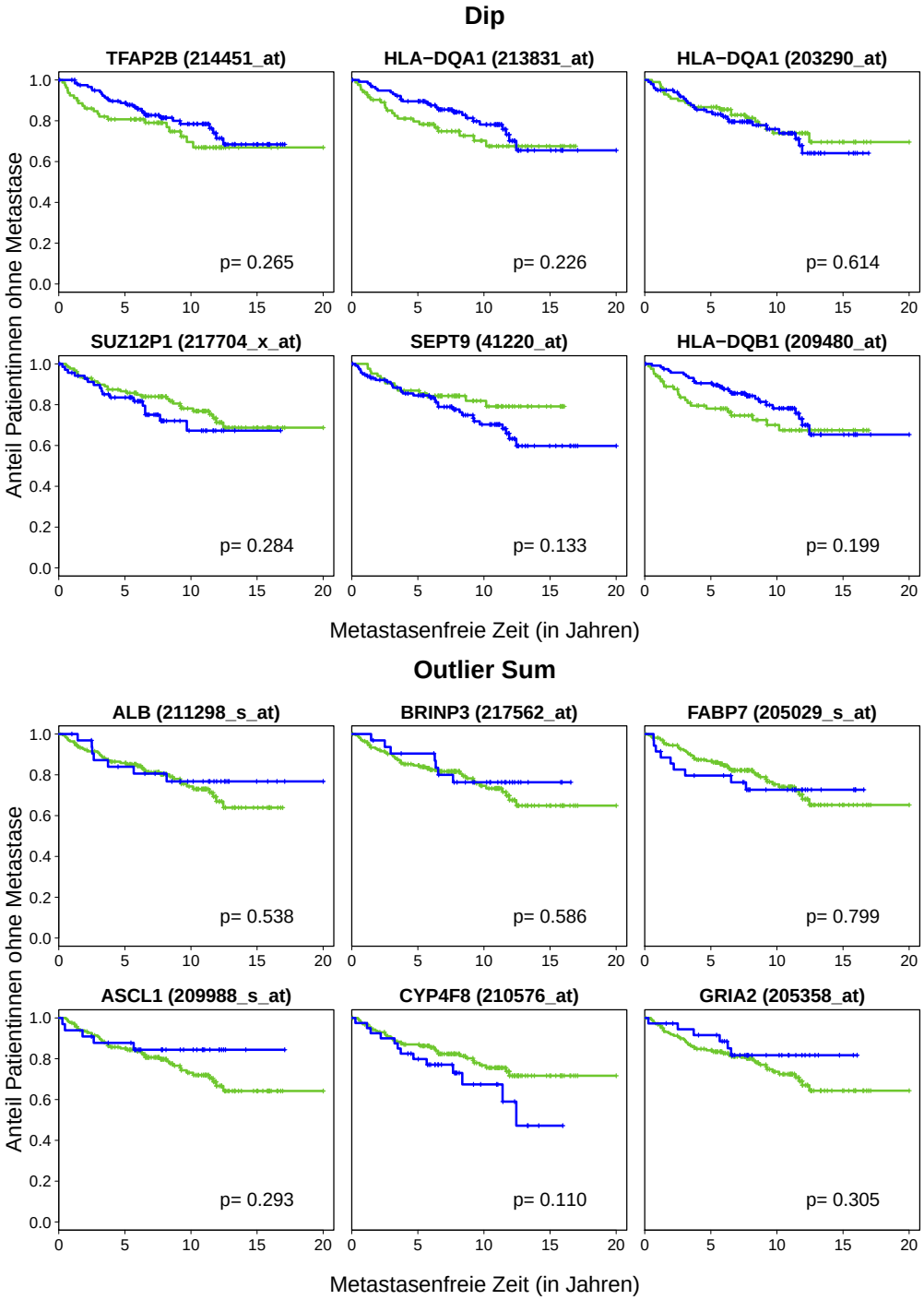

Abbildung 4.5: Kaplan-Meier-Kurven für die Top6-Gene der Scores dip-Statistik und Outlier-Sum-Statistik. Die grüne Kurve gehört jeweils zur Gruppe mit den niedrigen Expressionswerten, die blaue zu der Gruppe mit den hohen Expressionswerten des Gens.

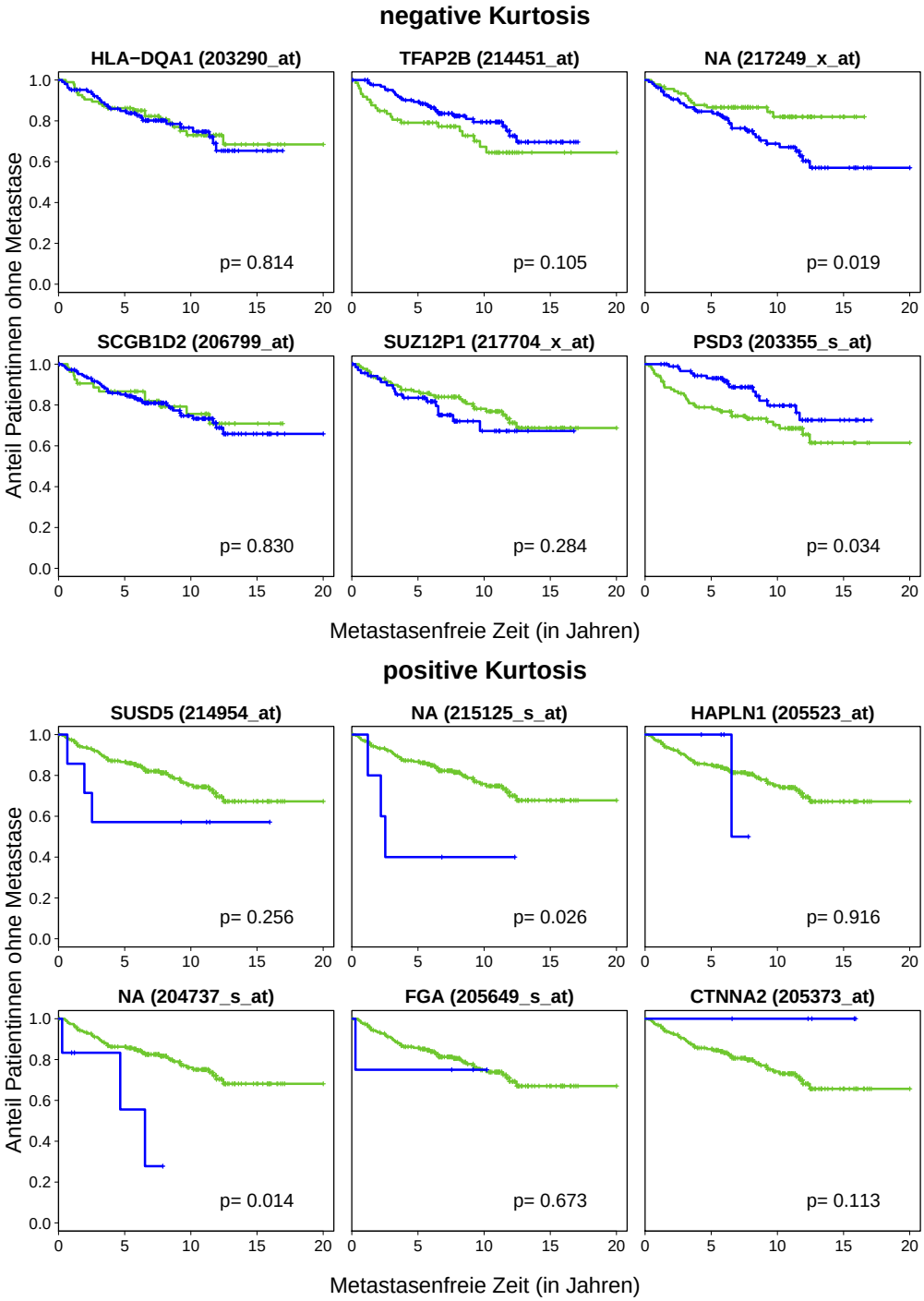

Abbildung 4.6: Kaplan-Meier-Kurven für die Top6-Gene der Scores negative und positive Kurtosis. Die grüne Kurve gehört jeweils zur Gruppe mit den niedrigen Expressionswerten, die blaue zu der Gruppe mit den hohen Expressionswerten des Gens.

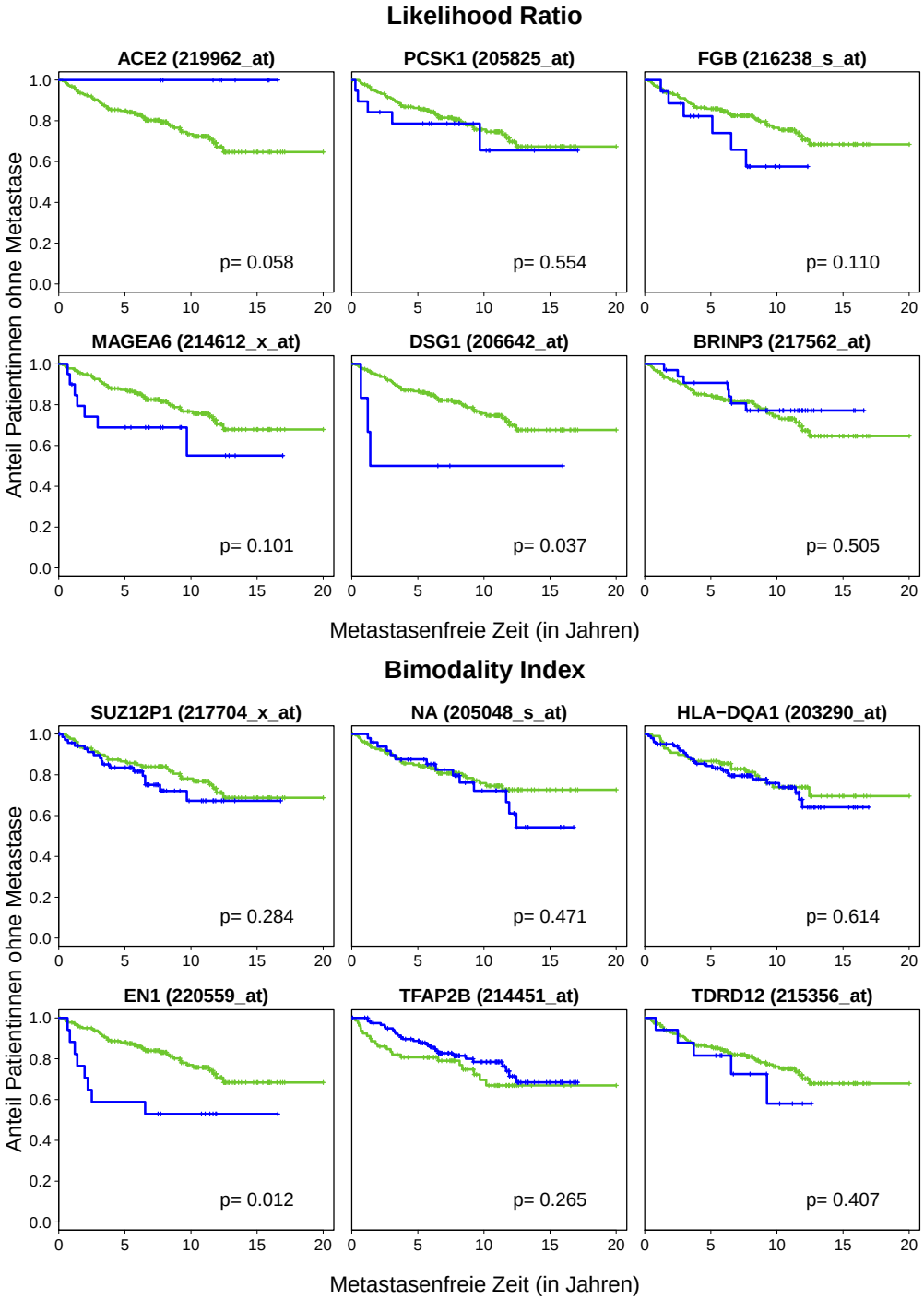

Abbildung 4.7: Kaplan-Meier-Kurven für die Top6-Gene der Scores Likelihood Ratio und Bimodality Index. Die grüne Kurve gehört jeweils zur Gruppe mit den niedrigen Expressionswerten, die blaue zu der Gruppe mit den hohen Expressionswerten des Gens.

Tabelle 4.2: Top10-Gene der 8 Bimodalitäts-Scores mit p-Werten des Log-Rank-Tests.

VRS				**WVRS**			
Affy ID	Gensymbol	p	p.adj	Affy ID	Gensymbol	p	p.adj
206898_at	*CDH19*	0.772	0.929	214954_at	*SUSD5*	0.487	0.793
217704_x_at	*SUZ12P1*	0.284	0.658	215125_s_at	*unbekannt*	0.431	0.762
205048_s_at	*unbekannt*	0.471	0.785	205523_at	*HAPLN1*	0.167	0.536
214954_at	*SUSD5*	0.487	0.793	204737_s_at	*unbekannt*	0.167	0.536
203290_at	*HLA-DQA1*	0.614	0.861	205649_s_at	*FGA*	0.583	0.845
206339_at	*CARTPT*	0.364	0.723	205373_at	*CTNNA2*	0.431	0.762
220559_at	*EN1*	0.026	0.246	219915_s_at	*SLC16A10*	0.386	0.734
206642_at	*DSG1*	<0.001	0.026	217561_at	*CALCA*	0.386	0.734
214451_at	*TFAP2B*	0.265	0.644	217704_x_at	*SUZ12P1*	0.284	0.658
209942_x_at	*unbekannt*	0.011	0.173	214566_at	*SMR3A*	0.207	0.586
dip-Statistik				**Outlier Sum**			
214451_at	*TFAP2B*	0.265	0.644	211298_s_at	*ALB*	0.538	0.853
213831_at	*HLA-DQA1*	0.226	0.605	217562_at	*BRINP3*	0.586	0.876
203290_at	*HLA-DQA1*	0.614	0.861	205029_s_at	*FABP7*	0.799	0.953
217704_x_at	*SUZ12P1*	0.284	0.658	209988_s_at	*ASCL1*	0.293	0.685
41220_at	*SEPT9*	0.133	0.490	210576_at	*CYP4F8*	0.110	0.419
209480_at	*HLA-DQB1*	0.199	0.580	205358_at	*GRIA2*	0.305	0.697
217249_x_at	*unbekannt*	0.019	0.216	207802_at	*CRISP3*	0.424	0.791
201424_s_at	*CUL4A*	0.377	0.730	209942_x_at	*unbekannt*	0.098	0.394
218341_at	*PPCS*	0.531	0.819	205916_at	*S100A7*	0.406	0.779
212999_x_at	*HLA-DQB1*	0.237	0.616	204712_at	*WIF1*	0.532	0.850
negative Kurtosis				**positive Kurtosis**			
203290_at	*HLA-DQA1*	0.814	0.943	214954_at	*SUSD5*	0.256	0.600
214451_at	*TFAP2B*	0.105	0.390	215125_s_at	*unbekannt*	0.026	0.200
217249_x_at	*unbekannt*	0.019	0.171	205523_at	*HAPLN1*	0.916	0.980
206799_at	*SCGB1D2*	0.830	0.949	204737_s_at	*unbekannt*	0.014	0.148
217704_x_at	*SUZ12P1*	0.284	0.630	205649_s_at	*FGA*	0.673	0.888
203355_s_at	*PSD3*	0.034	0.227	205373_at	*CTNNA2*	0.113	0.405
213664_at	*SLC1A1*	0.038	0.237	219915_s_at	*SLC16A10*	0.775	0.931
213831_at	*HLA-DQA1*	0.122	0.419	217561_at	*CALCA*	0.915	0.979
212094_at	*PEG10*	0.940	0.987	214566_at	*SMR3A*	0.373	0.710
204320_at	*COL11A1*	0.632	0.870	207174_at	*GPC5*	0.519	0.810
Likelihood Ratio				**Bimodality Index**			
219962_at	*ACE2*	0.058	0.288	217704_x_at	*SUZ12P1*	0.284	0.637
205825_at	*PCSK1*	0.554	0.829	205048_s_at	*unbekannt*	0.471	0.783
216238_s_at	*FGB*	0.110	0.398	203290_at	*HLA-DQA1*	0.614	0.863
214612_x_at	*MAGEA6*	0.101	0.382	220559_at	*EN1*	0.012	0.140
206642_at	*DSG1*	0.037	0.235	214451_at	*TFAP2B*	0.265	0.617
217562_at	*BRINP3*	0.505	0.802	215356_at	*TDRD12*	0.407	0.739
209942_x_at	*unbekannt*	0.154	0.472	206373_at	*ZIC1*	0.635	0.874
204885_s_at	*MSLN*	0.037	0.236	205358_at	*GRIA2*	0.095	0.372
205029_s_at	*FABP7*	0.836	0.951	208358_s_at	*UGT8*	0.019	0.170
219612_s_at	*FGG*	0.012	0.141	205916_at	*S100A7*	0.973	0.991

Tabelle 4.3 enthält für jede der vier Methoden der Gruppeneinteilung die Anzahl der Gene mit einem p-Wert des Log-Rank-Tests kleiner 5 %. Für k-means beobachten wir die kleinste Anzahl von signifikanten Testergebnissen, für die anderen Methoden ähnlich viele. Nach dem Adjustieren mit der Methode von Benjamini-Hochberg bleiben bei der Outlier-Sum-Statistik jedoch nur 291 p-Werte kleiner 5 %, wohingegen es bei den modellbasierten Clustermethoden deutlich mehr sind (437 bzw. 444). Schaut man sich den Überlapp der Gene an (siehe Venn-Diagramme im Anhang: Abbildungen G.1 und G.2), so sieht man, dass die beiden Mclust-Methoden einen großen Überlapp haben. Das ist damit zu begründen, dass bei einer Vielzahl von Genen auch beim modellbasierten Clustern ohne Annahme der Gleichheit der Varianzen das Modell mit gleichen Varianzen gewählt wurde und die Gruppeneinteilungen deshalb bei beiden Methoden gleich sind (334 von 374 Genen im Überlapp). Bei k-means (115) und Outlier-Sum-Statistik (125) gibt es einen großen Anteil an Genen, die nur bei jeweils einer der beiden Methoden eine Gruppeneinteilung mit signifikant unterschiedlicher Prognose liefern.

Tabelle 4.3: Anzahl p-Werte des Log-Rank-Tests kleiner 5% in Abhängigkeit von Methode der Gruppeneinteilung.

Methode	unadjustiert	FDR-adjustiert
k-means	3427	219
Outlier Sum	4050	291
Mclust	4130	437
Mclust gleiche Varianzen	4158	444

Für die Analyse der Genlisten auf Enrichment mit prognostischen Genen definieren wir zunächst die $M = 200$ Gene mit den kleinsten p-Werten als prognostisch. Zu beachten ist dabei, dass sich die prognostischen Gene für Scores, die unterschiedliche Gruppeneinteilungen verwenden, unterscheiden. Tabelle 4.4 enthält die Ergebnisse der Fisher-Tests, wobei als Cutoff für die Scores 200 bzw. 1000 gewählt wurde, und den p-Wert des einseitigen Kolmogorov-Smirnov-Tests.

Beim Fisher-Test mit Cutoff 200 ist nur bei der dip-Statistik ein zum Niveau 0.05 signifikantes Ergebnis zu beobachten (p-Wert unadjustiert $p = 0.0095$). 17 der 200 prognostischen Gene sind unter den ersten 1000 Genen des Scores VRS. Dies führt zu einem signifikanten Testergebnis (p-Wert unadjustiert $p = 0.0089$). Bei der dip-Statistik

sind es 15 der 200 Gene (p-Wert unadjustiert $p = 0.0364$). Für alle anderen Bimodalitätsmaße kann kein signifikantes Enrichment der ersten 1000 Gene mit prognostischen Genen beobachtet werden. Für die Interpretation der Ergebnisse ist es wichtig darauf hinzuweisen, dass alle p-Werte unadjustiert sind. Führt man eine Adjustierung für das multiple Testen mit der Methode von Bonferroni-Holm durch, so kann die Nullhypothese für keinen der Tests abgelehnt werden.

Tabelle 4.4: Ergebnisse der Tests auf Enrichment mit $M = 200$ prognostischen Genen.

Score	Fisher				Kolmogorov-Smirnov
	200 Gene	p-Wert	1000 Gene	p-Wert	p-Wert
VRS	4	0.1062	17	0.0089	0.0005
WVRS	3	0.2676	11	0.2877	0.8368
Dip	6	0.0095	15	0.0364	0.5439
Outlier Sum	0	1.0000	12	0.1892	$< 10^{-7}$
negative Kurtosis	1	0.8366	10	0.4092	0.5101
positive Kurtosis	3	0.2676	6	0.8894	0.4930
Likelihood Ratio	0	1.0000	6	0.8894	0.0002
Bimodality Index	1	0.8366	14	0.0672	$< 10^{-8}$

Beim Kolmogorov-Smirnov-Test kann für vier Methoden ein signifikantes Ergebnis beobachtet werden: für VRS (p-Wert unadjustiert $p = 0.0005$), Outlier-Sum-Statistik (p-Wert unadjustiert $p < 10^{-7}$), Likelihood Ratio (p-Wert unadjustiert $p = 0.0002$) und Bimodality Index (p-Wert unadjustiert $p < 10^{-7}$). Die Testergebnisse bleiben auch nach der Adjustierung mit Bonferroni-Holm signifikant. Betrachtet man die Plots der Running-Sum-Statistiken der Bimodalitätsmaße (Abbildung 4.8), so sieht man, dass für diese Scores die Kurve deutlich oberhalb der horizontalen Linie bei 0 verläuft. Die Kurven für Outlier-Sum-Statistik und Bimodality Index ähneln sich, bei beiden Maßen sind unter den ersten 5000 Genen besonders viele prognostische Gene. Die Kurven für VRS und Likelihood Ratio verlaufen flacher, wobei bei VRS unter den ersten 2000 Genen besonders viele prognostische Gene sind, wohingegen beim Likelihood Ratio die Kurve erst ab etwa Rang 2500 ansteigt. Bei dip-Statistik und Kurtosis sind die Kurven nahe bei der Nulllinie. Die Kurve für WVRS weicht nach unten von der Nulllinie ab. Das bedeutet,

dass bei diesem Score gerade unter den Genen mit den größten Werten, die für große Varianz innerhalb der Cluster sprechen, viele prognostische Gene sind.

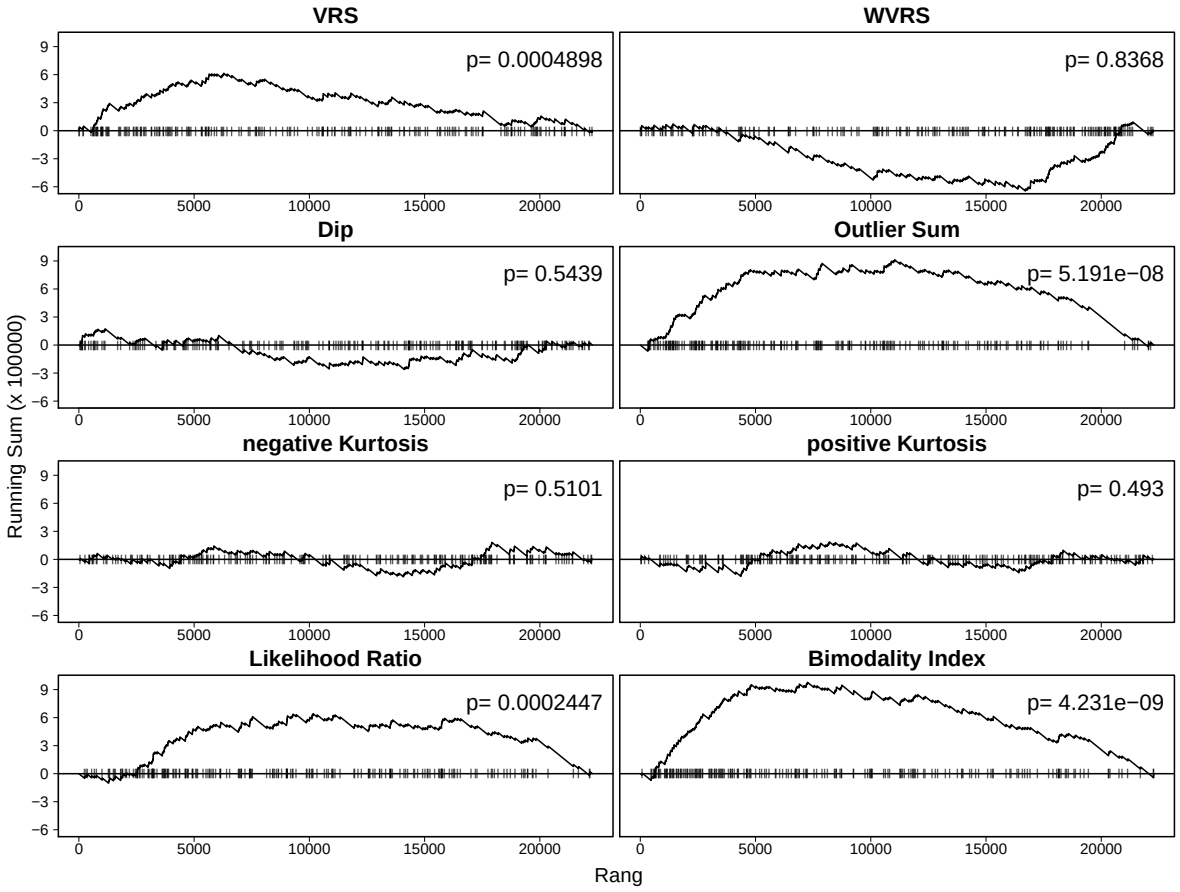

Abbildung 4.8: Running-Sum-Statistik für die acht Bimodalitäts-Scores bei $M = 200$ prognostischen Genen. Die Genlisten wurden auf Basis des jeweiligen Maßes geordnet. Jeder Strich repräsentiert den Rang eines der prognostischen Gene.

Insgesamt gibt es bei den Scores VRS, Outlier-Sum-Statistik, Likelihood Ratio und Bimodality Index unter den Genen am Anfang der Genlisten besonders viele Gene, deren Expressionswerte Gruppen mit unterschiedlicher Prognose generieren. Allerdings sind es jeweils nicht die Gene mit den größten Werten dieser Scores, die prognostisch sind, sondern es gibt ein Enrichment mit prognostischen Genen unter den Top5000 (Outlier-Sum-Statistik und Bimodality Index) bzw. der ersten Hälfte der geordneten Genlisten (dip-Statistik und Kurtosis).

4.4 Untersuchung der Top-Gene in den Validierungskohorten

Um zu untersuchen, ob die Top-Gene der jeweiligen Bimodalitäts-Scores auch in unabhängigen Kohorten eine bimodale Expressionsverteilung aufweisen, wurden die Scores auch für die Rotterdam- und die Transbig-Kohorte berechnet. Betrachtet man die Spearman-Korrelationen der Bimodalitäts-Scores für die Mainz-Kohorte und die beiden Validierungskohorten (Tabelle F.2), so fällt auf, dass die Werte der Bimodalitäts-Scores positiv korreliert sind. Eine Ausnahme bildet dabei die dip-Statistik, die bei beiden Validierungskohorten den kleinsten Korrelationskoeffizienten hat. Dieser liegt nahe bei 0, was darauf schließen lässt, dass es keinen Zusammenhang zwischen den Werten der dip-Statistik für die verschiedenen Kohorten gibt. Der größte Wert der Korrelationskoeffizienten kann jeweils für das Likelihood Ratio beobachtet werden (0.485 bei Rotterdam, 0.451 bei Transbig). Insgesamt liegt keine starke Korrelation vor.

Abbildung 4.9 zeigt paarweise Streudiagramme für die verschiedenen Bimodalitätsmaße im Vergleich der Mainz-Kohorte (x-Achse) mit der Rotterdam-Kohorte (y-Achse). Jeder Punkt repräsentiert ein Gen, die 100 Gene mit den auffälligsten Werten des jeweiligen Scores sind grün dargestellt. Bei WVRS und Likelihood Ratio wurden die Werte $\log_{10}$-transformiert, damit die interessanten Punkte besser zu sehen sind. Bei der negativen Kurtosis wurden die Werte ebenfalls transformiert (durch Verschiebung und log-Transformation), die Achsen sind hier logarithmisch.

Für alle Maße ist der Wertebereich in beiden Kohorten ähnlich. Die Top-Gene von VRS der Mainz-Kohorte haben in der Rotterdam-Kohorte überwiegend auch kleine Werte. Für einige Gene streuen die VRS-Werte allerdings. Bei WVRS streuen die Werte der Top-Gene aus der Mainz-Kohorte im gesamten Wertebereich des Scores für die Rotterdam-Kohorte. Bei der dip-Statistik sind vier von den 10 Genen mit den größten Werten auch unter den Top10-Genen für die Rotterdam-Kohorte. Von den 30 Genen, deren Verteilung sich unadjustiert signifikant von einer unimodalen Verteilung unterscheidet, sind 8 auch unter den Top30-Genen für die Rotterdam-Kohorte. Viele der Top100-Gene liegen jedoch bei beiden Kohorten in einem unauffälligen Bereich, es ist keine systematische Streuung zu beobachten. Bei Outlier Sum, Likelihood Ratio und Bimodality Index haben viele Gene

mit den größten Scores in der Mainz-Kohorte auch große Werte in der Rotterdam-Kohorte, einige haben in der Rotterdam-Kohorte aber auch Werte, die nicht im auffälligen Bereich liegen. Auffällig ist, dass es einige Gene gibt, die einen großen Wert für den Bimodality Index in der Mainz-Kohorte besitzen, der Wert in der Rotterdam-Kohorte aber nahe bei 0 liegt. Dieses Phänomen ist auch umgekehrt zu sehen. Diese Gene besitzen in beiden Kohorten eine Expressionsverteilung mit einer unimodalen Hauptverteilung und einigen zusätzlichen Ausreißern mit großer Varianz. Bei der Kurtosis gibt es keine große Übereinstimmung zwischen den Kohorten. Bei den großen positiven Werten streuen die Werte der Top100-Gene aus der Mainz-Kohorte in der Rotterdam-Kohorte im gesamten Wertebereich. Bei den Top100 der negativen Kurtosis-Werte ist ebenfalls eine große Streuung in der Rotterdam-Kohorte zu beobachten. Die meisten Werte (80) liegen aber auch bei dieser Kohorte im negativen Bereich.

Die gleichen Streudiagramme wurden auch für den Vergleich von Mainz-Kohorte und Transbig-Kohorte erstellt (Abbildung 4.10). Hier zeigt sich ein ähnliches Bild. Auffällig ist, dass es bei der Transbig-Kohorte deutlich mehr Gene mit großen Werten der dip-Statistik gibt. Eine Erklärung für dieses Phänomen kann man in Abbildung G.3 finden. In den Streudiagrammen der Genexpressionswerte der 10 Gene mit den größten Werten der dip-Statistik kann man sehen, dass bei sechs Genen nur deswegen eine bimodale Verteilung beobachtet wird, weil die Expressionswerte des Gens in den beiden Kohorten, aus denen die Transbig-Kohorte zusammengesetzt ist, auf unterschiedlichen Levels liegen. Auch beim Bimodality Index lässt sich diese Beobachtung machen.

Insgesamt gibt es bei beiden Kohorten im Vergleich mit Mainz-Kohorte große Unterschiede in den Werten der Bimodalitäts-Scores. Für viele der Top100-Gene der Scores Outlier-Sum-Statistik, Likelihood Ratio und Bimodality Index kann die charakteristische Form der Expressionsverteilung jedoch auch auf den Validierungskohorten bestätigt werden.

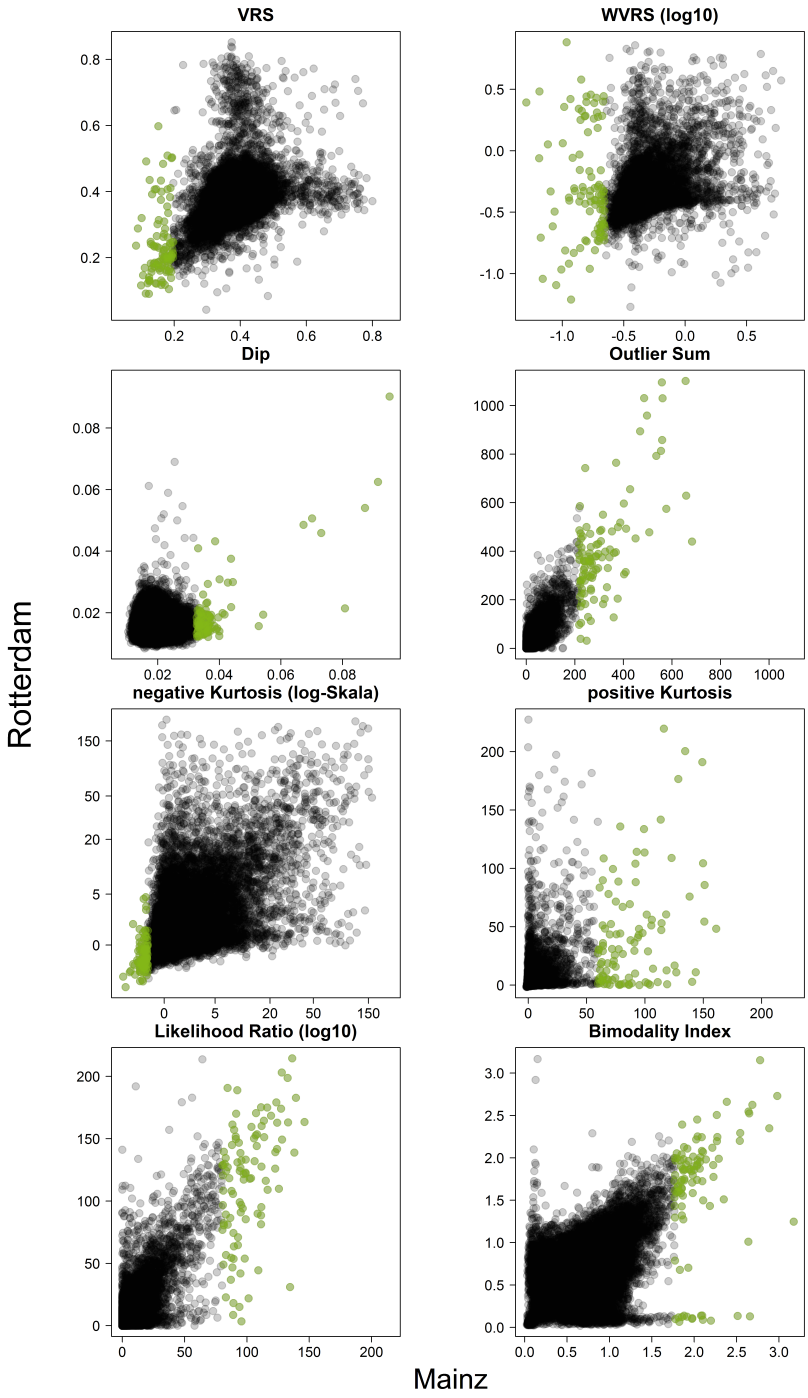

Abbildung 4.9: Streudiagramme der Bimodalitäts-Scores für Mainz- und Rotterdam-Kohorte. Jeder Punkt repräsentiert ein Gen, die 100 Gene mit den auffälligsten Werten des jeweiligen Scores sind grün dargestellt.

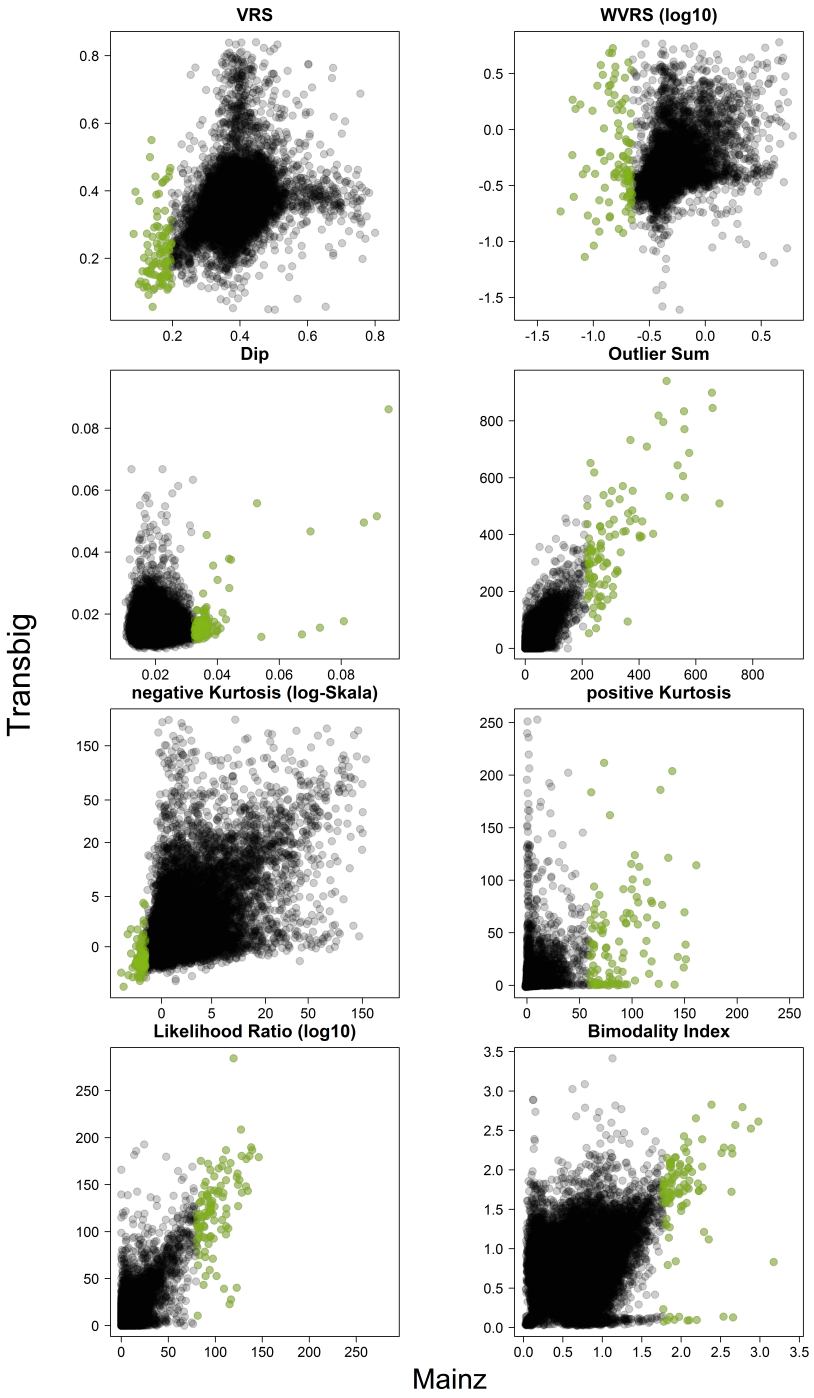

Abbildung 4.10: Streudiagramme der Bimodalitäts-Scores für Mainz- und Transbig-Kohorte. Jeder Punkt repräsentiert ein Gen, die 100 Gene mit den auffälligsten Werten des jeweiligen Scores sind grün dargestellt.

4.5 Vergleich der Ergebnisse mit den Ergebnissen für RMA-normalisierte Daten

Wie am Anfang des Ergebnisteils erläutert, wurden die in den bisherigen Kapiteln vorgestellten Analysen bereits in ähnlicher Form in Hellwig et al. (2010) beschrieben. Die Analysen wurden in der Publikation ebenfalls auf den Daten der Mainz-Kohorte durchgeführt, wobei es in dieser Arbeit vier Unterschiede zu der Vorgehensweise in der früheren Arbeit gibt:

1. Wir verwenden in dieser Arbeit die Genexpressionsdaten von 200 Patientinnen. Für die Analysen von Hellwig et al. (2010) standen Daten von lediglich 194 Patientinnen zur Verfügung.
2. Die Genexpressionsdaten wurden mit fRMA normalisiert, nicht mit RMA.
3. Es wird eine Korrektur der Clusterergebnisse wie in Kapitel 3.1.1 durchgeführt. Dieser Aspekt wurde in Hellwig et al. (2010) nicht beachtet.
4. Für die Überlebenszeitanalyse verwenden wir in dieser Arbeit das Metastasis Free Survival. In Hellwig et al. (2010) wurde die prognostische Relevanz in Bezug auf das Disease Free Survival (DFS) untersucht. Bei diesem ist neben dem Auftreten einer Fernmetastase und dem Tod durch Brustkrebs auch das Wiederauftreten der Erkrankung ein Zielereignis.

Es stellt sich die Frage, ob die Unterschiede einen Einfluss auf die Ergebnisse bezüglich der Bimodalitätsmaße haben. Um diese Fragestellung zu untersuchen, wird zunächst die Spearman-Korrelation der Bimodalitäts-Scores zwischen den beiden Datensätzen betrachtet (Tabelle F.2). Für alle Maße bis auf die dip-Statistik sind die Werte stark korreliert, die Korrelation liegt zwischen 0.692 (Bimodality Index) und 0.786 (Kurtosis). Der Wert für die dip-Statistik hingegen ist mit 0.089 nahe bei 0.

Zusätzlich werden Streudiagramme der Bimodalitäts-Scores für die beiden Ansätze betrachtet (Abbildung 4.11). Die 100 Gene mit den auffälligsten Bimodalitäts-Scores für die Mainz-Kohorte mit 200 Patientinnen und fRMA-normalisierten Genexpressionsdaten sind grün eingefärbt. Bei Outlier Sum (52), positiver Kurtosis (73) und Likelihood Ratio (80) sind viele Top100-Gene auch unter den Top100-Genen der Mainz-Kohorte mit 194

Patientinnen und RMA-normalisierten Daten zu finden. Die Punkte streuen in den interessanten Bereichen um eine Gerade mit positiver Steigung.

Bei der dip-Statistik fällt auf, dass es bei den fRMA-normalisierten Daten mehr Gene mit auffälligen Werten gibt als bei den RMA-normalisierten Daten. Es gibt offenbar Gene, deren Expressionsverteilungen nur bei den fRMA-normalisierten Daten bimodal im Sinne der dip-Statistik sind, Histogramme der Expressionswerte zweier dieser Gene sind in Abbildung G.4 im Anhang zu finden. Besonders auffällig sind dabei die Histogramme für das Gen *SUZ12P1*, das bei den fRMA-normalisierten Daten eine klare bimodale Struktur im Expressionsbereich von etwa 6 bis 8 zeigt, wohingegen die Verteilung der RMA-normalisierten Daten eher unimodal mit einigen Ausreißern im hohen Expressionsbereich ist. Die Expressionswerte liegen hier im Rauschbereich. Betrachtet man Streudiagramme der Expressionswerte für die 194 Patientinnen, für die beide Arten von Daten vorliegen, so ist zu sehen, dass bei *SUZ12P1* die Expressionswerte keine Korrelation aufweisen (oder wenn sogar eine negative). Die Hauptursache für die unterschiedlichen Ergebnisse ist vermutlich, dass bei fRMA auch Probe-spezifische Gewichte verwendet werden.

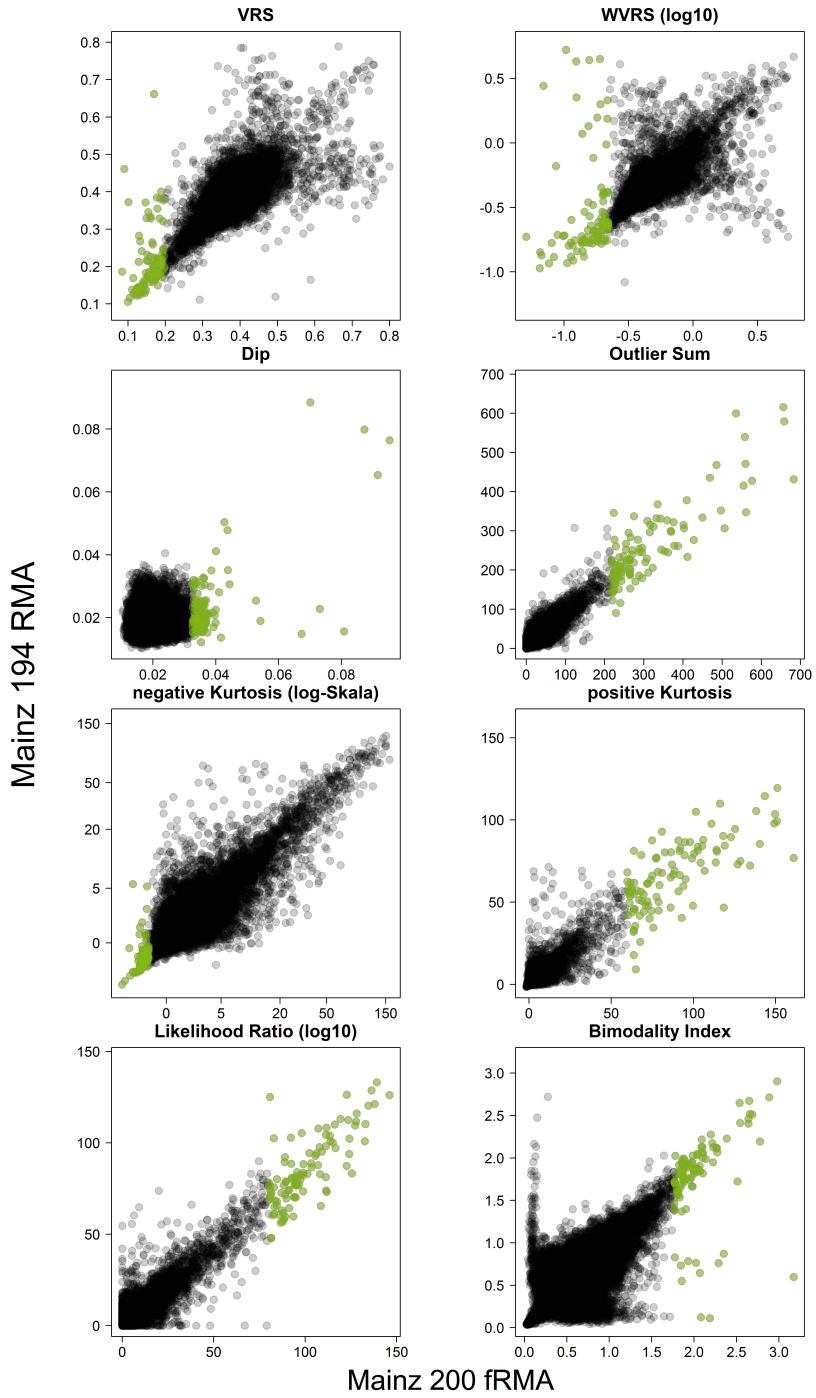

Abbildung 4.11: Vergleich der Bimodalitäts-Scores für die Mainz-Kohorte mit 200 Patientinnen und fRMA-normalisierten Daten (x-Achse) und der Mainz-Kohorte mit 194 Patientinnen und RMA-normalisierten Daten (y-Achse).

5 Klassifikation der Brustkrebspatientinnen

Auf Basis eines Klassifikators sollen die Patientinnen in eine Gruppe mit früher Metastase ($T+$, Metastase innerhalb der ersten 5 Jahre) und eine metastasenfreie Gruppe ($T-$, mindestens 5 Jahre beobachtet und keine Metastase) eingeteilt werden. Dazu werden nur die Patientinnen verwendet, die einer dieser Gruppen zugeordnet werden können. Patientinnen, die eine Metastase nach über 5 Jahren bekommen haben, und Patientinnen, die eine zensierte Beobachtungszeit von weniger als 5 Jahren haben, werden bei der Analyse nicht betrachtet. Bei der Mainz-Kohorte werden insgesamt 164 Patientinnen für die Klassifikation verwendet, 28 mit früher Metastase ($D+$) und 136 ohne Metastase ($D-$). Bei der Rotterdam-Kohorte können insgesamt 263 Patientinnen verwendet werden (95 $D+$, 168 $D-$), bei der Transbig-Kohorte 232 (52 $D+$, 180 $D-$). Die Klassifikationsgüte wird mit dem negativ prädiktiven Wert (NPV) und der Spezifität (TNR) beurteilt. Unter den Patientinnen, die der Klassifikator in die Gruppe mit guter Prognose einteilt, sollen möglichst viele Patientinnen sein, die tatsächlich keine Metastase bekommen, dies entspricht einem hohen Wert für NPV. Gleichzeitig soll der Klassifikator viele metastasenfreie Patientinnen richtig identifizieren, was einer hohen Spezifität entspricht.

Es werden verschiedene Ansätze verwendet. In Kapitel 5.1 werden zunächst die Ergebnisse der etablierten Klassifikatoren auf den drei Kohorten mit nodal-negativen Brustkrebspatientinnen betrachtet und verglichen. In den folgenden Kapiteln werden Methoden betrachtet Klassifikatoren mit Hilfe der Genexpressionsdaten zu bilden. Dabei wird zunächst versucht einen Klassifikator zu erstellen, der sowohl gut interpretierbar ist als auch eine gute Klassifikationsgüte besitzt. Dazu verwenden wir die vorausgewählten Gene mit charakteristischer Expressionsverteilung und kombinieren sie mit Klassifikationsbäumen (Kapitel 5.2). Im darauffolgenden Kapitel untersuchen wir die Fragestellung, ob bessere Klassifikationsergebnisse erreicht werden können, wenn nicht mehr vorausgesetzt wird, dass der Klassifikator leicht interpretierbar sein muss. Es werden Random Forests mit den vorausgewählten Genen aufgestellt (Kapitel 5.3). In Kapitel 5.4 wird schließlich versucht mit dem R-Paket `mlr` Random Forests bezüglich zweier Parameter zu optimieren, wobei nicht mehr die anhand der Bimodalitätsmaße ausgewählten Gene verwendet werden. Dies

soll dem Vergleich mit den Methoden mit vorausgewählten Genen dienen. Daher sind die Ergebnisse dieser Analysen in den Grafiken für die Baummodelle und Random Forests bereits enthalten (gekennzeichnet mit mlr.th für die Random Forests mit optimiertem Threshold und mlr.cw für Random Forests mit optimierten Klassengewichten).

Die Modelle werden jeweils auf der Mainz-Kohorte als Trainingskohorte aufgestellt und es wird anschließend versucht die Ergebnisse auf der Rotterdam- und der Transbig-Kohorte zu validieren.

5.1 Ergebnisse der bekannten Klassifikatoren auf den drei Kohorten mit unbehandelten Patientinnen

Um die Ergebnisse der neu-entwickelten Baummodelle mit denen etablierter Klassifikatoren vergleichen zu können, wurden für die drei Kohorten mit nodal-negativen unbehandelten Patientinnen Vorhersagen mit dem Bioconductor Paket `genefu` erstellt (Gendoo et al. 2015). Klasseneinteilungen nach dem 70-Gen-Klassifikator, dem 76-Gen-Klassifikator und Oncotype DX konnten für alle 3 Kohorten vorgenommen werden, da bei diesen Klassifikatoren nur die Genexpressionsdaten bzw. Genexpressionsdaten und immunhistochemisch bestimmter ER-Status (für den 76-Gen-Klassifikator) verwendet werden. Diese Informationen sind für alle drei Kohorten verfügbar, wobei für fünf Patientinnen der Transbig-Kohorte die Information über den ER-Status fehlt. Für den Genomic Grade Index (GGI) wird neben den Expressionswerten der histologische Tumorgrad benötigt. Dieser liegt uns für die Rotterdam-Kohorte nicht vor, daher kann der GGI für diese Kohorte nicht bestimmt werden. Für 14 Patientinnen der Transbig-Kohorte gibt es keine Informationen über den histologischen Grad (vgl. Tabelle A.5).

Für die Mainz-Kohorte kann zusätzlich zu den Gensignaturen auch die Risikogruppe nach der NNBC-3-Studie bestimmt werden. Oncotype DX und NNBC-3 teilen die Patientinnen jeweils in drei Gruppen mit niedrigem, mittlerem oder hohem Risiko ein. Die anderen Klassifikatoren teilen nur in zwei Gruppen (niedriges und hohes Risiko) ein. Um die Ergebnisse der Klassifikatoren zu vergleichen werden bei diesen Klassifikatoren die Patientinnen der Gruppe mit mittlerem Risiko einmal der Gruppen mit niedrigem und

einmal der Gruppe mit hohem Risiko zugeordnet, sodass es für diese Klassifikatoren jeweils zwei Ergebnisse gibt. Die Ergebnisse sind in Tabelle 5.1 zusammengefasst, die zugehörigen Vierfeldertafeln sind im Anhang (Kapitel C) zu finden.

Tabelle 5.1: NPV und Spezifität (TNR) für die Ergebnisse der bekannten Klassifikatoren für die drei Kohorten mit nodal-negativen unbehandelten Patientinnen.

		Mainz		Rotterdam		Transbig	
Klassifikator	Vergleich	NPV	TNR	NPV	TNR	NPV	TNR
70-Gen		0.957	0.324	0.847	0.363	0.940	0.350
76-Gen		1.000	0.037	0.966	0.167	0.947	0.102
Oncotype DX	*low* vs. *intermediate/high*	0.878	0.316	0.833	0.208	0.961	0.272
Oncotype DX	*low/intermediate* vs. *high*	0.920	0.588	0.752	0.452	0.893	0.511
GGI		0.928	0.662	-	-	0.909	0.611
NNBC-3	*low* vs. *intermediate/high*	0.965	0.417	-	-	-	-
NNBC-3	*low/intermediate* vs. *high*	0.948	0.553	-	-	-	-

Bei der Mainz-Kohorte wird der größte NPV-Wert für den Klassifikator NNBC-3 beobachtet, wenn man Patientinnen mit mittlerem und hohem Risiko zusammenfasst. 96.5 % der Patientinnen, die als low risk klassifiziert werden, bleiben metastasenfrei. Dabei werden 41.7 % der metastasenfreien Patientinnen auch als niedrig-Risiko Patientinnen erkannt. Ordnet man die Patientinnen mit mittlerem Risiko der low-risk-Gruppe zu, sinkt der NPV-Wert auf 0.948 und die Spezifität steigt auf 0.553. Den zweitgrößten NPV-Wert hat der 70-Gen-Klassifikator (0.957). Dieser identifiziert jedoch nur knapp ein Drittel (32.4 %) der metastasenfreien Patientinnen als low risk. Beim 76-Gen-Klassifikator wird ein NPV-Wert von 100 % erreicht, jedoch liegt der zugehörige Spezifitätswert bei nur 3.7 %. Das bedeutet, es werden nur 5 der 136 Patientinnen ohne Metastase in die low-risk-Gruppe eingeteilt. Auch GGI (0.928) und Oncotype DX (0.920) können hohe NPV-Werte mit Spezifität über 50 % erreichen. Die Werte bei Oncotype DX sind schlechter, wenn man die Patientinnen mit mittlerem Risiko der Gruppe mit hohem Risiko zuordnet. Insgesamt ist die Klassifikationsleistung bei der Mainz-Kohorte in Bezug auf die festgelegten Kriterien gut, wobei der 76-Gen-Klassifikator eine Ausnahme bildet.

Bei der Rotterdam-Kohorte kann nur der 76-Gen-Klassifikator einen NPV-Wert über 90 % erreichen (0.966). Allerdings ist auch hier wie bei der Mainz-Kohorte die Spezifität sehr

niedrig (0.167), es werden also nur wenige Patientinnen der low-risk-Gruppe zugeordnet. Bei der Transbig-Kohorte können wir für alle vier Gensignaturen NPV-Werte über 90 % beobachten. Den höchsten Wert hat Oncotype DX (niedrig vs. mittel/hoch, 0.961), wobei hier nur 27.2 % der low-risk-Patientinnen identifiziert werden. Insgesamt ist keine der Gensignaturen bei allen drei Kohorten die beste Methode. Beim 76-Gen-Klassifikator bleiben die als low risk klassifizierten Patientinnen bei allen drei Kohorten auch sehr sicher metastasenfrei, jedoch werden sehr wenige Patientinnen in die low-risk-Gruppe eingeordnet.

Bei der Interpretation der Ergebnisse ist zu beachten, dass die Klassifikation bei den Gensignaturen auch von der Vorverarbeitungsmethode der Expressionsdaten abhängt. Wählt man statt fRMA RMA als Vorverarbeitungsmethode, so erhält man andere Klassifikationsergebnisse.

5.2 Ergebnisse der Klassifikationsbäume

Um die von den verschiedenen Methoden identifizierten Gene zu einem Klassifikator zu kombinieren verwenden wir Klassifikationsbäume (vgl. Kapitel 3.4.2). Auf Basis des Klassifikators wollen wir die Patientinnen in eine Gruppe mit früher Metastase ($T+$, Metastase innerhalb der ersten 5 Jahre) und eine metastasenfreie Gruppe ($T-$, mindestens 5 Jahre beobachtet und keine Metastase) einteilen. Dazu verwenden wir nur die Patientinnen der Mainz-Kohorte, die einer dieser Gruppen zugeordnet werden können. Dies sind insgesamt 164 Patientinnen, 28 mit früher Metastase ($D+$), 136 ohne Metastase ($D-$). Die Klassifikationsgüte wird mit dem negativ prädiktiven Wert (NPV) und der Spezifität (TNR) beurteilt. Unter den Patientinnen, die der Klassifikator in die Gruppe mit guter Prognose einteilt, sollen möglichst viele Patientinnen sein, die tatsächlich keine Metastase bekommen, dies entspricht einem hohen Wert für NPV. Gleichzeitig soll der Klassifikator viele metastasenfreie Patientinnen richtig identifizieren, was einer hohen Spezifität entspricht. Anzumerken ist, dass aufgrund der stark unterschiedlich großen Gruppen die triviale Lösung alle Patientinnen als metastasenfrei zu klassifizieren zu wenigen Klassifikationsfehlern und einem NPV-Wert von 0.829 und Spezifität von 1.000

führt. Deshalb könnte der Algorithmus diese Lösung bevorzugen. Wir untersuchen daher den Einfluss unterschiedlicher Verlust-Werte auf das Klassifikationsergebnis.

Tabelle 5.2 enthält eine Übersicht der verschiedenen Parametereinstellungen, die verwendet wurden. Für jeden Score wählen wir die 100 Gene mit den extremsten Werten aus und fügen diese in Gruppen von 10 Genen dem Algorithmus hinzu. Der Verlust-Wert für die Zuordnung zur metastasenfreien Gruppe ($T-$) bleibt gleich, der Wert für die Zuordnung zur Gruppe mit früher Metastase ($T+$) wird im Bereich $2^{-10}, 2^{-9}, \ldots, 2^{10}$ verschoben. Der CART-Algorithmus wählt an jedem Knoten die beste Variable mit dem besten Cutoff aus. Das kann zu einer hohen Varianz der Klassifikationsbäume in einer Kreuzvalidierung führen. Durch die verschiedenen verwendeten Clusterverfahren haben wir für die Gene bereits eine Gruppeneinteilung vorgegeben. Daher sollen auch mit den dichotomisierten Werten Klassifikationsbäume erzeugt werden. Zu erwarten ist, dass diese Bäume eine stabilere Klassifikation erlauben. Darüber hinaus wäre die Interpretation dieser Baummodelle einfacher. In der klinischen Praxis haben sich wichtige

Tabelle 5.2: Verwendete Parametereinstellungen bei `rpart`.

Parameter	Einstellungen
Top-Gene	$10, 20, \ldots, 100$
Verlust für $T+$-Gruppe	$2^{-10}, 2^{-9}, \ldots, 2^{10}$
Verlust für $T-$-Gruppe	1
minsplit	20, 5
Komplexitätsparameter *cp*	0.01
Cutoff	fest vorgegeben, frei wählbar
klinische Variablen	nein, ja

klinische Einflussfaktoren etabliert: das Alter bei der Diagnose, pT-Stage, Tumorgrad, ER-Status und HER2-Status. Daher sollen alle Ergebnisse der Klassifikationsbäume, die nur die Gene als Variablen enthalten, mit einem Baummodell, dass nur die klinischen Variablen enthält und einem Modell, das sowohl klinische Variablen als auch Gene enthält, verglichen werden. Die Komplexität des Baumes wird durch den Komplexitätsparameter *cp* und das Argument *minsplit* kontrolliert, dieses gibt an, wie viele Elemente ein Knoten mindestens enthalten muss, damit ein Split versucht wird.

Für jede Kombination der Parametereinstellungen teilen wir den Datensatz zufällig in 10 Teildatensätze auf und bilden ein Baummodell auf 9/10 (Trainingsmenge) des Datensatzes. Mit dem Modell wird dann eine Vorhersage für das 1/10 (Testmenge) gemacht, das nicht für die Modellbildung verwendet wurde. Das Verfahren wird wiederholt, sodass jeder Teildatensatz einmal Testmenge ist. Insgesamt erhalten wir so für jede Probe eine Vorhersage für die Gruppenzugehörigkeit. Vorhersage und wahre Gruppenzugehörigkeit können verglichen werden und damit Maßzahlen zur Vorhersagegüte berechnet werden. Da die Gruppengrößen sich stark unterscheiden, wird das Randomisieren auf die Teildatensätze stratifiziert nach prognostischer Gruppe durchgeführt. Um die Reliabilität der Ergebnisse beurteilen zu können wurde dieses Verfahren jeweils $k = 100$ Mal durchgeführt.

Konstruktion der vergleichenden Streudiagramme

Um die Ergebnisse der Baummodelle für einen Score mit den verschiedenen Parametereinstellungen zu vergleichen, kann man die NPV- und Spezifitätswerte in einem Streudiagramm gegeneinander abtragen (vgl. Abbildung 5.1). In jedem dieser Streudiagramme ist der negativ prädiktive Wert auf der x-Achse und die Spezifität auf der y-Achse abgetragen. Für jede Kombination von Anzahl der Top-Gene und Verlust-Wert wurde der Median der Maßzahlen über die 100 Läufe berechnet und als Punkt in die Grafik eingetragen. Dabei haben die Punkte, die zur gleichen Anzahl von Top-Genen gehören, die gleiche Farbe und sind in der Reihenfolge der Verlust-Werte mit Linien der gleichen Farbe verbunden. Dem verwendeten Symbol kann man entnehmen, ob der zugehörige Verlustwert größer 1 (Dreieck mit Spitze nach oben), gleich 1 (Quadrat) oder kleiner 1 (Dreieck mit Spitze nach unten) ist. Jede Grafik enthält außerdem eine Kurve (dunkelgrau), die die Ergebnisse der Baummodelle mit den fünf klinischen Variablen darstellt. Die hellgrauen Punkte in der Grafik zeigen für jedes der 22 283 auf dem Microarray vorhandenen Gene die optimal erreichbare Kombination von NPV und Spezifität. Bei frei wählbarem Cutoff bedeutet das, dass für alle möglichen Cutoffs, die auf Basis der Expressionswerte des Gens möglich sind, NPV und Spezifität berechnet wurden und dann die Aufteilung so gewählt wurde, dass der Wert für NPV maximal ist. Bei den Grafiken für die Baummodelle mit fest vorgegebenem Cutoff gibt der hellgraue Punkt NPV-Wert und Spezifität an, die zu der Gruppeneinteilung gehören, die sich durch die dichotomisierten Expressionswerte ergibt. Für jedes Gen sind dabei zwei Einteilungen

möglich, je nachdem ob man Patientinnen mit hohen oder mit niedrigen Expressionswerten der Metastasen-Gruppe zuordnet. Gewählt wurde die Aufteilung mit dem größeren NPV-Wert. Dadurch, dass wir vier unterschiedliche Methoden zur Gruppeneinteilung verwenden, unterscheiden sich die hellgrauen Punkte in den Grafiken für Baummodelle mit fest vorgegebenem Cutoff für Maße, die zu unterschiedlichen Methoden gehören.

Die Ergebnisse der bekannten Klassifikatoren (vgl. Kapitel 5.1) sind ebenfalls in der Grafik dargestellt, sowie die Ergebnisse der Analysen aus Kapitel 5.4. Für Oncotype DX und NNBC-3 gibt es jeweils zwei Punkte, da man zwei Wertepaare erhält, je nachdem, ob man die Gruppe mit mittlerem Risiko der high-risk-Gruppe (l/ih) oder der low-risk-Gruppe (li/h) zuordnet (vgl. Kapitel 5.1). Die gestrichelte Linie markiert den Wert für den NPV, der bei Zuordnung aller Patientinnen in die metastasenfreie Gruppe erreicht werden würde (0.829). Ein guter Klassifikator nach unseren Kriterien würde in der Grafik rechts oben liegen. Betrachten wir die Punkte für die einzelnen Gene, so ist zu sehen, dass sowohl bei optimierten als auch bei den durch die Expressionsverteilung vorgegebenen Cutoffs, kein Gen eine Kombination von hohem NPV- und hohem Spezifitätswert erreicht. Die Gene mit NPV-Werten von über 0.95 haben maximal eine Spezifität von etwa 0.60. Für jeden Score sind die Plots für die vier Kombinationen von Art des Cutoffs und Verwendung von klinischen Variablen zusammen dargestellt. Dabei sind auf der linken Seite jeweils die Ergebnisse der Modelle nur mit Genen dargestellt und auf der rechten Seite die Ergebnisse der Modelle mit Genen und klinischen Variablen. In der oberen Reihe ist der Cutoff für die Gene frei wählbar, in der unteren fest vorgegeben.

Ergebnisse bei Wahl von minsplit = 20

Als erstes werden die Ergebnisse der Analysen bei Wahl von *minsplit* = 20 betrachtet. Diese Parametereinstellung bewirkt, dass nur Knoten mit mindestens 20 Elementen für einen Split zur Verfügung stehen. Die resultierenden Bäume sind weniger komplex als Bäume, die mit der Wahl eines kleineren Wertes für *minsplit* gebildet werden. Die NPV- und Spezifitätswerte der Baummodelle mit nur klinischen Variablen in Abhängigkeit des Verlustwertes für $T+$ sind in Tabelle 5.3 zusammengefasst. Für einen Verlustwert von 2^3 und größer wählt der Baum in allen Fällen die triviale Lösung, das heißt alle Patientinnen werden der metastasenfreien Gruppe zugeordnet. Bei Verlustwerten kleiner 1 steigen die NPV-Werte, wobei der maximale Wert von 0.917 bei einem Verlustwert von 2^{-6} erreicht

wird. Bei dieser Parametereinstellung sind also 91.7 % der als metastasenfrei klassifizierten Patientinnen tatsächlich metastasenfrei. Der zugehörige Wert der Spezifität ist jedoch nur 0.507, was bedeutet, dass nur etwa die Hälfte der metastasenfreien Patientinnen als metastasenfrei klassifiziert werden.

Tabelle 5.3: Ergebnisse der Baummodelle mit den klinischen Variablen Alter, pT-Stage, Tumorgrad, ER- und HER2-Status bei Wahl von *minsplit* = 20.

	NPV			Spezifität		
Verlust für $T+$	Median	IQR	Spannweite	Median	IQR	Spannweite
2^{-10}	**0.911**	0.022	0.073	**0.507**	0.046	0.140
2^{-9}	**0.912**	0.016	0.059	**0.507**	0.037	0.118
2^{-8}	**0.909**	0.023	0.079	**0.504**	0.039	0.140
2^{-7}	**0.915**	0.019	0.062	**0.500**	0.031	0.132
2^{-6}	**0.917**	0.021	0.066	**0.507**	0.037	0.125
2^{-5}	**0.912**	0.019	0.078	**0.507**	0.037	0.140
2^{-4}	**0.909**	0.019	0.087	**0.500**	0.037	0.110
2^{-3}	**0.899**	0.022	0.065	**0.551**	0.044	0.199
2^{-2}	**0.885**	0.009	0.035	**0.724**	0.051	0.191
2^{-1}	**0.891**	0.012	0.040	**0.912**	0.015	0.066
2^{0}	**0.869**	0.009	0.035	**0.934**	0.022	0.066
2^{1}	**0.862**	0.010	0.029	**0.971**	0.015	0.037
2^{2}	**0.832**	0.005	0.013	**0.985**	0.002	0.044
2^{3}	**0.829**	0.000	0.002	**1.000**	0.000	0.015
2^{4}	**0.829**	0.000	0.002	**1.000**	0.000	0.015
2^{5}	**0.829**	0.000	0.002	**1.000**	0.000	0.015
2^{6}	**0.829**	0.000	0.002	**1.000**	0.000	0.015
2^{7}	**0.829**	0.000	0.002	**1.000**	0.000	0.015
2^{8}	**0.829**	0.000	0.002	**1.000**	0.000	0.015
2^{9}	**0.829**	0.000	0.002	**1.000**	0.000	0.015
2^{10}	**0.829**	0.000	0.002	**1.000**	0.000	0.015

Die Grafiken für die verschiedenen Scores sind teilweise sehr ähnlich. Daher werden im Folgenden nur die Plots für die negative Kurtosis und das Likelihood Ratio (Abbildungen 5.1 und 5.2) betrachtet, die typische Verläufe zeigen. Die Grafiken für alle anderen untersuchten Kombinationen sind im Anhang (Kapitel D) zu finden. Bei der negativen Kurtosis (Abbildung 5.1) werden bei den Modellen mit ausschließlich Genen und frei wählbarem Cutoff für Verlust-Werte größer 1 keine großen Abweichungen von der trivialen Lösung beobachtet. Dieses Ergebnis ist zu erwarten, da mit größeren Werten

für den Verlust die Entscheidung für $T+$ stärker bestraft wird, wodurch eine Zuordnung aller Patientinnen in die Gruppe mit niedrigem Risiko mit weniger Kosten verbunden ist. Das Baummodell bevorzugt daher die Lösung alle Patientinnen als $T-$ zu klassifizieren. Diese Beobachtung lässt sich auch für die anderen Scores machen. Bei allen Varianten fällt darüber hinaus auf, dass für sehr kleine Verlustwerte keine starken Veränderungen der Werte mehr stattfinden. Am Ende jeder Kurve streuen die Werte nur noch gering. Auch dies ist für die anderen Scores zu beobachten.

Bei der negativen Kurtosis verlaufen die Kurven bis auf die Kurve für 10 Gene sehr ähnlich. Bei einem Verlustwert von 1 ist der mittlere NPV-Wert nicht viel höher als der NPV-Wert der trivialen Lösung, der Median der Spezifität liegt bei etwa 85 %. Für die Verlustwerte 2^{-1} und 2^{-2} steigt der NPV-Wert dann deutlich an, während die Spezifität sich nicht stark ändert. Dabei haben Modelle mit weniger Genen - das heißt es wurden dem Baum weniger Variablen zur Verfügung gestellt - höhere NPV-Werte. Werden zusätzlich zu den Genen die klinischen Variablen hinzugenommen, so sehen die Kurven anders aus. Bei Verlust-Werten in der Nähe der 1 sind die NPV-Werte größer als bei der trivialen Lösung, während die Spezifität nicht stark sinkt. Die Kurven folgen in diesem Bereich dem Verlauf der Kurve für das Baummodell nur mit klinischen Variablen. Das lässt die Vermutung zu, dass das Baummodell bei diesen Verlustwerten nur klinische Variablen verwendet. Die Kurven für 10-30 Top-Gene und kleinere Verlustwerte bewegen sich deutlich von der Kurve für die klinischen Variablen weg. Es werden NPV-Werte größer 90 % erreicht.

Werden die Cutoffs für die genetischen Variablen fest vorgegeben, so ist kein großer Einfluss der Anzahl der Top-Gene zu beobachten. Bei den Modellen, die nur Gene enthalten, bewegen sich die Kurven für kleine Verlust-Werte etwas von der trivialen Lösung weg, hohe NPV-Werte werden jedoch nicht erreicht. Bei den Modellen mit klinischen Variablen streuen die Kurven um die Kurve für das Modell mit nur klinischen Variablen.

Die Kurven vom Likelihood Ratio bei ausschließlicher Verwendung der Top-Gene mit frei wählbarem Cutoff zeigen alle einen ähnlichen Verlauf, wobei sie sehr nahe beieinander liegen. Zunächst sinkt die Spezifität bei gleichbleibendem NPV-Wert, dann bleibt die Spezifität gleich und der NPV-Wert steigt. Der größte NPV-Wert wird für das Baummodell

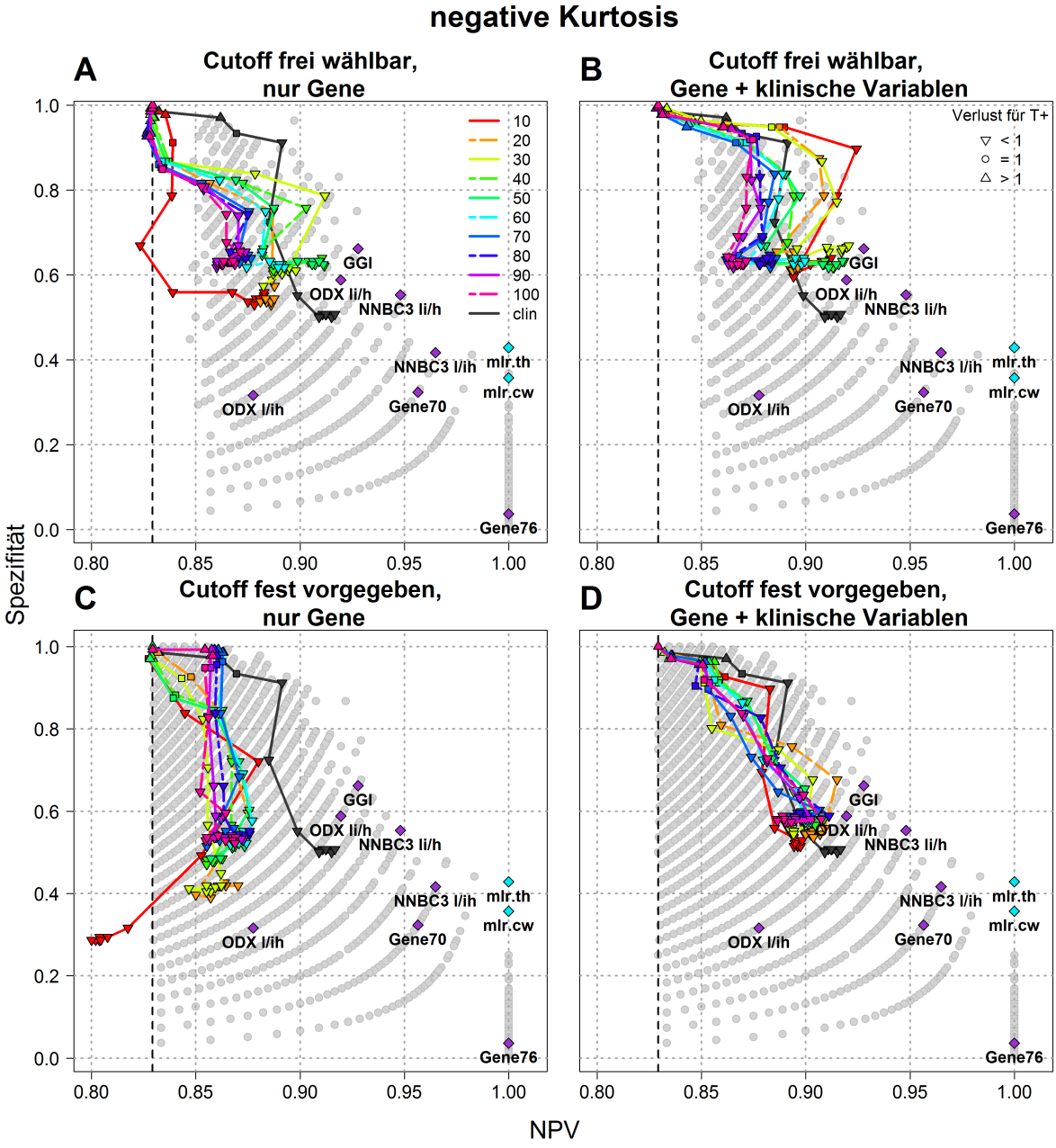

Abbildung 5.1: NPV und Spezifität der Klassifikationsbäume mit minsplit=20 für die Top-Gene der negativen Kurtosis. Links sind jeweils die Ergebnisse der Modelle nur mit Genen dargestellt, rechts die Ergebnisse der Modelle mit Genen und klinischen Variablen. In der oberen Reihe ist der Cutoff für die Gene frei wählbar, in der unteren fest vorgegeben.

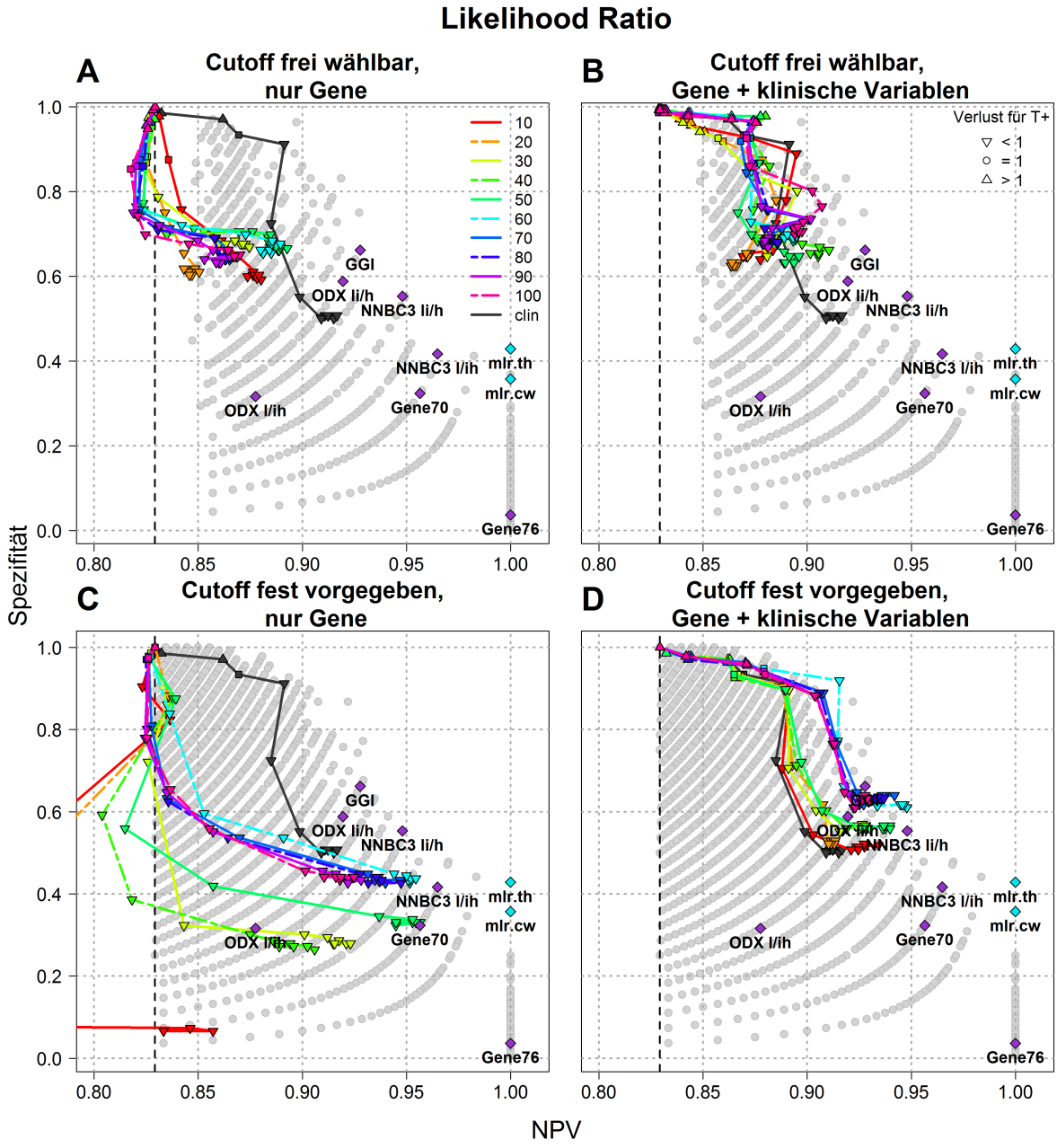

Abbildung 5.2: NPV und Spezifität der Klassifikationsbäume mit *minsplit* = 20 für die Top-Gene von Likelihood Ratio. Links sind jeweils die Ergebnisse der Modelle nur mit Genen dargestellt, rechts die Ergebnisse der Modelle mit Genen und klinischen Variablen. In der oberen Reihe ist der Cutoff für die Gene frei wählbar, in der unteren fest vorgegeben.

mit 50 Genen beobachtet (0.893, Spezifität 0.665). Werden zusätzlich zu den Top-Genen die klinische Variablen für die Konstruktion der Baummodelle verwendet, so kann man beobachten, dass die zugehörigen Kurven um die Kurve der für die klinischen Variablen streuen. Bei fest vorgegebenem Cutoff erhält man für weniger als 50 Gene sehr unterschiedliche Kurvenverläufe. Teilweise liegen die NPV-Werte unter dem Wert der trivialen Lösung. Die Kurven für 60 bis 90 Gene verlaufen nahezu aufeinander. Das ist damit zu begründen, dass bei den zugehörigen Modellen häufig die gleichen Gene verwendet werden. Die Gene, die der Klassifikationsbaum der Top90-Gene im Vergleich zu dem Baum der Top60-Gene zusätzlich zur Verfügung gestellt bekommt, werden für die Konstruktion der Klassifikationsbäume nicht verwendet. Werden zusätzlich zu den genetischen Variablen die klinischen Variablen verwendet, so liegen die Kurven für bis zu 50 Gene auf der Kurve für nur klinische Variablen. Die Kurven für 60 und mehr Gene liegen oberhalb dieser Kurve, der Verlauf sieht aber ähnlich aus.

Zum Vergleich der Ergebnisse der verschiedenen Scores betrachten wir Tabellen, die für jeden Score die Werte der drei Bäume mit den größten NPV-Werten enthalten (Tabelle 5.4 bis 5.7). Bei freier Wahl des Cutoffs und nur genetischen Variablen (Tabelle 5.4) ist die negative Kurtosis das einzige Bimodalitätsmaß, das einen NPV-Wert von über 0.9 erreicht. Bei den anderen Scores liegen die NPV-Werte zwischen 0.862 und 0.893. Die Streuung der NPV-Werte ist bei allen Maßen ähnlich, nur bei der Outlier-Sum-Statistik ist eine etwas größere Streuung zu beobachten. Auffällig ist, das nur bei VRS alle drei Modelle zu der gleichen Genanzahl gehören. Bei den Modellen mit klinischen Variablen haben die meisten Modelle einen NPV-Wert größer 0.9 (Tabelle 5.5). Die größten NPV-Werte werden für VRS und die negative Kurtosis beobachtet, die kleinsten für den Bimodality Index. Auch hier ist bei den meisten Scores die Anzahl der Top-Gene unterschiedlich.

Die Modelle mit fest vorgegebenem Cutoff und nur genetischen Variablen liefern für die unterschiedlichen Scores sehr unterschiedliche Ergebnisse (Tabelle 5.6). Hier fällt auf, dass bei der positiven Kurtosis die Top3-Klassifikatoren alle den NPV-Wert 1.000 haben, wobei die Spannweite der Werte ebenfalls 1.000 ist, was bedeutet, dass bei mindestens einem Durchlauf keine Patientin der metastasenfreien Gruppe zugeordnet wird. Der Interquartilsabstand beträgt 0, die Werte streuen also kaum. Die zugehörige Spezifität ist allerdings sehr niedrig (0.029), es werden also insgesamt sehr wenige Patientinnen in die Gruppe ohne Metastase klassifiziert. Die nächstgrößten NPV-Werte werden für das

Likelihood Ratio beobachtet, die drei besten Modelle haben einen NPV-Wert größer 0.95. Die zugehörige Spezifität liegt für zwei Modelle bei 0.331, für das dritte bei 0.438. Die kleinsten NPV-Werte haben die Modelle mit den Top-Genen von negativer Kurtosis und Bimodality Index.

Bei zusätzlicher Verwendung der klinischen Variablen (Tabelle 5.7) haben die Top3-Modelle mit den Top-Genen von Likelihood Ratio die größten NPV-Werte. Diese liegen bei etwas unter 0.95 und sind damit etwas niedriger als bei den Top-Modellen ohne klinische Variablen. Die zugehörige Spezifität ist aber höher (0.610, 0.618 und 0.618). Die Modelle aller Scores erreichen einen NPV-Wert über 0.9 und zeigen damit in Bezug auf den NPV-Wert eine bessere Klassifikationsleistung als das Modell, das nur klinische Variablen enthält (vgl. Tabelle 5.3).

Insgesamt zeigt sich für die acht Bimodalitätsmaße, dass bei Verwendung von dichotomisierten Daten höhere NPV-Werte erreicht werden. Das Hinzunehmen der klinischen Variablen verbessert die Klassifikationsleistung in Bezug auf die Spezifität und bei einigen Maßen auch in Bezug auf den NPV-Wert.

Tabelle 5.4: Bäume aus Top-Genen der Bimodalitäts-Scores ohne klinische Variablen mit den größten NPV-Werten bei frei wählbaren Cutoffs.

			NPV			Spezifität		
Score	Top	Verlust	Median	IQR	Spannweite	Median	IQR	Spannweite
negative Kurtosis	30	2^{-2}	**0.912**	0.019	0.073	**0.787**	0.037	0.154
	40	2^{-9}	**0.912**	0.022	0.086	**0.621**	0.037	0.147
	50	2^{-9}	**0.911**	0.021	0.088	**0.625**	0.029	0.154
Likelihood Ratio	50	2^{-9}	**0.893**	0.019	0.090	**0.665**	0.037	0.169
	60	2^{-6}	**0.889**	0.022	0.074	**0.676**	0.059	0.176
	50	2^{-7}	**0.889**	0.021	0.085	**0.669**	0.044	0.169
positive Kurtosis	100	2^{-3}	**0.875**	0.027	0.088	**0.699**	0.039	0.191
	90	2^{-3}	**0.874**	0.023	0.096	**0.691**	0.053	0.228
	100	2^{-5}	**0.872**	0.027	0.071	**0.713**	0.037	0.176
Bimodality Index	50	2^{-1}	**0.873**	0.022	0.092	**0.757**	0.044	0.147
	70	2^{0}	**0.871**	0.016	0.057	**0.853**	0.029	0.103
	80	2^{-3}	**0.871**	0.028	0.126	**0.643**	0.061	0.206
WVRS	80	2^{-2}	**0.876**	0.021	0.095	**0.746**	0.044	0.176
	80	2^{-3}	**0.871**	0.021	0.074	**0.713**	0.044	0.191
	100	2^{-3}	**0.869**	0.024	0.082	**0.684**	0.046	0.162
Dip	70	2^{-2}	**0.883**	0.023	0.092	**0.735**	0.044	0.169
	80	2^{-1}	**0.870**	0.021	0.070	**0.816**	0.037	0.118
	90	2^{-3}	**0.869**	0.025	0.115	**0.699**	0.061	0.213
Outlier Sum	70	2^{-3}	**0.876**	0.034	0.109	**0.676**	0.046	0.184
	80	2^{-3}	**0.869**	0.029	0.109	**0.684**	0.051	0.191
	30	2^{-4}	**0.862**	0.032	0.112	**0.577**	0.051	0.147
VRS	10	2^{-2}	**0.869**	0.023	0.088	**0.699**	0.040	0.169
	10	2^{-5}	**0.863**	0.029	0.113	**0.588**	0.053	0.169
	10	2^{-9}	**0.862**	0.031	0.116	**0.603**	0.051	0.147

Tabelle 5.5: Bäume aus Top-Genen der Bimodalitäts-Scores und klinischen Variablen mit den größten NPV-Werten bei frei wählbaren Cutoffs.

			NPV			Spezifität		
Score	Top	Verlust	Median	IQR	Spannweite	Median	IQR	Spannweite
negative Kurtosis	10	2^{-1}	**0.924**	0.014	0.052	**0.897**	0.022	0.066
	30	2^{-6}	**0.921**	0.018	0.078	**0.669**	0.046	0.250
	30	2^{-8}	**0.920**	0.020	0.081	**0.665**	0.044	0.147
positive Kurtosis	10	2^{-10}	**0.917**	0.014	0.055	**0.728**	0.029	0.140
	10	2^{-7}	**0.917**	0.015	0.046	**0.728**	0.037	0.118
	10	2^{-9}	**0.916**	0.012	0.052	**0.728**	0.029	0.162
Dip	10	2^{-8}	**0.915**	0.017	0.064	**0.732**	0.046	0.162
	70	2^{-2}	**0.913**	0.021	0.063	**0.846**	0.031	0.140
	10	2^{-5}	**0.913**	0.017	0.060	**0.739**	0.039	0.140
WVRS	10	2^{-3}	**0.915**	0.026	0.093	**0.640**	0.044	0.154
	10	2^{-10}	**0.912**	0.011	0.045	**0.724**	0.037	0.125
	10	2^{-7}	**0.911**	0.012	0.055	**0.721**	0.037	0.132
Outlier Sum	20	2^{-3}	**0.918**	0.029	0.081	**0.669**	0.029	0.154
	10	2^{-7}	**0.907**	0.026	0.090	**0.640**	0.037	0.125
	10	2^{-8}	**0.906**	0.023	0.091	**0.640**	0.029	0.132
Likelihood Ratio	40	2^{-6}	**0.910**	0.018	0.078	**0.662**	0.037	0.125
	100	2^{-3}	**0.907**	0.029	0.091	**0.765**	0.037	0.221
	40	2^{-5}	**0.906**	0.024	0.088	**0.669**	0.046	0.176
VRS	10	2^{-2}	**0.911**	0.011	0.041	**0.827**	0.029	0.096
	10	2^{-4}	**0.905**	0.016	0.070	**0.662**	0.044	0.184
	20	2^{-2}	**0.899**	0.010	0.038	**0.816**	0.037	0.132
Bimodality Index	30	2^{-3}	**0.906**	0.035	0.085	**0.676**	0.037	0.125
	10	2^{-3}	**0.895**	0.020	0.067	**0.603**	0.051	0.132
	80	2^{-2}	**0.895**	0.019	0.092	**0.779**	0.044	0.147

Tabelle 5.6: Bäume aus Top-Genen der Bimodalitäts-Scores ohne klinische Variablen mit den größten NPV-Werten bei fest vorgegebenen Cutoffs.

			NPV			Spezifität		
Score	Top	Verlust	Median	IQR	Spannweite	Median	IQR	Spannweite
positive Kurtosis	40	2^{-10}	**1.000**	0.000	1.000	**0.029**	0.015	0.059
	40	2^{-9}	**1.000**	0.000	1.000	**0.029**	0.015	0.059
	40	2^{-8}	**1.000**	0.000	1.000	**0.029**	0.022	0.059
Likelihood Ratio	50	2^{-9}	**0.957**	0.042	0.140	**0.331**	0.029	0.118
	50	2^{-7}	**0.955**	0.053	0.143	**0.331**	0.031	0.103
	60	2^{-6}	**0.955**	0.031	0.110	**0.438**	0.044	0.140
WVRS	90	2^{-10}	**0.929**	0.026	0.103	**0.485**	0.051	0.154
	90	2^{-6}	**0.929**	0.031	0.099	**0.478**	0.039	0.154
	90	2^{-9}	**0.928**	0.026	0.106	**0.485**	0.044	0.213
Outlier Sum	70	2^{-8}	**0.918**	0.049	0.169	**0.316**	0.031	0.088
	70	2^{-6}	**0.909**	0.043	0.137	**0.324**	0.029	0.147
	70	2^{-9}	**0.908**	0.044	0.129	**0.316**	0.037	0.125
Dip	10	2^{-5}	**0.907**	0.028	0.105	**0.294**	0.037	0.140
	10	2^{-4}	**0.901**	0.036	0.124	**0.316**	0.029	0.140
	20	2^{-4}	**0.897**	0.033	0.081	**0.471**	0.059	0.169
VRS	50	2^{-10}	**0.901**	0.037	0.141	**0.346**	0.044	0.154
	50	2^{-9}	**0.898**	0.038	0.143	**0.353**	0.039	0.132
	50	2^{-5}	**0.897**	0.038	0.132	**0.353**	0.046	0.176
negative Kurtosis	10	2^{-2}	**0.880**	0.021	0.066	**0.721**	0.044	0.199
	60	2^{-3}	**0.877**	0.023	0.124	**0.577**	0.059	0.228
	80	2^{-6}	**0.876**	0.025	0.103	**0.551**	0.051	0.206
Bimodality Index	80	2^{-6}	**0.877**	0.040	0.140	**0.482**	0.051	0.213
	100	2^{-10}	**0.877**	0.025	0.113	**0.500**	0.046	0.184
	100	2^{-9}	**0.873**	0.028	0.134	**0.500**	0.051	0.221

Tabelle 5.7: Bäume aus Top-Genen der Bimodalitäts-Scores und klinischen Variablen mit den größten NPV-Werten bei fest vorgegebenen Cutoffs.

			NPV			Spezifität		
Score	Top	Verlust	Median	IQR	Spannweite	Median	IQR	Spannweite
Likelihood Ratio	60	2^{-9}	**0.948**	0.017	0.080	**0.610**	0.037	0.125
	60	2^{-8}	**0.947**	0.021	0.075	**0.618**	0.037	0.132
	60	2^{-7}	**0.947**	0.024	0.075	**0.618**	0.037	0.110
Outlier Sum	10	2^{-3}	**0.949**	0.023	0.085	**0.588**	0.037	0.147
	20	2^{-3}	**0.943**	0.033	0.077	**0.566**	0.037	0.125
	30	2^{-3}	**0.935**	0.018	0.074	**0.654**	0.037	0.118
Bimodality Index	60	2^{-4}	**0.938**	0.022	0.098	**0.654**	0.029	0.147
	60	2^{-3}	**0.935**	0.022	0.083	**0.647**	0.029	0.132
	70	2^{-4}	**0.934**	0.022	0.088	**0.603**	0.037	0.169
WVRS	70	2^{-4}	**0.932**	0.029	0.086	**0.559**	0.022	0.103
	80	2^{-4}	**0.921**	0.030	0.104	**0.566**	0.037	0.125
	70	2^{-5}	**0.920**	0.031	0.103	**0.574**	0.029	0.132
VRS	70	2^{-7}	**0.921**	0.019	0.097	**0.585**	0.044	0.140
	70	2^{-10}	**0.920**	0.018	0.082	**0.581**	0.037	0.154
	70	2^{-6}	**0.919**	0.019	0.074	**0.588**	0.046	0.191
Dip	30	2^{-2}	**0.922**	0.011	0.053	**0.735**	0.044	0.140
	50	2^{-3}	**0.920**	0.019	0.092	**0.673**	0.037	0.147
	30	2^{-3}	**0.919**	0.017	0.067	**0.662**	0.044	0.199
positive Kurtosis	50	2^{-6}	**0.919**	0.020	0.068	**0.493**	0.037	0.169
	80	2^{-9}	**0.918**	0.019	0.083	**0.493**	0.029	0.140
	30	2^{-6}	**0.918**	0.017	0.088	**0.507**	0.029	0.125
negative Kurtosis	20	2^{-3}	**0.915**	0.023	0.073	**0.676**	0.037	0.125
	80	2^{-4}	**0.911**	0.027	0.090	**0.588**	0.059	0.147
	20	2^{-4}	**0.909**	0.018	0.075	**0.559**	0.040	0.169

Ergebnisse bei Wahl von minsplit = 5

Bei der Wahl von *minsplit* = 5 dürfen alle Knoten, die mindestens 5 Elemente enthalten, für einen Split verwendet werden. Die resultierenden Baummodelle sind komplexer als bei der Wahl von *minsplit* = 5, das heißt sie bestehen aus mehr Knoten. Komplexere Klassifikationsbäume sind stärker an die Trainingsdaten angepasst, was zu einer schlechteren Performance auf den Testdaten führen kann. Bei den NPV- und Spezifitätswerten für die Baummodelle aus klinischen Variablen bei Wahl von *minsplit* = 5 (Tabelle 5.8) fällt auf, dass sie für Verlust-Werte größer 1 sehr ähnlich sind. Den größten NPV-Wert beobachtet man für einen Verlust-Wert von 2^{-1} (NPV 0.894, Spezifität 0.860). Für kleinere Verlustwerte sinkt der NPV-Wert wieder, wobei die Spezifität gleich bleibt.

Tabelle 5.8: Ergebnisse der Baummodelle mit den klinischen Variablen Alter, pT-Stage, Tumorgrad, ER- und HER2-Status bei Wahl von *minsplit* = 5.

	NPV			Spezifität		
Verlust für $T+$	Median	IQR	Spannweite	Median	IQR	Spannweite
2^{-10}	**0.871**	0.013	0.055	**0.765**	0.029	0.103
2^{-9}	**0.872**	0.014	0.063	**0.765**	0.029	0.110
2^{-8}	**0.871**	0.014	0.055	**0.765**	0.031	0.088
2^{-7}	**0.873**	0.017	0.060	**0.765**	0.029	0.096
2^{-6}	**0.885**	0.017	0.068	**0.765**	0.024	0.096
2^{-5}	**0.887**	0.014	0.053	**0.757**	0.037	0.103
2^{-4}	**0.888**	0.018	0.064	**0.765**	0.029	0.096
2^{-3}	**0.886**	0.011	0.057	**0.757**	0.029	0.125
2^{-2}	**0.888**	0.011	0.059	**0.765**	0.044	0.147
2^{-1}	**0.894**	0.014	0.049	**0.860**	0.037	0.103
2^{0}	**0.875**	0.020	0.055	**0.919**	0.015	0.059
2^{1}	**0.847**	0.011	0.052	**0.934**	0.015	0.059
2^{2}	**0.839**	0.012	0.036	**0.941**	0.015	0.059
2^{3}	**0.843**	0.012	0.040	**0.941**	0.015	0.051
2^{4}	**0.844**	0.011	0.049	**0.941**	0.015	0.051
2^{5}	**0.843**	0.012	0.049	**0.941**	0.015	0.066
2^{6}	**0.844**	0.011	0.042	**0.941**	0.015	0.059
2^{7}	**0.848**	0.014	0.038	**0.941**	0.017	0.066
2^{8}	**0.844**	0.009	0.045	**0.941**	0.015	0.059
2^{9}	**0.843**	0.011	0.047	**0.949**	0.009	0.059
2^{10}	**0.842**	0.012	0.034	**0.941**	0.009	0.066

Die Grafiken für die Ergebnisse der Baummodelle für die verschiedenen Parametereinstellungen ähneln sich für die einzelnen Scores stark. Exemplarisch sollen daher nur die Grafiken für Likelihood Ratio betrachtet werden (Abbildung 5.3), die übrigen Grafiken sind im Anhang zu finden. Im Vergleich mit den Ergebnissen für *minsplit* = 20 fällt auf, dass es bei frei wählbarem Cutoff keine große Streuung der Werte für die verschiedenen Verlust-Werte und die unterschiedlichen Anzahlen von Top-Genen gibt. Für sehr kleine Verlust-Werte liegt der NPV-Wert nur knapp über dem NPV-Wert der trivialen Lösung. Auch bei zusätzlicher Verwendung der klinischen Variablen ist keine große Varianz der Ergebnisse zu beobachten. Die Kurven sind im Vergleich zu den Kurven mit nur genetischen Variablen etwas nach rechts verschoben. Wird der Cutoff für die genetischen Variablen fest vorgegeben, so unterscheiden sich die mittleren NPV- und Spezifitätswerte für die unterschiedlichen Anzahlen von Top-Genen ebenfalls nicht stark. Für große Verlust-Werte verlaufen die Kurven parallel zueinander. Für Verlust-Werte kleiner 2^{-3} fällt der NPV-Wert der Modelle mit 10 oder 20 Top-Genen unter den Wert von 0.829 (bei 10 Genen auf ca. 0.72, bei 20 Genen auf ca. 0.77), wobei die Spezifität sehr niedrig ist (bei 10 Genen ca. 0.13, bei 20 Genen ca. 0.38). Bei dieser Parametereinstellung werden demnach insgesamt wenige Patientinnen als metastasenfrei klassifiziert und darunter sind etwa 1/4 der Patientinnen mit früher Metastase. Bei zusätzlicher Verwendung von klinischen Variablen sind keine großen Abweichungen von der Kurve der Modelle mit nur klinischen Parametern zu beobachten. Die Verwendung der klinischen Variablen bringt hier also keinen Zusatznutzen. Lediglich das Modell mit 60 Top-Genen und Verlust 2^{-2} hat einen größeren NPV-Wert (0.912, Spezifität 0.824).

Bei den Modellen mit nur genetischen Variablen mit fest vorgegebenem Cutoff für die Top-Gene von VRS, WVRS und positiver Kurtosis (Abbildungen D.7, D.8 und D.12) fällt auf, dass für sehr kleine Verlust-Werte die Spezifität sehr stark fällt, der NPV-Wert aber bei etwa 0.83 bleibt. Insgesamt können bei der Parametereinstellung *minsplit* = 5 keine großen NPV-Werte erreicht werden. Das könnte damit begründet werden, dass die Klassifikationsbäume in den verschiedenen Iterationen zu stark an den Trainingsdatensatz angepasst sind (Overfitting). Die Ergebnisse sind für die verschiedenen Scores sehr ähnlich.

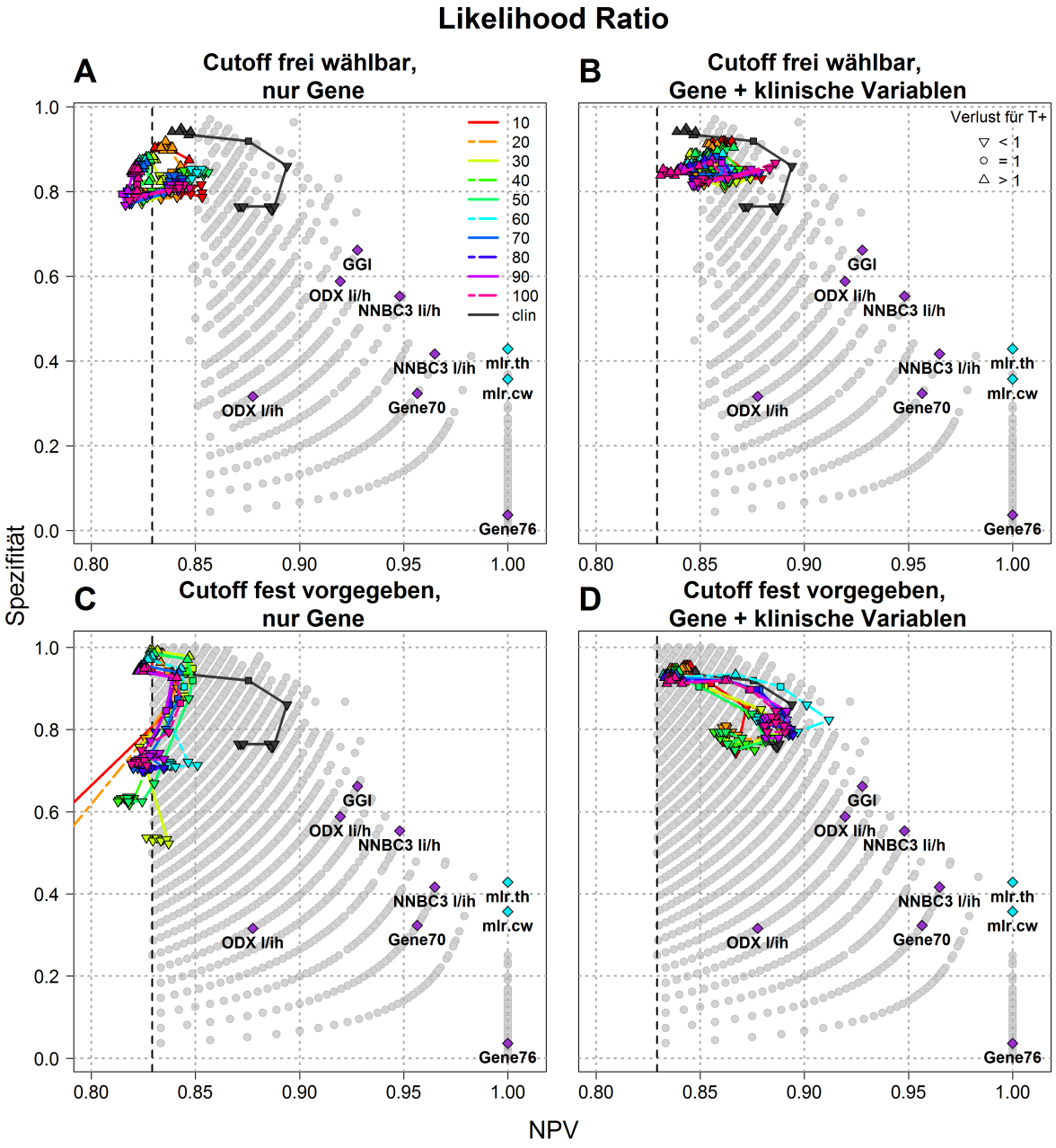

Abbildung 5.3: NPV und Spezifität der Klassifikationsbäume mit *minsplit* = 5 für die Top-Gene von Likelihood Ratio. Links sind jeweils die Ergebnisse der Modelle nur mit Genen dargestellt, rechts die Ergebnisse der Modelle mit Genen und klinischen Variablen. In der oberen Reihe ist der Cutoff für die Gene frei wählbar, in der unteren fest vorgegeben.

Beispiele für Klassifikationsbäume

In Abbildung 5.4 ist der Klassifikationsbaum für die Top30-Gene der negativen Kurtosis bei frei wählbarem Cutoff dargestellt. Dies ist der Baum mit der besten Kombination von NPV und Spezifität bei ausschließlicher Verwendung von Genen und freier Wahl des Cutoffs beim Splitten (vgl. Tabelle 5.4). An jedem Knoten ist ein Histogramm der Expressionswerte des entsprechenden Gens abgebildet, wobei der vom Modell gewählte Cutoff als vertikale rote Linie eingezeichnet ist. Der Baum besteht aus fünf Knoten mit fünf unterschiedlichen Genen. Dabei gibt es vier Terminalknoten. An der Wurzel des Baumes steht das Gen *JCHAIN*. Patientinnen mit hohen Expressionswerten von *JCHAIN* werden als metastasenfrei klassifiziert. Patientinnen mit niedrigen Expressionswerten von *JCHAIN* aber hohen Werten von *PSD3* werden ebenfalls der Gruppe ohne Metastase zugeordnet. Für Patientinnen, die für beide Gene niedrige Expressionswerte aufweisen, geschieht die weitere Aufteilung in Abhängigkeit von der Expression des Gens *DHRS2*. Patientinnen mit niedriger *DHRS2*-Expression und niedriger *CUL4A*-Expression werden als metastasenfrei klassifiziert und Patientinnen mit niedriger *DHRS2*-Expression und hoher *CUL4A*-Expression werden der Gruppe mit Metastase zugeordnet. Analog werden Patientinnen mit hoher *DHRS2*-Expression und hoher *STC2*-Expression der metastasenfreien Gruppe zugeordnet und Patientinnen mit hoher *DHRS2*-Expression und niedriger *STC2*-Expression der Gruppe mit Metastase. Betrachtet man die zugehörigen Histogramme, so kann man sehen, dass der vom Modell gewählte Cutoff für die Gene *JCHAIN*, *PSD3* und *DHRS2* gut zu der bimodalen Expressionsverteilung dieser Gene passt. Die Cutoffs für *CUL4A* und *STC2* passen dagegen nicht so gut.

Für die in diesem Baummodell enthaltenen Genen wurde zum Teil bereits ein Zusammenhang mit Brustkrebs beschrieben. Das Gen *PSD3* (*Pleckstrin And Sec7 Domain Containing* 3, Probe Set 218613_at) wurde von Thomassen et al. (2009) als Kandidat für ein Tumor-Supressor-Gen in Brustkrebs identifiziert. Field et al. (2012) untersuchten differentielle Genexpression in Tumoren von kaukasischen Frauen im Vergleich zu afroamerikanischen Frauen, da Tumore afroamerikanischer Frauen häufig schlechtere pathologische Charakteristika aufweisen und die Überlebensrate von Frauen dieser ethnischen Gruppe niedriger ist. In der Studie wurde *PSD3* als signifikant höher exprimiert in Tumoren kaukasischer Frauen identifiziert. Auch im hier betrachteten Baummodell werden Patientinnen mit höheren Werten als metastasenfrei klassifiziert. Dabei bekommen

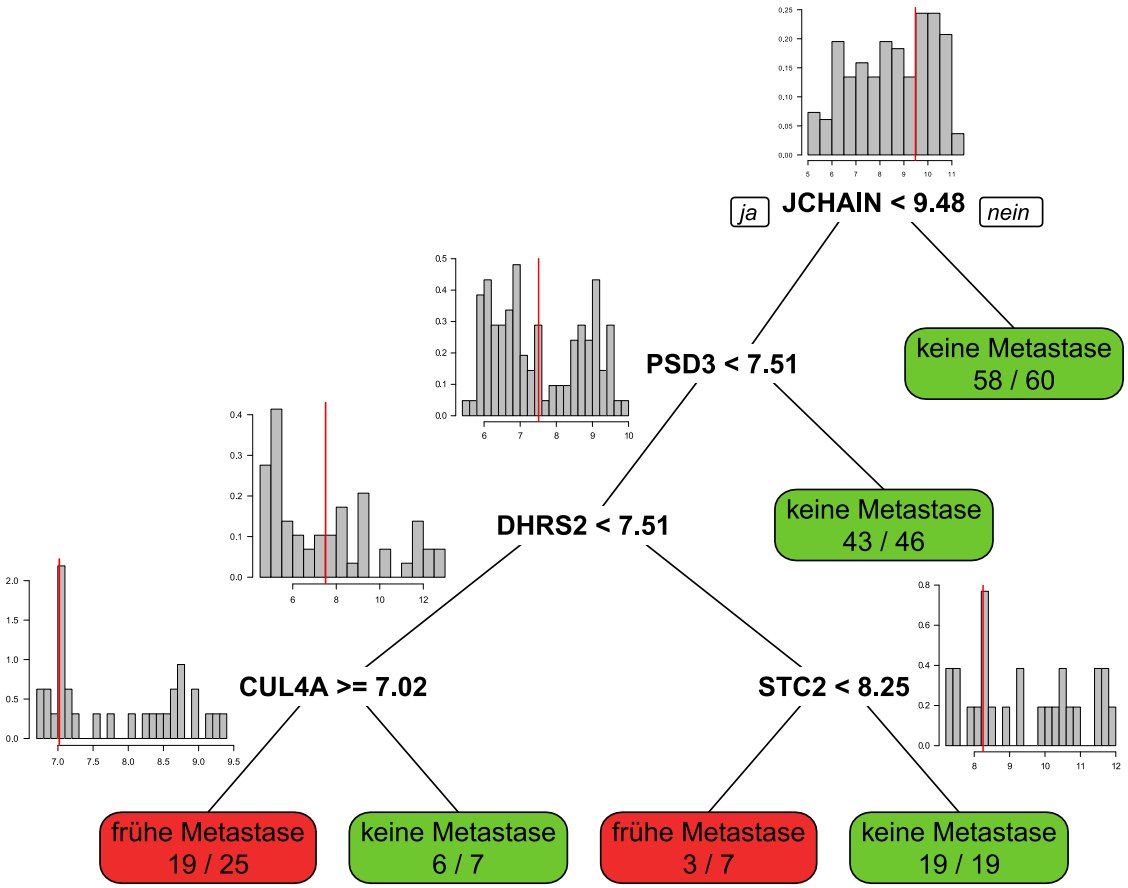

Abbildung 5.4: Klassifikationsbaum mit Top30-Genen der negativen Kurtosis bei frei wählbarem Cutoff und Verlust-Wert 2^{-2}.

43 von 46 als metastasenfrei klassifizierte Patientinnen auch tatsächlich keine Metastase. Das Gen *DHRS2* (*Dehydrogenase/Reductase 2*, auch *HEP27*, Probeset 214079_at) wurde in einer Studie von Deisenroth et al. (2010) als mögliches c-Myb-Target für die Stabilisierung der Aktivität des Onkogens *p53* identifiziert. Das Gen *CUL4A* (*Cullin 4A*, Probeset 201424_s_at) war in einer Studie von Chen et al. (1998) in Brustkrebszellen überexprimiert und wurde in Schindl et al. (2007) mit kürzerem Überleben in Verbindung gebracht. Todd et al. (2016) haben eine Assoziation von niedriger Expression von *STC2* (*Stanniocalcin 2*, Probeset 203438_at) mit einer kürzeren Überlebenszeit beobachtet. Das passt zum betrachteten Baummodell in dem Patientinnen mit hoher *CUL4A*- bzw. niedriger *STC2*-Expression der Gruppe mit früher Metastase zugeordnet werden. Für das Gen *JCHAIN* gibt es in der Literatur keine Hinweise auf einen Zusammenhang mit der Metastasenbildung bei Brustkrebs.

Als zweites Modell wird das beste Baummodell aus dichotomisierten Daten und klinischen Kovariablen betrachtet. Der Klassifikationsbaum für die Top60-Gene von Likelihood Ratio bei fest vorgegebenem Cutoff für die genetischen Variablen und zusätzlicher Verwendung der klinischen Variablen (Abbildung 5.5) besteht aus acht Knoten, wobei neben vier Genen (*ACE2*, *CYP2C8*, *TRH*, *ASCL1*) auch die Variablen Tumorgrad, Alter und pT-Staging vorkommen. Den Wurzelknoten bildet der Tumorgrad, da alle 36 Patientinnen mit Tumorgrad 1 metastasenfrei blieben, findet man diesen ersten Split in einer Vielzahl der Bäume mit klinischen Variablen. Patientinnen mit Tumorgrad 2 oder 3 und hohen Expressionswerten der Gene *ACE2*, *CYP2C8* oder *TRH* werden als metastasenfrei klassifiziert. Patientinnen mit niedriger Expression dieser drei Gene werden nochmals anhand des Tumorgrades aufgeteilt und Patientinnen mit Tumorgrad 3 der Gruppe mit Metastase zugeordnet. Patientinnen mit Tumorgrad 2 und einem Alter von über 74.3 Jahren werden als metastasenfrei klassifiziert. Die Gruppe der jüngeren Patientinnen wird anhand des pT-Stages aufgesplittet, wobei die Patientinnen mit Stage 2 der Gruppe mit Metastase zugeordnet werden. Für die übrigen Patientinnen wird die Expression des Gens *ASCL1* betrachtet. Patientinnen mit hoher *ASCL1*-Werten als metastasenfrei klassifiziert und Patientinnen mit niedrigen Expressionswerten der Gruppe der Metastasen-Gruppe zugeordnet.

Für zwei der Gene – *ACE2* und *CYP2C8* – wurde bereits ein Zusammenhang mit Brustkrebs beschrieben. Yu et al. (2016) haben beobachtet, dass eine Downregulierung

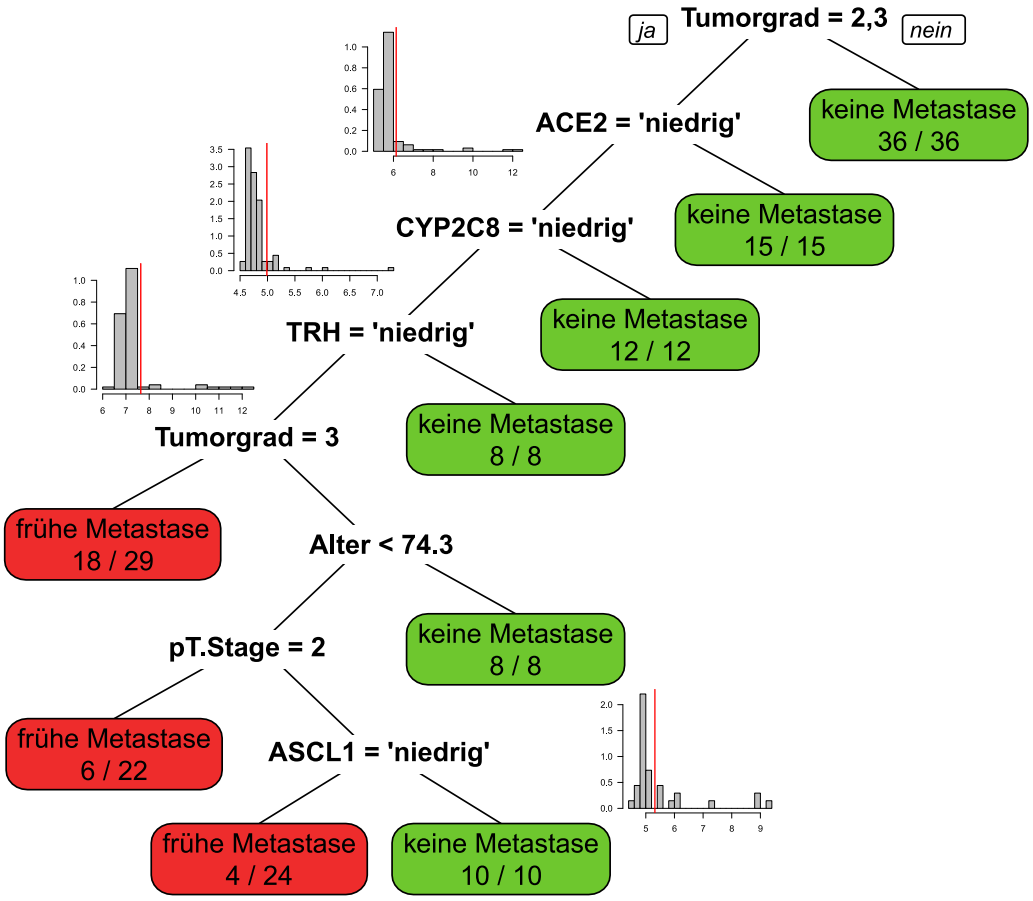

Abbildung 5.5: Klassifikationsbaum mit den Top60-Genen von Likelihood Ratio, bei fest vorgegebenem Cutoff und zusätzlicher Verwendung von klinischen Variablen und Verlust-Wert 2^{-9}.

von *ACE2* (*Angiotensin I Converting Enzyme 2*, Probeset 222257_s_at) die Metastasenbildung bei Brustkrebs begünstigt. In der Studie von Mitra et al. (2011) wurde beobachtet, dass *CYP2C8* (*Cytochrome P450 Family 2 Subfamily C Member 8*, Probe Set 2081247_s_at) das Wachstum von Brustkrebs-Zelllinien blockiert. Im Klassifikationsbaum werden die Patientinnen mit hohen Werten dieser Gene der metastasenfreien Gruppe zugeordnet. Für die Gene *TRH* und *ASCL1* hingegen ist in der Literatur kein Hinweis auf einen Zusammenhang mit Krebs zu finden.

5.2.1 Validierung der Klassifikationsbäume

In Kapitel 5.2 wurden für jedes der Bimodalitätsmaße die drei Klassifikationsbäume mit den höchsten NPV-Werten identifiziert. Nun stellt sich die Frage, wie gut die Klassifikatoren auf unabhängigen Datensätzen funktionieren. Um diese Frage zu beantworten gehen wir wie folgt vor: Es werden Klassifikationsbäume mit den Parametern der Top3-Modelle jedes Scores (vgl. Tabelle 5.4 bis 5.7) auf dem Gesamtdatensatz der Mainz-Kohorte angepasst. Mit dem jeweiligen Klassifikationsbaum wird dann eine Vorhersage für die Proben der Rotterdam- und der Transbig-Kohorte vorgenommen. Da für die Rotterdam-Kohorte außer dem ER-Status keine klinischen Variablen vorliegen, werden die Bäume, die klinische Variablen enthalten, nur auf der Transbig-Kohorte validiert. Die Ergebnisse sind in Tabelle 5.9 bis 5.12 dargestellt.

Für die Klassifikation mit fest vorgegebenen Cutoffs für die genetischen Variablen müssen die Expressionsdaten der beiden Validierungskohorten dichotomisiert werden. Die Frage, wie die Cutoffs für die Kohorten dabei gewählt werden sollten, ist nicht trivial. Eine einfache Möglichkeit ist es, den Cutoff der Mainz-Kohorte zu verwenden. Bei den modellbasierten Clusterverfahren ist der Cutoff durch den Schnittpunkt der Normalverteilungsdichten vorgegeben. Da allerdings bei ungleichen Varianzen der beiden Komponenten zwei Schnittpunkte existieren, muss einer der beiden ausgewählt werden (vgl. Kapitel 3.1.1). Auch bei der Outlier-Sum-Statistik können wir den Cutoff der Mainz-Kohorte leicht verwenden. Bei den Methoden, die die k-means-Clustereinteilungen verwenden, kann für jeden Expressionswert eines Gens der Abstand zu den beiden Clusterzentren berechnet werden. Die Beobachtung wird dann der Gruppe zugeordnet, zu deren Zentrum der Abstand kleiner ist. Allerdings ist anzumerken, dass der Cutoff der

Mainz-Kohorte nicht für die Validierungskohorten passen muss, da die Expressionswerte der Gene verschoben sein könnten. Trotzdem entscheiden wir uns für dieses Vorgehen, da es dem Vorgehen entspricht, das man in der Praxis für einzelne neue Messungen anwenden würde. Bei der Interpretation der Ergebnisse ist zu beachten, dass die trivialen NPV-Werte, die in den beiden Validierungskohorten durch Klassifizierung aller Patientinnen in die metastasenfreien Gruppe erreicht werden können, niedriger sind als in der Mainz-Kohorte (0.644 für Rotterdam und 0.776 für Transbig).

Der größte NPV-Wert bei den Baummodellen mit nur klinischen Variablen aus Kapitel 5.2 wurde für einen Verlust-Wert von 2^{-4} beobachtet. Bei der Validierung auf der Transbig-Kohorte erhalten wir einen NPV-Wert von 0.915 und eine Spezifität von 0.361. Damit ist die Klassifikationsgüte in Bezug auf NPV weniger gut als bei den etablierten Klassifikatoren 70-Gen-Klassifikator, 76-Gen-Klassifikator und Oncotype DX.

Bei den Baummodellen, die nur genetische Variablen enthalten, bei denen der Cutoff bei der Konstruktion frei gewählt werden darf (Tabelle 5.9), sind die NPV-Werte in der Rotterdam-Kohorte nahe an dem Wert, den man bei Zuordnung aller Patientinnen in die Gruppe ohne Metastase erreichen kann, oder sogar darunter. Insgesamt ist die Klassifikationsgüte nicht hoch. In der Transbig-Kohorte sind teilweise NPV-Werte von über 80 % zu beobachten (bei negativer Kurtosis, Likelihood Ratio, Bimodality Index, dip und VRS). Diese Werte sind aber für die klinische Praxis nicht vertretbar, da zu viele der Patientinnen, die eine Metastase bekommen, nicht erkannt werden.

Bei zusätzlicher Verwendung der klinischen Variablen sind die NPV-Werte in der Transbig-Kohorte ähnlich wie bei den Modellen ohne klinische Variablen (Tabelle 5.10). Bei negativer und positiver Kurtosis sind die NPV-Werte etwas größer. Der größte NPV-Wert (0.842) wird für ein Modell mit den Top-Genen des Bimodality Index beobachtet.

Bei ausschließlicher Verwendung von genetischen Variablen und fest vorgegebenen Cutoffs (Tabelle 5.11) werden in der Rotterdam-Kohorte für die Top3-Klassifikationsbäume ebenfalls nur sehr niedrige NPV-Werte erreicht. In der Transbig-Kohorte sind die Ergebnisse heterogen. Die Top3-Baummodelle der positiven Kurtosis erreichen nur einen NPV-Wert von 0.688, die Modelle für die Outlier-Sum-Statistik einen NPV-Wert von 0.912. Auffällig ist, dass bei diesen Maßen die jeweiligen Top3-Baummodelle dieselben NPV- und Spezifitätswerte haben. Das ist damit zu begründen, dass bei der Modellbildung

auf der gesamten Mainz-Kohorte unabhängig von dem Verlust-Wert dasselbe Modell gebildet wird. Das gleiche ist auch bei den Top3-Modellen von Likelihood Ratio und VRS mit vorgegebenem Cutoff und Verwendung von klinischen Variablen zu beobachten (Tabelle 5.11). Bei den Modellen mit klinischen Variablen sind die NPV-Werte größer als bei den Modellen ohne klinische Variablen. Für die Spezifität sind die Ergebnisse nicht einheitlich.

Insgesamt sind die Ergebnisse der Validierung der Top-Klassifikationsbäume nicht zufriedenstellend. Auf der Rotterdam-Kohorte können sowohl mit freiem als auch mit fest vorgegebenem Cutoff nur kleine bis mittlere NPV-Werte erreicht werden. Auf der Transbig-Kohorte werden im Allgemeinen höhere Werte erzielt, allerdings liegen auch diese bis auf wenige Ausnahmen nicht in einem für die klinische Praxis tolerierbaren Bereich. Es gibt keinen Klassifikationsbaum, der auf beiden Validierungskohorten gute Ergebnisse liefert. Das liegt aber zum Teil auch daran, dass für die Bäume mit klinischen Variablen nur die Transbig-Kohorte als Validierungskohorte zur Verfügung steht.

Tabelle 5.9: Validierung der Top3-Klassifikationsbäume nur mit Genen bei frei wählbaren Cutoffs.

			Rotterdam		Transbig	
Score	Top-Gene	Gewicht	NPV	Spezifität	NPV	Spezifität
negative Kurtosis	30	2^{-2}	0.663	0.810	0.795	0.778
	40	2^{-9}	0.652	0.613	0.809	0.633
	50	2^{-9}	0.652	0.613	0.809	0.633
Likelihood Ratio	50	2^{-9}	0.657	0.685	0.801	0.694
	60	2^{-6}	0.642	0.714	0.794	0.706
	50	2^{-7}	0.657	0.685	0.801	0.694
positive Kurtosis	100	2^{-3}	0.664	0.482	0.763	0.483
	90	2^{-3}	0.664	0.482	0.763	0.483
	100	2^{-5}	0.607	0.542	0.765	0.689
Bimodality Index	50	2^{-1}	0.622	0.440	0.826	0.661
	70	2^{0}	0.678	0.690	0.840	0.672
	80	2^{-3}	0.624	0.315	0.844	0.572
WVRS	80	2^{-2}	0.610	0.494	0.775	0.556
	80	2^{-3}	0.615	0.494	0.774	0.572
	100	2^{-3}	0.648	0.548	0.781	0.556
Dip	70	2^{-2}	0.612	0.488	0.783	0.561
	80	2^{-1}	0.687	0.339	0.794	0.472
	90	2^{-3}	0.639	0.631	0.809	0.517
Outlier Sum	70	2^{-3}	0.632	0.714	0.754	0.717
	80	2^{-3}	0.632	0.714	0.754	0.717
	30	2^{-4}	0.626	0.577	0.758	0.628
VRS	10	2^{-2}	0.669	0.673	0.790	0.689
	10	2^{-5}	0.664	0.601	0.777	0.639
	10	2^{-9}	0.681	0.673	0.817	0.572

Tabelle 5.10: Validierung der Top3-Klassifikationsbäume mit Genen und klinischen Variablen bei frei wählbaren Cutoffs.

			Transbig	
Score	Top-Gene	Verlust	NPV	Spezifität
negative Kurtosis	10	2^{-1}	0.821	0.739
	30	2^{-6}	0.821	0.639
	30	2^{-8}	0.829	0.756
positive Kurtosis	10	2^{-10}	0.807	0.767
	10	2^{-7}	0.807	0.767
	10	2^{-9}	0.807	0.767
Dip	10	2^{-8}	0.828	0.750
	70	2^{-2}	0.829	0.806
	10	2^{-5}	0.828	0.750
WVRS	10	2^{-3}	0.767	0.183
	10	2^{-10}	0.816	0.689
	10	2^{-7}	0.816	0.689
Outlier Sum	20	2^{-3}	0.807	0.628
	10	2^{-7}	0.802	0.472
	10	2^{-8}	0.802	0.472
Likelihood Ratio	40	2^{-6}	0.774	0.533
	100	2^{-3}	0.823	0.594
	40	2^{-5}	0.774	0.533
VRS	10	2^{-2}	0.782	0.756
	10	2^{-4}	0.814	0.461
	20	2^{-2}	0.780	0.750
Bimodality Index	30	2^{-3}	0.833	0.611
	10	2^{-3}	0.838	0.661
	80	2^{-2}	0.842	0.650

Tabelle 5.11: Validierung der Top3-Klassifikationsbäume nur mit Genen bei fest vorgegebenen Cutoffs.

			Rotterdam		Transbig	
Score	Top-Gene	Gewicht	NPV	Spezifität	NPV	Spezifität
positive Kurtosis	40	2^{-10}	0.455	0.030	0.688	0.061
	40	2^{-9}	0.455	0.030	0.688	0.061
	40	2^{-8}	0.455	0.030	0.688	0.061
Likelihood Ratio	50	2^{-9}	0.507	0.202	0.849	0.250
	50	2^{-7}	0.506	0.238	0.812	0.289
	60	2^{-6}	0.552	0.286	0.860	0.272
WVRS	90	2^{-10}	0.653	0.369	0.869	0.478
	90	2^{-6}	0.653	0.369	0.869	0.478
	90	2^{-9}	0.699	0.304	0.866	0.467
Outlier Sum	70	2^{-8}	0.476	0.119	0.912	0.172
	70	2^{-6}	0.476	0.119	0.912	0.172
	70	2^{-9}	0.476	0.119	0.912	0.172
Dip	10	2^{-5}	0.686	0.286	0.813	0.411
	10	2^{-4}	0.686	0.286	0.813	0.411
	20	2^{-4}	0.622	0.577	0.828	0.400
VRS	50	2^{-10}	0.594	0.226	0.763	0.250
	50	2^{-9}	0.594	0.226	0.763	0.250
	50	2^{-5}	0.594	0.226	0.763	0.250
negative Kurtosis	10	2^{-2}	0.656	0.637	0.814	0.583
	60	2^{-3}	0.672	0.536	0.847	0.617
	80	2^{-6}	0.651	0.565	0.788	0.578
Bimodality Index	80	2^{-6}	0.679	0.214	0.800	0.178
	100	2^{-10}	0.645	0.357	0.773	0.322
	100	2^{-9}	0.645	0.357	0.773	0.322

Tabelle 5.12: Validierung der Top3-Klassifikationsbäume mit Genen und klinischen Variablen bei fest vorgegebenen Cutoffs.

			Transbig	
Score	Top-Gene	Verlust	NPV	Spezifität
Likelihood Ratio	60	2^{-9}	0.864	0.422
	60	2^{-8}	0.864	0.422
	60	2^{-7}	0.864	0.422
Outlier Sum	10	2^{-3}	0.926	0.417
	20	2^{-3}	0.926	0.417
	30	2^{-3}	0.852	0.578
Bimodality Index	60	2^{-4}	0.845	0.333
	60	2^{-3}	0.845	0.333
	70	2^{-4}	0.828	0.400
WVRS	70	2^{-4}	0.885	0.256
	80	2^{-4}	0.894	0.233
	70	2^{-5}	0.885	0.256
VRS	70	2^{-7}	0.830	0.461
	70	2^{-10}	0.830	0.461
	70	2^{-6}	0.830	0.461
Dip	30	2^{-2}	0.827	0.583
	50	2^{-3}	0.835	0.533
	30	2^{-3}	0.835	0.533
positive Kurtosis	50	2^{-6}	0.852	0.383
	80	2^{-9}	0.852	0.383
	30	2^{-6}	0.915	0.361
negative Kurtosis	20	2^{-3}	0.864	0.422
	80	2^{-4}	0.889	0.489
	20	2^{-4}	0.911	0.400

5.3 Ergebnisse der Random Forests

In diesem Abschnitt soll die Frage beantwortet werden, ob wir bessere Klassifikationsergebnisse erhalten können, wenn wir nicht mehr fordern, dass der Klassifikator gut interpretierbar sein soll. Da Random Forests mit den Baummodellen verwandt sind und im Allgemeinen gute Klassifikationsergebnisse liefern (Fernández-Delgado et al. 2014), haben wir uns für dieses Klassifikationsverfahren entschieden.

Analog zu dem Vorgehen aus Kapitel 5.2 wurden für die Top-Gene der einzelnen Bimodalitätsmaße Random Forests mit unterschiedlichen Parametereinstellungen gebildet, wobei für jede Kombination der Parametereinstellungen 100 Replikationen erstellt wurden. Eine Übersicht der Parametereinstellungen ist in Tabelle 5.13 zu finden. Statt eines Verlust-Wertes für die Zuordnung in die Metastasen-Gruppe wie bei den Bäumen wird eine Sequenz von Klassengewichten verwendet. Diese Klassengewichte werden analog zu den Verlust-Werten bei den Klassifikationsbäumen bei der Konstruktion der einzelnen Bäume als Gewichte zum Verändern der Klassenwahrscheinlichkeiten zum Identifizieren des besten Splits verwendet. Alle Analysen werden zusätzlich einmal mit und einmal ohne Down-Sampling durchgeführt. Bei der Verwendung von Down-Sampling wird innerhalb eines Schrittes der 10-fachen Kreuzvalidierung die Bootstrap-Stichprobe so gezogen, dass die Gruppengrößen gleich sind. Dabei entspricht der Stichprobenumfang pro Gruppe der Anzahl der Patientinnen mit früher Metastase im jeweiligen Trainingsdatensatz.

Tabelle 5.13: Verwendete Parametereinstellungen bei `randomForest`.

Parameter	Einstellungen
Top-Gene	$10, 20, \ldots, 100$
Gewicht für $T+$-Gruppe	$2^{-22}, 2^{-20}, \ldots, 2^{10}$
Gewicht für $T-$-Gruppe	1
Down-Sampling	nein, ja
nodesize	1
Cutoff	fest vorgegeben, frei wählbar
klinische Variablen	nein, ja

Es wurden ebenfalls analog zum Vorgehen bei den Klassifikationsbäumen zusätzlich Random Forests erstellt, die nur die klinischen Variablen enthalten. Dabei wurden die

Gewichte für die Gruppe mit Metastasen variiert und ebenfalls Modelle mit und ohne Down-Sampling aufgestellt (vgl. Tabelle 5.14). Die NPV-Werte sind nur bei sehr kleinen Gewichten (2^{-20} ohne Down-Sampling, 2^{-18} mit Down-Sampling) deutlich über 90 %, wobei mit Down-Sampling ein höherer Wert erreicht wird (0.975 gegenüber 0.929). Die zugehörigen Spezifitätswerte sind relativ klein (0.287 ohne Down-Sampling, 0.265 mit Down-Sampling). Bei den Modellen ohne Down-Sampling sind NPV- und Spezifitätswerte für die Gewichte zwischen 2^{-16} und 2^{-4} sehr ähnlich, der NPV liegt zwischen 0.807 und 0.818 und die Spezifität zwischen 0.493 und 0.507. Für Gewichte größer 1 sind die Werte nah an denen der trivialen Lösung.

Bei den Modellen mit Down-Sampling fällt auf, dass bei einem Gewicht von 2^{-20} der Median der NPV-Werte 0 ist. Die Spannweite der Werte ist 1.000 und der zugehörige Median der Spezifitätswerte 0.000 (mit Spannweite 0.007). Bei diesem kleinen Gewicht werden die Patientinnen in 98 von 100 Durchläufen alle der Gruppe mit Metastase zugeordnet. Für die anderen Gewichte ist keine große Streuung der NPV-Werte zu beobachten (0.865 bis 0.906). Mit größerem Gewicht steigt der NPV-Wert dabei tendenziell etwas und pendelt sich bei ca. 0.890 ein. Die zugehörige Spezifität wird deutlich höher und erreicht einen Wert von ca. 90 % (das heißt es werden mehr Patientinnen der metastasenfreien Gruppe zugeordnet).

Die Ergebnisse der Random Forests mit den Top-Genen der jeweiligen Scores werden wie bei den Klassifikationsbäumen in Streudiagrammen dargestellt (vgl. Kapitel 5.2). Dabei entsprechen die Punkte dem Median von NPV und Spezifität über die 100 Replikationen der jeweiligen Parameterkombinationen. Für eine bessere Vergleichbarkeit der Auswirkungen des Down-Samplings auf das Klassifikationsergebnis sind die vier Grafiken, die zu einem Datentyp (Originaldaten, dichotomisierte Daten) eines Bimodalitätsmaßes gehören, zusammen in einer Abbildung gruppiert. Pro Bimodalitätsmaß gibt es folglich zwei Abbildungen mit vier Streudiagrammen. In den Abbildungen sind auf der linken Seite die Ergebnisse der Modelle nur mit Genen dargestellt und auf der rechten Seite die Ergebnisse der Modelle mit Genen und klinischen Variablen. Bei den Modellen in der oberen Reihe wurde kein Down-Sampling verwendet, bei den Modellen in der unteren Reihe wurde Down-Sampling verwendet. Bei Verwendung eines sehr kleinen Klassengewichtes ist der mittlere NPV-Wert häufig 0, d. h. alle Patientinnen werden in die Gruppe mit

Tabelle 5.14: NPV und Spezifität der Random Forests mit ausschließlich klinischen Variablen.

ohne Down-Sampling						
	NPV			Spezifität		
Gewicht	Median	IQR	Spannweite	Median	IQR	Spannweite
2^{-22}	**0.000**	0.000	1.000	**0.000**	0.000	0.015
2^{-20}	**0.929**	0.039	0.104	**0.287**	0.022	0.059
2^{-18}	**0.859**	0.024	0.084	**0.449**	0.029	0.132
2^{-16}	**0.815**	0.018	0.058	**0.493**	0.015	0.081
2^{-14}	**0.813**	0.018	0.069	**0.500**	0.022	0.074
2^{-12}	**0.818**	0.018	0.100	**0.493**	0.029	0.081
2^{-10}	**0.813**	0.023	0.070	**0.493**	0.022	0.125
2^{-8}	**0.816**	0.028	0.094	**0.493**	0.024	0.081
2^{-6}	**0.812**	0.022	0.079	**0.493**	0.024	0.103
2^{-4}	**0.807**	0.021	0.090	**0.507**	0.024	0.081
2^{-2}	**0.849**	0.014	0.069	**0.669**	0.029	0.103
2^{0}	**0.866**	0.013	0.047	**0.860**	0.015	0.081
2^{2}	**0.844**	0.010	0.035	**0.949**	0.007	0.044
2^{4}	**0.828**	0.005	0.015	**0.963**	0.007	0.037
2^{6}	**0.826**	0.005	0.015	**0.967**	0.007	0.029
2^{8}	**0.826**	0.006	0.019	**0.967**	0.007	0.037
2^{10}	**0.829**	0.005	0.014	**0.971**	0.007	0.029
mit Down-Sampling						
	NPV			Spezifität		
Gewicht	Median	IQR	Spannweite	Median	IQR	Spannweite
2^{-22}	**0.000**	0.000	0.000	**0.000**	0.000	0.000
2^{-20}	**0.000**	0.000	1.000	**0.000**	0.000	0.007
2^{-18}	**0.975**	0.027	0.071	**0.265**	0.007	0.037
2^{-16}	**0.868**	0.031	0.106	**0.331**	0.029	0.103
2^{-14}	**0.865**	0.026	0.114	**0.324**	0.022	0.096
2^{-12}	**0.873**	0.028	0.102	**0.331**	0.024	0.096
2^{-10}	**0.875**	0.032	0.126	**0.331**	0.015	0.081
2^{-8}	**0.874**	0.025	0.105	**0.331**	0.022	0.081
2^{-6}	**0.868**	0.031	0.098	**0.331**	0.022	0.088
2^{-4}	**0.865**	0.020	0.110	**0.331**	0.029	0.081
2^{-2}	**0.870**	0.030	0.097	**0.331**	0.022	0.088
2^{0}	**0.881**	0.017	0.060	**0.412**	0.029	0.096
2^{2}	**0.906**	0.010	0.039	**0.676**	0.024	0.088
2^{4}	**0.893**	0.011	0.039	**0.868**	0.015	0.074
2^{6}	**0.890**	0.007	0.040	**0.897**	0.015	0.051
2^{8}	**0.887**	0.009	0.035	**0.897**	0.015	0.066
2^{10}	**0.890**	0.009	0.038	**0.904**	0.015	0.059

Metastase klassifiziert. Die zugehörigen Punkte werden aus Gründen der Übersichtlichkeit in der Grafik nicht dargestellt.

Bei frei wählbarem Cutoff ist ohne Down-Sampling bei keinem der acht Scores ein großer Einfluss des Gewichtes zu beobachten. Abbildung 5.6 zeigt exemplarisch die Kurven für das Likelihood Ratio. Die übrigen Grafiken sind in Kapitel E.1.1 zu finden. Die Werte streuen um die Werte der trivialen Lösung. Nur bei dem kleinsten verwendeten Gewicht und 10 oder 20 Top-Genen wird ein NPV-Wert von 1.000 erreicht. Die zugehörigen Spezifitätswerte sind allerdings sehr klein (0.022 bei den Top10 bzw. 0.007 bei den Top20), es werden also nur sehr wenige Patientinnen als metastasenfrei klassifiziert. Auch bei Hinzunahme der klinischen Variablen gibt es keine großen Veränderungen. Die Spezifitätswerte für die Modelle mit den Top10- und Top20-Genen erhöhen sich ein wenig (auf 0.029 bzw. 0.051) und das Modell mit den 30 Top-Genen erreicht ebenfalls einen NPV-Wert von 1.000 mit Spezifität 0.007. Bei Verwendung von Down-Sampling erhält man deutlich andere Ergebnisse. Insgesamt haben die Werte, die zu einer bestimmte Genanzahl gehören, unabhängig von dem Gewicht einen ähnlichen NPV-Wert. Die größten Spezifitätswerte sind dabei für ein Gewicht von 2^2 zu beobachten. Den größten NPV-Wert hat das Modell mit den Top10-Genen und einem Gewicht von 2^{-18}. Auch hier ist bei zusätzlicher Verwendung der klinischen Variablen keine große Veränderung zu beobachten, die NPV-Werte der Modelle werden leicht größer. Allerdings erreichen die Modelle mit den 10 Top-Genen zum Teil NPV-Werte größer 90 %.

Die Ergebnisse der Random Forests auf dichotomisierten Daten unterscheiden sich deutlich von denen auf stetigen Daten (vgl. Abbildung 5.7 für das Likelihood Ratio und Kapitel E.1.2 für die übrigen Scores). Ohne Down-Sampling sind bei den Top-Genen des Likelihood Ratios für die kleinsten Klassengewichte die größten NPV-Werte zu beobachten. Für die anderen Gewichte ist die Streuung der NPV- und Spezifitätswerte sehr klein. Das Hinzunehmen von klinischen Parametern hat insgesamt keinen starken Einfluss auf den NPV-Wert. Allerdings ist die Spezifität der Modelle insgesamt etwas höher. Bei Verwendung von Down-Sampling werden die größten NPV-Werte ebenfalls bei den kleinsten Klassengewichten erreicht. Für die besten Modelle aus den 40-80 Genen mit dem größten Likelihood Ratio kann ein NPV-Wert von 1.000 beobachtet werden. Der zugehörige Spezifitätswert ist dabei jedoch sehr klein. Das bedeutet, es werden nur einzelne Patientinnen als metastasenfrei klassifiziert. Die Modelle mit großen Klassengewichten

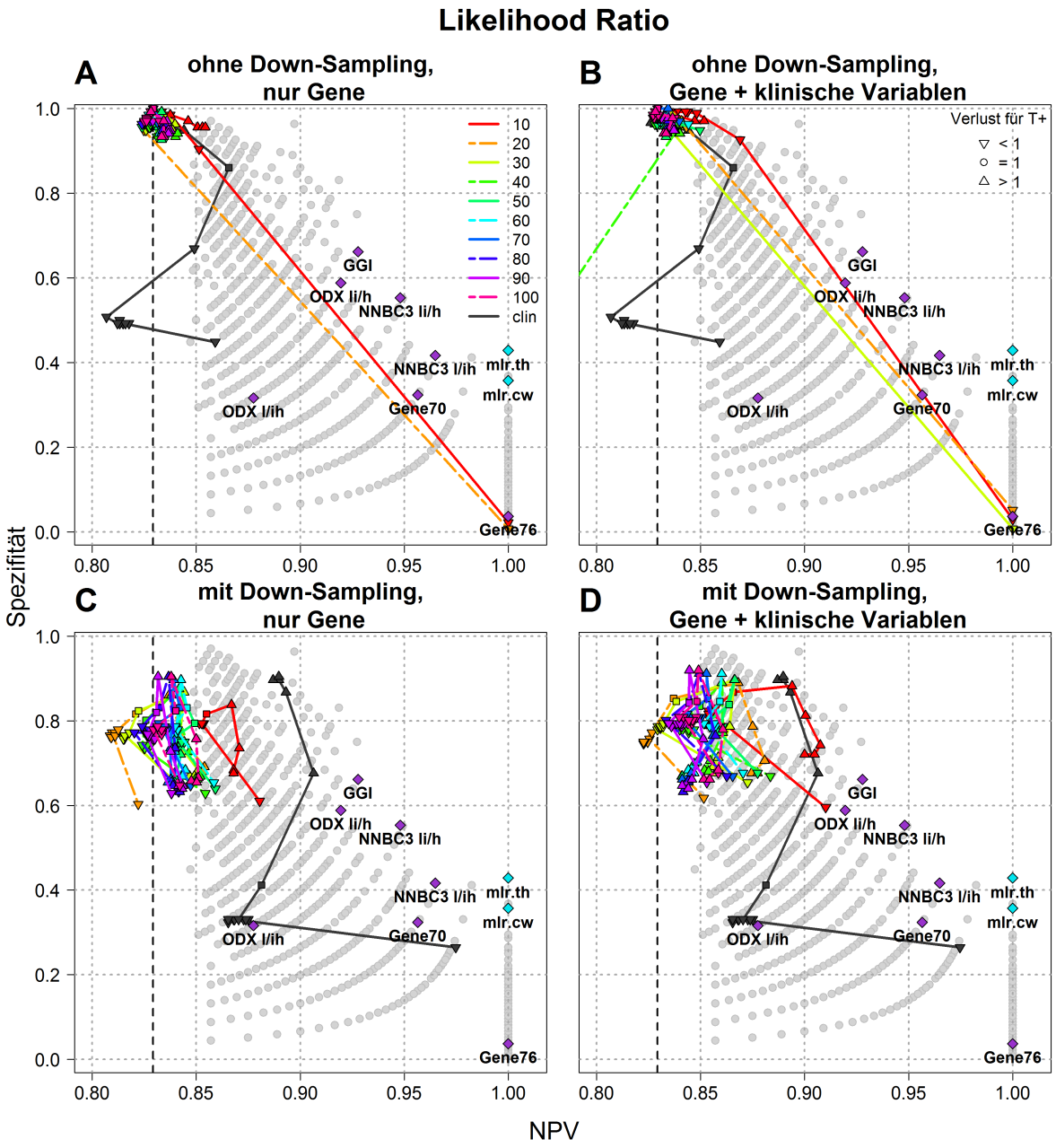

Abbildung 5.6: NPV und Spezifität der Random Forests mit frei wählbarem Cutoff für die Top-Gene von Likelihood Ratio. Links sind die Ergebnisse der Modelle nur mit Genen dargestellt, rechts die Ergebnisse der Modelle mit Genen und klinischen Variablen. Bei den Modellen in der oberen Reihe wurde kein Down-Sampling verwendet, bei den Modellen in der unteren Reihe wurde Down-Sampling verwendet.

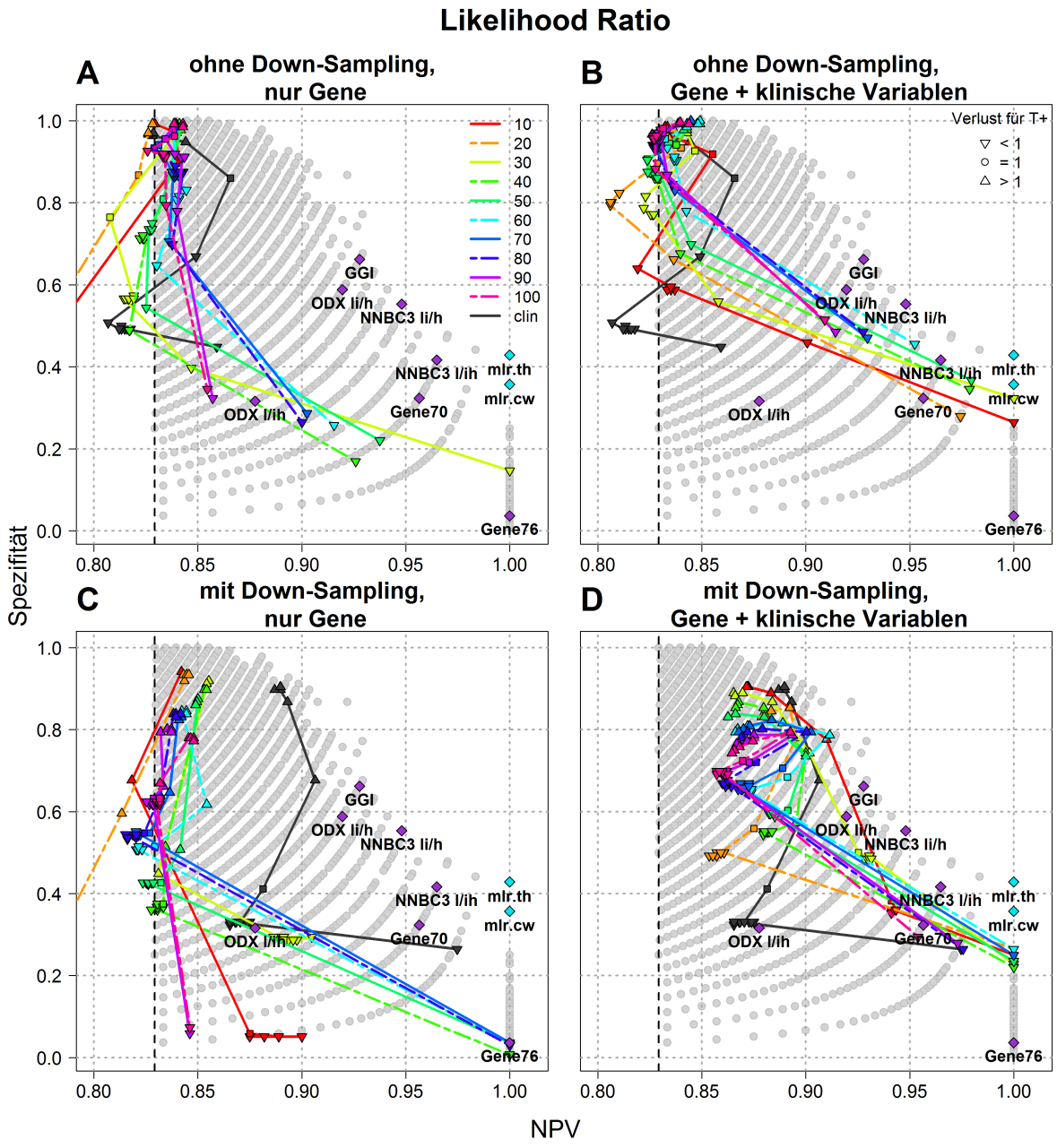

Abbildung 5.7: NPV und Spezifität der Random Forests mit fest vorgegebenem Cutoff für die Top-Gene von Likelihood Ratio. Links sind die Ergebnisse der Modelle nur mit Genen dargestellt, rechts die Ergebnisse der Modelle mit Genen und klinischen Variablen. Bei den Modellen in der oberen Reihe wurde kein Down-Sampling verwendet, bei den Modellen in der unteren Reihe wurde Down-Sampling verwendet.

haben überwiegend einen NPV-Wert nahe dem der trivialen Lösung, wobei die Spezifität bei Verwendung von kleineren Gewichten kleiner ist als bei größeren Gewichten. Bei Hinzunahme von klinischen Variablen erreichen die Modelle höhere Spezifitätswerte. Für ein sehr kleines Klassengewicht können auch hier NPV-Werte von 1.000 beobachtet werden. Die Spezifität liegt bei diesen Modellen bei etwa 25 %, es wird also etwa 1/4 der Patientinnen ohne Metastase von Random Forests richtig klassifiziert.

Wie in Kapitel 5.2 wurden pro Score die drei Modelle mit den größten NPV-Werten in jedem der acht Szenarien ausgewählt, um die Ergebnisse für die verschiedenen Bimodalitätsmaße zu vergleichen. In diesem Kapitel werden exemplarisch nur die beiden Tabellen für die Random Forests nur mit Genen und Genen und klinischen Variablen bei Verwendung von festen Cutoffs und Down-Sampling betrachtet (Tabelle 5.15 und 5.16). Die übrigen Tabellen sind im Anhang (Kapitel E.2) zu finden. Bei den Random Forests ohne klinische Variablen haben alle Modelle mit Ausnahme zweier Modelle aus den Top-Genen des Bimodality Index einen NPV-Wert größer 0.9 (Tabelle 5.15). Die Top3 Random Forests aus den Top-Genen von positiver Kurtosis und Likelihood Ratio, sowie jeweils ein Modell aus den Top-Genen von dip-Statistik und Outlier-Sum-Statistik, erreichen einen NPV-Wert von 1.000. Die Spezifitätswerte dieser Modelle sind allerdings auffällig klein (zwischen 0.007 und 0.044). Lediglich das Modell aus den Top-Genen der dip-Statistik hat einen höheren Spezifitätswert (0.228). Insgesamt unterscheiden sich die Ergebnisse für die verschiedenen Scores in Bezug auf NPV- und Spezifitätswerte.

In der Tabelle für die Modelle mit Genen und klinischen Variablen (Tabelle 5.16) kann für vier Bimodalitäts-Scores (positive Kurtosis, VRS, WVRS und Likelihood Ration) bei allen drei Top-Modellen ein NPV-Wert von 1.000 beobachtet werden. Auch der beste Random Forest für die Outlier-Sum-Statistik hat einen NPV-Wert von 1.000. Die zugehörige Spezifität liegt zwischen 0.199 und 0.287. Die Top3-Modelle aller Scores haben relativ hohe NPV-Werte, der kleinste NPV-Wert (0.938) wird für einen Random Forest aus den Top-Genen des Bimodality Index beobachtet. Die größte Spezifität haben die Random Forests aus den Top-Genen der dip-Statistik.

Insgesamt können bei den Random Forests bei zusätzlicher Verwendung von klinischen Variablen höhere NPV- und Spezifitätswerte erreicht werden.

Tabelle 5.15: Random Forests aus Top-Genen der Bimodalitäts-Scores mit den größten NPV-Werten bei fest vorgegebenen Cutoffs und Verwendung von Down-Sampling.

			NPV			Spezifität		
Score	Top	Gewicht	Median	IQR	Spannweite	Median	IQR	Spannweite
positive Kurtosis	10	2^{2}	**1.000**	0.000	0.200	**0.044**	0.015	0.184
	20	2^{-2}	**1.000**	0.000	1.000	**0.015**	0.007	0.029
	20	2^{-16}	**1.000**	0.000	1.000	**0.007**	0.007	0.029
Likelihood Ratio	50	2^{-18}	**1.000**	0.000	0.250	**0.037**	0.015	0.051
	70	2^{-18}	**1.000**	0.000	0.286	**0.037**	0.007	0.059
	60	2^{-18}	**1.000**	0.000	0.250	**0.033**	0.007	0.037
Dip	20	2^{-18}	**1.000**	0.000	0.065	**0.228**	0.029	0.074
	100	2^{-18}	**0.941**	0.019	0.076	**0.478**	0.029	0.103
	90	2^{-18}	**0.940**	0.024	0.070	**0.456**	0.029	0.118
negative Kurtosis	20	2^{-18}	**0.924**	0.033	0.071	**0.228**	0.022	0.096
	40	2^{-18}	**0.923**	0.014	0.073	**0.360**	0.024	0.118
	60	2^{-18}	**0.922**	0.007	0.058	**0.441**	0.022	0.081
Outlier Sum	30	2^{-18}	**1.000**	0.333	1.000	**0.022**	0.015	0.044
	80	2^{-18}	**0.923**	0.039	0.131	**0.213**	0.029	0.096
	90	2^{-18}	**0.903**	0.033	0.129	**0.250**	0.029	0.088
WVRS	90	2^{-12}	**0.923**	0.026	0.095	**0.276**	0.022	0.096
	90	2^{-6}	**0.921**	0.028	0.122	**0.279**	0.031	0.096
	90	2^{-8}	**0.918**	0.037	0.104	**0.279**	0.029	0.110
VRS	50	2^{-2}	**0.914**	0.032	0.089	**0.360**	0.029	0.118
	50	2^{-10}	**0.912**	0.034	0.110	**0.360**	0.024	0.110
	50	2^{-8}	**0.911**	0.025	0.096	**0.368**	0.029	0.110
Bimodality Index	100	2^{-18}	**0.922**	0.017	0.062	**0.368**	0.022	0.088
	80	2^{-18}	**0.892**	0.029	0.141	**0.235**	0.029	0.132
	90	2^{-18}	**0.889**	0.017	0.082	**0.294**	0.022	0.081

Tabelle 5.16: Random Forests aus Top-Genen der Bimodalitäts-Scores und klinischen Variablen mit den größten NPV-Werten bei fest vorgegebenen Cutoffs und Verwendung von Down-Sampling.

			NPV			Spezifität		
Score	Top	Gewicht	Median	IQR	Spannweite	Median	IQR	Spannweite
positive Kurtosis	20	2^{0}	**1.000**	0.023	0.106	**0.287**	0.015	0.051
	30	2^{0}	**1.000**	0.026	0.053	**0.272**	0.002	0.022
	10	2^{-16}	**1.000**	0.000	0.053	**0.265**	0.000	0.007
VRS	10	2^{-18}	**1.000**	0.000	0.030	**0.250**	0.007	0.059
	50	2^{-18}	**1.000**	0.033	0.108	**0.213**	0.015	0.074
	40	2^{-18}	**1.000**	0.000	0.074	**0.199**	0.029	0.081
WVRS	10	2^{-16}	**1.000**	0.000	0.000	**0.265**	0.007	0.037
	10	2^{-12}	**1.000**	0.000	0.000	**0.265**	0.007	0.044
	10	2^{-10}	**1.000**	0.000	0.029	**0.265**	0.007	0.029
Likelihood Ratio	20	2^{-18}	**1.000**	0.000	0.028	**0.265**	0.007	0.037
	60	2^{-18}	**1.000**	0.026	0.030	**0.265**	0.015	0.066
	10	2^{-18}	**1.000**	0.000	0.029	**0.250**	0.015	0.029
Outlier Sum	10	2^{-18}	**1.000**	0.000	0.054	**0.257**	0.015	0.051
	30	2^{-18}	**0.976**	0.026	0.071	**0.294**	0.015	0.066
	40	2^{-18}	**0.958**	0.024	0.093	**0.324**	0.029	0.096
Dip	20	2^{-18}	**0.967**	0.003	0.047	**0.434**	0.022	0.103
	30	2^{-18}	**0.957**	0.017	0.063	**0.342**	0.022	0.118
	90	2^{-18}	**0.950**	0.013	0.050	**0.566**	0.022	0.125
negative Kurtosis	10	2^{-18}	**0.972**	0.025	0.056	**0.272**	0.015	0.074
	20	2^{-18}	**0.953**	0.003	0.033	**0.449**	0.015	0.059
	30	2^{-18}	**0.942**	0.016	0.049	**0.419**	0.022	0.074
Bimodality Index	10	2^{-18}	**0.972**	0.024	0.079	**0.257**	0.007	0.051
	30	2^{-18}	**0.952**	0.041	0.114	**0.287**	0.022	0.081
	90	2^{-18}	**0.938**	0.018	0.054	**0.419**	0.022	0.081

5.3.1 Validierung der Random Forests

Die Ergebnisse der Random Forests aus Kapitel 5.3 sollen ebenfalls auf den zwei unabhängigen Kohorten von nodal-negativen unbehandelten Patientinnen validiert werden. Dazu wurden analog zu dem Vorgehen aus Kapitel 5.2.1 mit den Parametereinstellungen der Top3-Modelle der verschiedenen Bimodalitätsmaße Modelle auf der gesamten Mainz-Kohorte aufgestellt und aus diesen Modellen Vorhersagen für die Patientinnen der anderen beiden Kohorten erstellt. Ergebnisse dieser Analysen für die Tabellen aus Kapitel 5.3 sind in Tabelle 5.17 und 5.18 zusammengefasst. Die Ergebnisse für die übrigen Parameterkombinationen sind in Kapitel E.3 zu finden.

Die Top3 Random Forests der verschiedenen Scores nur mit Genen bei fest vorgegebenen Cutoffs und Verwendung von Down-Sampling Tabelle 5.17 erreichen in der Rotterdam-Kohorte sehr unterschiedliche NPV-Werte. Der höchste NPV-Wert wird für einen Random Forest aus den Top80-Genen des Bimodality Index beobachtet (0.889). Die übrigen Werte liegen deutlich darunter. Die Spezifitätswerte sind für einige Scores (positive Kurtosis, Likelihood Ratio, Outlier Sum und WVRS) sehr niedrig, es werden also nur sehr wenige der Patientinnen ohne Metastase als metastasenfrei klassifiziert. Auch in der Transbig-Kohorte kann man diese Beobachtung machen. Hier werden durch die Top3 Random Forests für Likelihood Ratio nur Patientinnen als metastasenfrei klassifiziert, die tatsächlich keine Metastase bekommen (NPV=1.000). Jedoch werden insgesamt nur sehr wenige Patientinnen in die Gruppe ohne Metastase klassifiziert (Spezifität 0.017, 0.022 und 0.033).

Die Random Forests mit Genen und klinischen Variablen haben in der Transbig-Kohorte überwiegend NPV-Werte größer 0.9. Die zugehörigen Spezifitätswerte liegen zwischen 0.164 und 0.333 und sind damit höher als bei den Modellen ohne klinische Variablen. Insgesamt zeigen die Random Forests mit vorausgewählten Genen auf der Rotterdam-Kohorte eine niedrige Klassifikationsgüte. Auf der Transbig-Kohorte sind die Ergebnisse in Bezug auf den NPV-Wert besser. Die Spezifitätswerte sind jedoch niedriger als die der bekannten Klassifikatoren.

Tabelle 5.17: Validierung der Top3 Random Forests nur mit Genen bei fest vorgegebenen Cutoffs und Verwendung von Down-Sampling.

			Rotterdam		Transbig	
Score	Top-Gene	Gewicht	NPV	Spezifität	NPV	Spezifität
positive Kurtosis	10	2^{2}	0.667	0.071	0.850	0.094
	20	2^{-2}	0.500	0.024	0.778	0.039
	20	2^{-16}	0.444	0.024	0.778	0.039
Likelihood Ratio	50	2^{-18}	0.286	0.012	1.000	0.022
	70	2^{-18}	0.600	0.036	1.000	0.039
	60	2^{-18}	0.500	0.024	1.000	0.033
Dip	20	2^{-18}	0.702	0.238	0.839	0.144
	100	2^{-18}	0.656	0.351	0.951	0.217
	90	2^{-18}	0.651	0.333	0.912	0.172
negative Kurtosis	20	2^{-18}	0.756	0.185	0.941	0.178
	40	2^{-18}	0.694	0.256	0.919	0.317
	60	2^{-18}	0.750	0.232	0.901	0.356
Outlier Sum	30	2^{-18}	0.000	0.000	0.750	0.017
	80	2^{-18}	0.480	0.071	0.886	0.172
	90	2^{-18}	0.536	0.089	0.875	0.194
WVRS	90	2^{-12}	0.667	0.071	0.933	0.078
	90	2^{-6}	0.609	0.083	0.958	0.128
	90	2^{-8}	0.606	0.119	0.939	0.172
VRS	50	2^{-2}	0.618	0.202	0.830	0.244
	50	2^{-10}	0.574	0.185	0.804	0.206
	50	2^{-8}	0.566	0.179	0.816	0.222
Bimodality Index	100	2^{-18}	0.667	0.143	0.952	0.222
	80	2^{-18}	0.889	0.048	0.933	0.078
	90	2^{-18}	0.760	0.113	0.880	0.122

Tabelle 5.18: Validierung der Top3 Random Forests mit Genen und klinischen Variablen bei fest vorgegebenen Cutoffs und Verwendung von Down-Sampling.

			Transbig	
Score	Top-Gene	Verlust	NPV	Spezifität
positive Kurtosis	20	2^{0}	0.909	0.242
	30	2^{0}	0.929	0.236
	10	2^{-16}	0.905	0.230
VRS	10	2^{-18}	0.925	0.224
	50	2^{-18}	0.971	0.200
	40	2^{-18}	0.947	0.218
WVRS	10	2^{-16}	0.925	0.224
	10	2^{-12}	0.927	0.230
	10	2^{-10}	0.914	0.194
Likelihood Ratio	20	2^{-18}	0.927	0.230
	60	2^{-18}	0.902	0.224
	10	2^{-18}	0.923	0.218
Outlier Sum	10	2^{-18}	0.929	0.236
	30	2^{-18}	0.936	0.267
	40	2^{-18}	0.933	0.255
Dip	20	2^{-18}	0.915	0.327
	30	2^{-18}	0.929	0.236
	90	2^{-18}	0.887	0.285
negative Kurtosis	10	2^{-18}	0.946	0.212
	20	2^{-18}	0.948	0.333
	30	2^{-18}	0.945	0.315
Bimodality Index	10	2^{-18}	0.923	0.218
	30	2^{-18}	0.896	0.261
	90	2^{-18}	0.911	0.248

5.4 Random Forests ohne vorausgewählte Gene

Um die Frage zu beantworten, ob man bessere Klassifikationsergebnisse erzielen kann, wenn man von dem Klassifikator weder verlangt, dass er bimodale Gene enthalten muss, noch dass er leicht interpretierbar sein muss, wurde das R-Paket `mlr` (Bischl et al. 2016) verwendet, um Random Forests (auf Basis des Paketes `ranger`) mit optimierten Parametern aufzustellen. Das Paket bietet ein Framework für Machine Learning in R. Es stellt eine einheitliche Schnittstelle für über 160 Lernverfahren bereit. Mit Hilfe des Paketes können Parameter eines Lernverfahrens mittels geschachtelter Kreuzvalidierung optimiert und die Vorhersagegüte beurteilt werden.

Bei der Klassifikation mit Genexpressionsdaten steht man vor dem Problem, dass es eine sehr große Anzahl von Parametern gibt, die zu einem großen Teil Rauschen enthalten. Vor der Klassifikation ist deswegen ein Filterschritt sinnvoll. Ein einfacher und häufig verwendeter Filter ist der Varianzfilter. Bei diesem werden für die Klassifikation nur die p Variablen mit der größten Varianz im Datensatz verwendet. Für die Analysen wurden die 2000 Gene mit der größten Varianz ausgewählt.

Random Forests können eine Wahrscheinlichkeit ausgeben, mit der eine Beobachtung zu einer der Klassen gehört. Der Threshold für die Klassifikation legt fest, ab welchem Wahrscheinlichkeitswert eine Beobachtung der Gruppe mit Metastasen zugeordnet werden soll. Die Standardeinstellung ist 0.5. Ein niedrigerer Threshold kann dafür sorgen, dass mehr Beobachtungen der Metastasen-Gruppe zugeordnet werden.

Mit dem Parameter *case.weights* können Gewichte für die Beobachtungen festgelegt werden. Beobachtungen mit höheren Gewichten werden mit höherer Wahrscheinlichkeit in die Bootstrap-Stichprobe aufgenommen. Im Fall von zwei Gruppen kann ein einzelner Wert angegeben werden, der dem Gewicht für eine Beobachtung der positiven Klasse entspricht. In unserem Fall gewichten wir die Gruppe mit Metastasen höher, da der Anteil der Patientinnen mit Metastase kleiner ist. Dies kann dazu beitragen das Problem der unterschiedlich großen Gruppen zu lösen. Dieses Vorgehen funktioniert ähnlich wie ein Up-Sampling der kleineren Gruppe.

Für die Klassifikation der Brustkrebspatientinnen wurden dann zwei Ansätze verfolgt:

1. Optimierung des Thresholds für Maximierung eines eigenen Performance-Maßes (I(NPV ≥ 0.95) + TNR).
2. Optimierung der *case.weights* für Maximierung des NPV-Wertes.

Es wurde auch versucht den Threshold direkt auf Maximierung des NPV-Wertes zu optimieren (Analysen in dieser Arbeit nicht dargestellt). Dieser Ansatz führte jedoch nur zu der trivialen Lösung. Daher wurde ein anderes Maß zur Optimierung verwendet. Es wurde jeweils eine geschachtelte Kreuzvalidierung mit einer äußeren 10-fachen Kreuzvalidierung und einer inneren 5-fachen Kreuzvalidierung durchgeführt. Um die Stabilität der Ergebnisse beurteilen zu können, wurden 100 Wiederholungen durchgeführt. Abbildung 5.8 zeigt Boxplots der getunten Parameter über die 100 Durchläufe hinweg. Horizontale Linien markieren den Median (durchgezogen) und das arithmetische Mittel (gestrichelt). Beim Parameter case.weight liegen Median (44) und arithmetisches Mittel (43.7) nah beieinander. Die Verteilung des Thresholds hingegen ist stark rechtsschief. Der Median liegt mit 0.106 deutlich unter dem arithmetischen Mittel (0.176). Beide Parameter weisen eine große Varianz auf. Die Ergebnisse der einzelnen Durchläufe sind sehr unterschiedlich (vgl. Abbildung G.5 auf Seite 225).

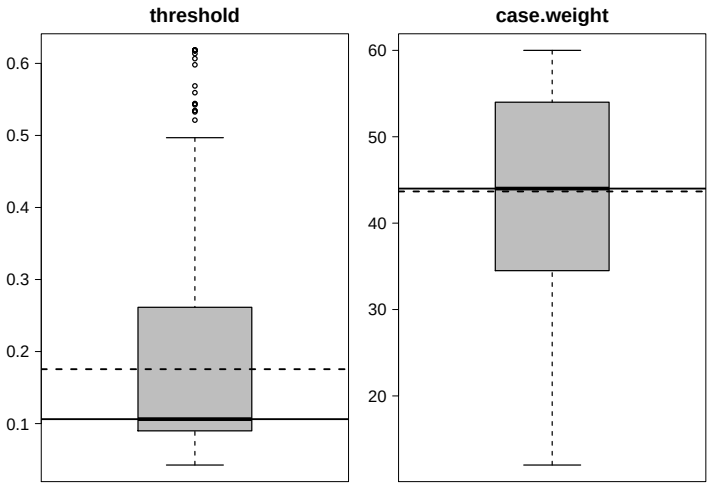

Abbildung 5.8: Verteilung der Werte der Tuning-Parameter in den 100 Durchläufen.

Mit den getunten Parametern wurden Modelle auf dem gesamten Trainingsdatensatz aufgestellt und aus diesen Modellen Vorhersagen für die Validierungskohorten erstellt (Ta-

belle 5.19). Das Modell mit dem getunten Threshold führt in beiden Validierungskohorten zu Vorhersagen mit NPV-Werten über 0.8 (Rotterdam 0.833, Transbig 0.842). Allerdings werden jeweils nur sehr wenige Patientinnen als metastasenfrei klassifiziert, was zu sehr niedrigen Spezifitätswerten führt (Rotterdam 0.060, Transbig 0.089). Dieses Ergebnis war insofern zu erwarten, als der getunte Threshold sehr niedrig ist, was bedeutet, dass eine Patientin mit hoher Wahrscheinlichkeit der Gruppe mit früher Metastase zugeordnet wird. Im Modell mit getunten Gewichten für die Beobachtungen ist der NPV-Wert für Rotterdam niedriger (0.789) und der zugehörige Wert der Spezifität 0.268. In der Transbig-Kohorte wird ein ganz guter NPV-Wert erreicht (0.891), wobei die Spezifität mit 0.456 im mittleren Bereich liegt. Etwas weniger als die Hälfte der Patientinnen ohne Metastase werden von dem Modell richtig zugeordnet. Verglichen mit den Ergebnissen der bekannten Klassifikatoren (Tabelle 5.1) sind die Ergebnisse aus diesen Modellen jedoch schlechter. Lediglich der Klassifikator Oncotype DX mit Zuordnung der Patientinnen mit mittlerem Risiko zu der Gruppe mit niedrigem Risiko liefert ähnliche Werte.

Tabelle 5.19: NPV und Spezifität der Random Forests in der Trainingskohorte (Mainz, kreuzvalidiert) und den Validierungskohorten (Rotterdam, Transbig).

getunter Parameter		Mainz	Rotterdam	Transbig
threshold = 0.106	NPV	1.000	0.833	0.842
	Spezifität	0.429	0.060	0.089
case.weight = 44	NPV	1.000	0.789	0.891
	Spezifität	0.357	0.268	0.456

5.5 Vergleich der Ergebnisse der verschiedenen Ansätze

Zum abschließenden Vergleich der Ergebnisse der verschiedenen Ansätze wurde für den Score Likelihood Ratio jeweils das Modell mit dem größten NPV-Wert ausgewählt. Die NPV- und Spezifitätswerte dieser Modelle auf den drei Kohorten sind in Tabelle 5.20 zusammengefasst. Zum Vergleich enthält die Tabelle die Ergebnisse der bekannten Klassifikatoren (aus Kapitel 5.1) und der Random Forests ohne vorausgewählte Gene aber mit Parameteroptimierung. Bei der Mainz-Kohorte entsprechen NPV und Spezifität bei

den Klassifikationsbäumen und den Random Forests dem Median über die entsprechenden Werte der 100 Durchläufe. Für die Rotterdam- und Transbig-Kohorte wurden NPV und Spezifität aus den Vorhersagen bestimmt, die aus einem Modell, das auf der gesamten Mainz-Kohorte gebildet wurde, erstellt wurden. Das Maß Likelihood Ratio wurde für den Vergleich ausgewählt, weil es beim Vergleich der Ergebnisse der unterschiedlichen Scores in Kapitel 5.2 und 5.3 in vielen Szenarien mit fest vorgegebenem Cutoff eine gute Klassifikationsleistung in Bezug auf NPV und Spezifität lieferte.

Bei den Klassifikationsbäumen haben die Modelle mit dichotomisierten Daten auf der Mainz-Kohorte größere NPV-Werte als die Modelle auf stetigen Daten. Dies gilt auch bei der Validierung auf der Transbig-Kohorte. Auf der Rotterdam-Kohorte ist die gegenteilige Beobachtung zu machen, allerdings sind beide NPV-Werte niedrig (0.657 mit stetigen Daten, 0.507 mit dichotomisierten Daten). Bei den Random Forests haben die Modelle mit festem Cutoff alle einen NPV-Wert von 1.000. Bei den Modellen mit freier Wahl des Cutoffs liefern nur die Modelle ohne Down-Sampling einen NPV-Wert von 1.000. Diese Modelle haben eine sehr niedrige Spezifität. Die NPV-Werte der Modelle mit frei wählbarem Cutoff und Down-Sampling sind deutlich niedriger (0.880 ohne klinische Variablen, 0.910 mit klinischen Variablen). Bei fest vorgegebenem Cutoff haben die Random Forests mit klinischen Variablen eine höhere Spezifität als die Random Forests ohne klinische Variablen. Für die Modelle mit vorgegebenem Cutoff sind auch auf der Transbig-Kohorte hohe NPV-Werte zu beobachten. Die Modelle ohne klinische Variablen haben dabei höhere NPV-Werte als die Modelle ohne klinische Variablen, dafür ist aber die Spezifität deutlich niedriger. Auf der Rotterdam-Kohorte erreichen alle Random Forests nur niedrige NPV-Werte. Für die Random Forests ohne vorausgewählte Gene werden auf der Mainz-Kohorte im Vergleich zu den Random Forests mit vorausgewählten Genen höhere Spezifitätswerte beobachtet. Gleichzeitig werden auch hier NPV-Werte von 1.000 erreicht. Die Klassifikationsgüte in der Rotterdam-Kohorte ist in Bezug auf den NPV-Wert bei beiden Modellen höher als bei allen Modellen mit vorausgewählten Genen. Die Klassifikationsgüte ist jedoch nicht so hoch wie bei den etablierten Klassifikatoren. Auf der Transbig-Kohorte sind die Ergebnisse schlechter als die der Random Forests mit vorausgewählten Genen und fest vorgegebenem Cutoff.

Die NPV- und Spezifitätswerte in der Mainz-Kohorte wurden zusätzlich in einem Streudiagramm gegeneinander abgetragen (Abbildung 5.9). Hier ist zu sehen, dass Modelle

Tabelle 5.20: Vergleich der Ergebnisse verschiedener Klassifikatoren. NPV und Spezifität (TNR) für die Ergebnisse der bekannten Klassifikatoren, der besten Baummodelle und Random Forests mit den Top-Genen von Likelihood Ratio und Random Forests mit getunten Parametern.

Klassifikator		Mainz		Rotterdam		Transbig	
		NPV	TNR	NPV	TNR	NPV	TNR
bekannte Klassifikatoren							
70-Gen		0.957	0.324	0.847	0.363	0.940	0.350
76-Gen		1.000	0.037	0.966	0.167	0.947	0.102
Oncotype DX	*low* vs. *intermediate/high*	0.878	0.316	0.833	0.208	0.961	0.272
	low/intermediate vs. *high*	0.920	0.588	0.752	0.452	0.893	0.511
GGI		0.928	0.662	-	-	0.909	0.611
NNBC-3	*low* vs. *intermediate/high*	0.965	0.417	-	-	-	-
	low/intermediate vs. *high*	0.948	0.553	-	-	-	-
Klassifikationsbäume							
Cutoff frei	Gene	0.893	0.665	0.657	0.685	0.801	0.694
Cutoff frei	Gene + klin. Variablen	0.910	0.662	-	-	0.774	0.533
Cutoff fest	Gene	0.957	0.331	0.507	0.202	0.849	0.250
Cutoff fest	Gene + klin. Variablen	0.948	0.610	-	-	0.864	0.422
Random Forests							
Cutoff frei	Gene	1.000	0.022	0.667	0.024	0.500	0.006
Cutoff frei	Gene + klin. Variablen	1.000	0.051	-	-	0.833	0.030
Cutoff frei, DS	Gene	0.880	0.610	0.641	0.542	0.774	0.683
Cutoff frei, DS	Gene + klin. Variablen	0.910	0.596	-	-	0.831	0.685
Cutoff fest	Gene	1.000	0.147	0.577	0.089	0.947	0.100
Cutoff fest	Gene + klin. Variablen	1.000	0.324	-	-	0.900	0.273
Cutoff fest, DS	Gene	1.000	0.037	0.286	0.012	1.000	0.022
Cutoff fest, DS	Gene + klin. Variablen	1.000	0.265	-	-	0.927	0.230
Random Forests mit getunten Parametern (mlr)							
	threshold = 0.106	1.000	0.429	0.833	0.060	0.842	0.089
	case.weight = 44	1.000	0.357	0.789	0.268	0.891	0.456

mit einem NPV-Wert von 1.000 nur niedrige bis mittlere Spezifität erreichen. Die besten Kombinationen von NPV und Spezifität haben der Genomic Grade Index, der Klassifikationsbaum mit dichotomisierten Expressionsdaten und klinischen Kovariablen, und der Random Forest mit optimiertem Threshold. Diese Punkte liegen auf einer Art Pareto-Front.

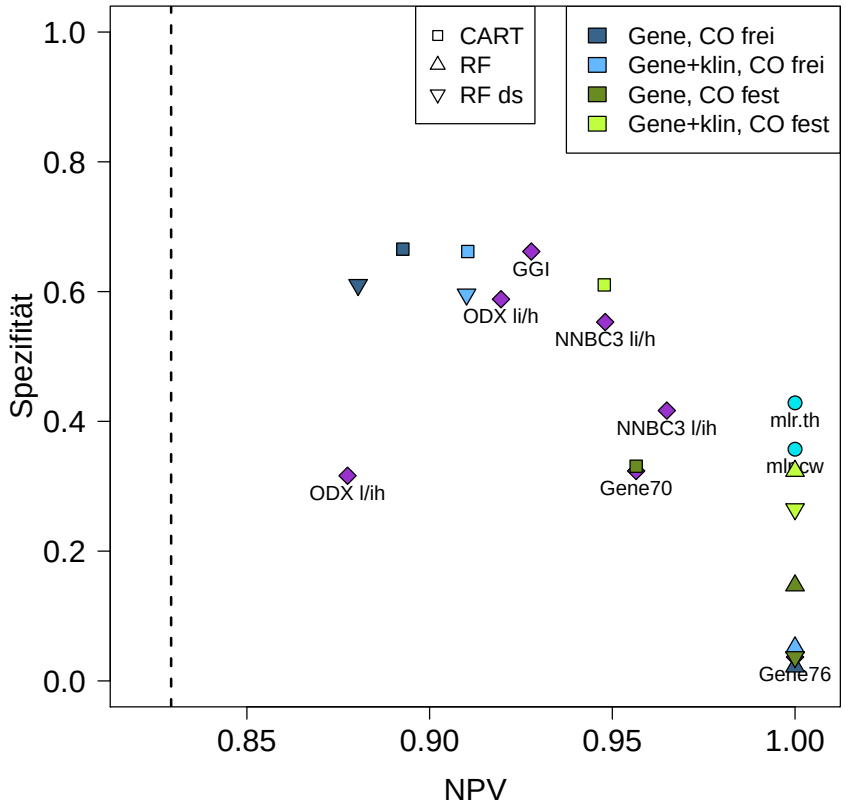

Abbildung 5.9: NPV und Spezifität für die Ergebnisse der bekannten Klassifikatoren, der besten Baummodelle und Random Forests mit den Top-Genen von Likelihood Ratio und Random Forests mit getunten Parametern.

Die Erkenntnisse sind mit Vorsicht zu bewerten, weil die Auswahl des besten Modells für die Top-Gene von Likelihood Ratio auf Basis des maximalen NPV-Wertes ziemlich willkürlich ist. Zum einen gibt es teilweise mehr als ein Modell mit dem gleichen NPV-Wert. Zum anderen kann es ein Modell geben, das eine bessere Kombination von NPV und Spezifität liefert oder sich besser auf den unabhängigen Datensätzen validieren lässt.

6 Zusammenfassung und Ausblick

Ziel dieser Arbeit war es mit Hilfe von Genexpressionsdaten Klassifikatoren für Brustkrebspatientinnen zu erstellen, mit denen vorhergesagt werden soll, ob eine Patientin in den ersten fünf Jahren nach der Operation eine Fernmetastase bekommt oder metastasenfrei bleibt. Diese Fragestellung ist für die klinische Praxis relevant, da etwa zwei Drittel der Patientinnen mit nodal-negativem Brustkrebs auch ohne eine Chemotherapie metastasenfrei bleiben. Diese Patientengruppe sicher zu identifizieren ist damit von großem Interesse, da man den Frauen die häufig starken Nebenwirkungen einer Chemotherapie ersparen könnte. Die Anforderungen an den Klassifikator sind, dass er eine hohe prognostische Güte besitzt und gleichzeitig gut interpretierbar ist. Die Idee war daher Gene mit Expressionsverteilungen zu identifizieren, die klar zwischen einer Gruppe mit niedriger und einer Gruppe mit hoher Expression unterscheiden, und diese dann zur Konstruktion von Klassifikatoren zu verwenden. Dabei boten sich Klassifikationsbäume an, da sie leicht zu interpretieren sind.

Im Kontext der Problemstellung ist es sinnvoll die Klassifikationsgüte nicht mit der Fehlklassifikationsrate, sondern mit dem negativ prädiktiven Wert (NPV) und der Spezifität zu bewerten. Ein guter Klassifikator im Sinne der Problemstellung hat einen hohen NPV-Wert und eine hohe Spezifität. Ein hoher NPV-Wert sagt aus, dass ein großer Anteil der Patientinnen, die der Klassifikator in die metastasenfreie Gruppe einordnet, auch wirklich metastasenfrei bleibt. Das ist von großer Bedeutung, weil es wichtig ist, dass Patientinnen, denen man sagt, dass sie keine Chemotherapie benötigen, diese auch wirklich nicht benötigen. Eine hohe Spezifität sagt aus, dass viele der Patientinnen, die metastasenfrei bleiben, auch vom Klassifikator der Gruppe ohne Metastase zugeordnet werden. Der NPV-Wert ist das wichtigere Bewertungskriterium für die klinische Praxis, jedoch ist auch eine hohe Spezifität wichtig für die Beurteilung der Relevanz des Klassifikators für den klinischen Einsatz.

In Kapitel 2 wurde der medizinische und biologische Hintergrund dieser Arbeit beschrieben. Dabei wurden insbesondere der biologische Hintergrund der Genexpressionsmessung

erklärt und die Affymetrix-Technologie zur Messung der Genexpression vorgestellt sowie verschiedene Methoden zur Vorverarbeitung dieser Daten beschrieben. Anschließend folgte eine ausführliche Beschreibung bekannter Gensignaturen, die in dieser Arbeit als Referenz zur Beurteilung der Klassifikationsgüte der neu-entwickelten Klassifikatoren dienen. Dies sind der 70-Gen-Klassifikator, der 76-Gen-Klassifikator, Oncotype DX und der Genomic Grade Index. Diese Gensignaturen bestehen zum Teil aus mehreren Dutzend Genen und sind biologisch schwer zu interpretieren. Aus statistischer Sicht gehen in die Konstruktion der Klassifikatoren zum Teil heuristische Entscheidungen ein. Die in der Arbeit verwendeten Gensignaturen wurden umfangreich auf unabhängigen Datensätzen und in klinischen Studien evaluiert. Kommerzielle Arrays, die auf den Signaturen basieren, werden in der klinischen Praxis eingesetzt.

Zur Identifikation von Genen mit charakteristischer Expressionsverteilung wurden in dieser Arbeit verschiedene Scores verwendet. Diese wurden in Kapitel 3 ausführlich beschrieben. Im Wesentlichen gibt es zwei Hauptansätze um Gene zu finden, deren Expressionsverteilungen zwei Gruppen definieren. Beim ersten Ansatz werden die Expressionswerte eines Gens mit Hilfe von Clusterverfahren in zwei Gruppen eingeteilt. Dann können Scores verwendet werden, die bewerten, wie gut die Aufteilung ist. Diese Scores können zum Beispiel auf der Zerlegung der Total Sum of Squares (wie bei VRS und WVRS), einem standardisierten Abstand zwischen den Mittelwerten (Bimodality Index) oder dem Likelihood Ratio basieren. Teschendorff et al. (2006) verwenden die Kurtosis zur Identifizierung von Genen mit bimodaler Expressionsverteilung. Ein alternativer Ansatz wird bei der Outlier-Sum-Statistik von Tibshirani und Hastie (2007) verwendet. Hier werden alle Expressionswerte, die größer als ein bestimmter Wert sind, als Ausreißer bewertet und die Patienten auf Basis dieses Cutoffs in zwei Gruppen eingeteilt. Der Wert der Outlier-Sum-Statistik berechnet sich als Summe über die Werte der Ausreißer und ist damit groß, wenn es viele Ausreißer gibt oder wenige Ausreißer mit großen Werten. Ein ganz anderer Ansatz ist der dip-Test auf Unimodalität von Hartigan und Hartigan (1985).

Die vorgestellten Bimodalitätsmaße wurden auf die Expressionsdaten einer Kohorte von 200 nodal-negativen Brustkrebspatientinnen angewendet. Die Korrelation der Werte der Scores wurde betrachtet, um die Frage zu beantworten, ob die Scores auf globaler Ebene die gleiche Art von Expressionsverteilung finden. Dabei stellte sich heraus, dass die Maße

Outlier Sum und Likelihood Ratio stark korreliert sind. Außerdem ist die Rangkorrelation zwischen WVRS und Kurtosis groß. Das Betrachten der Expressionsverteilungen der Top-Gene der verschiedenen Scores mit Hilfe von Histogrammen zeigte, dass die verschiedenen Bimodalitätsmaße sehr unterschiedliche Expressionsverteilungen finden. Dabei gibt es drei Haupttypen: klare bimodale Verteilungen (dip, negative Kurtosis, Bimodality Index), unimodale Verteilungen mit einem oder wenigen Ausreißern (VRS, WVRS, positive Kurtosis) und unimodale Verteilungen mit vielen Ausreißern mit großer Varianz (Outlier Sum, Likelihood Ratio). Bei der Untersuchung der prognostischen Relevanz der Top-Gene mit Hilfe des Log-Rank-Tests zeigte sich, dass sich nur in wenigen durch die Genexpressionsverteilungen generierten Subgruppen die metastasenfreie Überlebenszeit signifikant unterscheidet. Auf globaler Ebene konnte mit dem Kolmogorov-Smirnov-Test nachgewiesen werden, dass prognostische Gene unter den Top-Genen der Maße VRS, Outlier Sum, Likelihood Ratio und Bimodality Index überrepräsentiert sind. Dies stimmt im Wesentlichen damit überein, was in Hellwig et al. (2010) veröffentlicht wurde.

Kapitel 5 beschäftigt sich mit der Klassifikation der Brustkrebspatientinnen. Mit Hilfe der Klassifikatoren sollte vorhergesagt werden, welche Patientinnen in den ersten 5 Jahren nach der Operation eine Metastase bekommen und welche Patientinnen mindestens 5 Jahre beobachtet werden und keine Metastase bekommen. Dabei konnten 164 von 200 Patientinnen der Trainingskohorte verwendet werden. Als Referenz dienen die Ergebnisse der etablierten Klassifikatoren auf dem verwendeten Datenmaterial. Aus den Expressionsdaten wurden Klassifikatoren gebildet, wobei als erstes Klassifikationsbäume mit vorausgewählten Genen aufgestellt wurden. Im nächsten Schritt wurden statt Klassifikationsbäumen Random Forests mit vorausgewählten Genen zur Modellbildung verwendet. Dabei wurden verschiedene Parametereinstellungen untersucht, wobei insbesondere in der Verwendung der genetischen Variablen unterschieden wurde (stetig oder dichotomisiert). Außerdem wurden Modelle nur mit Genen und Modelle mit Genen und klinischen Parametern (Tumorgrad, pT-Stage, Alter, ER- und HER2-Status) aufgestellt. Als Schwierigkeit bei der Klassifikation stellte sich heraus, dass die Daten unbalanciert sind (28 Patientinnen mit Metastase, 136 ohne Metastase), wodurch die Lösung, alle Patientinnen als metastasenfrei zu klassifizieren, von den Klassifikatoren bevorzugt wurde, da sie insgesamt wenig Fehler erzeugt (NPV 0.829, Spezifität 1). Als möglicher Ausweg wurden unterschiedliche Kosten für die Fehlklassifikation verwendet und bei den Random Forests

wurde ein Down-Sampling durchgeführt. Die Ergebnisse der verschiedenen Szenarien für einen Score wurden dabei mit Hilfe von Streudiagrammen verglichen, in denen der NPV-Wert gegen die Spezifität abgetragen wird. Abschließend wurde das R-Paket `mlr` verwendet um Random Forests mit optimierten Parametern zu erzeugen, wobei auf eine Vorauswahl der Gene verzichtet wurde. Zur Validierung der Modelle wurden zwei unabhängige Kohorten (Rotterdam (n=286), Transbig (n=280)) von nodal-negativen unbehandelten Patientinnen verwendet. Da für die Rotterdam-Kohorte keine klinischen Variablen zur Verfügung standen, konnte für die Validierung von Klassifikatoren, die klinische Kovariablen enthalten, nur die Transbig-Kohorte verwendet werden.

Insgesamt konnten bei den Klassifikationsbäumen in allen Szenarien NPV-Werte von über 0.9 erreicht werden. Die größten NPV-Werte wurden für kleine Verlust-Werte für falsche Klassifikation einer Patientin in die Gruppe mit Metastase beobachtet. Es zeigte sich beim Vergleich der Ergebnisse, dass bei Verwendung von dichotomisierten Daten höhere NPV-Werte erreicht werden. Das Hinzunehmen der klinischen Variablen verbessert die Klassifikationsleistung in Bezug auf die Spezifität und bei einigen Maßen auch in Bezug auf den NPV-Wert. Zwei der besten Klassifikationsbäume wurden in Bezug auf die enthaltenen Gene betrachtet. Dabei stellte sich heraus, dass für einige der enthaltenen Gene in der Literatur bereits ein Zusammenhang mit Brustkrebs beschrieben wurde. Die Ergebnisse der Validierung der Top-Klassifikationsbäume war hingegen nicht zufriedenstellend. Auf der Rotterdam-Kohorte können sowohl mit freiem als auch mit fest vorgegebenem Cutoff nur kleine bis mittlere NPV-Werte erreicht werden. Auf der Transbig-Kohorte werden im Allgemeinen höhere Werte erzielt, allerdings liegen auch diese bis auf wenige Ausnahmen nicht in einem für die klinische Praxis tolerierbaren Bereich.

Bei den Random Forests mit vorausgewählten Genen werden die Ergebnisse wesentlich davon beeinflusst, ob stetige oder dichotomisierte Expressionsdaten verwendet werden. Bei frei wählbarem Cutoff werden nur einzelne Ergebnisse mit hohen NPV-Werten aber niedriger Spezifität beobachtet. Das Hinzunehmen von klinischen Variablen hat keinen großen Einfluss auf die Klassifikationsgüte. Bei Verwendung von fest vorgegebenen Cutoffs werden höhere NPV-Werte erreicht, einige Modelle haben einen NPV-Wert von 1.000, dafür aber niedrige Spezifitätswerte. Sehr hohe NPV-Werte sind fast ausschließlich nur für die kleinsten verwendeten Klassengewichte zu beobachten. Insgesamt erhalten wir die

beste Kombination aus NPV und Spezifität, wenn Gene mit festem Cutoff und zusätzlich klinische Kovariablen verwendet werden. Die Verwendung von Down-Sampling hat keinen großen Einfluss auf die Ergebnisse. Bei der Validierung werden auf der Rotterdam-Kohorte wie bei den Klassifikationsbäumen fast ausschließlich niedrige und mittlere NPV-Werte erreicht. In der Transbig-Kohorte sind die Ergebnisse besser. Der NPV-Wert liegt zum Teil bei über 0.9, allerdings ist die Spezifität auch bei Verwendung von klinischen Kovariablen nur im unteren Bereich; es wird höchstens ein Drittel der metastasenfreien Patientinnen vom Klassifikator erkannt.

Zwei Ansätze von Random Forests ohne vorausgewählte Gene, aber mit Parameteroptimierung, liefern Modelle mit einem NPV-Wert von 1.000 und Spezifität 0.429 und 0.357. Diese erreichen auf den Validierungskohorten jedoch auch nur NPV-Werte unter 0.9. Insgesamt gibt es bei den in der Arbeit aufgestellten Modellen kein Modell, dass sowohl einen hohen NPV-Wert, als auch eine hohe Spezifität erreicht. Die gleiche Beobachtung ist auch für die verwendeten bekannten Klassifikatoren zu machen. Die entwickelten Modelle schneiden insbesondere bei der Validierung auf der Rotterdam- und Transbig-Kohorte schlechter ab als die etablierten Klassifikatoren. Hohe NPV-Werte werden nur bei Verwendung von unterschiedlichen Fehlklassifikationskosten beobachtet. Ohne Verwendung von Verlusten oder Klassengewichten wählen die Klassifikatoren häufig die triviale Lösung, alle Patientinnen als metastasenfrei zu klassifizieren.

Bei der Interpretation der Ergebnisse dieser Arbeit ist zu beachten, dass bei den Analysen Hunderte von Modellen mit verschiedenen Parametereinstellungen aufgestellt wurden. Durch die Wahl der in Bezug auf die Klassifikationsgüte besten Modelle wird Optimismus eingeführt.

Bei weiteren Analysen sollte berücksichtigt werden, dass Brustkrebs eine sehr heterogene Erkrankung ist. Der Ansatz vieler Studien molekulare Subgruppen zu identifizieren, die eine unterschiedliche Prognose aufweisen, und dann unterschiedliche Modelle für die verschiedenen Subtypen zu bilden, könnte die Klassifikationsgüte verbessern. Insbesondere eine Stratifizierung nach ER- und HER2-Status bildet die wichtigsten Subtypen ab. In dieser Arbeit wurde keine Stratifizierung vorgenommen, da der Stichprobenumfang nicht ausreichend war. Eine Erhöhung des Stichprobenumfangs könnte durch Poolen mehrerer Datensätze erreicht werden, dabei steht man jedoch wieder vor dem Problem des Batch-

Effektes. Wie in Kapitel 4.4 für die Transbig-Kohorte gezeigt, kann es vorkommen, dass die Expressionswerte einiger Gene trotz der Verwendung von fRMA auf unterschiedlichen Leveln liegen. Eine Begründung dafür kann technisches Rauschen sein. Wird dieser Punkt nicht beachtet, erhält man bei der Bewertung der Expressionsverteilungen mit Bimodalitätsmaßen falsch-positive Ergebnisse. Ein Ausweg könnte das Standardisieren der Expressionswerte pro Datensatz sein. Allerdings kann durch eine Standardisierung auch ein tatsächlicher biologischer Unterschied in den Daten verloren gehen.

Die Grenze von 5 Jahren für das Auftreten einer Metastase wird in der klinischen Praxis häufig verwendet, wenn Aussagen über die Prognose der Patientin getroffen werden sollen. Aus der Sicht des Statistikers ist diese Grenze recht willkürlich und die Dichotomisierung der Zielvariablen führt zu einem Informationsverlust. Ein möglicher Ausweg könnte das Verwenden von Survival Trees oder Random Survival Forests sein.

Weiter sollte beachtet werden, dass insbesondere modellbasiertes Clustern für einige Gene nicht geeignet ist, da ihre Expressionsverteilungen eher unimodal sind. Durch das Clustern werden künstlich zwei Gruppen eingeführt, die biologisch keinen Sinn machen. Denkbar wäre den Analysen einen Filterschritt vorzulagern, bei denen Gene mit eher unimodaler Verteilung herausgefiltert werden. Einen solchen Ansatz verwenden zum Beispiel Teschendorff et al. (2006) bei ihrer PACK-Methode. Modellbasiertes Clustern wird mit BIC verwendet um die optimale Anzahl von Clustern zu bestimmen. Zusätzlich wird die Voraussetzung getroffen, dass der kleinere Cluster mindestens fünf Beobachtungen enthält.

Boulesteix et al. (2008) beschäftigen sich mit dem Problem des sogenannten *additional predictive value*. Klinische Kovariablen sind als Prädiktoren verfügbar und wurden umfangreich untersucht und validiert. Klassifikatoren, die auf Genexpressionsdaten basieren, sind in der Regel schwerer zu messen. Die Frage, welchen zusätzlichen Nutzen die genetischen Daten für die Vorhersage liefern, sollte untersucht werden.

Auch in Zukunft wird es Ziel der Forschung bleiben Genexpressionsdaten zu verwenden um Patienten in Bezug auf Therapiewahl und Prognose möglichst besser zu unterscheiden. Die bereits etablierten Gensignaturen helfen zwar bereits heute bei der Therapieentscheidung, jedoch können sie in Bezug auf die Spezifität noch verbessert werden.

Literaturverzeichnis

Aalen, O. (1978). Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics* 6.4, S. 701–726.

Affymetrix (2002). *Statistical Algorithms Description Document*. Technischer Bericht. URL: `http://media.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf`.

Affymetrix (2005). *Guide to Probe Logarithmic Intensity Error (PLIER) Estimation*. Technischer Bericht. URL: `http://media.affymetrix.com/support/technical/technotes/plier_technote.pdf`.

Albain, K. S., Barlow, W. E., Shak, S., Hortobagyi, G. N., Livingston, R. B., Yeh, I.-T., Ravdin, P., Bugarini, R., Baehner, F. L., Davidson, N. E., Sledge, G. W., Winer, E. P., Hudis, C., Ingle, J. N., Perez, E. A., Pritchard, K. I., Shepherd, L., Gralow, J. R., Yoshizawa, C., Allred, D. C., Osborne, C. K., Hayes, D. F. und Breast Cancer Intergroup of North America (2010). Prognostic and Predictive Value of the 21-Gene Recurrence Score Assay in Postmenopausal Women with Node-Positive, Oestrogen-Receptor-Positive Breast Cancer on Chemotherapy: A Retrospective Analysis of a Randomised Trial. *Lancet Oncol* 11.1, S. 55–65.

Benjamini, Y. und Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, S. 289–300.

Bessarabova, M., Kirillov, E., Shi, W., Bugrim, A., Nikolsky, Y. und Nikolskaya, T. (2010). Bimodal Gene Expression Patterns in Breast Cancer. *BMC Genomics* 11 Suppl 1, S8. DOI: `10.1186/1471-2164-11-S1-S8`.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G. und Jones, Z. M. (2016). mlr: Machine Learning in R. *J. Mach. Learn. Res.* 17.1, S. 5938–5942.

Boes, T. (2007). Auswirkungen der Low-Level-Analyse auf die Ergebnisse von Genexpressionsdaten der Firma Affymetrix. Dissertation. Universität Dusiburg-Essen.

Bolstad, B. M., Irizarry, R. A., Åstrand, M. und Speed, T. P. (2003). A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics* 19.2, S. 185–193. DOI: 10.1093/bioinformatics/19.2.185.

Boulesteix, A.-L., Porzelius, C. und Daumer, M. (2008). Microarray-Based Classification and Clinical Predictors: On Combined Classifiers and Additional Predictive Value. *Bioinformatics* 24.15, S. 1698–1706. DOI: 10.1093/bioinformatics/btn262.

Breiman, L. (2001). Random Forests. *Machine Learning* 45.1, S. 5–32. DOI: 10.1023/A:1010933404324.

Breiman, L., Friedman, J., Stone, C. J. und Olshen, R. A. (1984). *Classification and Regression Trees*. Taylor & Francis.

Buyse, M., Loi, S., van't Veer, L., Viale, G., Delorenzi, M., Glas, A. M., d'Assignies, M. S., Bergh, J., Lidereau, R., Ellis, P., Harris, A., Bogaerts, J., Therasse, P., Floore, A., Amakrane, M., Piette, F., Rutgers, E., Sotiriou, C., Cardoso, F., Piccart, M. J. und TRANSBIG Consortium (2006). Validation and Clinical Utility of a 70-Gene Prognostic Signature for Women with Node-Negative Breast Cancer. *Journal of the National Cancer Institute* 98.17, S. 1183–1192. DOI: 10.1093/jnci/djj329.

Buyse, M., Michiels, S., Sargent, D. J., Grothey, A., Matheson, A. und de Gramont, A. (2011). Integrating Biomarkers in Clinical Trials. *Expert Review of Molecular Diagnostics* 11.2, S. 171–182. DOI: 10.1586/erm.10.120.

Cardoso, F., Piccart-Gebhart, M., Veer, L. und Rutgers, E. (2007). The MINDACT Trial: The First Prospective Clinical Validation of a Genomic Tool. *Molecular Oncology* 1.3, S. 246–251. DOI: 10.1016/j.molonc.2007.10.004.

Cardoso, F., van't Veer, L. J., Bogaerts, J., Slaets, L., Viale, G., Delaloge, S., Pierga, J.-Y., Brain, E., Causeret, S., DeLorenzi, M., Glas, A. M., Golfinopoulos, V., Goulioti, T., Knox, S., Matos, E., Meulemans, B., Neijenhuis, P. A., Nitz, U., Passalacqua, R., Ravdin, P., Rubio, I. T., Saghatchian, M., Smilde, T. J., Sotiriou, C., Stork, L., Straehle, C., Thomas, G., Thompson, A. M., van der Hoeven, J. M., Vuylsteke, P., Bernards, R., Tryfonidis, K., Rutgers, E. und Piccart, M. (2016). 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. *N Engl J Med* 375.8, S. 717–729. DOI: 10.1056/NEJMoa1602253.

Carlson, M. (2016). *hgu133a.db: Affymetrix Human Genome U133 Set Annotation Data (Chip hgu133a)*.

Chen, C., Liaw, A. und Breiman, L. (2004). *Using Random Forest to Learn Imbalanced Data*. Technischer Bericht. University of California at Berkeley, Berkeley, California: Statistics Department. URL: http://stat-reports.lib.berkeley.edu/accessPages/666.html.

Chen, L.-C., Manjeshwar, S., Lu, Y., Moore, D., Ljung, B.-M., Kuo, W.-L., Dairkee, S. H., Wernick, M., Collins, C. und Smith, H. S. (1998). The Human Homologue for the *Caenorhabditis Elegans Cul-4* Gene Is Amplified and Overexpressed in Primary Breast Cancers. *Cancer Research* 58.16, S. 3677–3683.

Conover, W. J. (1999). *Practical Nonparametric Statistics*. 3. Auflage. Wiley series in probability and statistics : Applied probability and statistics section. John Wiley, New York.

Deisenroth, C., Thorner, A. R., Enomoto, T., Perou, C. M. und Zhang, Y. (2010). Mitochondrial Hep27 Is a C-Myb Target Gene That Inhibits Mdm2 and Stabilizes P53. *Molecular and Cellular Biology* 30.16, S. 3981–3993. DOI: 10.1128/MCB.01284-09.

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., D'Assignies, M. S., Bergh, J., Lidereau, R., Ellis, P., Harris, A. L., Klijn, J. G. M., Foekens, J. A., Cardoso, F., Piccart, M. J., Buyse, M. und Sotiriou, C. (2007). Strong Time Dependence of the 76-Gene Prognostic Signature for Node-Negative Breast Cancer Patients in the TRANSBIG Multicenter Independent Validation Series. *Clin Cancer Res* 13.11, S. 3207–3214.

Dowsett, M., Cuzick, J., Wale, C., Forbes, J., Mallon, E. A., Salter, J., Quinn, E., Dunbier, A., Baum, M., Buzdar, A., Howell, A., Bugarini, R., Baehner, F. L. und Shak, S. (2010). Prediction of Risk of Distant Recurrence Using the 21-Gene Recurrence Score in Node-Negative and Node-Positive Postmenopausal Patients with Breast Cancer Treated with Anastrozole or Tamoxifen: A TransATAC Study. *J Clin Oncol* 28.11, S. 1829–1834. DOI: 10.1200/JCO.2009.24.4798.

Edgar, R., Domrachev, M. und Lash, A. E. (2002). Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository. *Nucleic Acids Research* 30.1, S. 207–210. DOI: 10.1093/nar/30.1.207.

Efron, B. (1967). The Two Sample Problem with Censored Data. *Proceedings of the fifth Berkeley symposium on*, S. 831–853.

Elston, C. W. und Ellis, I. O. (1991). Pathological Prognostic Factors in Breast Cancer. I. The Value of Histological Grade in Breast Cancer: Experience from a Large Study

with Long-Term Follow-Up. *Histopathology* 19.5, S. 403–410. DOI: 10.1111/j.1365-2559.1991.tb00229.x.

Ertel, A. und Tozeren, A. (2008). Switch-like Genes Populate Cell Communication Pathways and Are Enriched for Extracellular Proteins. *BMC Genomics* 9, S. 3. DOI: 10.1186/1471-2164-9-3.

Fernández-Delgado, M., Cernadas, E., Barro, S. und Amorim, D. (2014). Do We Need Hundreds of Classifiers to Solve Real World Classification Problems? *J. Mach. Learn. Res.* 15.1, S. 3133–3181.

Field, L. A., Love, B., Deyarmin, B., Hooke, J. A., Shriver, C. D. und Ellsworth, R. E. (2012). Identification of Differentially Expressed Genes in Breast Tumors from African American Compared with Caucasian Women. *Cancer* 118.5, S. 1334–1344. DOI: 10.1002/cncr.26405.

Filipits, M., Rudas, M., Jakesz, R., Dubsky, P., Fitzal, F., Singer, C. F., Dietze, O., Greil, R., Jelen, A., Sevelda, P., Freibauer, C., Müller, V., Jänicke, F., Schmidt, M., Kölbl, H., Rody, A., Kaufmann, M., Schroth, W., Brauch, H., Schwab, M., Fritz, P., Weber, K. E., Feder, I. S., Hennig, G., Kronenwett, R., Gehrmann, M. und Gnant, M. (2011). A New Molecular Predictor of Distant Recurrence in ER-Positive, HER2-Negative Breast Cancer Adds Independent Information to Conventional Clinical Risk Factors. *Clinical Cancer Research* 17.18, S. 6012–6020. DOI: 10.1158/1078-0432.CCR-11-0926.

Filipits, M., Nielsen, T. O., Rudas, M., Greil, R., Stöger, H., Jakesz, R., Bago-Horvath, Z., Dietze, O., Regitnig, P., Gruber-Rossipal, C., Müller-Holzner, E., Singer, C. F., Mlineritsch, B., Dubsky, P., Bauernhofer, T., Hubalek, M., Knauer, M., Trapl, H., Fesl, C., Schaper, C., Ferree, S., Liu, S., Cowens, J. W., Gnant, M., Group, f. t.A. B. und Study, C. C. (2014). The PAM50 Risk-of-Recurrence Score Predicts Risk for Late Distant Recurrence after Endocrine Therapy in Postmenopausal Women with Endocrine-Responsive Early Breast Cancer. *Clinical Cancer Research* 20.5, S. 1298–1305. DOI: 10.1158/1078-0432.CCR-13-1845.

Foekens, J. A., Atkins, D., Zhang, Y., Sweep, F. C.G. J., Harbeck, N., Paradiso, A., Cufer, T., Sieuwerts, A. M., Talantov, D., Span, P. N., Tjan-Heijnen, V. C. G., Zito, A. F., Specht, K., Hoefler, H., Golouh, R., Schittulli, F., Schmitt, M., Beex, L. V.A. M., Klijn, J. G. M. und Wang, Y. (2006). Multicenter Validation of a Gene Expression-Based Prognostic Signature in Lymph Node-Negative Primary Breast Cancer. *J Clin Oncol* 24.11, S. 1665–1671. DOI: 10.1200/JCO.2005.03.9115.

Foulkes, A. S. (2009). *Applied Statistical Genetics with R*. 1. Auflage. Springer, Heidelberg, London, New York.

Fraley, C. und Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis and Density Estimation. *Journal of the American Statistical Association* 97, S. 611–631.

Fraley, C. und Raftery, A. E. (2006). *Mclust Version 3 for R: Normal Mixture Modeling and Model-Based Clustering*. Technical Report. Department of Statistics: University of Washington.

Gendoo, D. M. A., Ratanasirigulchai, N., Schroder, M. S., Pare, L., Parker, J. S., Prat, A. und Haibe-Kains, B. (2015). *Genefu: Computation of Gene Expression-Based Signatures in Breast Cancer*. R package version 2.4.2.

Gill, R. D. (1980). Censoring and Stochastic Integrals. *Mathematical Centre, Amsterdam (MC Tract 124)*.

Glas, A. M., Floore, A., Delahaye, L. J.M. J., Witteveen, A. T., Pover, R. C. F., Bakx, N., Lahti-Domenici, J. S. T., Bruinsma, T. J., Warmoes, M. O., Bernards, R., Wessels, L. F. A. und Van't Veer, L. J. (2006). Converting a Breast Cancer Microarray Signature into a High-Throughput Diagnostic Test. *BMC Genomics* 7, S. 278. DOI: 10.1186/1471-2164-7-278.

Göhlmann, H. und Talloen, W. (2009). *Gene Expression Studies Using Affymetrix Microarrays*. Mathematical and Computational Biology. Chapman & Hall/CRC.

Habel, L. A., Shak, S., Jacobs, M. K., Capra, A., Alexander, C., Pho, M., Baker, J., Walker, M., Watson, D., Hackett, J., Blick, N. T., Greenberg, D., Fehrenbacher, L., Langholz, B. und Quesenberry, C. P. (2006). A Population-Based Study of Tumor Gene Expression and Risk of Breast Cancer Death among Lymph Node-Negative Patients. *Breast Cancer Res* 8.3, R25. DOI: 10.1186/bcr1412.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J. und Sotiriou, C. (2012). A Three-Gene Model to Robustly Identify Breast Cancer Molecular Subtypes. *J Natl Cancer Inst* 104.4, S. 311–325. DOI: 10.1093/jnci/djr545.

Hartigan, J. A. und Hartigan, P. M. (1985). The Dip Test of Unimodality. *The Annals of Statistics* 113.1, S. 70–84.

Hartigan, P. M. (1985). Algorithm AS 217: Computation of the Dip Statistic to Test for Unimodality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 34.3, S. 320–325. DOI: 10.2307/2347485.

Hastie, T. J., Tibshirani, R. J. und Friedman, J. H. (2009). *The Elements of Statistical Learning The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. Auflage. Springer series in statistics. Springer, New York.

Hedges, L. V. und Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press.

Hellwig, B., Hengstler, J. G., Schmidt, M., Gehrmann, M. C., Schormann, W. und Rahnenführer, J. (2010). Comparison of Scores for Bimodality of Gene Expression Distributions and Genome-Wide Evaluation of the Prognostic Relevance of High-Scoring Genes. *BMC Bioinformatics* 11.1, S. 276.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, S. 65–70.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., Pagès, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L. und Morgan, M. (2015). Orchestrating High-Throughput Genomic Analysis with Bioconductor. *Nature Methods* 12.2, S. 115–121. DOI: 10.1038/nmeth.3252.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. und Speed, T. P. (2003). Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics* 4.2, S. 249–264. DOI: 10.1093/biostatistics/4.2.249.

Kaplan, E. L. und Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* 53.282, S. 457–481.

Kellerer, A. M. und Chmelevsky, D. (1983). Small-Sample Properties of Censored-Data Rank Tests. *Biometrics* 39.3, S. 675–682.

Klein, J. P. (1991). Small Sample Moments of Some Estimators of the Variance of the Kaplan-Meier and Nelson-Aalen Estimators. *Scandinavian Journal of Statistics* 18.4, S. 333–340.

Klein, J. P. und Moeschberger, M. L. (2005). *Survival Analysis*. 2. Auflage. Springer.

Knapp, T. R. (2007). Bimodality Revisited. *Journal of Modern Applied Statistical Methods* 6.1. DOI: 10.22237/jmasm/1177992120.

Korn, E. L., Troendle, J. F., McShane, L. M. und Simon, R. (2004). Controlling the Number of False Discoveries: Application to High-Dimensional Genomic Data. *Journal*

of Statistical Planning and Inference 124.2, S. 379 –398. DOI: http://dx.doi.org/10.1016/S0378-3758(03)00211-8.

Latta, R. B. (1981). A Monte Carlo Study of Some Two-Sample Rank Tests with Censored Data. *Journal of the American Statistical Association* 76.375, S. 713–719.

Li, C. und Hung Wong, W. H. (2001). Model-Based Analysis of Oligonucleotide Arrays: Model Validation, Design Issues and Standard Error Application. *Genome Biology* 2, research0032. DOI: 10.1186/gb-2001-2-8-research0032.

Liaw, A. und Wiener, M. (2002). Classification and Regression by randomForest. *R News* 2.3, S. 18–22.

Liedtke, C., Hatzis, C., Symmans, W. F., Desmedt, C., Haibe-Kains, B., Valero, V., Kuerer, H., Hortobagyi, G. N., Piccart-Gebhart, M., Sotiriou, C. und Pusztai, L. (2009). Genomic Grade Index Is Associated With Response to Chemotherapy in Patients With Breast Cancer. *Journal of Clinical Oncology* 27.19, S. 3185–3191. DOI: 10.1200/JCO.2008.18.5934.

Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R. und Lockhart, D. J. (1999). High Density Synthetic Oligonucleotide Arrays. *Nature Genetics* 21, S. 20–24. DOI: 10.1038/4447.

Loi, S., Haibe-Kains, B., Desmedt, C., Lallemand, F., Tutt, A. M., Gillet, C., Ellis, P., Harris, A., Bergh, J., Foekens, J. A., Klijn, J. G. M., Larsimont, D., Buyse, M., Bontempi, G., Delorenzi, M., Piccart, M. J. und Sotiriou, C. (2007). Definition of Clinically Distinct Molecular Subtypes in Estrogen Receptor-Positive Breast Carcinomas through Genomic Grade. *J Clin Oncol* 25.10, S. 1239–1246. DOI: 10.1200/JCO.2006.07.1522.

Ma, X.-J., Salunga, R., Dahiya, S., Wang, W., Carney, E., Durbecq, V., Harris, A., Goss, P., Sotiriou, C., Erlander, M. und Sgroi, D. (2008). A Five-Gene Molecular Grade Index and *HOXB13:IL17BR* Are Complementary Prognostic Factors in Early Stage Breast Cancer. *Clinical Cancer Research* 14.9, S. 2601. DOI: 10.1158/1078-0432.CCR-07-5026.

MacQueen, J. B. (1967). Some Methods of Classification and Analysis of Multivariate Observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, S. 281–297.

Mächler, M. (2009). *Diptest: Hartigan's Dip Test Statistic for Unimodality - Corrected Code*.

Mantel, N. (1966). Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration. *Cancer Chemother Rep* 50.3, S. 163–170.

McCall, M. N., Bolstad, B. M. und Irizarry, R. A. (2010). Frozen Robust Multiarray Analysis (fRMA). *Biostatistics* 11.2, S. 242–253. DOI: 10.1093/biostatistics/kxp059.

Mitra, R., Guo, Z., Milani, M., Mesaros, C., Rodriguez, M., Nguyen, J., Luo, X., Clarke, D., Lamba, J., Schuetz, E., Donner, D. B., Puli, N., Falck, J. R., Capdevila, J., Gupta, K., Blair, I. A. und Potter, D. A. (2011). CYP3A4 Mediates Growth of Estrogen Receptor-Positive Breast Cancer Cells in Part by Inducing Nuclear Translocation of Phospho-Stat3 through Biosynthesis of (±)-14,15-Epoxyeicosatrienoic Acid (EET). *The Journal of Biological Chemistry* 286.20, S. 17543–17559. DOI: 10.1074/jbc.M110.198515.

Nelson, W. (1972). Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics* 14.4, S. 945–966.

Nielsen, T. O., Parker, J. S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., Davies, S. R., Snider, J., Stijleman, I. J., Reed, J., Cheang, M. C. U., Mardis, E. R., Perou, C. M., Bernard, P. S. und Ellis, M. J. (2010). A Comparison of PAM50 Intrinsic Subtyping with Immunohistochemistry and Clinical Prognostic Factors in Tamoxifen-Treated Estrogen Receptor–Positive Breast Cancer. *Clinical Cancer Research* 16.21, S. 5222–5232. DOI: 10.1158/1078-0432.CCR-10-1282.

Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F. L., Walker, M. G., Watson, D., Park, T., Hiller, W., Fisher, E. R., Wickerham, D. L., Bryant, J. und Wolmark, N. (2004). A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *N Engl J Med* 351.27, S. 2817–2826. DOI: 10.1056/NEJMoa041588.

Paik, S., Tang, G., Shak, S., Kim, C., Baker, J., Kim, W., Cronin, M., Baehner, F. L., Watson, D., Bryant, J., Costantino, J. P., Geyer Jr, C. E., Wickerham, D. L. und Wolmark, N. (2006). Gene Expression and Benefit of Chemotherapy in Women with Node-Negative, Estrogen Receptor-Positive Breast Cancer. *J Clin Oncol* 24.23, S. 3726–3734. DOI: 10.1200/JCO.2005.04.7985.

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J. F., Stijleman, I. J., Palazzo, J., Marron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M. und Bernard, P. S. (2009). Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology* 27.8, S. 1160–1167. DOI: 10.1200/JCO.2008.18.1370.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, Ø., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A.-L., Brown, P. O. und Botstein, D. (2000). Molecular Portraits of Human Breast Tumours. *Nature* 406.6797, S. 747–752. DOI: 10.1038/35021093.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ravdin, P. M., Siminoff, L. A., Davis, G. J., Mercer, M. B., Hewlett, J., Gerson, N. und Parker, H. L. (2001). Computer Program to Assist in Making Decisions About Adjuvant Therapy for Women With Early Breast Cancer. *Journal of Clinical Oncology* 19.4, S. 980–991. DOI: 10.1200/JCO.2001.19.4.980.

Robert Koch-Institut (RKI) und Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V., Hrsg. (2015). *Krebs in Deutschland 2011/2012*. 8. Auflage.

SAS Institute Inc (2008). *SAS/STAT 9.2 User's Guide*. 2. Auflage. SAS Institute Inc, Cary, NC. URL: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm.

Schindl, M., Gnant, M., Schoppmann, S. F., Horvat, R. und Birner, P. (2007). Overexpression of the Human Homologue for Caenorhabditis Elegans Cul-4 Gene Is Associated with Poor Outcome in Node-Negative Breast Cancer. *Anticancer Research* 27.2, S. 949–952.

Schmidt, M., Victor, A., Bratzel, D., Boehm, D., Cotarelo, C., Lebrecht, A., Siggelkow, W., Hengstler, J. G., Elsäßer, A., Gehrmann, M., Lehr, H.-A., Koelbl, H., Minckwitz, G. von, Harbeck, N. und Thomssen, C. (2009). Long-Term Outcome Prediction by Clinicopathological Risk Classification Algorithms in Node-Negative Breast Cancer—comparison between Adjuvant!, St Gallen, and a Novel Risk Algorithm Used in the Prospective Randomized Node-Negative-Breast Cancer-3 (NNBC-3) Trial. *Annals of Oncology* 20.2, S. 258–264. DOI: 10.1093/annonc/mdn590.

Schmidt, M., Böhm, D., von Törne, C., Steiner, E., Puhl, A., Pilch, H., Lehr, H.-A., Hengstler, J. G., Kölbl, H. und Gehrmann, M. (2008). The Humoral Immune System Has a Key Prognostic Impact in Node-Negative Breast Cancer. *Cancer Res* 68.13, S. 5405–5413. DOI: 10.1158/0008-5472.CAN-07-5206.

Schulze, A. und Downward, J. (2001). Navigating Gene Expression Using Microarrays — a Technology Review. *Nature Cell Biology* 3.8, E190–E195. DOI: 10.1038/35087138.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics* 6. DOI: 10.1214/aos/1176344136.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E. und Børresen-Dale, A.-L. (2001). Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications. *Proceedings of the National Academy of Sciences of the United States of America* 98.19, S. 10869–10874. DOI: 10.1073/pnas.191367098.

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., van de Vijver, M. J., Bergh, J., Piccart, M. und Delorenzi, M. (2006). Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis. *JNCI: Journal of the National Cancer Institute* 98.4, S. 262–272. DOI: 10.1093/jnci/djj052.

Sparano, J. A., Gray, R. J., Makower, D. F., Pritchard, K. I., Albain, K. S., Hayes, D. F., Geyer, C. E. J., Dees, E. C., Perez, E. A., Olson, J. A. J., Zujewski, J., Lively, T., Badve, S. S., Saphner, T. J., Wagner, L. I., Whelan, T. J., Ellis, M. J., Paik, S., Wood, W. C., Ravdin, P., Keane, M. M., Gomez Moreno, H. L., Reddy, P. S., Goggins, T. F., Mayer, I. A., Brufsky, A. M., Toppmeyer, D. L., Kaklamani, V. G., Atkins, J. N., Berenberg, J. L. und Sledge, G. W. (2015). Prospective Validation of a 21-Gene Expression Assay in Breast Cancer. *New England Journal of Medicine* 373.21, S. 2005–2014. DOI: 10.1056/NEJMoa1510764.

Teschendorff, A. E., Naderi, A., Barbosa-Morais, N. L. und Caldas, C. (2006). PACK: Profile Analysis Using Clustering and Kurtosis to Find Molecular Classifiers in Cancer. *Bioinformatics* 22.18, S. 2269–2275. DOI: 10.1093/bioinformatics/btl174.

Therneau, T., Atkinson, B. und Ripley, B. (2015). *Rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-10.

Thomassen, M., Tan, Q. und Kruse, T. A. (2009). Gene Expression Meta-Analysis Identifies Chromosomal Regions and Candidate Genes Involved in Breast Cancer Metastasis. *Breast Cancer Research and Treatment* 113.2, S. 239–249. DOI: 10.1007/s10549-008-9927-2.

Tibshirani, R. und Hastie, T. (2007). Outlier Sums for Differential Gene Expression Analysis. *Biostatistics* 8.1, S. 2–8. DOI: 10.1093/biostatistics/kxl005.

Tibshirani, R., Hastie, T., Narasimhan, B. und Chu, G. (2002). Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression. *Proc Natl Acad Sci U S A* 99.10, S. 6567–6572. DOI: 10.1073/pnas.082099299.

Todd, J. R., Ryall, K. A., Vyse, S., Wong, J. P., Natrajan, R. C., Yuan, Y., Tan, A.-C., Huang, P. H., Todd, J. R., Ryall, K. A., Vyse, S., Wong, J. P., Natrajan, R. C., Yuan, Y., Tan, A.-C. und Huang, P. H. (2016). Systematic Analysis of Tumour Cell-Extracellular Matrix Adhesion Identifies Independent Prognostic Factors in Breast Cancer. *Oncotarget* 7.39, S. 62939–62953. DOI: 10.18632/oncotarget.11307.

Toussaint, J., Sieuwerts, A. M., Haibe-Kains, B., Desmedt, C., Rouas, G., Harris, A. L., Larsimont, D., Piccart, M., Foekens, J. A., Durbecq, V. und Sotiriou, C. (2009). Improvement of the Clinical Applicability of the Genomic Grade Index through a qRT-PCR Test Performed on Frozen and Formalin-Fixed Paraffin-Embedded Tissues. *BMC Genomics* 10.1, S. 424. DOI: 10.1186/1471-2164-10-424.

Tsai, C.-A., Hsueh, H.-m. und Chen, J. J. (2003). Estimation of False Discovery Rates in Multiple Testing: Application to Gene Microarray Data. *Biometrics* 59.4, S. 1071–1081. DOI: 10.1111/j.0006-341X.2003.00123.x.

Wang, J., Wen, S., Symmans, W. F., Pusztai, L. und Coombes, K. R. (2009). The Bimodality Index: A Criterion for Discovering and Ranking Bimodal Signatures from Cancer Gene Expression Profiling Data. *Cancer Informatics* 7, S. 199–216.

Wang, Y., Klijn, J. G. M., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M.J. J., Atkins, D. und Foekens, J. A. (2005). Gene-Expression Profiles to Predict Distant Metastasis of Lymph-Node-Negative Primary Breast Cancer. *Lancet* 365.9460, S. 671–679. DOI: 10.1016/S0140-6736(05)17947-1.

Westfall, P. H. und Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. John Wiley & Sons.

Wright, M. und Ziegler, A. (2017). Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software, Articles* 77.1, S. 1–17. DOI: 10.18637/jss.v077.i01.

Wu, Z., Irizarry, R., Gentleman, R., Murillo, F. M. und Spencer, F. (2004). A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *Johns Hopkins University, Dept. of Biostatistics Working Papers*.

Yu, C., Tang, W., Wang, Y., Shen, Q., Wang, B., Cai, C., Meng, X. und Zou, F. (2016). Downregulation of ACE2/Ang-(1-7)/Mas Axis Promotes Breast Cancer Metastasis by Enhancing Store-Operated Calcium Entry. *Cancer Letters* 376.2, S. 268–277. DOI: 10.1016/j.canlet.2016.04.006.

Zhang, C., Mapes, B. E. und Soden, B. J. (2003). Bimodality in Tropical Water Vapour. *Quarterly Journal of the Royal Meteorological Society* 129.594, S. 2847–2866. DOI: 10.1256/qj.02.166.

Zhang, Y., Sieuwerts, A. M., McGreevy, M., Casey, G., Cufer, T., Paradiso, A., Harbeck, N., Span, P. N., Hicks, D. G., Crowe, J., Tubbs, R. R., Budd, G. T., Lyons, J., Sweep, F. C.G. J., Schmitt, M., Schittulli, F., Golouh, R., Talantov, D., Wang, Y. und Foekens, J. A. (2009). The 76-Gene Signature Defines High-Risk Patients That Benefit from Adjuvant Tamoxifen Therapy. *Breast Cancer Res Treat* 116.2, S. 303–309. DOI: 10.1007/s10549-008-0183-2.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H. und Bernards, R. (2002). A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *N Engl J Med* 347.25, S. 1999–2009. DOI: 10.1056/NEJMoa021967.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. und Friend, S. H. (2002). Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature* 415.6871, S. 530–536. DOI: 10.1038/415530a.

Anhang

A Zusätzliche Informationen zu den drei Brustkrebskohorten

Tabelle A.1: Zusammenhang zwischen Metastasis Free Survival (MFS) und Metastasenstatus in den drei Brustkrebskohorten. Für den in der Arbeit verwendeten Cutoff von 5 Jahren wurden Kontingenztafeln aufgestellt. Für die Klassifikation werden nur Patientinnen verwendet, die in den ersten 5 Jahren nach der Operation eine Metastase bekommen haben (MFS ≤ 5, Metastase), oder mindestens 5 Jahre beobachtet wurden und keine Metastase bekommen haben (MFS > 5, keine Metastase).

	Mainz		Rotterdam		Transbig	
	Metastase		Metastase		Metastase	
	ja	nein	ja	nein	ja	nein
MFS ≤ 5	17	28	11	95	18	52
MFS > 5	136	19	168	12	180	30

Tabelle A.2: Übersicht der klinischen Parameter der Mainz-Kohorte.

	alle Patientinnen (n=200)		**in Klassifikation verwendet (n=164)**		**frühe Metastase (n=28)**		**lange keine Metastase (n=136)**	
Klinische Parameter	n	%	n	%	n	%	n	%
Alter bei der Diagnose								
<50	49	24.5	35	27.4	10	35.7	35	25.7
≥50	151	75.5	119	72.6	18	64.3	101	74.3
pT-Stage								
1	112	56.0	91	55.5	12	42.9	79	58.1
2	85	42.5	73	45.5	16	57.1	57	41.9
3	3	1.5	0	0.0	0	0.0	0	0.0
Histologischer Grad								
G1	42	21.0	36	22.0	0	0.0	36	26.5
G2	109	54.5	87	53.0	10	35.7	77	56.6
G3	49	24.5	41	25.0	18	64.3	23	16.9
ER-Status (IHC)								
negativ	38	19.0	33	20.1	9	32.1	24	17.6
positiv	162	81.0	131	79.9	19	67.9	112	82.4
HER2-Status (IHC)								
negativ	169	84.5	141	86.0	21	75.0	120	88.2
positiv	31	15.5	23	14.0	7	25.0	16	11.8
Metastase								
Ja	47	23.5	28	20.6	28	100.0	0	0.0
Nein	153	76.5	136	79.4	0	0.0	136	100.0

Tabelle A.3: Übersicht der klinischen Parameter der Rotterdam-Kohorte.

	alle Patientinnen (n=286)		**in Klassifikation verwendet (n=261)**		**frühe Metastase (n=93)**		**lange keine Metastase (n=168)**	
Klinische Parameter	n	%	n	%	n	%	n	%
Alter bei der Diagnose								
<56	165	57.7	-	-	-	-	-	-
≥56	121	42.3	-	-	-	-	-	-
pT-Stage								
1	146	51.0	-	-	-	-	-	-
2	132	46.2	-	-	-	-	-	-
3 und 4	8	2.8	-	-	-	-	-	-
Histologischer Grad								
G1	7	2.4	-		-	-	-	-
G2	42	14.7	-	-	-	-	-	-
G3	148	51.7	-	-	-	-	-	-
unbekannt	89	31.1	-	-	-	-	-	-
ER-Status (IHC)								
negativ	77	26.9	72	27.6	27	29.0	45	26.8
positiv	209	73.1	189	72.4	66	71.0	123	73.2
Metastase								
Ja	107	37.4	93	35.6	93	100.0	0	0.0
Nein	179	62.6	168	64.4	0	0.0	168	100.0

Tabelle A.4: GSM-Nummern der für die Transbig-Kohorte verwendeten Samples.

GSE6532		**GSE7390**				
GSM65753	GSM65816	GSM177885	GSM177927	GSM177969	GSM178011	GSM178055
GSM65754	GSM65818	GSM177886	GSM177928	GSM177970	GSM178012	GSM178056
GSM65755	GSM65819	GSM177887	GSM177929	GSM177971	GSM178013	GSM178057
GSM65756	GSM65820	GSM177888	GSM177930	GSM177972	GSM178015	GSM178058
GSM65757	GSM65821	GSM177889	GSM177931	GSM177973	GSM178016	GSM178059
GSM65758	GSM65823	GSM177890	GSM177932	GSM177974	GSM178017	GSM178060
GSM65760	GSM65825	GSM177891	GSM177933	GSM177975	GSM178018	GSM178061
GSM65761	GSM65826	GSM177892	GSM177934	GSM177976	GSM178019	GSM178062
GSM65763	GSM65827	GSM177893	GSM177935	GSM177977	GSM178020	GSM178063
GSM65765	GSM65829	GSM177894	GSM177936	GSM177978	GSM178021	GSM178064
GSM65766	GSM65830	GSM177895	GSM177937	GSM177979	GSM178022	GSM178065
GSM65767	GSM65831	GSM177896	GSM177938	GSM177980	GSM178023	GSM178066
GSM65768	GSM65832	GSM177897	GSM177939	GSM177981	GSM178024	GSM178067
GSM65769	GSM65834	GSM177898	GSM177940	GSM177982	GSM178025	GSM178068
GSM65770	GSM65835	GSM177899	GSM177941	GSM177983	GSM178026	GSM178069
GSM65771	GSM65837	GSM177900	GSM177942	GSM177984	GSM178027	GSM178070
GSM65772	GSM65839	GSM177901	GSM177943	GSM177985	GSM178029	GSM178071
GSM65773	GSM65840	GSM177902	GSM177944	GSM177986	GSM178030	GSM178072
GSM65776	GSM65841	GSM177903	GSM177945	GSM177987	GSM178031	GSM178073
GSM65780	GSM65844	GSM177904	GSM177946	GSM177988	GSM178032	GSM178074
GSM65781	GSM65847	GSM177905	GSM177947	GSM177989	GSM178033	GSM178075
GSM65783	GSM65848	GSM177906	GSM177948	GSM177990	GSM178034	GSM178076
GSM65784	GSM65849	GSM177907	GSM177949	GSM177991	GSM178035	GSM178077
GSM65785	GSM65850	GSM177908	GSM177950	GSM177992	GSM178036	GSM178078
GSM65788	GSM65851	GSM177909	GSM177951	GSM177993	GSM178037	GSM178079
GSM65789	GSM65852	GSM177910	GSM177952	GSM177994	GSM178038	GSM178080
GSM65790	GSM65853	GSM177911	GSM177953	GSM177995	GSM178039	GSM178081
GSM65793	GSM65854	GSM177912	GSM177954	GSM177996	GSM178040	GSM178082
GSM65794	GSM65856	GSM177913	GSM177955	GSM177997	GSM178041	
GSM65796	GSM65858	GSM177914	GSM177956	GSM177998	GSM178042	
GSM65799	GSM65859	GSM177915	GSM177957	GSM177999	GSM178043	
GSM65802	GSM65860	GSM177916	GSM177958	GSM178000	GSM178044	
GSM65803	GSM65861	GSM177917	GSM177959	GSM178001	GSM178045	
GSM65804	GSM65864	GSM177918	GSM177960	GSM178002	GSM178046	
GSM65806	GSM65865	GSM177919	GSM177961	GSM178003	GSM178047	
GSM65807	GSM65866	GSM177920	GSM177962	GSM178004	GSM178048	
GSM65808	GSM65867	GSM177921	GSM177963	GSM178005	GSM178049	
GSM65810	GSM65868	GSM177922	GSM177964	GSM178006	GSM178050	
GSM65812	GSM65869	GSM177923	GSM177965	GSM178007	GSM178051	
GSM65813	GSM65875	GSM177924	GSM177966	GSM178008	GSM178052	
GSM65814	GSM65876	GSM177925	GSM177967	GSM178009	GSM178053	
GSM65815	GSM65880	GSM177926	GSM177968	GSM178010	GSM178054	

Tabelle A.5: Übersicht der klinischen Parameter der Transbig-Kohorte.

	alle Patientinnen (n=280)		**in Klassifikation verwendet (n=232)**		**frühe Metastase (n=52)**		**lange keine Metastase (n=180)**	
Klinische Parameter	n	%	n	%	n	%	n	%
Alter bei der Diagnose								
<50	158	56.4	130	56.0	30	57.7	100	55.6
≥50	122	43.6	102	44.0	22	42.3	80	44.4
pT-Stage								
1	149	53.2	130	56.0	23	44.2	107	59.4
2	130	46.4	102	44.0	29	55.8	73	40.6
3	1	0.4	0	0.0	0	0.0	0	0.0
Histologischer Grad								
G1	56	20.0	43	18.5	3	5.8	40	22.2
G2	109	38.9	86	37.1	20	38.5	66	36.7
G3	100	35.7	89	38.4	26	50.0	63	35.0
unbekannt	15	5.4	14	6.0	3	5.8	11	6.1
ER-Status (IHC)								
negativ	81	28.9	68	29.3	22	42.3	46	25.6
positiv	194	69.3	159	68.5	29	55.8	130	72.2
unbekannt	5	1.8	5	2.2	1	1.9	4	2.2
HER2-Status (RNA)								
negativ	245	87.5	199	85.8	43	82.7	156	86.7
positiv	35	12.5	33	14.2	9	17.3	24	13.3
Metastase								
Ja	82	29.3	52	22.4	52	100.0	0	0.0
Nein	198	70.7	180	77.6	0	0.0	180	100.0

B Schnittpunkte zweier Normalverteilungsdichten

Im Folgenden wird die Formel zur Berechnung der Schnittpunkte zweier Normalverteilungsdichten hergeleitet, die in Kapitel 3.1.1 verwendet wird. Sei dafür ϕ_θ die Dichte der Normalverteilung mit $\theta = (\mu, \sigma^2)$ und p ein Skalierungsfaktor. Wir betrachten zunächst den Fall von ungleichen Varianzen, das heißt $\sigma_1^2 \neq \sigma_2^2$. Setze

$$p \cdot \phi_1(x) = (1-p) \cdot \phi_2(x).$$

Einsetzen der Dichtefunktionen liefert:

$$p \cdot \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right) = (1-p)\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right).$$

Anwenden des Logarithmus:

$$\log\left(\frac{p}{\sqrt{2\pi\sigma_1^2}}\right) - \frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2 = \log\left(\frac{1-p}{\sqrt{2\pi\sigma_2^2}}\right) - \frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2.$$

Weitere Umformungen und Ausmultiplizieren:

$$\log\left(\frac{p}{\sqrt{2\pi\sigma_1^2}}\right) - \log\left(\frac{1-p}{\sqrt{2\pi\sigma_2^2}}\right) - \frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2 = 0$$

$$\log\left(\frac{p \cdot \sqrt{2\pi\sigma_2^2}}{(1-p) \cdot \sqrt{2\pi\sigma_1^2}}\right) - \frac{1}{2}\left(\frac{x^2 - 2\mu_1 x + \mu_1^2}{\sigma_1^2}\right) + \frac{1}{2}\left(\frac{x^2 - 2\mu_2 x + \mu_2^2}{\sigma_2^2}\right) = 0$$

$$2\sigma_1^2\sigma_2^2 \log\left(\frac{p \cdot \sigma_2}{(1-p) \cdot \sigma_1}\right) - \sigma_2^2(x^2 - 2\mu_1 x + \mu_1^2) + \sigma_1^2(x^2 - 2\mu_2 x + \mu_2^2) = 0$$

$$2\sigma_1^2\sigma_2^2 \log\left(\frac{p \cdot \sigma_2}{(1-p) \cdot \sigma_1}\right) + (\sigma_1^2 - \sigma_2^2)x^2 + 2(\mu_1\sigma_2^2 - \mu_2\sigma_1^2)x + \mu_2^2\sigma_1^2 - \mu_1^2\sigma_2^2 = 0$$

Umordnen der Summanden, sodass eine quadratische Gleichung entsteht:

$$x^2 + \frac{2(\mu_1\sigma_2^2 - \mu_2\sigma_1^2)}{\sigma_1^2 - \sigma_2^2}x + \frac{2\sigma_1^2\sigma_2^2 \log\left(\frac{p\cdot\sigma_2}{(1-p)\cdot\sigma_1}\right) + \mu_2^2\sigma_1^2 - \mu_1^2\sigma_2^2}{\sigma_1^2 - \sigma_2^2} = 0.$$

Lösung der quadratischen Gleichung mit Hilfe der p-q-Formel:

$$\begin{aligned}
x_{1,2} &= -\frac{p}{2} \pm \sqrt{\frac{p^2}{4} - q} \\
&= -\frac{2(\mu_1\sigma_2^2 - \mu_2\sigma_1^2)}{2(\sigma_1^2 - \sigma_2^2)} \pm \sqrt{\frac{4(\mu_1\sigma_2^2 - \mu_2\sigma_1^2)^2}{4(\sigma_1^2 - \sigma_2^2)^2} - \left(\frac{2\sigma_1^2\sigma_2^2 \log\left(\frac{p\cdot\sigma_2}{(1-p)\cdot\sigma_1}\right) + \mu_2^2\sigma_1^2 - \mu_1^2\sigma_2^2}{\sigma_1^2 - \sigma_2^2}\right)} \\
&= -\frac{\mu_2\sigma_1^2 - \mu_1\sigma_2^2}{\sigma_1^2 - \sigma_2^2} \\
&\quad \pm\sqrt{\frac{(\mu_1^2\sigma_2^4 - 2\mu_1\mu_2\sigma_1^2\sigma_2^2 + \mu_2^2\sigma_1^4) - (\sigma_1^2 - \sigma_2^2)(2\sigma_1^2\sigma_2^2 \log\left(\frac{p\cdot\sigma_2}{(1-p)\cdot\sigma_1}\right) + \mu_2^2\sigma_1^2 - \mu_1^2\sigma_2^2)}{(\sigma_1^2 - \sigma_2^2)^2}} \\
&= \frac{\mu_1\sigma_2^2 - \mu_2\sigma_1^2}{\sigma_1^2 - \sigma_2^2} \\
&\quad \pm\sqrt{\frac{\sigma_1^2\sigma_2^2(\mu_1^2 - 2\mu_1\mu_2 + \mu_2^2) - 2(\sigma_1^2 - \sigma_2^2)\log\left(\frac{p\cdot\sigma_2}{(1-p)\cdot\sigma_1}\right)}{(\sigma_1^2 - \sigma_2^2)^2}} \\
&= \frac{\mu_1\sigma_2^2 - \mu_2\sigma_1^2}{\sigma_1^2 - \sigma_2^2} \pm \frac{\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}\sqrt{(\mu_1 - \mu_2)^2 - 2(\sigma_1^2 - \sigma_2^2)\log\left(\frac{p\cdot\sigma_2}{(1-p)\cdot\sigma_1}\right)}.
\end{aligned}$$

Aus dieser Formel leitet sich die Bedingung ab, dass der Schnittpunkt der Normalverteilungsdichten nur definiert ist, wenn gilt:

$$(\mu_1 - \mu_2)^2 \geq 2(\sigma_1^2 - \sigma_2^2)\log\left(\frac{p\cdot\sigma_2}{(1-p)\cdot\sigma_1}\right).$$

Im Fall einer gemeinsamen Varianz σ^2 gibt es genau einen Schnittpunkt. Setze:

$$p\cdot\phi_1(x) = (1-p)\cdot\phi_2(x).$$

Einsetzen der Dichtefunktion liefert:

$$p\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2\right) = (1-p)\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma}\right)^2\right).$$

Anwenden des Logarithmus:

$$\log(p) - \frac{1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2 = \log(1-p) - \frac{1}{2}\left(\frac{x-\mu_2}{\sigma}\right)^2.$$

Weitere Umformungen:

$$(x-\mu_1)^2 - (x-\mu_2)^2 = 2\sigma^2 \log\left(\frac{p}{1-p}\right)$$

$$x^2 - 2\mu_1 x + \mu_1^2 - x^2 + 2\mu_2 x - \mu_2^2 = 2\sigma^2 \log\left(\frac{p}{1-p}\right)$$

$$2(\mu_2-\mu_1)x + (\mu_1^2 - \mu_2^2) = 2\sigma^2 \log\left(\frac{p}{1-p}\right)$$

$$2(\mu_2-\mu_1)x = 2\sigma^2 \log\left(\frac{p}{1-p}\right) + (\mu_2^2 - \mu_1^2)$$

Auflösen nach x liefert:

$$x = \frac{\sigma^2}{\mu_2-\mu_1} \log\left(\frac{p}{1-p}\right) + \frac{\mu_1+\mu_2}{2}.$$

C Kontingenztafeln für die Klassifikation mit den etablierten Gensignaturen auf den drei Brustkrebskohorten

Um die Ergebnisse der neu-entwickelten Baummodelle mit denen etablierter Klassifikatoren vergleichen zu können, wurden für die drei Kohorten mit nodal-negativen unbehandelten Patientinnen Vorhersagen mit dem Bioconductor Paket `genefu` erstellt (Gendoo et al. 2015). Klasseneinteilungen nach dem 70-Gen-Klassifikator, dem 76-Gen-Klassifikator und Oncotype DX konnten für alle 3 Kohorten vorgenommen werden, da bei diesen Klassifikatoren nur die Genexpressionsdaten bzw. Genexpressionsdaten und immunhistochemisch bestimmter ER-Status (für den 76-Gen-Klassifikator) verwendet werden. Diese Informationen sind für alle drei Kohorten verfügbar, wobei für fünf Patientinnen der Transbig-Kohorte die Information über den ER-Status fehlt. Für den Genomic Grade Index (GGI) wird neben den Expressionswerten der histologische Grad benötigt. Dieser liegt für die Rotterdam-Kohorte nicht vor, daher kann der GGI für diese Kohorte nicht bestimmt werden. Für 14 Patientinnen der Transbig-Kohorte gibt es ebenfalls keine Informationen über den histologischen Grad. Für die Mainz-Kohorte kann zusätzlich zu den Gensignaturen auch die Risikogruppe nach der NNBC-3-Studie bestimmt werden.

Oncotype DX und NNBC-3 teilen die Patientinnen jeweils in drei Gruppen mit niedrigem (low risk), mittlerem (intermediate risk) oder hohem Risiko (high risk) ein. Die anderen Klassifikatoren teilen nur in zwei Gruppen (niedriges und hohes Risiko) ein. Um die Ergebnisse der Klassifikatoren zu vergleichen, werden bei diesen Klassifikatoren die Patientinnen der Gruppe mit mittlerem Risiko einmal der Gruppen mit niedrigem und einmal der Gruppe mit hohem Risiko zugeordnet, sodass es für diese Klassifikatoren jeweils zwei Ergebnisse gibt. Die Kontingenztafeln für den Vergleich der vorgesagten Gruppenzugehörigkeit mit der wahren Gruppenzugehörigkeit sind im folgenden Abschnitt zu finden.

C.1 70-Gen-Klassifikator

Mainz

Tabelle C.1: Kontingenztafel für die Vorhersagen des 70-Gen-Klassifikators auf der Mainz-Kohorte.

	Metastase in den ersten 5 Jahren				
Klassifikation	ja	nein	$\sum$		
high	26	92	118	Sensitivität	0.929
low	2	44	46	PPV	0.220
$\sum$	28	136	164	Fehler	94

NPV	0.957
Spezifität	0.324

Rotterdam

Tabelle C.2: Kontingenztafel für die Vorhersagen des 70-Gen-Klassifikators auf der Rotterdam-Kohorte.

	Metastase in den ersten 5 Jahren				
Klassifikation	ja	nein	$\sum$		
high	82	107	189	Sensitivität	0.882
low	11	61	72	PPV	0.434
$\sum$	93	168	261	Fehler	118

NPV	0.847
Spezifität	0.363

Transbig

Tabelle C.3: Kontingenztafel für die Vorhersagen des 70-Gen-Klassifikators auf der Transbig-Kohorte.

	Metastase in den ersten 5 Jahren		
Klassifikation	ja	nein	$\sum$
high	48	117	165
low	4	63	67
$\sum$	52	180	232

NPV	0.940
Spezifität	0.350
Sensitivität	0.923
PPV	0.291
Fehler	121

C.2 76-Gen-Klassifikator

Mainz

Tabelle C.4: Kontingenztafel für die Vorhersagen des 76-Gen-Klassifikators auf der Mainz-Kohorte.

	Metastase in den ersten 5 Jahren				
Klassifikation	ja	nein	$\sum$		
high	28	131	159	Sensitivität	1.000
low	0	5	5	PPV	0.176
$\sum$	28	136	164	Fehler	131

NPV	1.000
Spezifität	0.037

Rotterdam

Tabelle C.5: Kontingenztafel für die Vorhersagen des 76-Gen-Klassifikators auf der Rotterdam-Kohorte.

	Metastase in den ersten 5 Jahren				
Klassifikation	ja	nein	$\sum$		
high	92	140	232	Sensitivität	0.989
low	1	28	29	PPV	0.397
$\sum$	93	168	261	Fehler	141

NPV	0.966
Spezifität	0.167

Transbig

Tabelle C.6: Kontingenztafel für die Vorhersagen des 76-Gen-Klassifikators auf der Transbig-Kohorte.

	Metastase in den ersten 5 Jahren				
Klassifikation	ja	nein	$\sum$	NPV	0.947
high	50	158	208	Spezifität	0.102
low	1	18	19	Sensitivität	0.980
$\sum$	51	176	227	PPV	0.240
				Fehler	159

C.3 Oncotype DX

Mainz

Tabelle C.7: Kontingenztafel für die Vorhersagen von Oncotype DX auf der Mainz-Kohorte.

	Metastase in den ersten 5 Jahren		
Klassifikation	ja	nein	$\sum$
high risk	21	56	77
intermediate risk	1	37	38
low risk	6	43	49
$\sum$	28	136	164

Tabelle C.8: Kontingenztafel für die Vorhersagen von Oncotype DX auf der Mainz-Kohorte (low vs. intermediate+high).

	Metastase in den ersten 5 Jahren				
Klassifikation	ja	nein	$\sum$	NPV	0.878
intermediate+high	22	93	115	Spezifität	0.316
low	6	43	49	Sensitivität	0.786
$\sum$	28	136	164	PPV	0.191
				Fehler	99

Tabelle C.9: Kontingenztafel für die Vorhersagen von Oncotype DX auf der Mainz-Kohorte (low+intermediate vs. high).

	Metastase in den ersten 5 Jahren				
Klassifikation	ja	nein	$\sum$	NPV	0.920
high	21	56	77	Spezifität	0.588
low+intermediate	7	80	87	Sensitivität	0.750
$\sum$	28	136	164	PPV	0.273
				Fehler	63

Rotterdam

Tabelle C.10: Kontingenztafel für die Vorhersagen von Oncotype DX auf der Rotterdam-Kohorte.

	Metastase in den ersten 5 Jahren		
Klassifikation	ja	nein	$\sum$
high risk	68	92	160
intermediate risk	18	41	59
low risk	7	35	42
$\sum$	93	168	261

Tabelle C.11: Kontingenztafel für die Vorhersagen von Oncotype DX auf der Rotterdam-Kohorte (low vs. intermediate+high).

	Metastase in den ersten 5 Jahren				
Klassifikation	ja	nein	$\sum$	NPV	0.833
intermediate+high	86	133	219	Spezifität	0.208
low	7	35	42	Sensitivität	0.925
$\sum$	93	168	261	PPV	0.393
				Fehler	140

Tabelle C.12: Kontingenztafel für die Vorhersagen von Oncotype DX auf der Rotterdam-Kohorte (low+intermediate vs. high).

	Metastase in den ersten 5 Jahren				
Klassifikation	ja	nein	$\sum$	NPV	0.752
high	68	92	160	Spezifität	0.452
low+intermediate	25	76	101	Sensitivität	0.731
$\sum$	93	168	261	PPV	0.425
				Fehler	117

Transbig

Tabelle C.13: Kontingenztafel für die Vorhersagen von Oncotype DX auf der Transbig-Kohorte.

	Metastase in den ersten 5 Jahren		
Klassifikation	ja	nein	$\sum$
high risk	41	88	129
intermediate risk	9	43	52
low risk	2	49	51
$\sum$	52	180	232

Tabelle C.14: Kontingenztafel für die Vorhersagen von Oncotype DX auf der Transbig-Kohorte (low vs. intermediate+high).

	Metastase in den ersten 5 Jahren				
Klassifikation	ja	nein	$\sum$	NPV	0.961
intermediate+high	50	131	181	Spezifität	0.272
low	2	49	51	Sensitivität	0.962
$\sum$	52	180	232	PPV	0.276
				Fehler	133

Tabelle C.15: Kontingenztafel für die Vorhersagen von Oncotype DX auf der Transbig-Kohorte (low+intermediate vs. high).

	Metastase in den ersten 5 Jahren				
Klassifikation	ja	nein	$\sum$	NPV	0.893
high	41	88	129	Spezifität	0.511
low+intermediate	11	92	103	Sensitivität	0.788
$\sum$	52	180	232	PPV	0.318
				Fehler	99

C.4 GGI

Mainz

Tabelle C.16: Kontingenztafel für die Vorhersagen des GGI auf der Mainz-Kohorte.

	Metastase in den ersten 5 Jahren				
Klassifikation	ja	nein	$\sum$	NPV	0.928
high	21	46	67	Spezifität	0.662
low	7	90	97	Sensitivität	0.750
$\sum$	28	136	164	PPV	0.313
				Fehler	53

Transbig

Tabelle C.17: Kontingenztafel für die Vorhersagen des GGI auf der Transbig-Kohorte.

	Metastase in den ersten 5 Jahren				
Klassifikation	ja	nein	$\sum$	NPV	0.909
high	41	70	111	Spezifität	0.611
low	11	110	121	Sensitivität	0.788
$\sum$	52	180	232	PPV	0.369
				Fehler	81

C.5 NNBC-3

Tabelle C.18: Kontingenztafel für die Vorhersagen von NNBC-3 auf der Mainz-Kohorte.

	Metastase in den ersten 5 Jahren		
Klassifikation	ja	nein	$\sum$
high risk	24	59	83
intermediate risk	2	18	20
low risk	2	55	57
$\sum$	28	132	160

Tabelle C.19: Kontingenztafel für die Vorhersagen von NNBC-3 auf der Mainz-Kohorte (low vs. intermediate+high).

	Metastase in den ersten 5 Jahren				
Klassifikation	ja	nein	$\sum$	NPV	0.965
intermediate+high	26	77	103	Spezifität	0.417
low	2	55	57	Sensitivität	0.929
$\sum$	28	132	160	PPV	0.252
				Fehler	79

Tabelle C.20: Kontingenztafel für die Vorhersagen von NNBC-3 auf der Mainz-Kohorte (low+intermediate vs. high).

	Metastase in den ersten 5 Jahren				
Klassifikation	ja	nein	$\sum$	NPV	0.948
high	24	59	83	Spezifität	0.553
low+intermediate	4	73	77	Sensitivität	0.857
$\sum$	28	132	160	PPV	0.289
				Fehler	63

D Ergebnisse der Klassifikationsbäume

D.1 Abbildungen für *minsplit* = 20

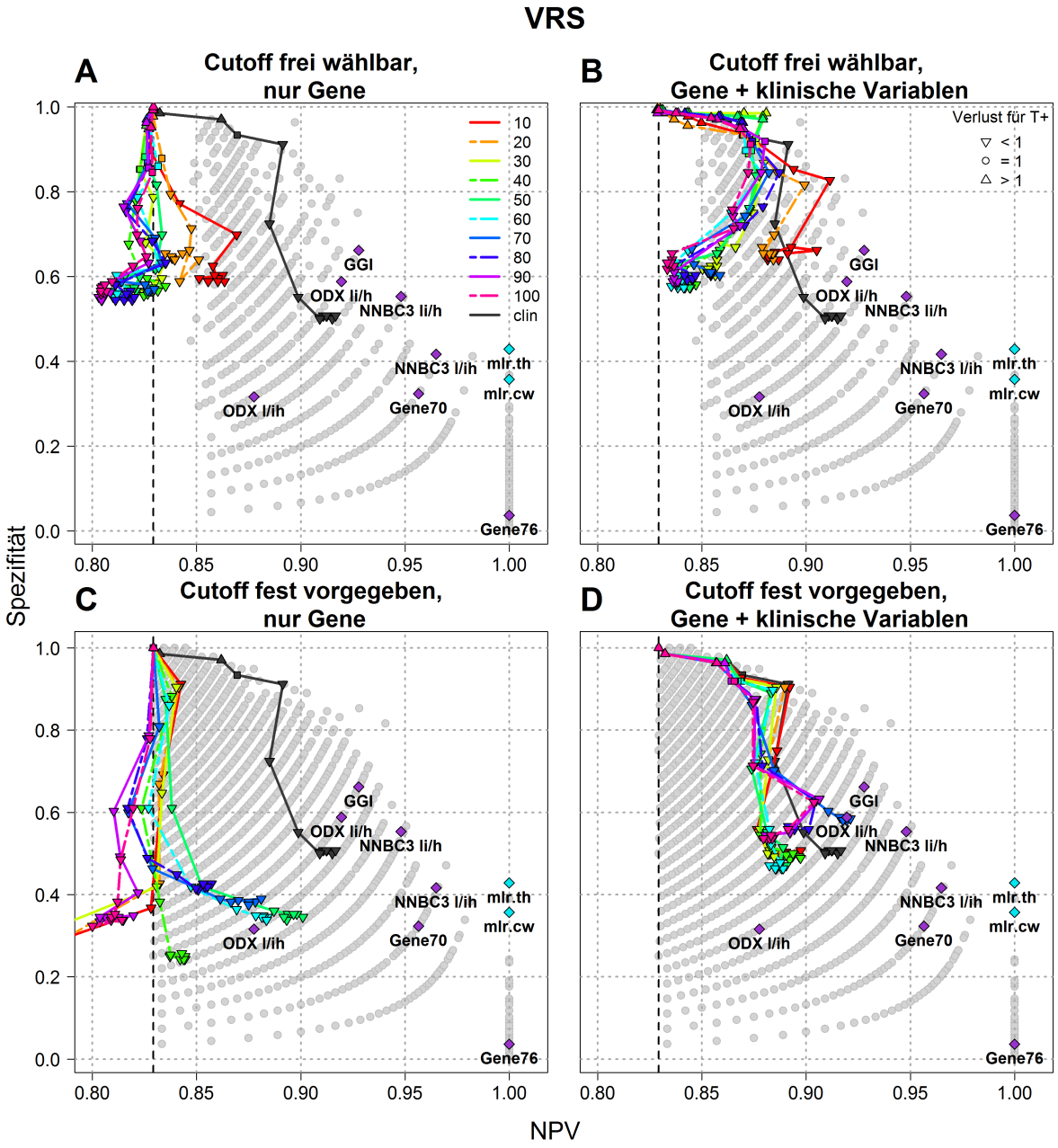

Abbildung D.1: NPV und Spezifität der Klassifikationsbäume mit *minsplit* = 20 für die Top-Gene von VRS.

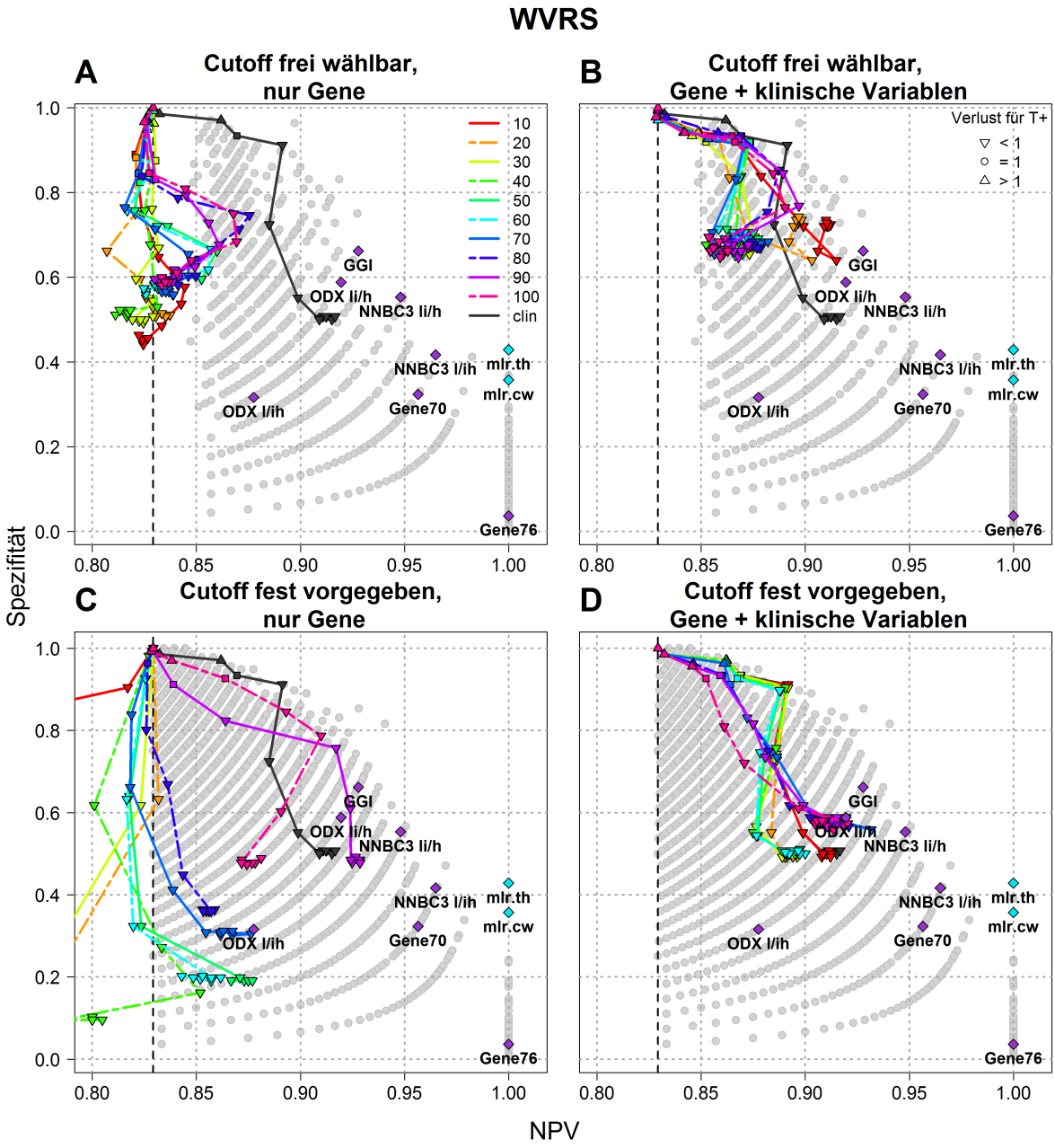

Abbildung D.2: NPV und Spezifität der Klassifikationsbäume mit *minsplit* = 20 für die Top-Gene von WVRS.

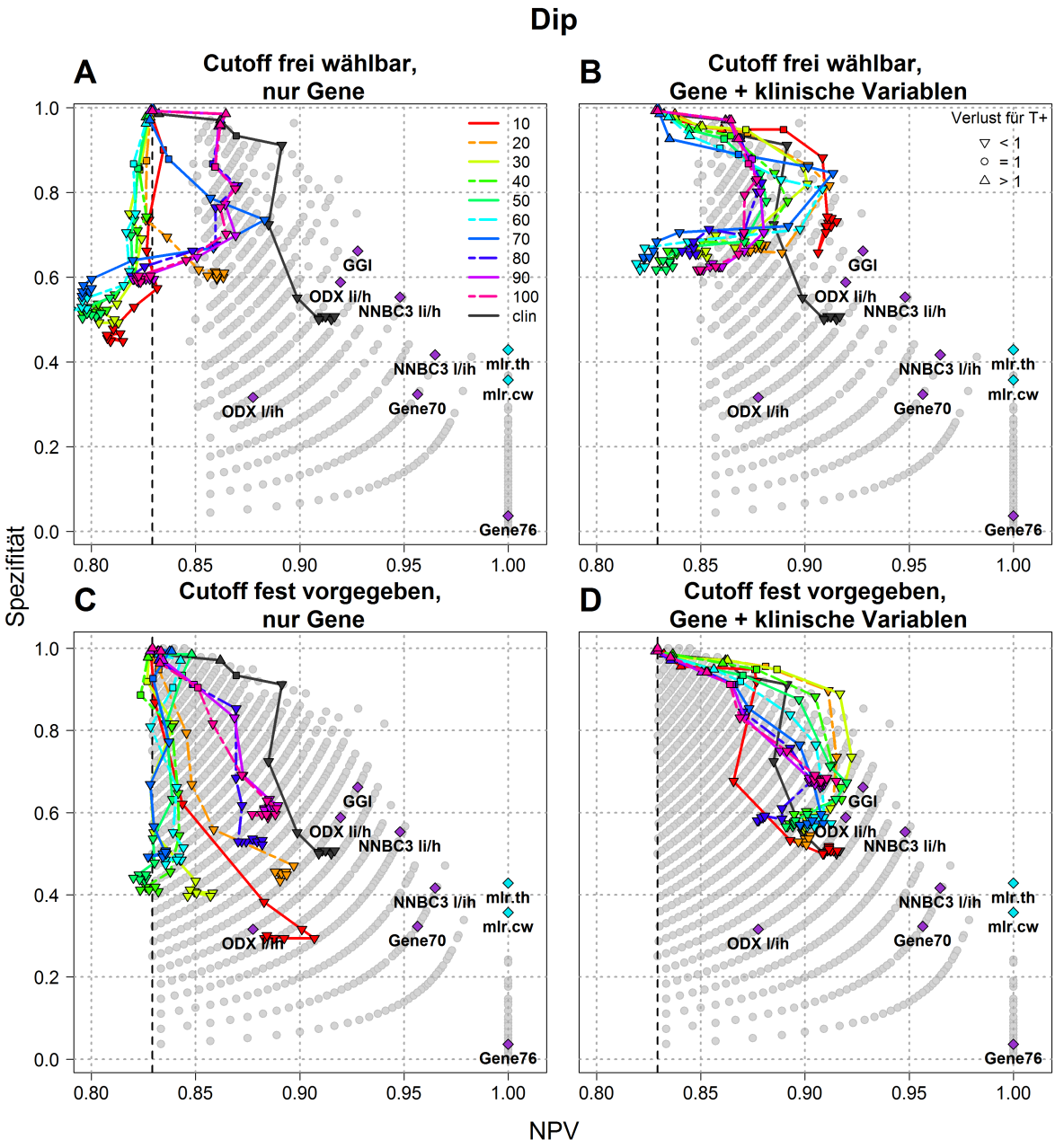

Abbildung D.3: NPV und Spezifität der Klassifikationsbäume mit *minsplit* = 20 für die Top-Gene der dip-Statistik.

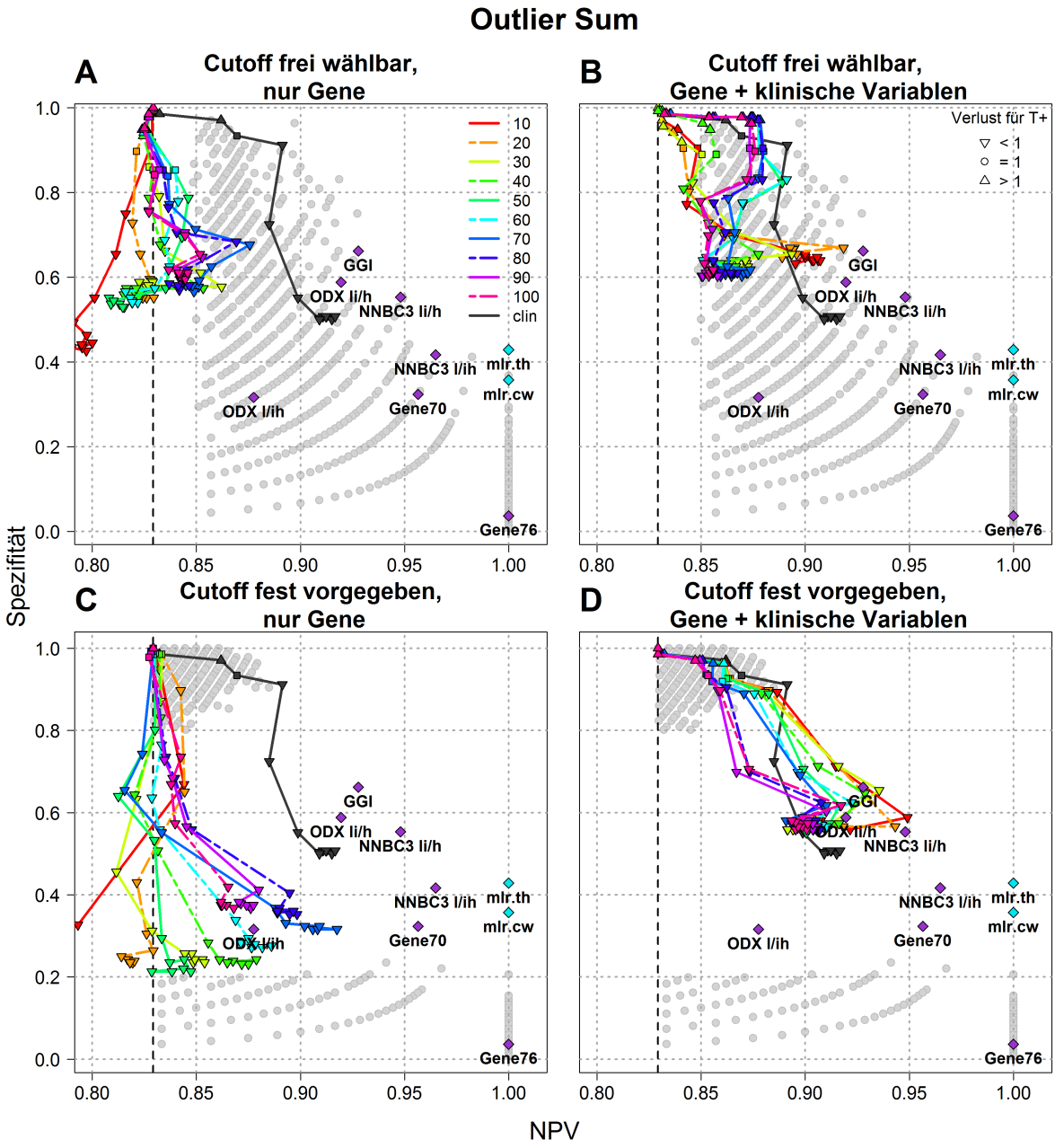

Abbildung D.4: NPV und Spezifität der Klassifikationsbäume mit *minsplit* = 20 für die Top-Gene der Outlier-Sum-Statistik.

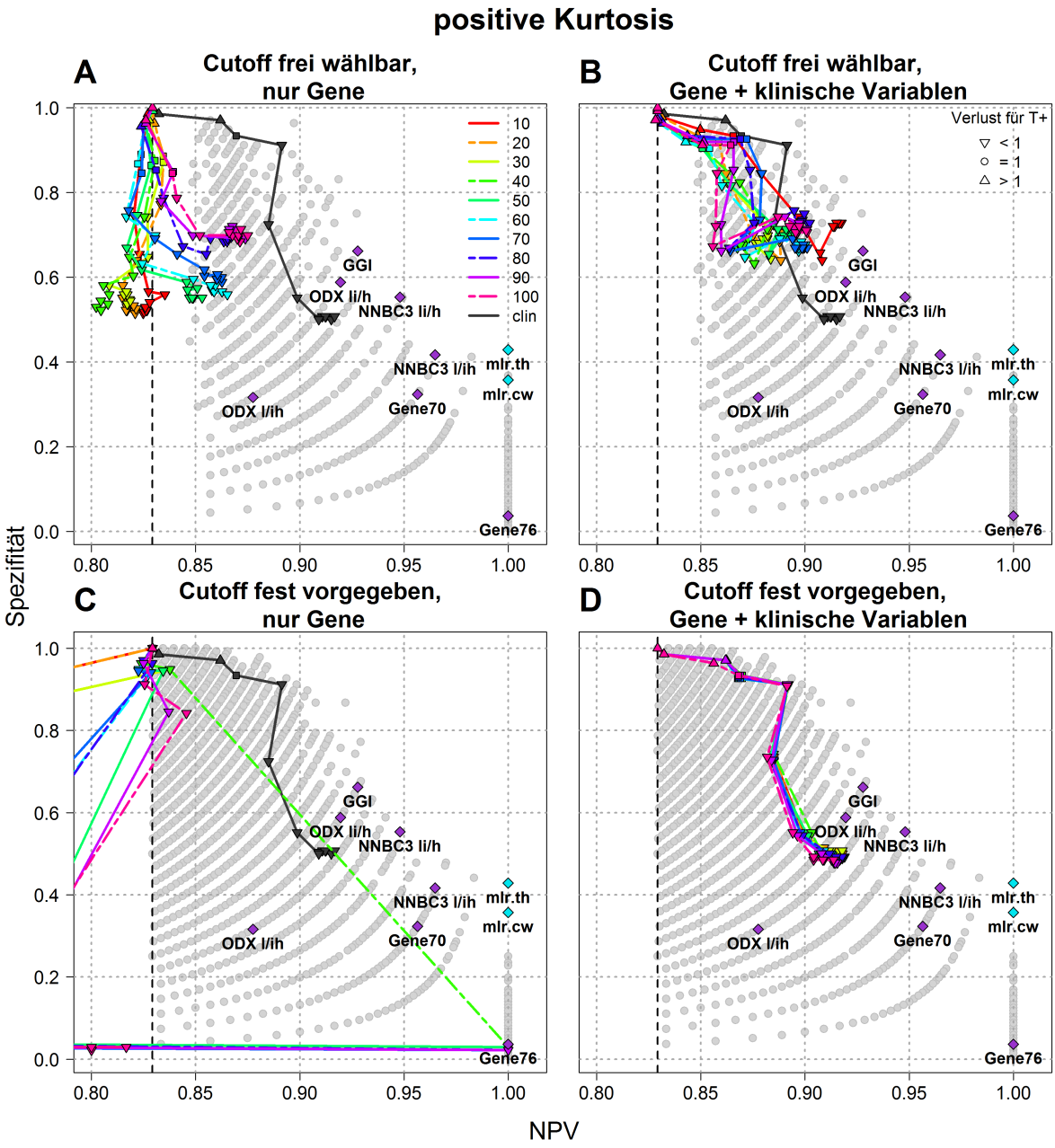

Abbildung D.5: NPV und Spezifität der Klassifikationsbäume mit *minsplit* = 20 für die Top-Gene der positiven Kurtosis.

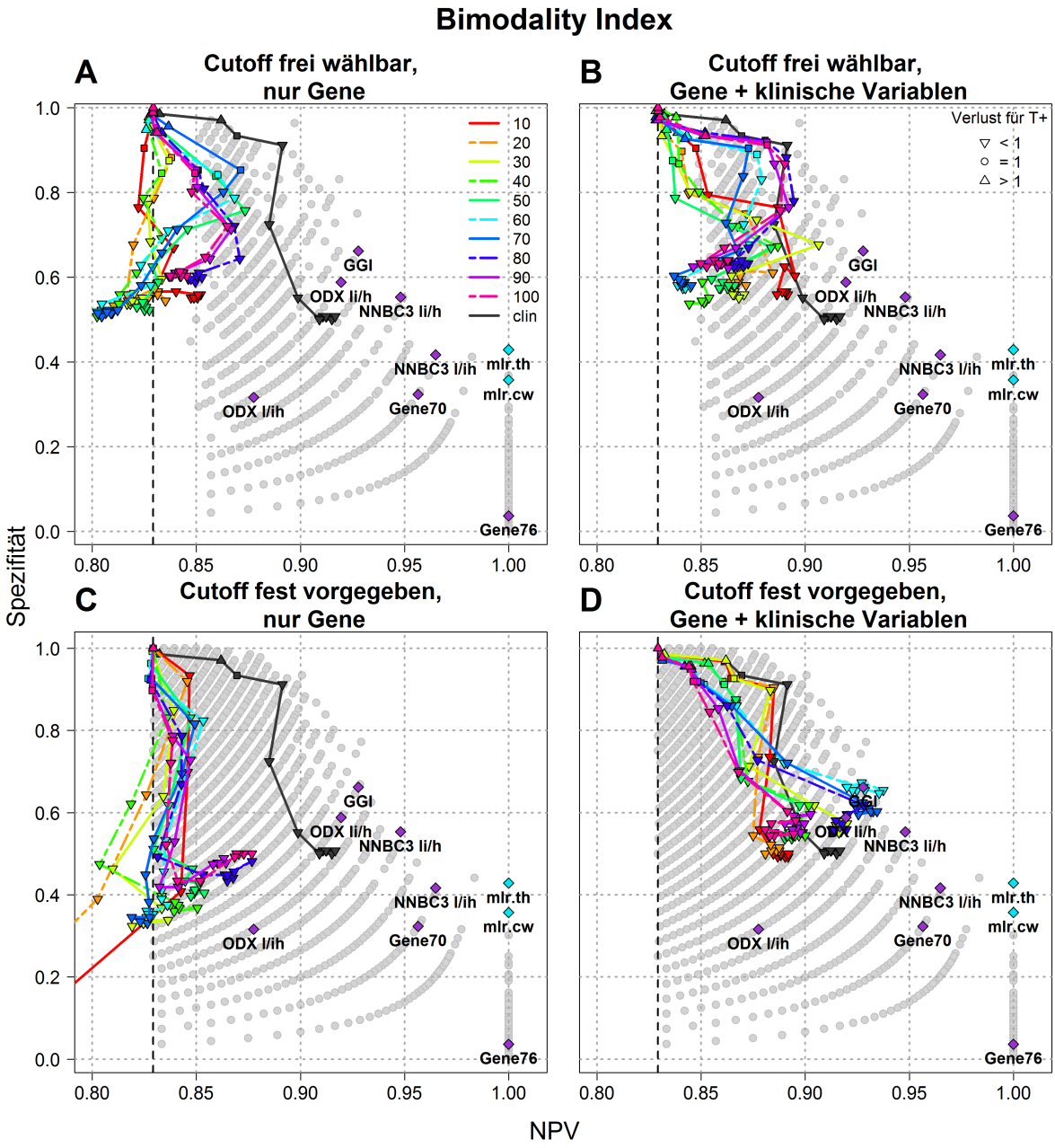

Abbildung D.6: NPV und Spezifität der Klassifikationsbäume mit *minsplit* = 20 für die Top-Gene des Bimodality Index.

D.2 Abbildungen für *minsplit* = 5

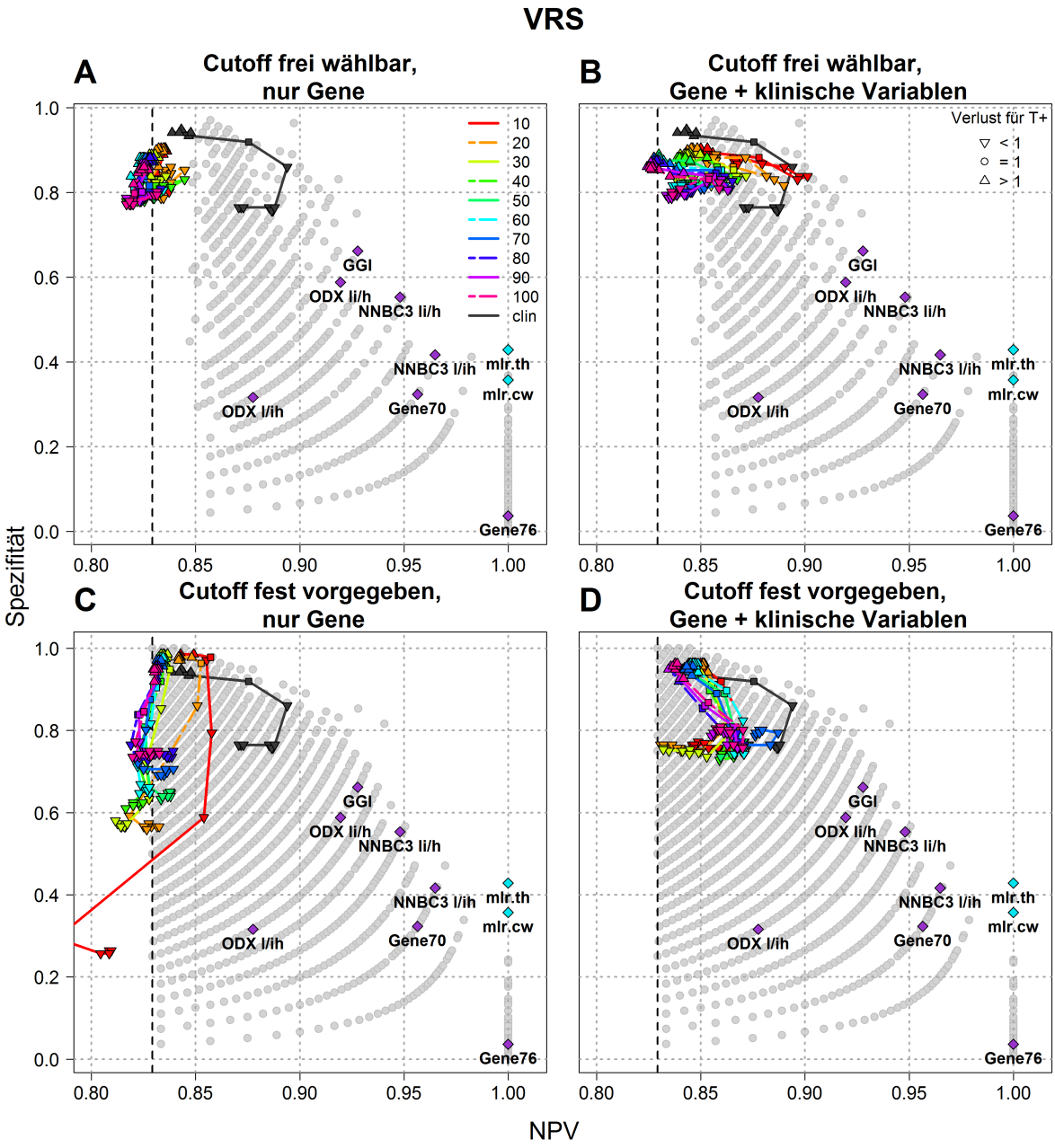

Abbildung D.7: NPV und Spezifität der Klassifikationsbäume mit *minsplit* = 5 für die Top-Gene von VRS.

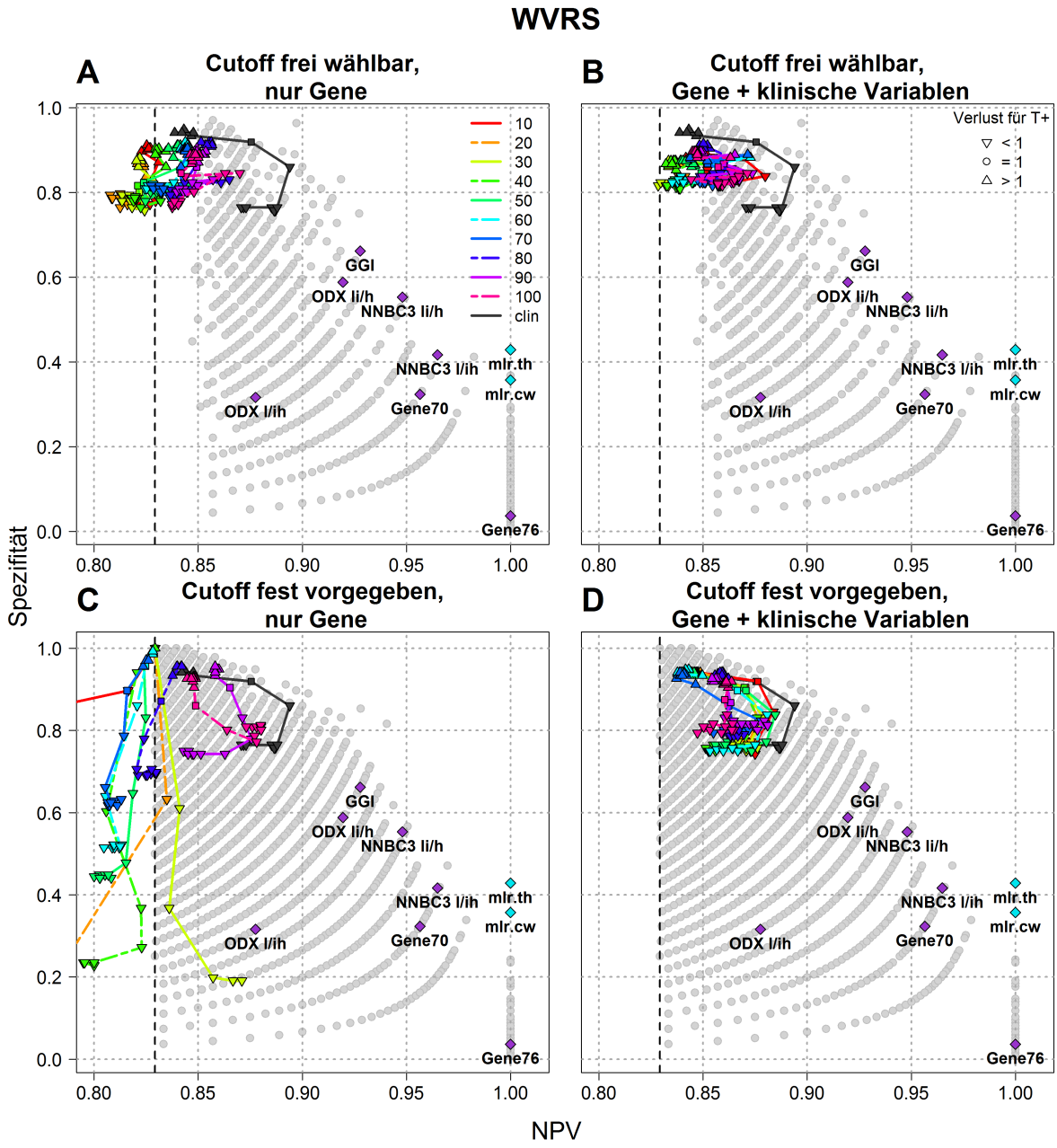

Abbildung D.8: NPV und Spezifität der Klassifikationsbäume mit *minsplit* = 5 für die Top-Gene von WVRS.

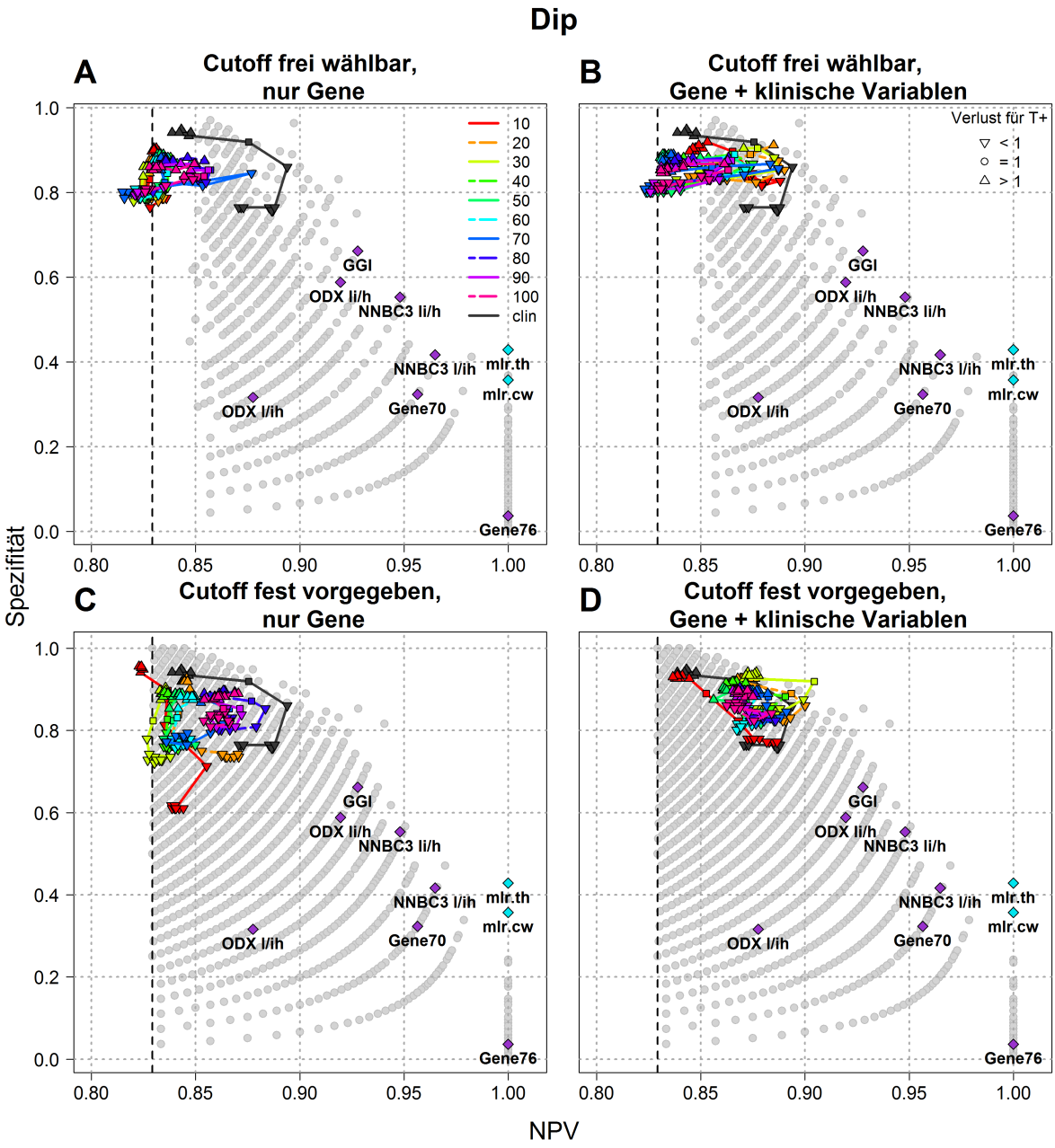

Abbildung D.9: NPV und Spezifität der Klassifikationsbäume mit *minsplit* = 5 für die Top-Gene der dip-Statistik.

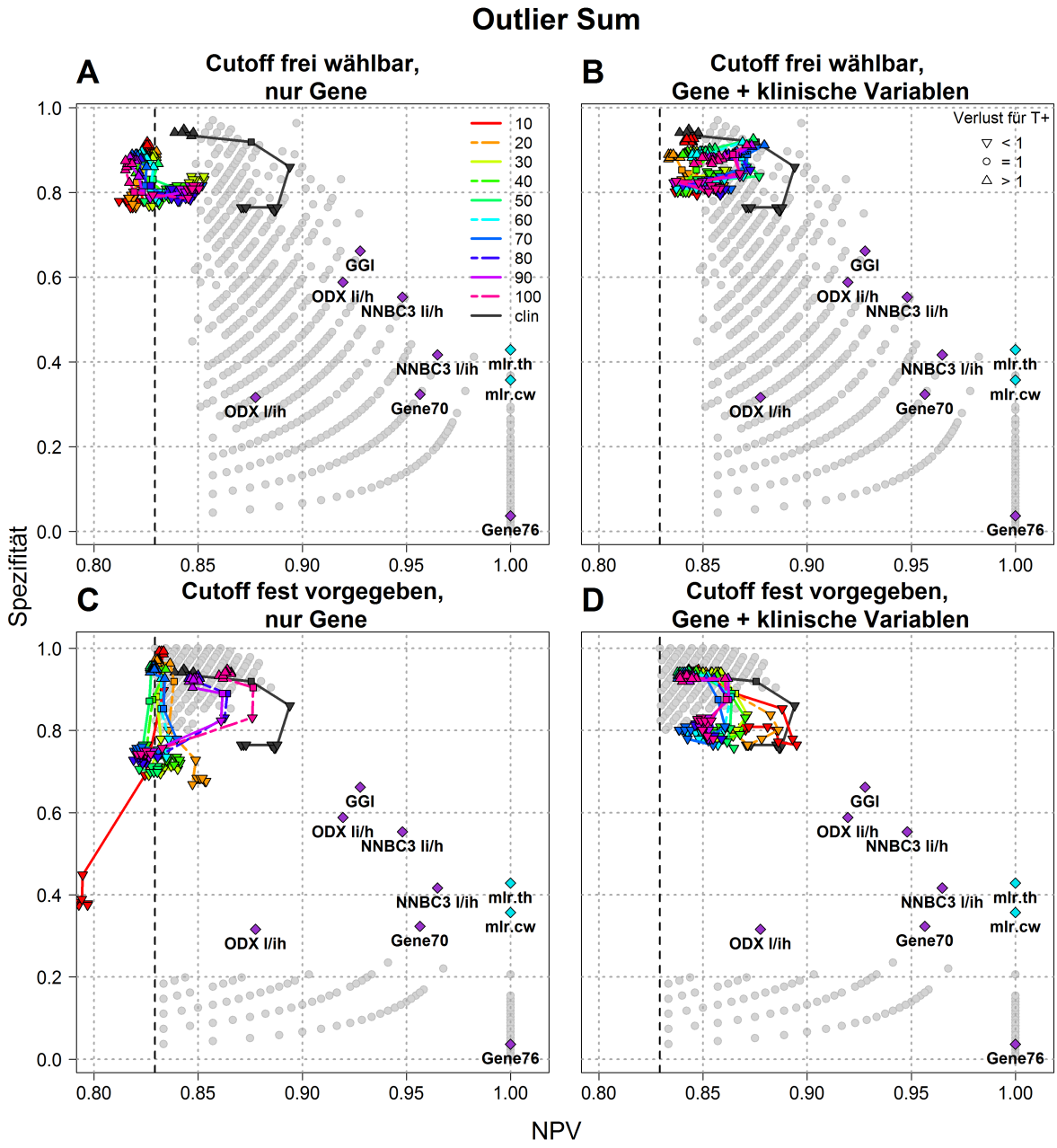

Abbildung D.10: NPV und Spezifität der Klassifikationsbäume mit *minsplit* = 5 für die Top-Gene der Outlier-Sum-Statistik.

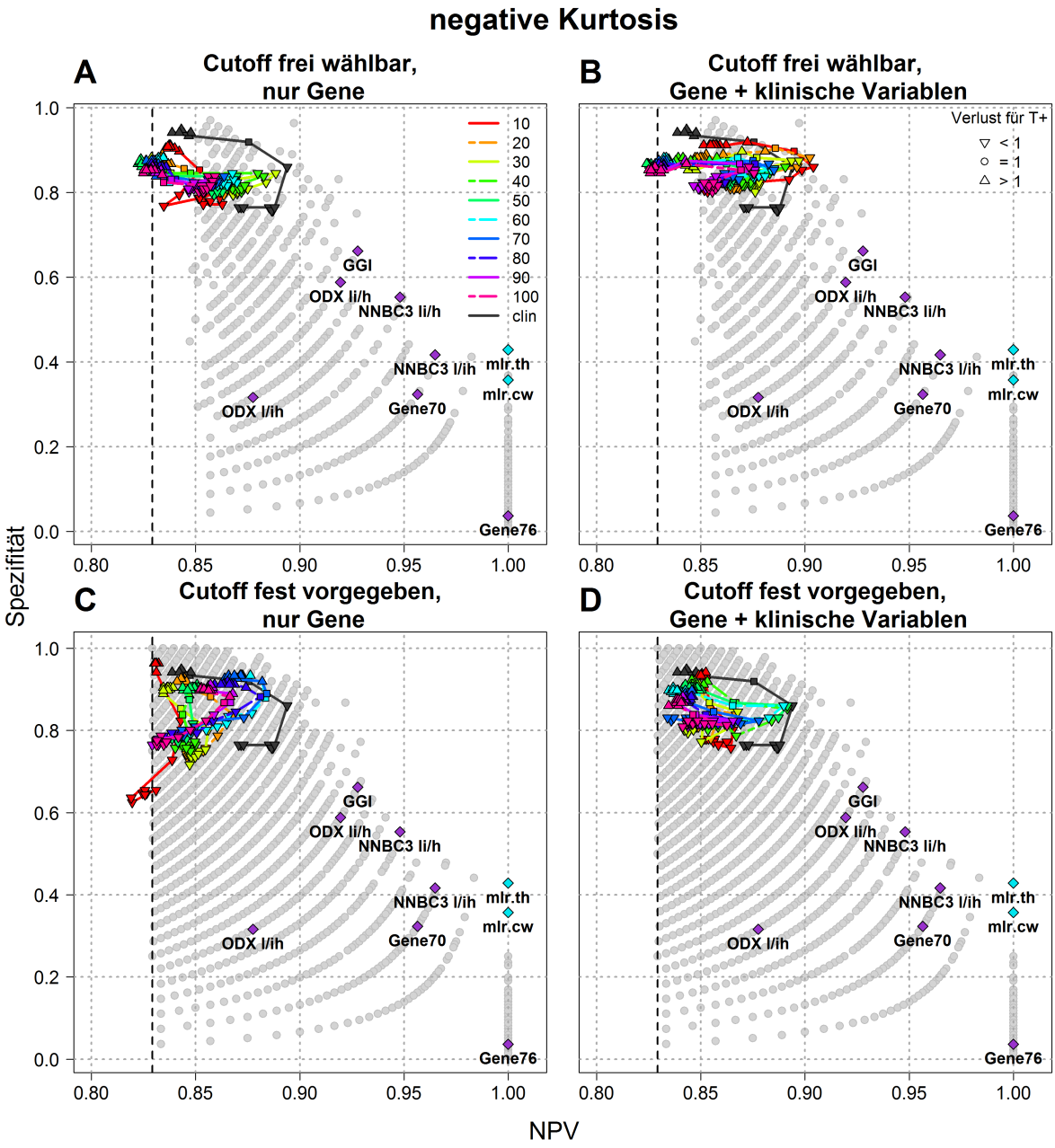

Abbildung D.11: NPV und Spezifität der Klassifikationsbäume mit *minsplit* = 5 für die Top-Gene der negativen Kurtosis.

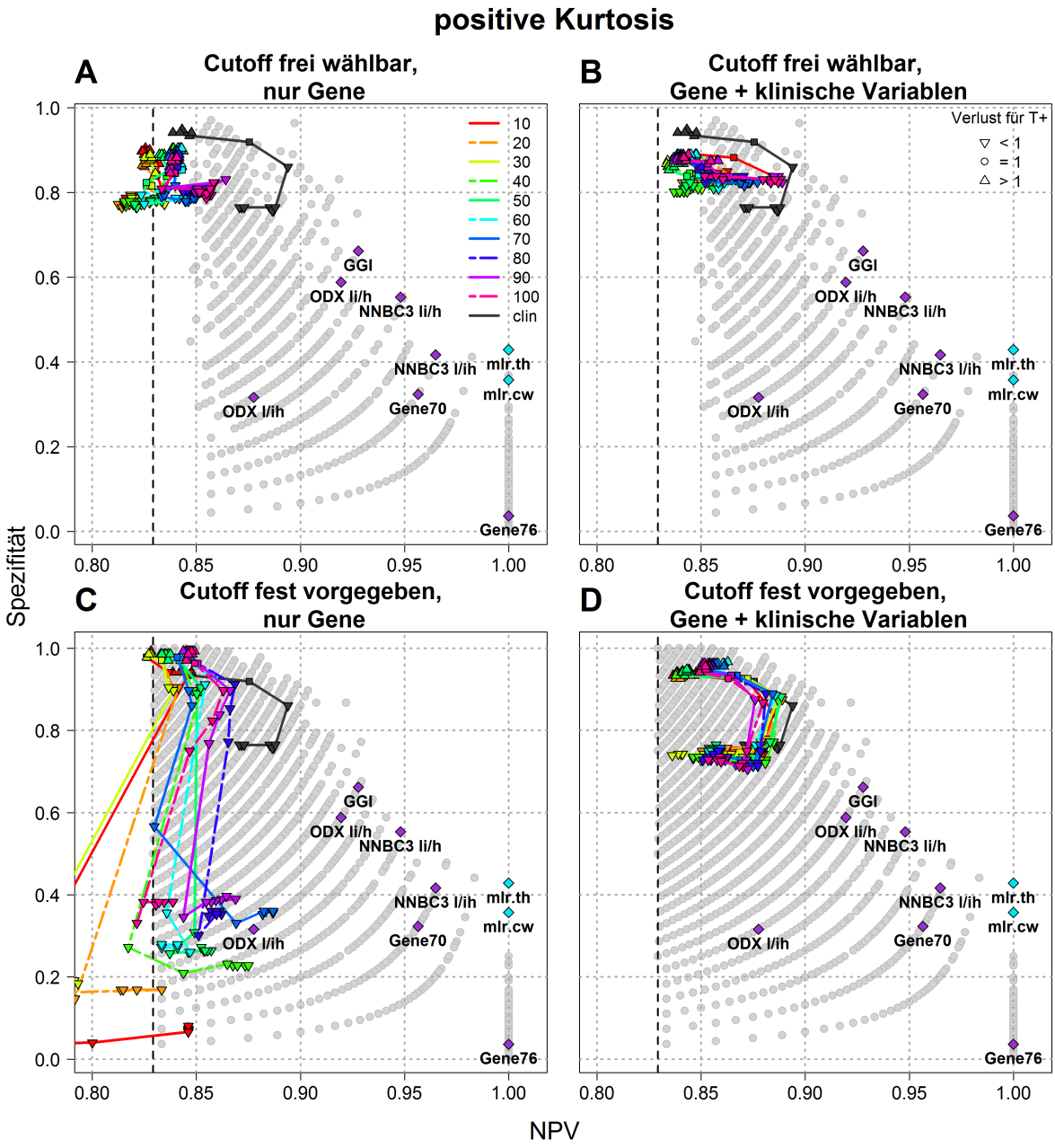

Abbildung D.12: NPV und Spezifität der Klassifikationsbäume mit *minsplit* = 5 für die Top-Gene der positiven Kurtosis.

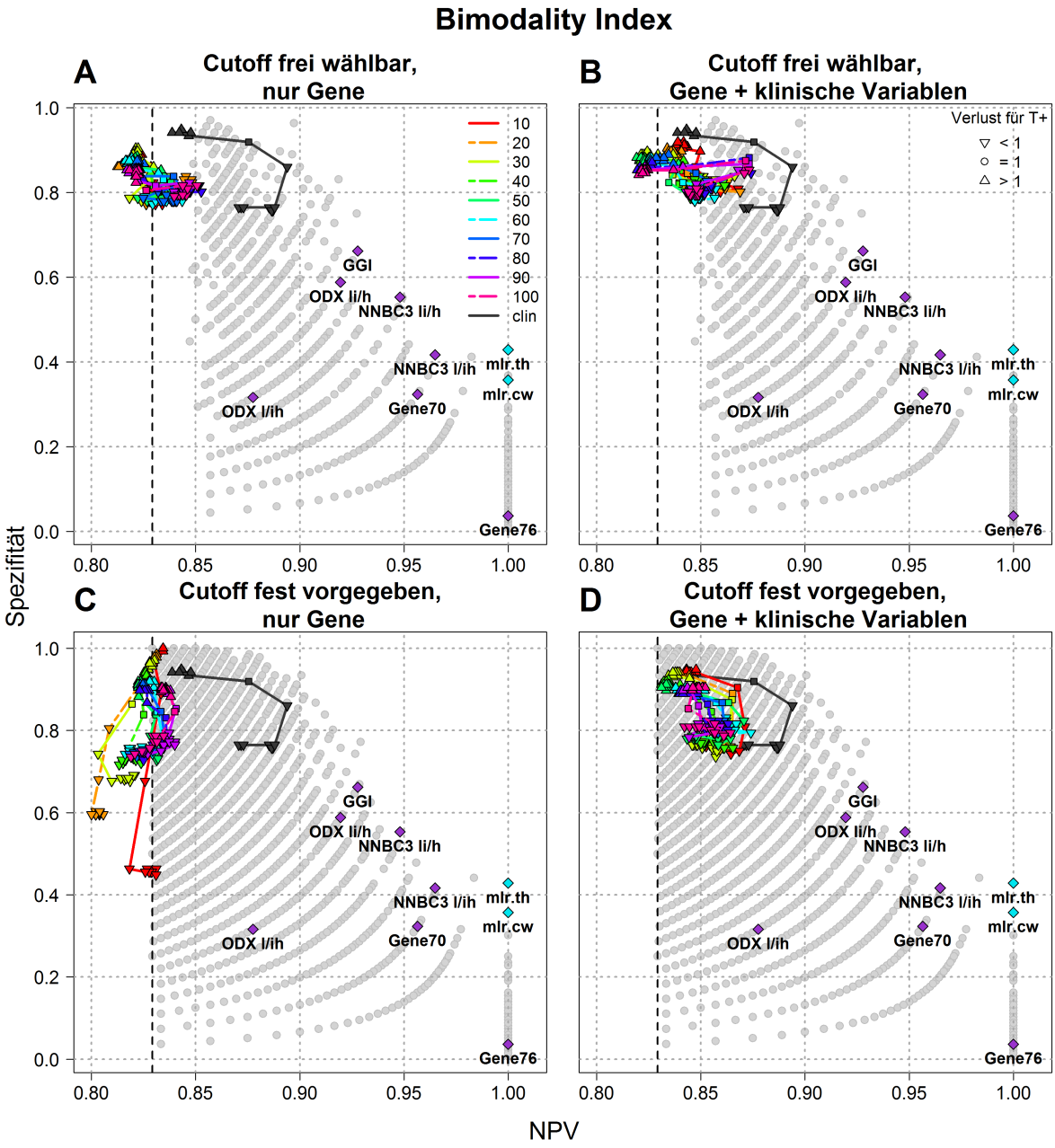

Abbildung D.13: NPV und Spezifität der Klassifikationsbäume mit *minsplit* = 5 für die Top-Gene des Bimodality Index.

E Ergebnisse der Random Forests

E.1 Streudiagramme

E.1.1 Mit frei wählbarem Cutoff

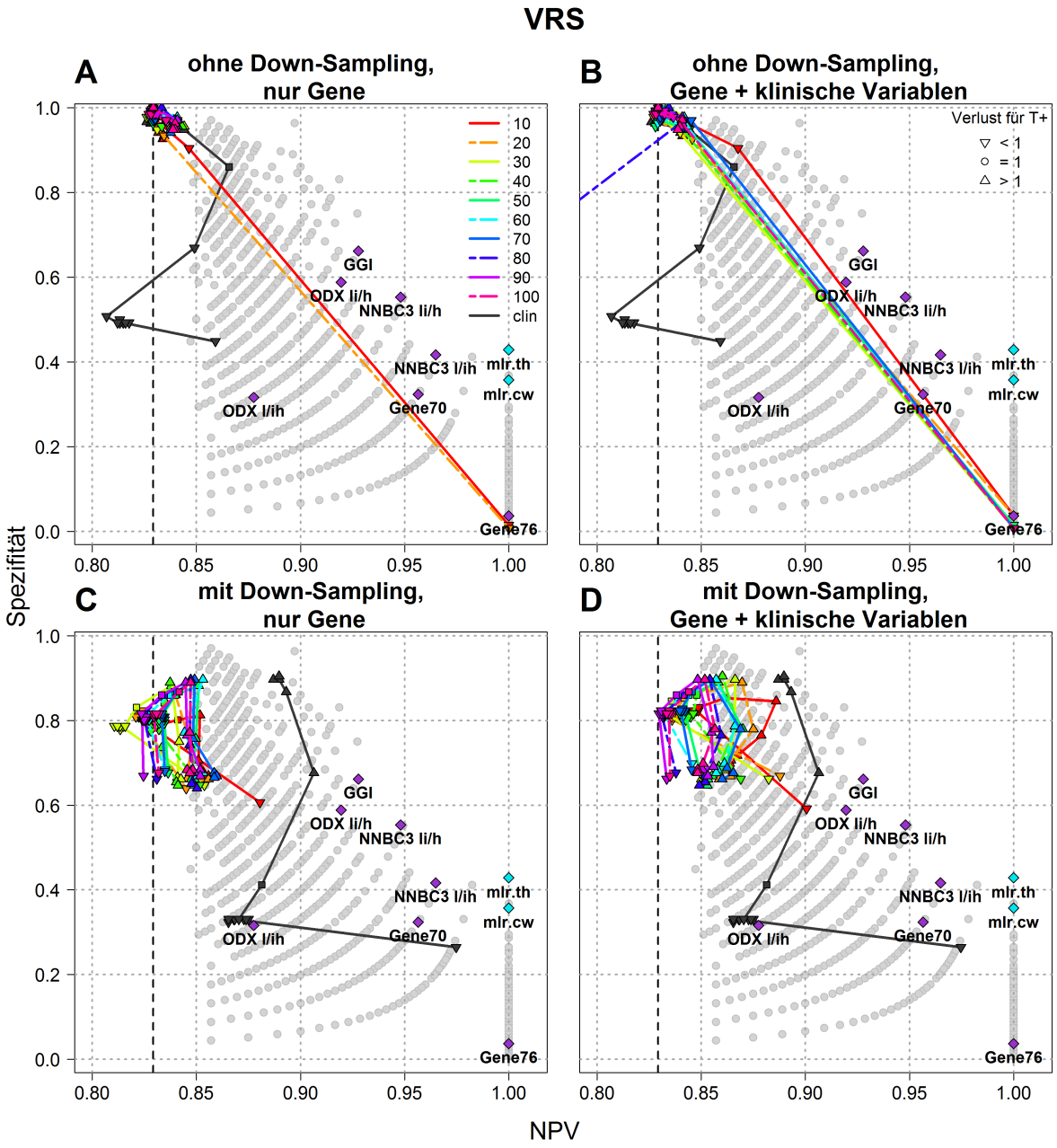

Abbildung E.1: NPV und Spezifität der Random Forests mit frei wählbarem Cutoff für die Top-Gene von VRS.

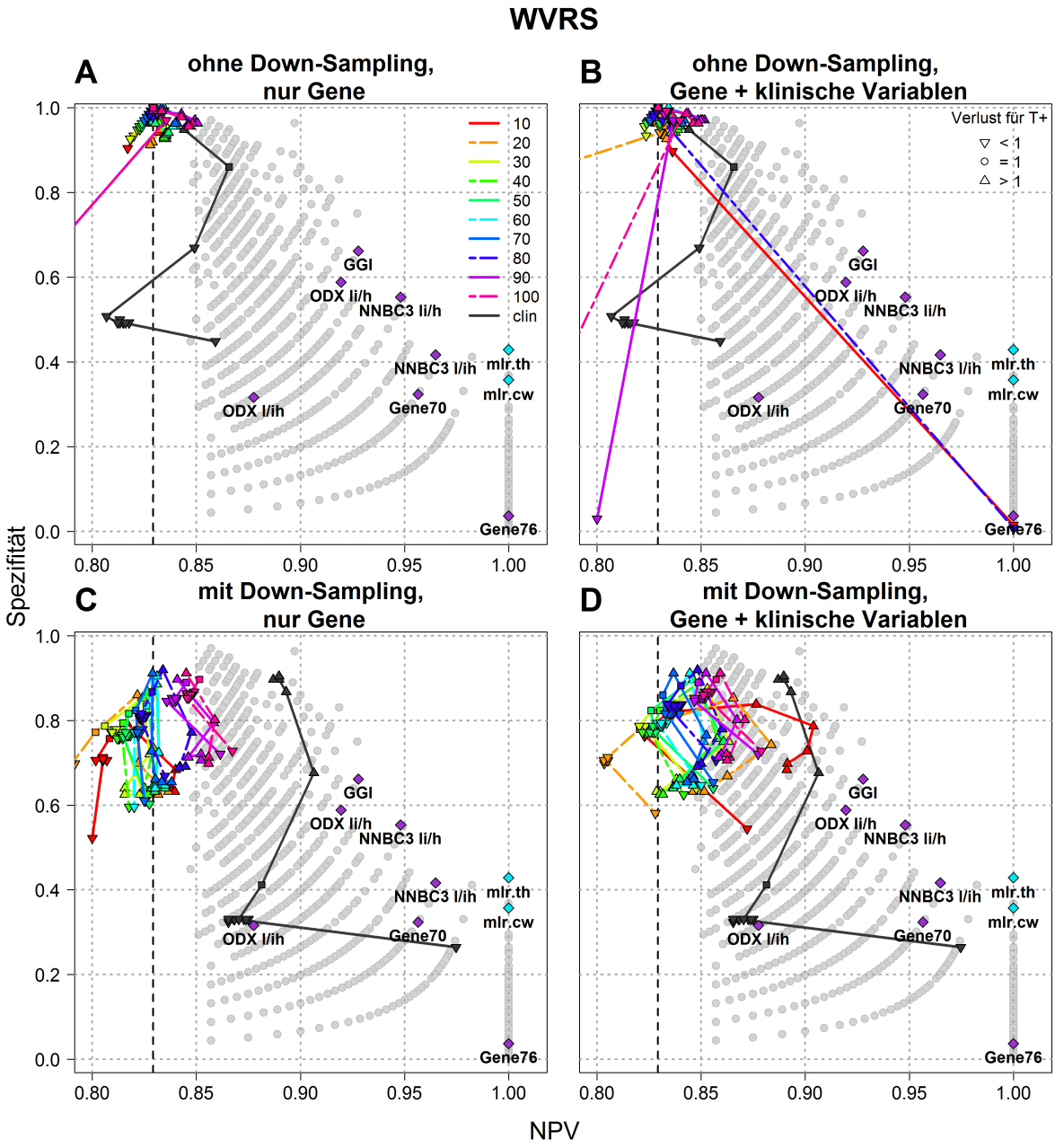

Abbildung E.2: NPV und Spezifität der Random Forests mit frei wählbarem Cutoff für die Top-Gene von WVRS.

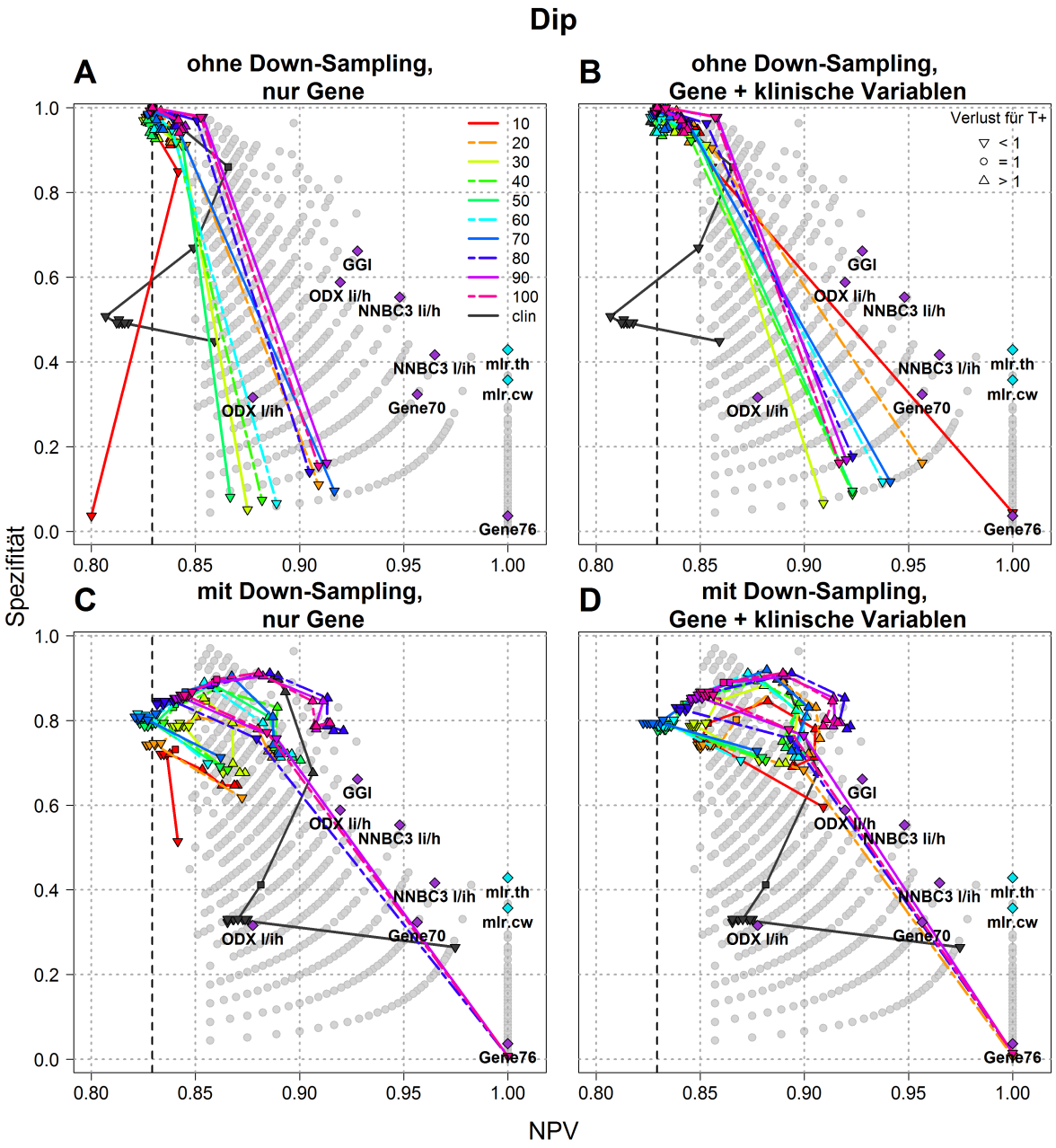

Abbildung E.3: NPV und Spezifität der Random Forests mit frei wählbarem Cutoff für die Top-Gene der dip-Statistik.

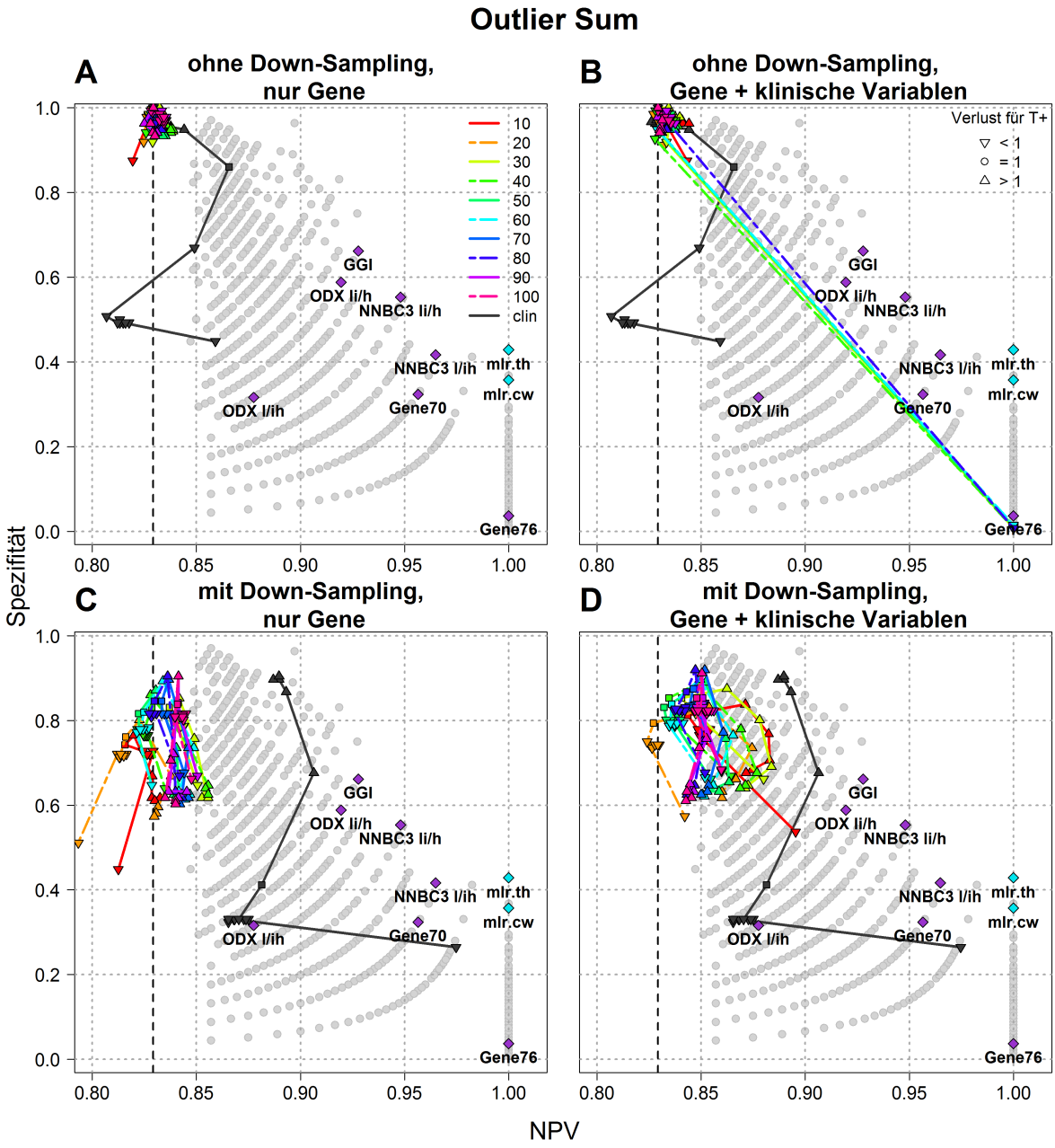

Abbildung E.4: NPV und Spezifität der Random Forests mit frei wählbarem Cutoff für die Top-Gene der Outlier-Sum-Statistik.

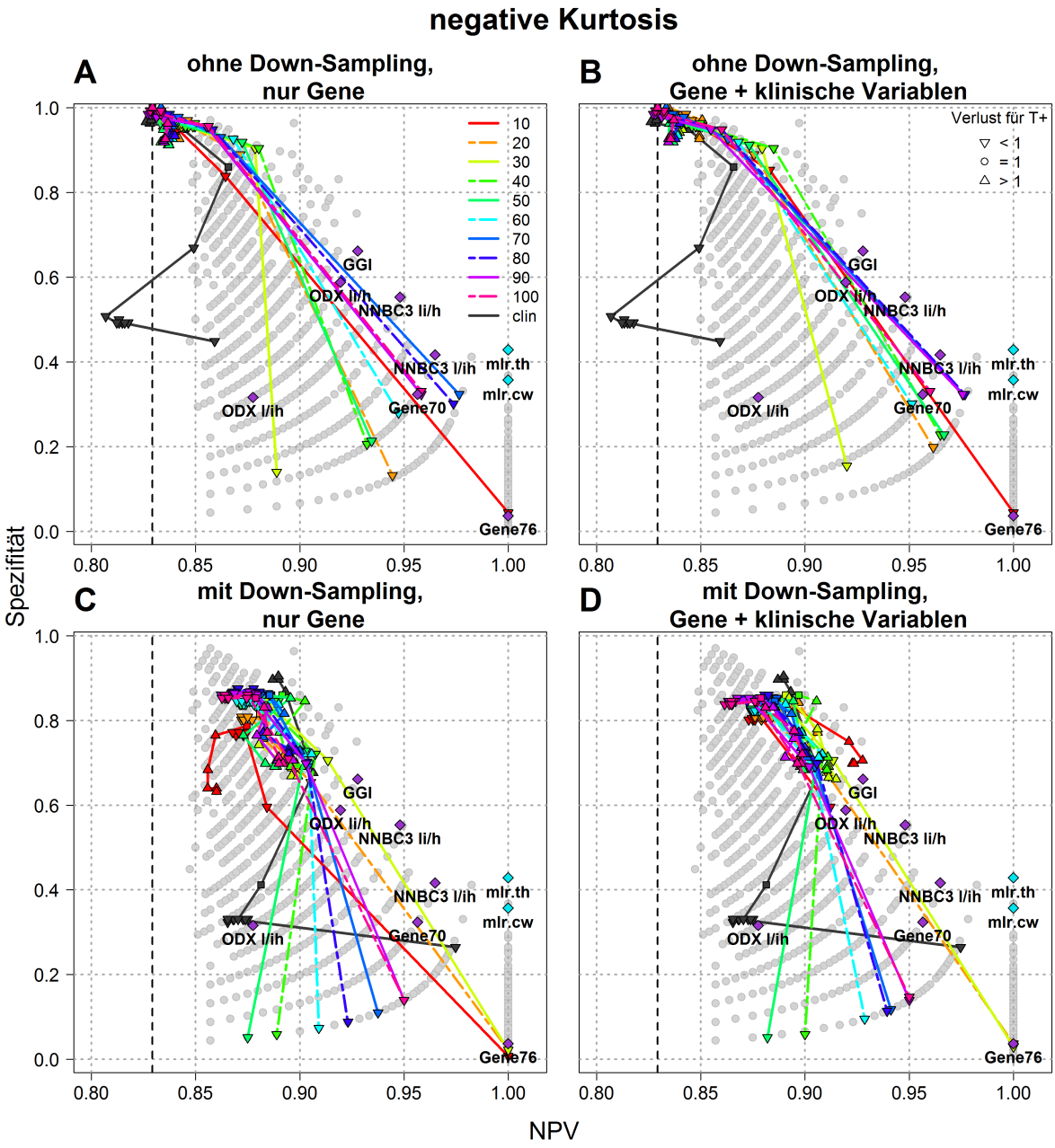

Abbildung E.5: NPV und Spezifität der Random Forests mit frei wählbarem Cutoff für die Top-Gene der negativen Kurtosis.

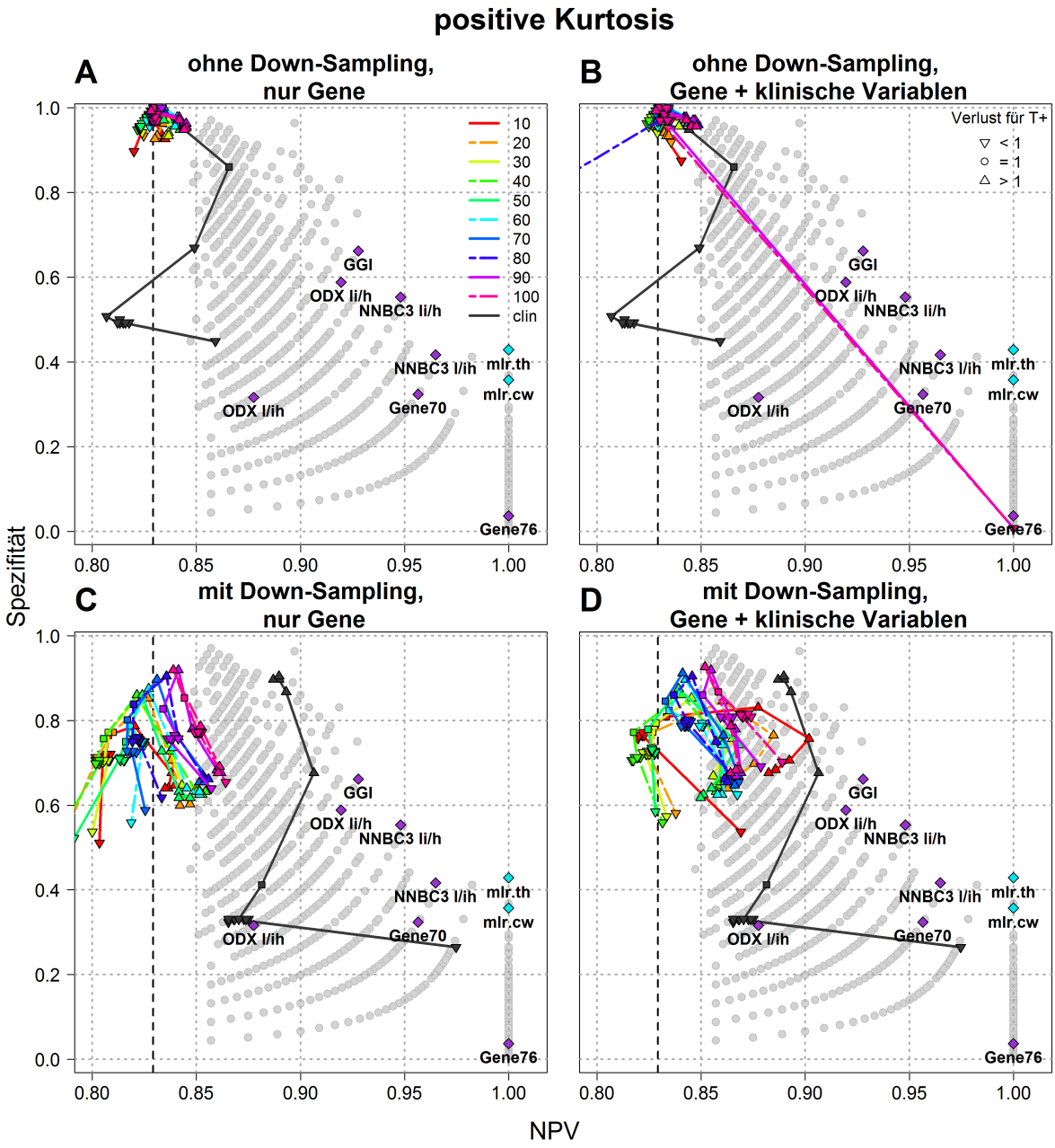

Abbildung E.6: NPV und Spezifität der Random Forests mit frei wählbarem Cutoff für die Top-Gene der positiven Kurtosis.

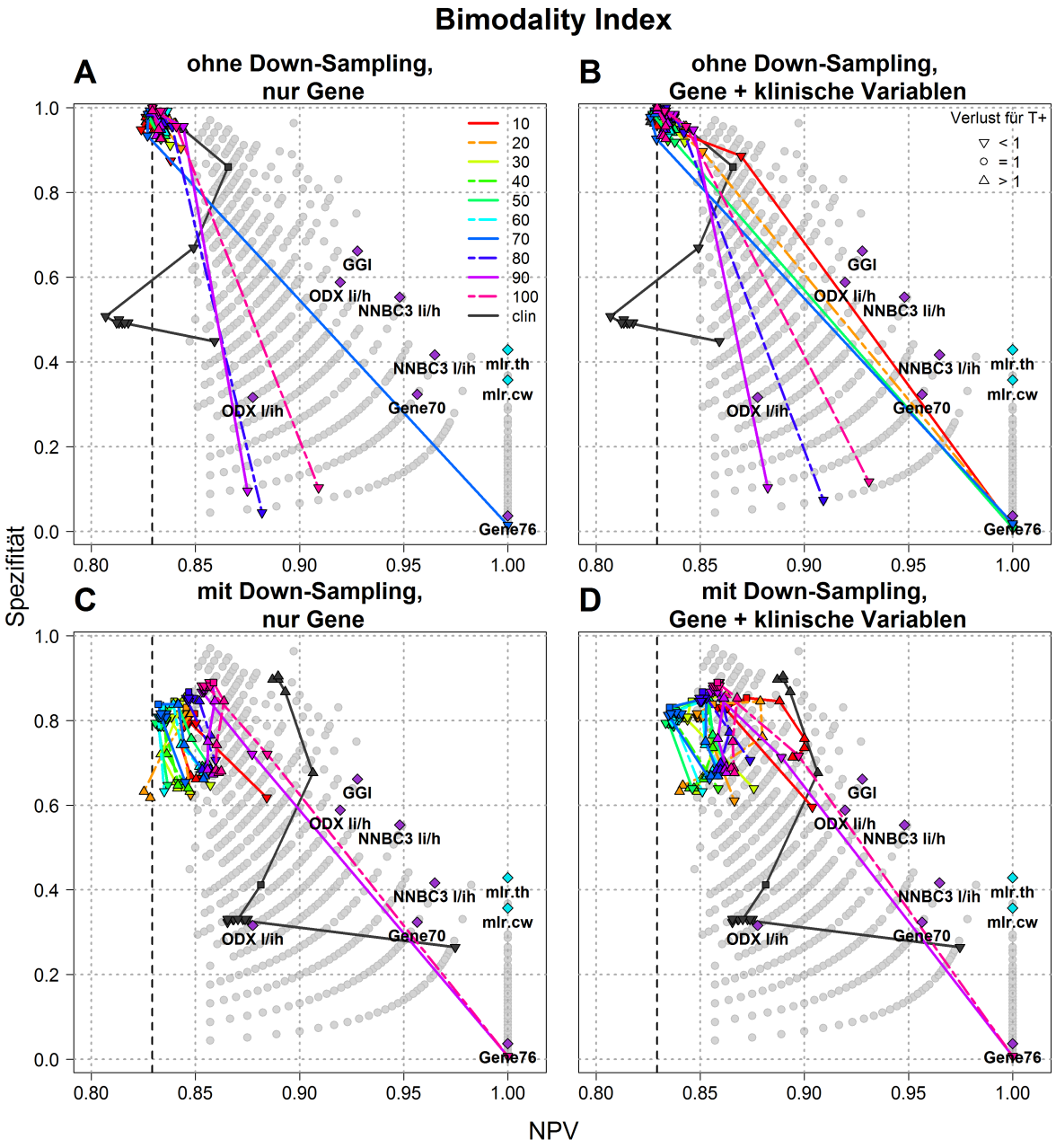

Abbildung E.7: NPV und Spezifität der Random Forests mit frei wählbarem Cutoff für die Top-Gene des Bimodality Index.

E.1.2 Mit fest vorgegebenem Cutoff

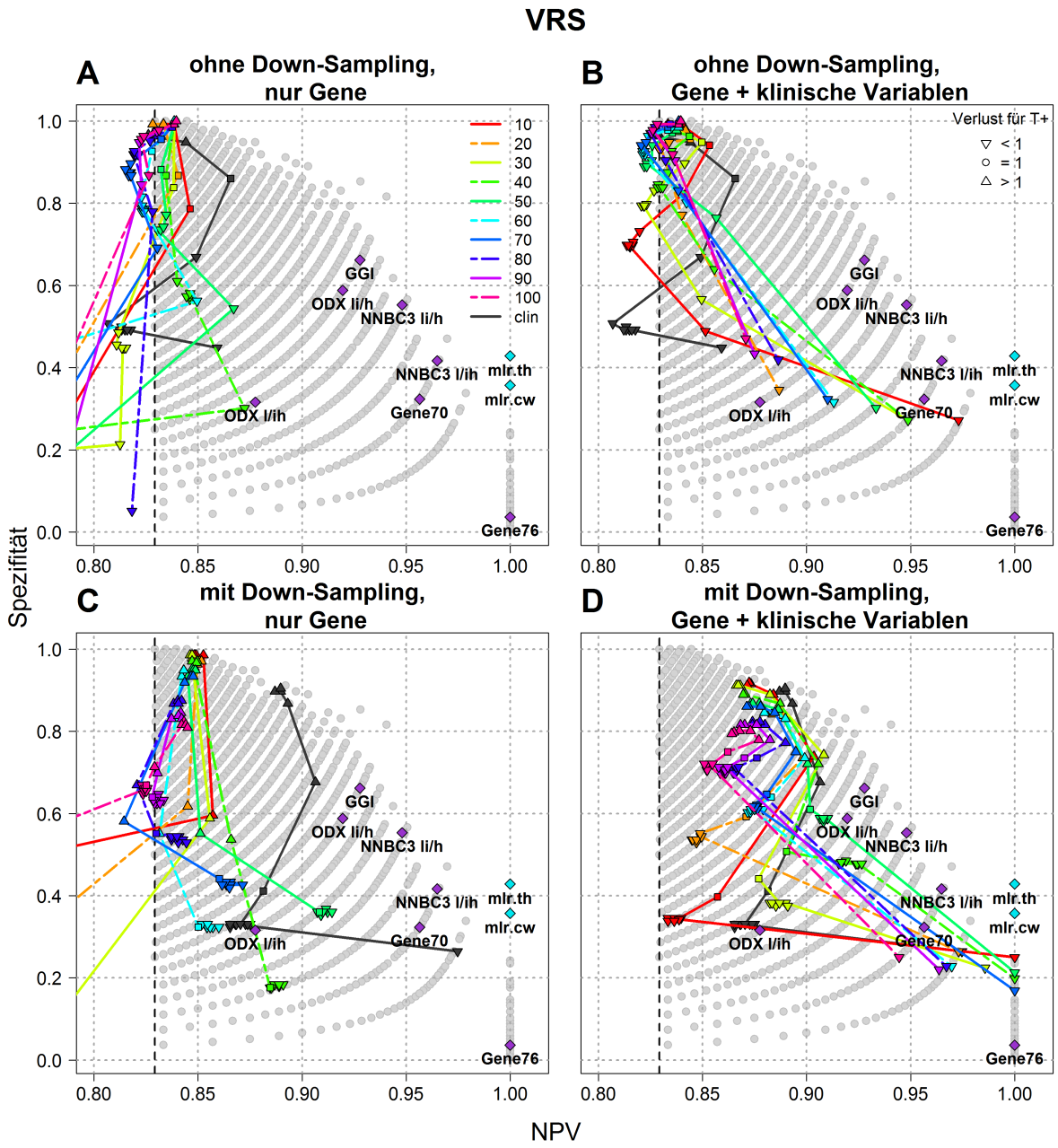

Abbildung E.8: NPV und Spezifität der Random Forests mit fest vorgegebenem Cutoff für die Top-Gene von VRS.

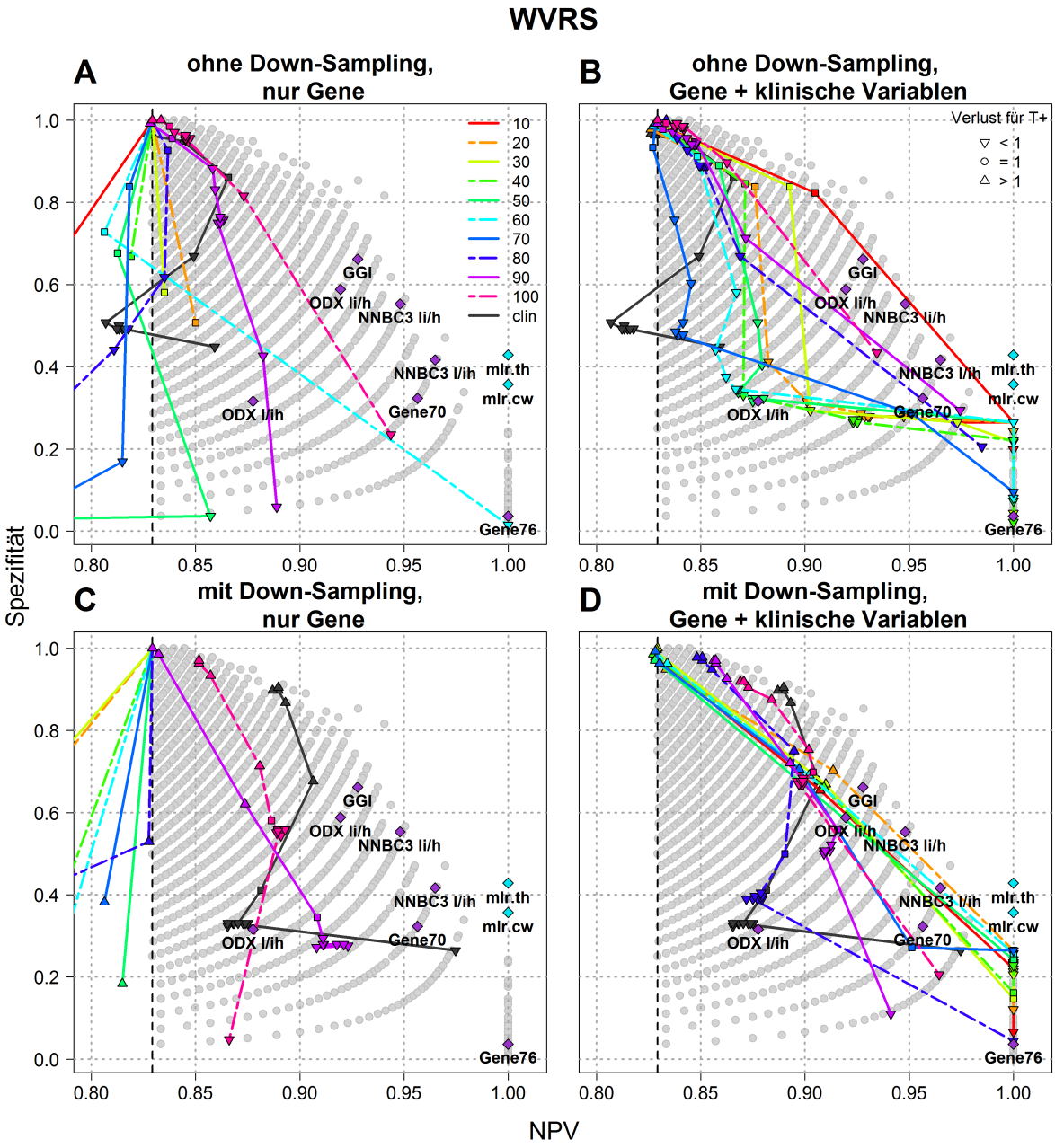

Abbildung E.9: NPV und Spezifität der Random Forests mit fest vorgegebenem Cutoff für die Top-Gene von WVRS.

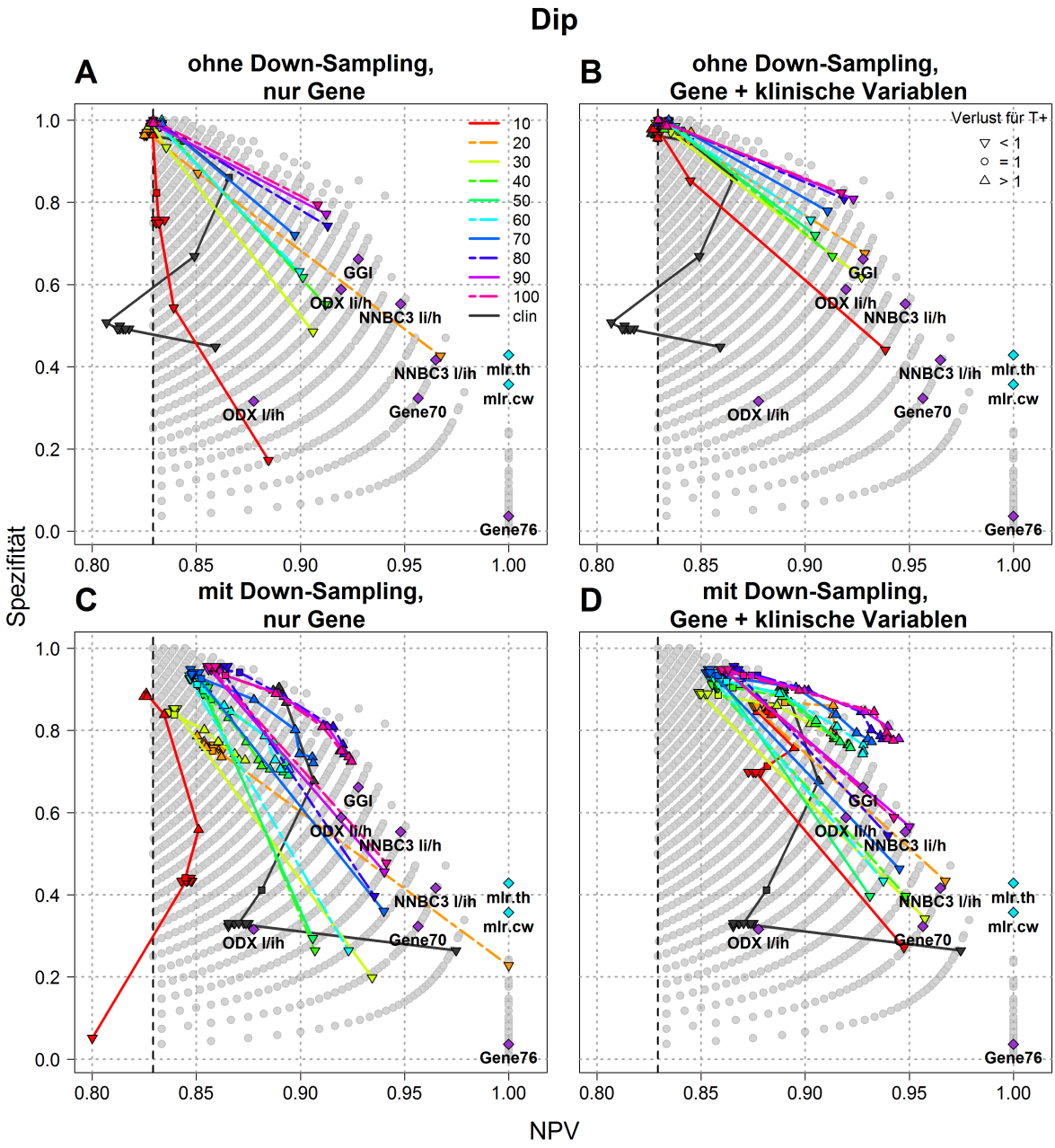

Abbildung E.10: NPV und Spezifität der Random Forests mit fest vorgegebenem Cutoff für die Top-Gene der dip-Statistik.

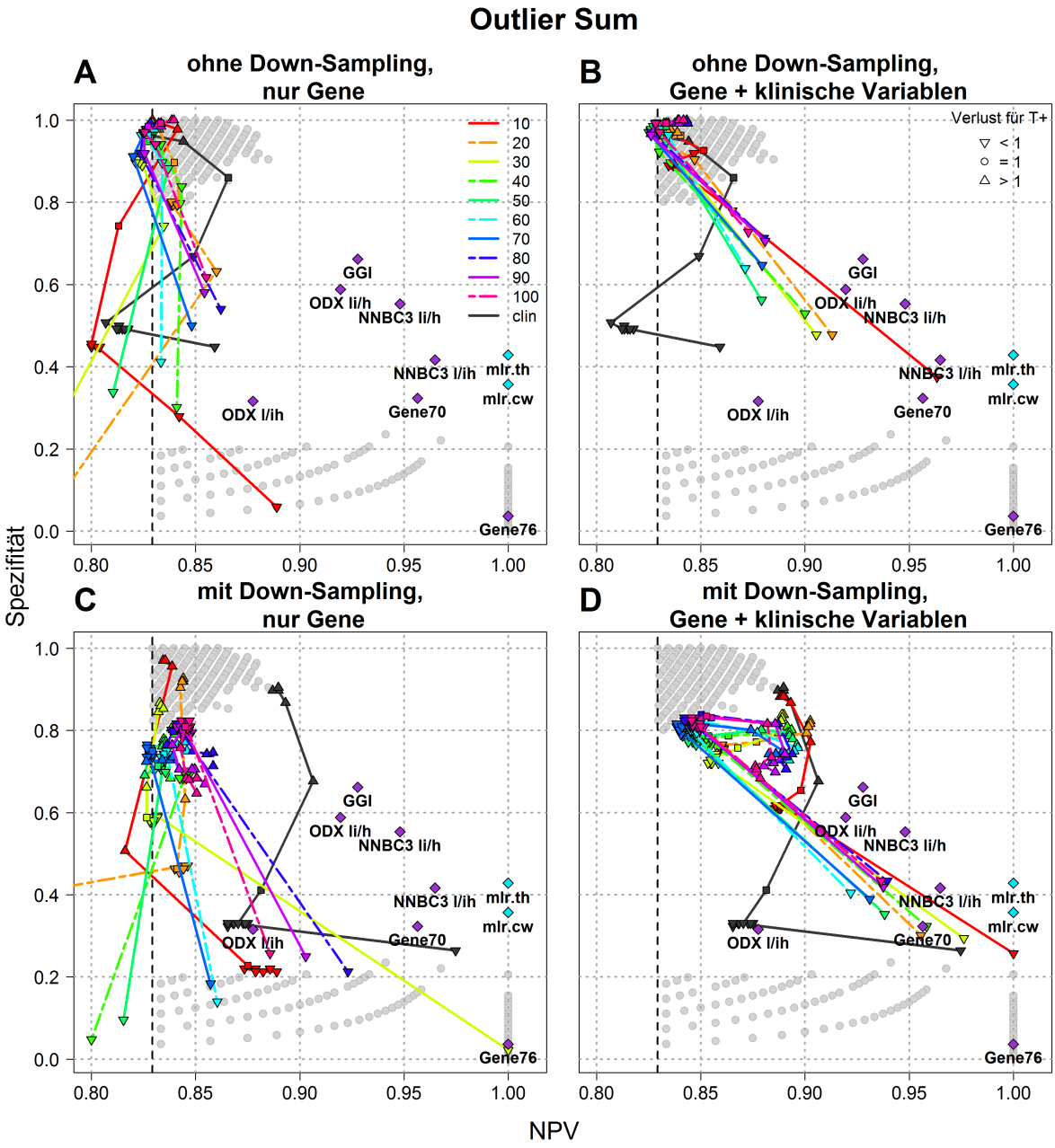

Abbildung E.11: NPV und Spezifität der Random Forests mit fest vorgegebenem Cutoff für die Top-Gene der Outlier-Sum-Statistik.

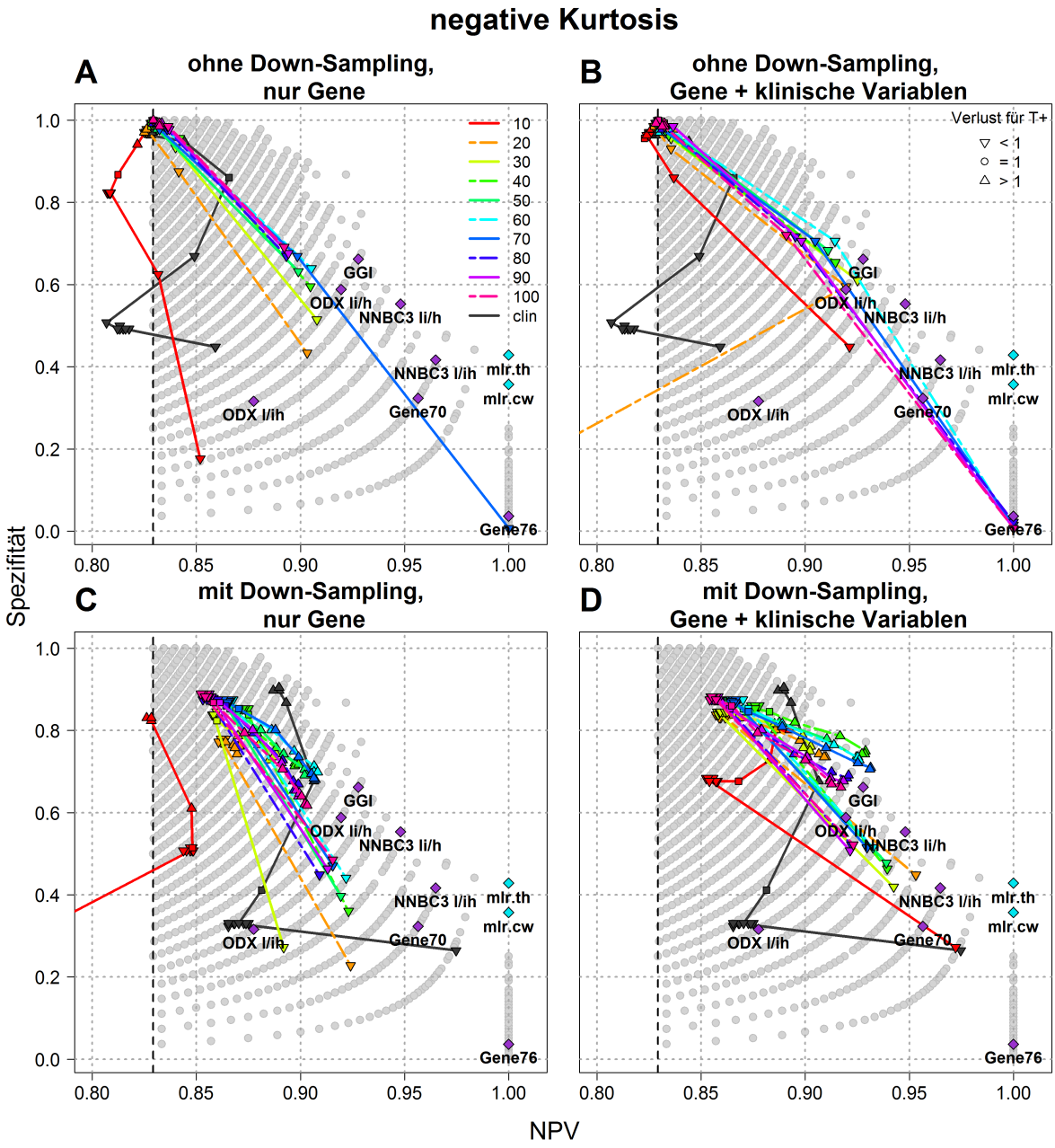

Abbildung E.12: NPV und Spezifität der Random Forests mit mit fest vorgegebenem Cutoff für die Top-Gene der negativen Kurtosis.

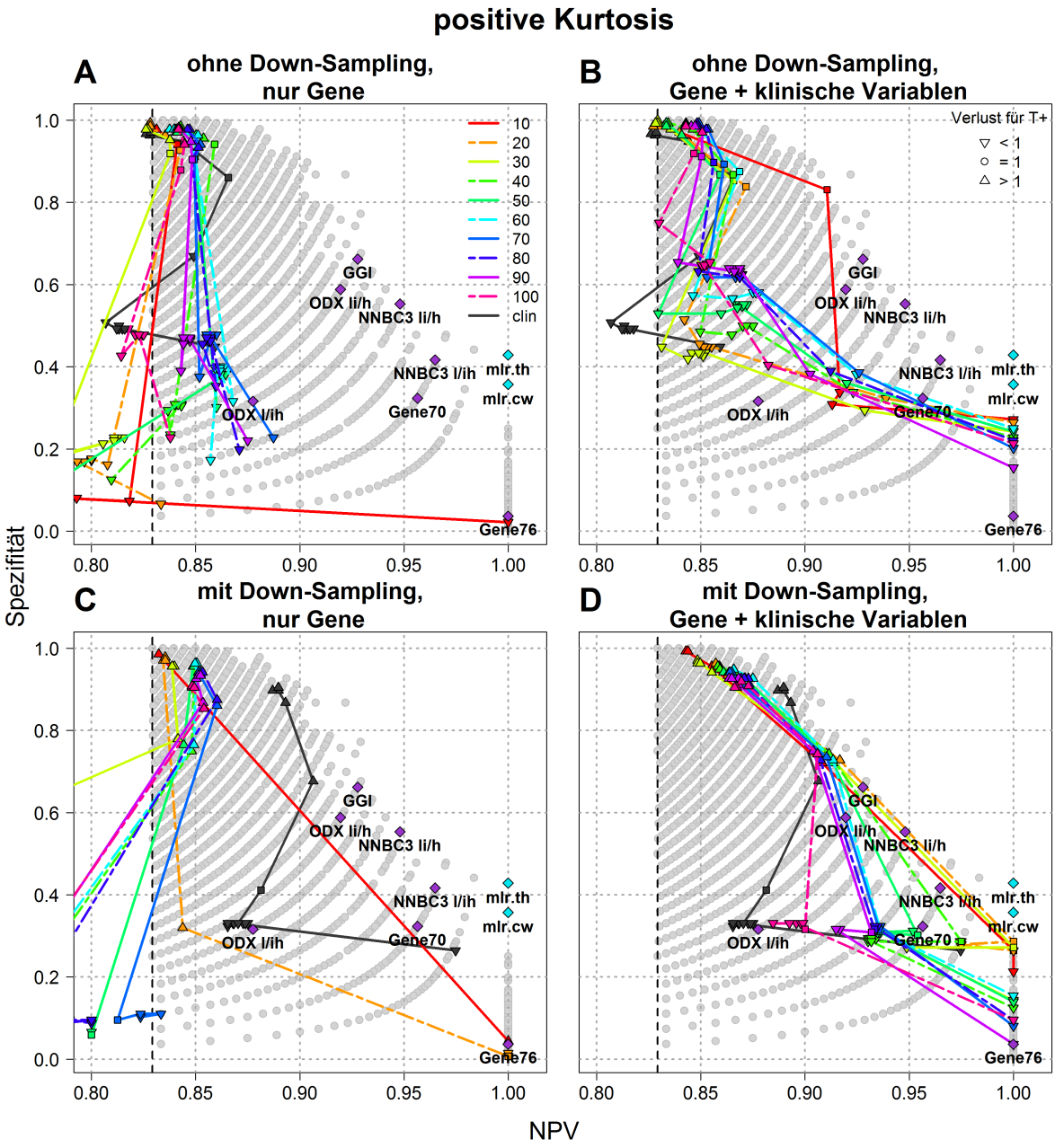

Abbildung E.13: NPV und Spezifität der Random Forests mit fest vorgegebenem Cutoff für die Top-Gene der positiven Kurtosis.

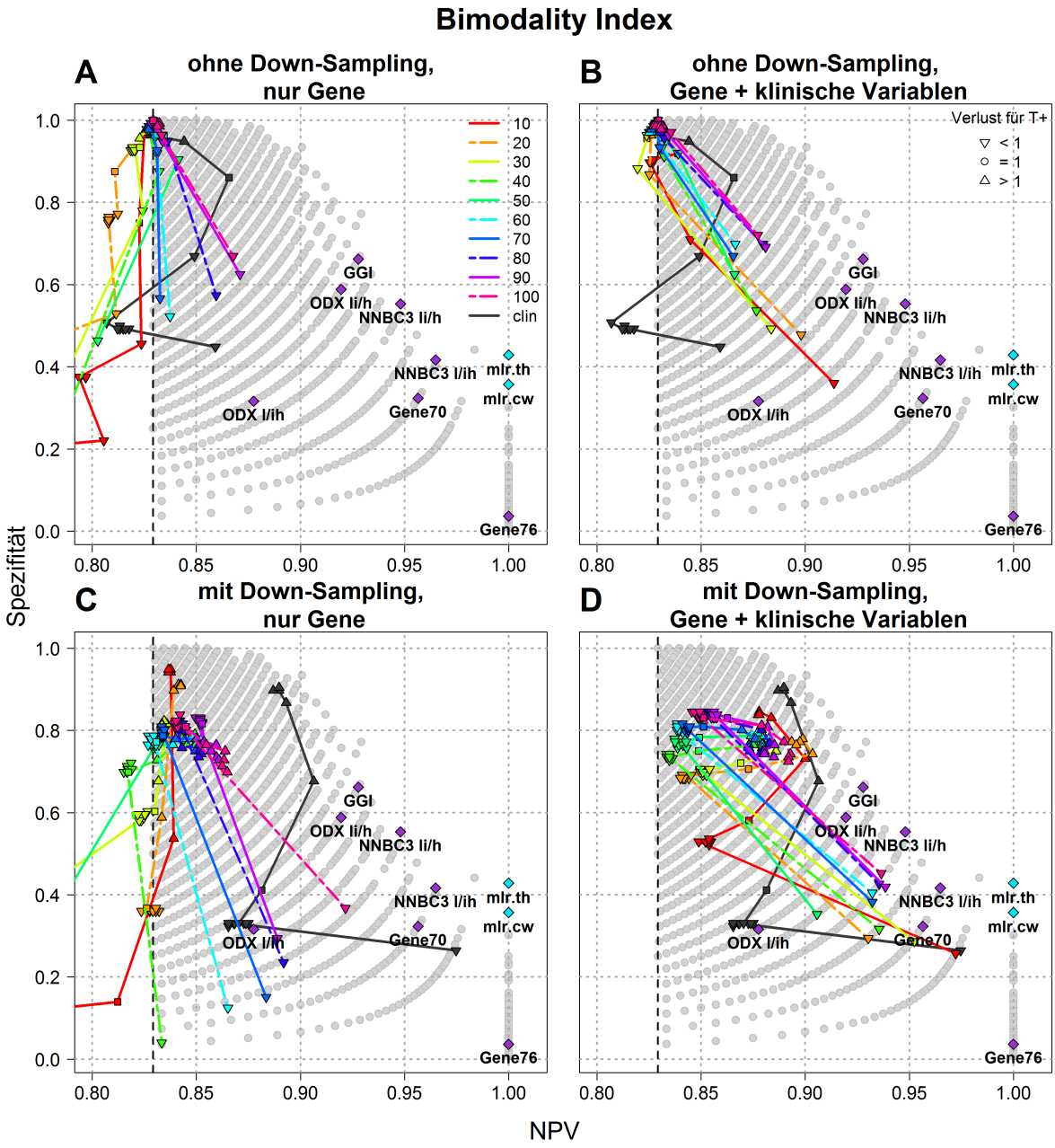

Abbildung E.14: NPV und Spezifität der Random Forest mit fest vorgegebenem Cutoff für die Top-Gene des Bimodality Index.

E.2 Tabellen mit Top3 Random Forests

E.2.1 Mit frei wählbarem Cutoff

Tabelle E.1: Random Forests aus Top-Genen der Bimodalitäts-Scores mit den größten NPV-Werten bei frei wählbaren Cutoffs.

			NPV			Spezifität		
Score	Top	Gewicht	Median	IQR	Spannweite	Median	IQR	Spannweite
negative Kurtosis	10	2^{-22}	**1.000**	0.000	0.250	**0.044**	0.015	0.059
	70	2^{-22}	**0.976**	0.022	0.075	**0.324**	0.022	0.081
	80	2^{-22}	**0.974**	0.024	0.095	**0.301**	0.022	0.081
Likelihood Ratio	10	2^{-22}	**1.000**	0.000	0.333	**0.022**	0.015	0.044
	20	2^{-22}	**1.000**	0.000	1.000	**0.007**	0.007	0.037
	10	2^{10}	**0.854**	0.009	0.023	**0.956**	0.015	0.051
VRS	10	2^{-22}	**1.000**	0.143	1.000	**0.015**	0.022	0.044
	20	2^{-22}	**1.000**	1.000	1.000	**0.007**	0.007	0.015
	10	2^{-20}	**0.846**	0.011	0.043	**0.904**	0.015	0.125
Bimodality Index	70	2^{-22}	**1.000**	0.271	1.000	**0.015**	0.007	0.044
	100	2^{-22}	**0.909**	0.058	0.221	**0.103**	0.029	0.096
	80	2^{-22}	**0.882**	0.167	0.333	**0.044**	0.017	0.066
Dip	70	2^{-22}	**0.917**	0.044	0.263	**0.096**	0.029	0.096
	90	2^{-22}	**0.913**	0.015	0.117	**0.162**	0.029	0.110
	100	2^{-22}	**0.909**	0.025	0.149	**0.154**	0.022	0.088
WVRS	90	2^{10}	**0.851**	0.006	0.015	**0.963**	0.015	0.044
	90	2^{8}	**0.850**	0.008	0.022	**0.963**	0.015	0.051
	90	2^{6}	**0.849**	0.005	0.022	**0.971**	0.022	0.044
positive Kurtosis	90	2^{6}	**0.845**	0.007	0.021	**0.963**	0.015	0.059
	90	2^{8}	**0.845**	0.009	0.021	**0.963**	0.015	0.044
	90	2^{10}	**0.845**	0.010	0.031	**0.963**	0.015	0.059
Outlier Sum	40	2^{6}	**0.839**	0.008	0.025	**0.949**	0.015	0.066
	30	2^{10}	**0.839**	0.007	0.021	**0.945**	0.022	0.059
	30	2^{8}	**0.838**	0.007	0.026	**0.949**	0.017	0.066

Tabelle E.2: Random Forests aus Top-Genen der Bimodalitäts-Scores und klinischen Variablen mit den größten NPV-Werten bei frei wählbaren Cutoffs.

			NPV			Spezifität		
Score	Top	Gewicht	Median	IQR	Spannweite	Median	IQR	Spannweite
Outlier Sum	60	2^{-22}	**1.000**	0.000	1.000	**0.015**	0.007	0.037
	50	2^{-22}	**1.000**	0.000	1.000	**0.011**	0.007	0.029
	30	2^{-22}	**1.000**	1.000	1.000	**0.007**	0.015	0.029
VRS	10	2^{-22}	**1.000**	0.000	0.250	**0.037**	0.017	0.059
	20	2^{-22}	**1.000**	0.000	0.286	**0.037**	0.015	0.066
	50	2^{-22}	**1.000**	0.000	1.000	**0.015**	0.015	0.037
Bimodality Index	70	2^{-22}	**1.000**	0.000	1.000	**0.018**	0.009	0.044
	20	2^{-22}	**1.000**	0.000	1.000	**0.015**	0.015	0.044
	10	2^{-22}	**1.000**	1.000	1.000	**0.007**	0.007	0.022
Likelihood Ratio	20	2^{-22}	**1.000**	0.000	0.200	**0.051**	0.022	0.051
	10	2^{-22}	**1.000**	0.000	0.000	**0.029**	0.009	0.051
	30	2^{-22}	**1.000**	1.000	1.000	**0.007**	0.015	0.029
WVRS	10	2^{-22}	**1.000**	0.000	1.000	**0.015**	0.007	0.029
	80	2^{-22}	**1.000**	1.000	1.000	**0.007**	0.015	0.029
	90	2^{6}	**0.852**	0.007	0.027	**0.971**	0.015	0.051
positive Kurtosis	90	2^{-22}	**1.000**	1.000	1.000	**0.007**	0.015	0.029
	100	2^{-22}	**1.000**	1.000	1.000	**0.007**	0.015	0.037
	80	2^{6}	**0.849**	0.006	0.026	**0.960**	0.015	0.051
negative Kurtosis	10	2^{-22}	**1.000**	0.000	0.200	**0.044**	0.015	0.044
	80	2^{-22}	**0.977**	0.022	0.065	**0.324**	0.022	0.074
	70	2^{-22}	**0.976**	0.022	0.068	**0.324**	0.022	0.081
Dip	10	2^{-22}	**1.000**	0.000	0.400	**0.044**	0.015	0.088
	20	2^{-22}	**0.957**	0.014	0.136	**0.162**	0.029	0.096
	70	2^{-22}	**0.941**	0.021	0.158	**0.118**	0.029	0.088

Tabelle E.3: Random Forests aus Top-Genen der Bimodalitäts-Scores mit den größten NPV-Werten bei frei wählbaren Cutoffs und Verwendung von Down-Sampling.

			NPV			Spezifität		
Score	Top	Gewicht	Median	IQR	Spannweite	Median	IQR	Spannweite
negative Kurtosis	20	2^{-20}	**1.000**	0.000	1.000	**0.022**	0.015	0.051
	30	2^{-20}	**1.000**	0.000	0.333	**0.022**	0.015	0.044
	10	2^{-20}	**1.000**	1.000	1.000	**0.007**	0.007	0.015
Dip	80	2^{-20}	**1.000**	0.500	1.000	**0.007**	0.007	0.029
	90	2^{-20}	**1.000**	1.000	1.000	**0.007**	0.015	0.022
	100	2^{-20}	**1.000**	1.000	1.000	**0.007**	0.015	0.029
Bimodality Index	90	2^{-20}	**1.000**	1.000	1.000	**0.007**	0.007	0.015
	100	2^{-20}	**1.000**	1.000	1.000	**0.007**	0.015	0.029
	100	2^{-18}	**0.884**	0.020	0.062	**0.721**	0.029	0.096
Likelihood Ratio	10	2^{-18}	**0.880**	0.020	0.063	**0.610**	0.029	0.118
	10	2^{4}	**0.871**	0.014	0.049	**0.735**	0.029	0.096
	10	2^{6}	**0.868**	0.013	0.048	**0.684**	0.024	0.088
positive Kurtosis	100	2^{-18}	**0.864**	0.018	0.068	**0.654**	0.031	0.125
	90	2^{10}	**0.862**	0.017	0.055	**0.676**	0.029	0.096
	100	2^{6}	**0.861**	0.014	0.040	**0.691**	0.022	0.074
VRS	10	2^{-18}	**0.880**	0.020	0.057	**0.607**	0.029	0.125
	70	2^{8}	**0.860**	0.014	0.047	**0.669**	0.022	0.096
	70	2^{10}	**0.859**	0.015	0.047	**0.665**	0.029	0.088
WVRS	100	2^{-18}	**0.867**	0.015	0.060	**0.728**	0.031	0.118
	90	2^{-18}	**0.861**	0.018	0.055	**0.721**	0.037	0.118
	100	2^{4}	**0.859**	0.011	0.039	**0.801**	0.017	0.074
Outlier Sum	40	2^{6}	**0.856**	0.015	0.052	**0.647**	0.029	0.081
	40	2^{10}	**0.856**	0.014	0.051	**0.625**	0.022	0.088
	30	2^{10}	**0.856**	0.012	0.046	**0.618**	0.015	0.066

Tabelle E.4: Random Forests aus Top-Genen der Bimodalitäts-Scores und klinischen Variablen mit den größten NPV-Werten bei frei wählbaren Cutoffs und Verwendung von Down-Sampling.

			NPV			Spezifität		
Score	Top	Gewicht	Median	IQR	Spannweite	Median	IQR	Spannweite
Dip	80	2^{-20}	**1.000**	0.000	1.000	**0.015**	0.007	0.059
	90	2^{-20}	**1.000**	0.000	1.000	**0.015**	0.015	0.044
	100	2^{-20}	**1.000**	0.000	1.000	**0.015**	0.007	0.037
negative Kurtosis	20	2^{-20}	**1.000**	0.000	0.200	**0.037**	0.015	0.051
	30	2^{-20}	**1.000**	0.000	0.250	**0.029**	0.007	0.044
	100	2^{-20}	**0.950**	0.009	0.073	**0.147**	0.029	0.088
Bimodality Index	90	2^{-20}	**1.000**	1.000	1.000	**0.007**	0.007	0.022
	100	2^{-20}	**1.000**	1.000	1.000	**0.007**	0.009	0.022
	10	2^{-18}	**0.904**	0.028	0.090	**0.596**	0.031	0.110
Likelihood Ratio	10	2^{-18}	**0.910**	0.018	0.079	**0.596**	0.037	0.118
	10	2^{6}	**0.907**	0.015	0.063	**0.743**	0.022	0.096
	10	2^{8}	**0.905**	0.014	0.063	**0.721**	0.029	0.081
WVRS	10	2^{4}	**0.904**	0.011	0.036	**0.787**	0.022	0.088
	10	2^{6}	**0.901**	0.006	0.029	**0.728**	0.022	0.110
	10	2^{8}	**0.891**	0.012	0.045	**0.699**	0.029	0.088
positive Kurtosis	10	2^{4}	**0.902**	0.011	0.046	**0.757**	0.022	0.081
	10	2^{6}	**0.892**	0.011	0.042	**0.706**	0.022	0.088
	100	2^{-18}	**0.889**	0.015	0.062	**0.702**	0.022	0.118
VRS	10	2^{-18}	**0.901**	0.019	0.057	**0.592**	0.037	0.118
	20	2^{-18}	**0.888**	0.015	0.075	**0.669**	0.022	0.125
	10	2^{2}	**0.886**	0.009	0.031	**0.846**	0.022	0.088
Outlier Sum	10	2^{-18}	**0.895**	0.022	0.072	**0.537**	0.029	0.103
	30	2^{6}	**0.884**	0.012	0.061	**0.691**	0.024	0.110
	10	2^{6}	**0.883**	0.012	0.048	**0.706**	0.024	0.088

E.2.2 Mit fest vorgegebenem Cutoff

Tabelle E.5: Random Forests aus Top-Genen der Bimodalitäts-Scores mit den größten NPV-Werten bei fest vorgegebenen Cutoffs.

			NPV			Spezifität		
Score	Top	Gewicht	Median	IQR	Spannweite	Median	IQR	Spannweite
Likelihood Ratio	30	2^{-20}	**1.000**	0.048	0.120	**0.147**	0.015	0.059
	50	2^{-20}	**0.938**	0.032	0.171	**0.221**	0.022	0.103
	40	2^{-20}	**0.926**	0.042	0.148	**0.169**	0.022	0.074
WVRS	60	2^{-2}	**1.000**	0.333	1.000	**0.015**	0.015	0.044
	100	2^{-20}	**0.944**	0.035	0.104	**0.235**	0.029	0.110
	90	2^{-20}	**0.889**	0.025	0.196	**0.059**	0.015	0.059
Dip	20	2^{-20}	**0.967**	0.031	0.068	**0.426**	0.022	0.081
	80	2^{-20}	**0.913**	0.012	0.049	**0.743**	0.029	0.088
	90	2^{-20}	**0.912**	0.011	0.047	**0.772**	0.029	0.096
negative Kurtosis	70	2^{-22}	**1.000**	1.000	1.000	**0.007**	0.015	0.029
	30	2^{-20}	**0.908**	0.018	0.054	**0.515**	0.029	0.132
	60	2^{-20}	**0.905**	0.012	0.050	**0.640**	0.022	0.088
positive Kurtosis	10	2^{-18}	**1.000**	0.000	1.000	**0.022**	0.015	0.037
	70	2^{-18}	**0.887**	0.043	0.112	**0.228**	0.022	0.088
	90	2^{-18}	**0.875**	0.021	0.088	**0.221**	0.029	0.088
Outlier Sum	10	2^{-20}	**0.889**	0.069	0.500	**0.059**	0.029	0.066
	80	2^{-20}	**0.862**	0.018	0.066	**0.540**	0.024	0.103
	20	2^{-18}	**0.860**	0.014	0.045	**0.632**	0.024	0.125
Bimodality Index	90	2^{-20}	**0.871**	0.015	0.067	**0.625**	0.029	0.110
	100	2^{-20}	**0.868**	0.017	0.067	**0.669**	0.022	0.096
	80	2^{-20}	**0.860**	0.023	0.093	**0.574**	0.029	0.096
VRS	40	2^{-18}	**0.872**	0.032	0.143	**0.301**	0.029	0.110
	50	2^{-18}	**0.867**	0.021	0.089	**0.544**	0.044	0.096
	60	2^{-18}	**0.849**	0.022	0.082	**0.562**	0.044	0.132

Tabelle E.6: Random Forests aus Top-Genen der Bimodalitäts-Scores und klinischen Variablen mit den größten NPV-Werten bei fest vorgegebenen Cutoffs.

			NPV			Spezifität		
Score	Top	Gewicht	Median	IQR	Spannweite	Median	IQR	Spannweite
negative Kurtosis	80	2^{-22}	**1.000**	0.000	1.000	**0.022**	0.007	0.044
	70	2^{-22}	**1.000**	0.000	1.000	**0.015**	0.015	0.051
	60	2^{-22}	**1.000**	1.000	1.000	**0.007**	0.015	0.037
positive Kurtosis	10	2^{-18}	**1.000**	0.000	0.071	**0.272**	0.009	0.022
	10	2^{-20}	**1.000**	0.000	0.053	**0.265**	0.007	0.037
	20	2^{-20}	**1.000**	0.000	0.026	**0.265**	0.000	0.015
WVRS	10	2^{-18}	**1.000**	0.000	0.026	**0.265**	0.007	0.029
	10	2^{-4}	**1.000**	0.000	0.051	**0.265**	0.000	0.015
	10	2^{-2}	**1.000**	0.000	0.053	**0.265**	0.000	0.022
Likelihood Ratio	30	2^{-20}	**1.000**	0.000	0.047	**0.324**	0.015	0.051
	10	2^{-20}	**1.000**	0.000	0.053	**0.265**	0.000	0.029
	50	2^{-20}	**0.980**	0.020	0.085	**0.368**	0.015	0.088
VRS	10	2^{-20}	**0.973**	0.026	0.098	**0.272**	0.007	0.029
	30	2^{-20}	**0.949**	0.003	0.052	**0.272**	0.015	0.037
	40	2^{-20}	**0.949**	0.003	0.087	**0.272**	0.007	0.044
Dip	10	2^{-20}	**0.938**	0.017	0.059	**0.441**	0.022	0.088
	20	2^{-20}	**0.929**	0.018	0.053	**0.676**	0.029	0.088
	30	2^{-20}	**0.927**	0.012	0.056	**0.618**	0.024	0.110
Outlier Sum	10	2^{-20}	**0.963**	0.021	0.078	**0.375**	0.029	0.110
	20	2^{-20}	**0.913**	0.025	0.083	**0.478**	0.029	0.103
	30	2^{-20}	**0.905**	0.025	0.090	**0.478**	0.022	0.088
Bimodality Index	10	2^{-20}	**0.914**	0.025	0.090	**0.360**	0.029	0.096
	20	2^{-20}	**0.898**	0.018	0.064	**0.478**	0.022	0.081
	30	2^{-20}	**0.884**	0.015	0.064	**0.493**	0.029	0.103

E.3 Validierung der Top3 Random Forests

E.3.1 Mit frei wählbarem Cutoff

Tabelle E.7: Validierung der Top3 Random Forests mit nur Genen bei frei wählbaren Cutoffs.

			Rotterdam		Transbig	
Score	Top-Gene	Gewicht	NPV	Spezifität	NPV	Spezifität
negative Kurtosis	10	2^{-22}	0.875	0.042	1.000	0.022
	70	2^{-22}	0.839	0.155	0.932	0.228
	80	2^{-22}	0.885	0.137	0.889	0.178
Likelihood Ratio	10	2^{-22}	0.667	0.024	0.500	0.006
	20	2^{-22}	0.750	0.018	0.000	0.000
	10	2^{10}	0.662	0.946	0.781	0.833
VRS	10	2^{-22}	0.000	0.000	0.000	0.000
	20	2^{-22}	0.333	0.006	0.000	0.000
	10	2^{-20}	0.665	0.899	0.799	0.928
Bimodality Index	70	2^{-22}	0.000	0.000	0.000	0.000
	100	2^{-22}	0.600	0.036	1.000	0.072
	80	2^{-22}	0.600	0.018	1.000	0.017
Dip	70	2^{-22}	0.800	0.071	1.000	0.011
	90	2^{-22}	0.800	0.119	0.600	0.017
	100	2^{-22}	0.792	0.113	0.750	0.033
WVRS	90	2^{10}	0.651	0.887	0.776	0.906
	90	2^{8}	0.651	0.899	0.777	0.889
	90	2^{6}	0.650	0.905	0.778	0.878
positive Kurtosis	90	2^{6}	0.652	0.905	0.770	0.817
	90	2^{8}	0.658	0.893	0.776	0.828
	90	2^{10}	0.652	0.893	0.773	0.833
Outlier Sum	40	2^{6}	0.658	0.940	0.788	0.783
	30	2^{10}	0.665	0.935	0.766	0.839
	30	2^{8}	0.662	0.935	0.766	0.839

Tabelle E.8: Validierung der Top3 Random Forests mit Genen und klinischen Variablen bei frei wählbaren Cutoffs.

			Transbig	
Score	Top-Gene	Verlust	NPV	Spezifität
Outlier Sum	60	2^{-22}	0.000	0.000
	50	2^{-22}	0.000	0.000
	30	2^{-22}	0.000	0.000
VRS	10	2^{-22}	0.833	0.061
	20	2^{-22}	0.750	0.036
	50	2^{-22}	0.000	0.000
Bimodality Index	70	2^{-22}	0.000	0.000
	20	2^{-22}	0.000	0.000
	10	2^{-22}	1.000	0.012
Likelihood Ratio	20	2^{-22}	0.833	0.030
	10	2^{-22}	0.000	0.000
	30	2^{-22}	0.000	0.000
WVRS	10	2^{-22}	1.000	0.018
	80	2^{-22}	0.000	0.000
	90	2^{6}	0.772	0.885
positive Kurtosis	90	2^{-22}	0.000	0.000
	100	2^{-22}	0.000	0.000
	80	2^{6}	0.757	0.776
negative Kurtosis	10	2^{-22}	0.857	0.036
	80	2^{-22}	0.900	0.218
	70	2^{-22}	0.872	0.206
Dip	10	2^{-22}	0.885	0.139
	20	2^{-22}	0.950	0.115
	70	2^{-22}	0.857	0.036

Tabelle E.9: Validierung der Top3 Random Forests nur mit Genen bei frei wählbaren Cutoffs und Verwendung von Down-Sampling.

			Rotterdam		Transbig	
Score	Top-Gene	Gewicht	NPV	Spezifität	NPV	Spezifität
negative Kurtosis	20	2^{-20}	0.500	0.006	0.000	0.000
	30	2^{-20}	1.000	0.006	1.000	0.006
	10	2^{-20}	0.000	0.000	0.000	0.000
Dip	80	2^{-20}	1.000	0.006	0.000	0.000
	90	2^{-20}	1.000	0.006	0.000	0.000
	100	2^{-20}	1.000	0.012	0.000	0.000
Bimodality Index	90	2^{-20}	0.000	0.000	0.000	0.000
	100	2^{-20}	0.000	0.000	0.000	0.000
	100	2^{-18}	0.627	0.530	0.807	0.744
Likelihood Ratio	10	2^{-18}	0.641	0.542	0.774	0.683
	10	2^{4}	0.671	0.690	0.747	0.411
	10	2^{6}	0.671	0.655	0.744	0.322
positive Kurtosis	100	2^{-18}	0.601	0.565	0.789	0.789
	90	2^{10}	0.656	0.238	0.791	0.189
	100	2^{6}	0.638	0.220	0.767	0.183
VRS	10	2^{-18}	0.650	0.708	0.822	0.772
	70	2^{8}	0.671	0.304	0.727	0.178
	70	2^{10}	0.670	0.435	0.772	0.244
WVRS	100	2^{-18}	0.610	0.762	0.791	0.861
	90	2^{-18}	0.600	0.571	0.810	0.783
	100	2^{4}	0.636	0.333	0.838	0.344
Outlier Sum	40	2^{6}	0.701	0.696	0.845	0.333
	40	2^{10}	0.685	0.673	0.851	0.350
	30	2^{10}	0.716	0.631	0.766	0.328

Tabelle E.10: Validierung der Top3 Random Forests mit Genen und klinischen Variablen bei frei wählbaren Cutoffs und Verwendung von Down-Sampling.

			Transbig	
Score	Top-Gene	Verlust	NPV	Spezifität
Dip	80	2^{-20}	0.000	0.000
	90	2^{-20}	0.000	0.000
	100	2^{-20}	0.000	0.000
negative Kurtosis	20	2^{-20}	0.000	0.000
	30	2^{-20}	0.000	0.000
	100	2^{-20}	1.000	0.061
Bimodality Index	90	2^{-20}	0.000	0.000
	100	2^{-20}	0.000	0.000
	10	2^{-18}	0.776	0.715
Likelihood Ratio	10	2^{-18}	0.831	0.685
	10	2^{6}	0.865	0.388
	10	2^{8}	0.855	0.358
WVRS	10	2^{4}	0.812	0.418
	10	2^{6}	0.815	0.321
	10	2^{8}	0.790	0.297
positive Kurtosis	10	2^{4}	0.826	0.461
	10	2^{6}	0.827	0.376
	100	2^{-18}	0.784	0.812
VRS	10	2^{-18}	0.829	0.648
	20	2^{-18}	0.795	0.800
	10	2^{2}	0.810	0.776
Outlier Sum	10	2^{-18}	0.828	0.497
	30	2^{6}	0.845	0.364
	10	2^{6}	0.837	0.467

E.3.2 Mit fest vorgegebenem Cutoff

Tabelle E.11: Validierung der Top3 Random Forests nur mit Genen bei fest vorgegebenen Cutoffs.

			Rotterdam		Transbig	
Score	Top-Gene	Gewicht	NPV	Spezifität	NPV	Spezifität
Likelihood Ratio	30	2^{-20}	0.577	0.089	0.947	0.100
	50	2^{-20}	0.523	0.137	0.829	0.161
	40	2^{-20}	0.594	0.113	0.952	0.111
WVRS	60	2^{-2}	1.000	0.006	0.750	0.017
	100	2^{-20}	0.571	0.071	0.909	0.111
	90	2^{-20}	1.000	0.012	1.000	0.011
Dip	20	2^{-20}	0.640	0.381	0.840	0.350
	80	2^{-20}	0.648	0.560	0.870	0.483
	90	2^{-20}	0.628	0.673	0.853	0.550
negative Kurtosis	70	2^{-22}	0.000	0.000	0.000	0.000
	30	2^{-20}	0.643	0.321	0.917	0.367
	60	2^{-20}	0.686	0.482	0.895	0.522
positive Kurtosis	10	2^{-18}	0.500	0.018	0.833	0.028
	70	2^{-18}	0.551	0.226	0.738	0.267
	90	2^{-18}	0.575	0.250	0.700	0.272
Outlier Sum	10	2^{-20}	0.706	0.071	0.842	0.089
	80	2^{-20}	0.619	0.357	0.849	0.439
	20	2^{-18}	0.644	0.560	0.763	0.572
Bimodality Index	90	2^{-20}	0.649	0.429	0.851	0.478
	100	2^{-20}	0.669	0.482	0.842	0.561
	80	2^{-20}	0.684	0.310	0.839	0.406
VRS	40	2^{-18}	0.519	0.161	0.778	0.233
	50	2^{-18}	0.578	0.351	0.826	0.422
	60	2^{-18}	0.600	0.357	0.810	0.356

Tabelle E.12: Validierung der Top3 Random Forests mit Genen und klinischen Variablen bei fest vorgegebenen Cutoffs.

			Transbig	
Score	Top-Gene	Verlust	NPV	Spezifität
negative Kurtosis	80	2^{-22}	1.000	0.018
	70	2^{-22}	1.000	0.006
	60	2^{-22}	1.000	0.012
positive Kurtosis	10	2^{-18}	0.907	0.236
	10	2^{-20}	0.943	0.200
	20	2^{-20}	0.927	0.230
WVRS	10	2^{-18}	0.927	0.230
	10	2^{-4}	0.927	0.230
	10	2^{-2}	0.927	0.230
Likelihood Ratio	30	2^{-20}	0.900	0.273
	10	2^{-20}	0.925	0.224
	50	2^{-20}	0.889	0.291
VRS	10	2^{-20}	0.925	0.224
	30	2^{-20}	0.923	0.218
	40	2^{-20}	0.927	0.230
Dip	10	2^{-20}	0.885	0.418
	20	2^{-20}	0.888	0.527
	30	2^{-20}	0.905	0.406
Outlier Sum	10	2^{-20}	0.905	0.345
	20	2^{-20}	0.867	0.394
	30	2^{-20}	0.882	0.364
Bimodality Index	10	2^{-20}	0.914	0.321
	20	2^{-20}	0.908	0.418
	30	2^{-20}	0.862	0.455

F Weitere Tabellen

Tabelle F.1: Berechnung von GRB7-, ER-, Proliferations- und Invasions-Scores für Oncotype DX.

Gruppe	Score
GRB7	$0.9 \cdot GRB7 + 0.1 \cdot HER2$ Scores kleiner 8 werden auf 8 gesetzt
ER	$(0.8 \cdot ER + 1.2 \cdot PGR + BCL2 + SCUBE2)/4$
Proliferation	$(Survivin + KI67 + MYBL2 + CCNB1 + STK15)/5$ Scores kleiner 6.5 werden auf 6.5 gesetzt
Invasion	$(CTSL2 + MMP11)/2$

Tabelle F.2: Spearman-Korrelation der Bimodalitätsmaße für die Mainz-Kohorte und die beiden Validierungskohorten bzw. die Mainz-Kohorte mit 194 Patientinnen und RMA-normalisierten Expressionsdaten.

Score	Rotterdam	Transbig	Mainz 194 RMA
VRS	0.333	0.266	0.702
WVRS	0.372	0.351	0.721
Dip	0.010	0.019	0.089
Kurtosis	0.434	0.421	0.786
Likelihood Ratio	0.485	0.451	0.748
Bimodality Index	0.366	0.286	0.692
Outlier Sum	0.400	0.381	0.709

G Weitere Abbildungen

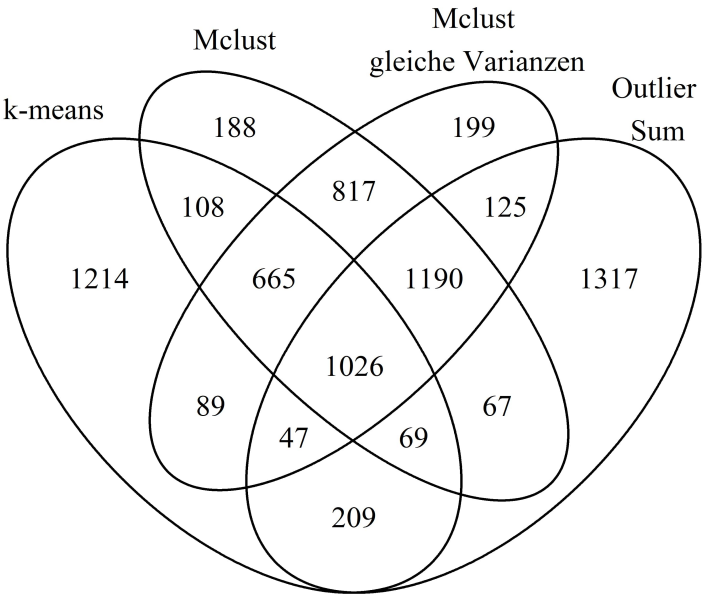

Abbildung G.1: Anzahl der p-Werte des Log-Rank-Tests kleiner 5 % für die vier Gruppeneinteilungen.

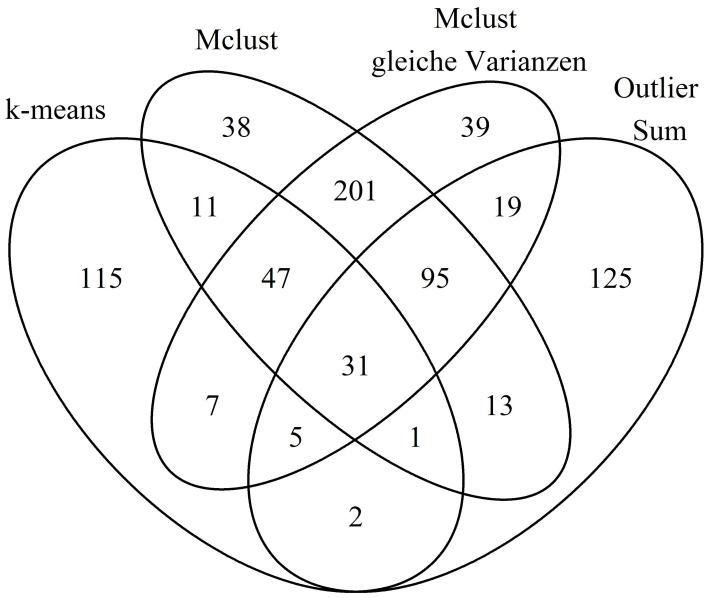

Abbildung G.2: Anzahl der FDR-adjustierten p-Werte des Log-Rank-Tests kleiner 5 % für die vier Gruppeneinteilungen.

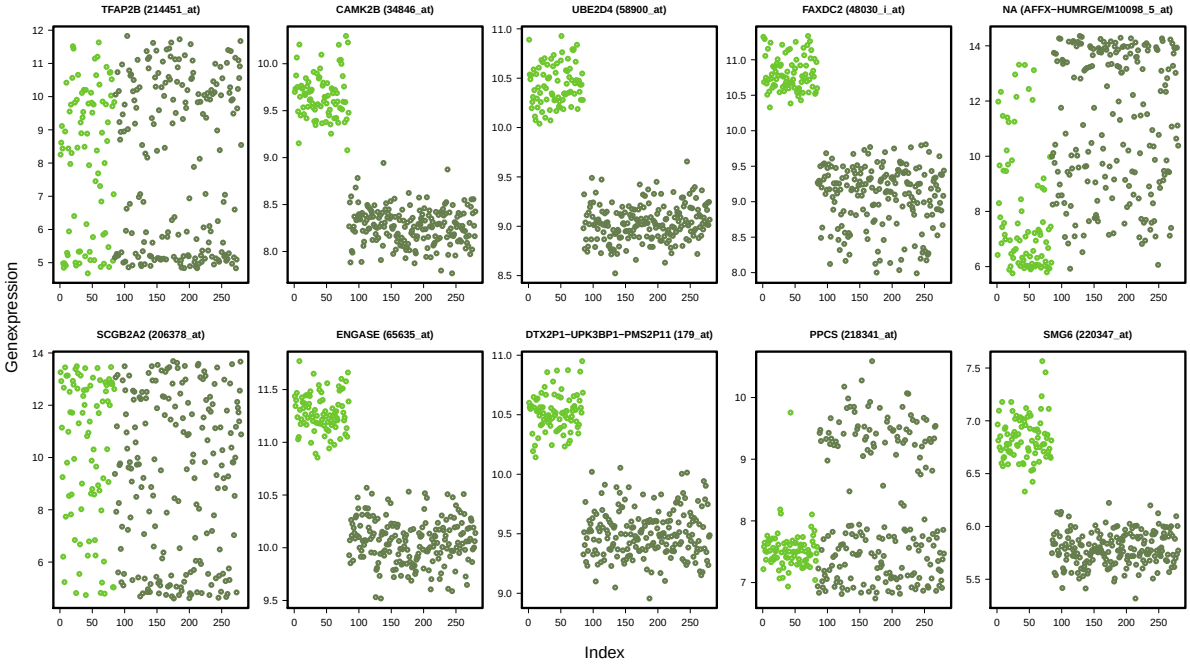

Abbildung G.3: Expressionswerte der Gene mit den größten Werten der dip-Statistik bei der Transbig-Kohorte. Die hellgrünen Punkte stellen die Werte der Proben aus dem Datensatz GSE6532 dar, die dunkelgrünen die Werte der Proben aus dem Datensatz GSE6532.

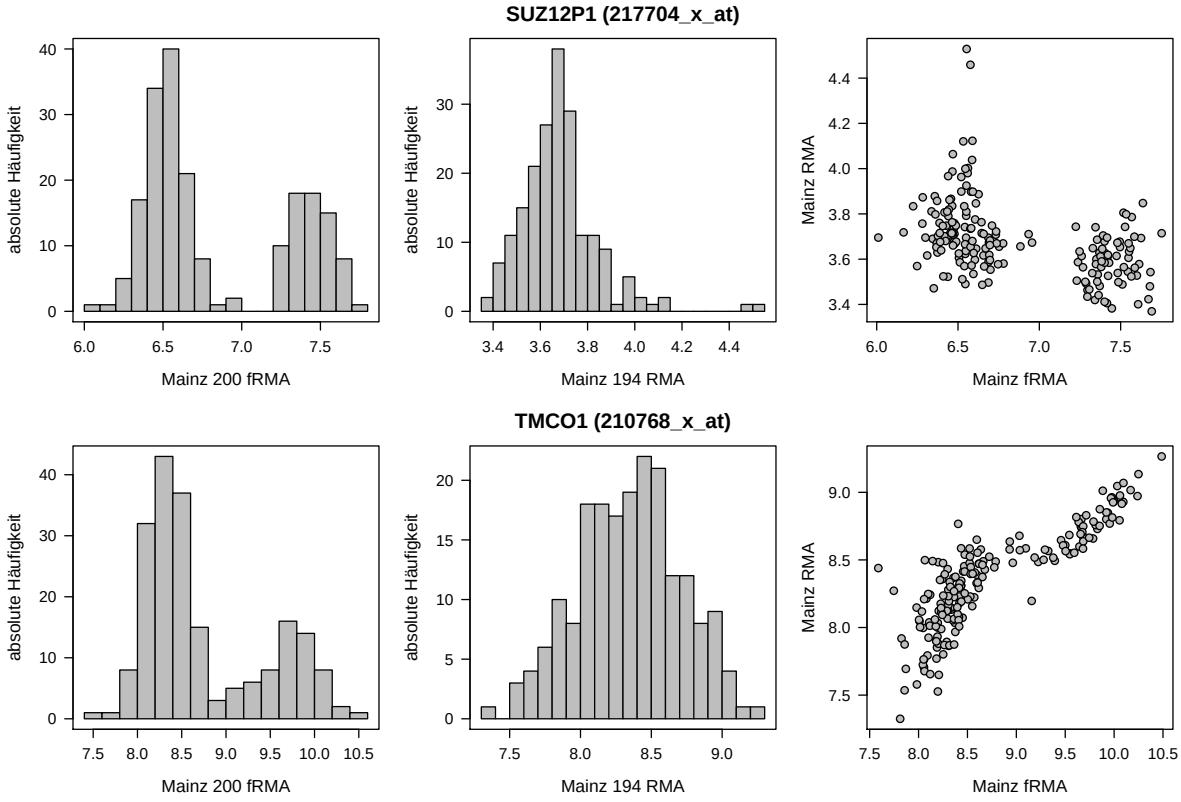

Abbildung G.4: Beispiele für Gene, die nur bei den fRMA-normalisierten Daten eine bimodale Expressionsverteilung im Sinne der dip-Statistik besitzen.

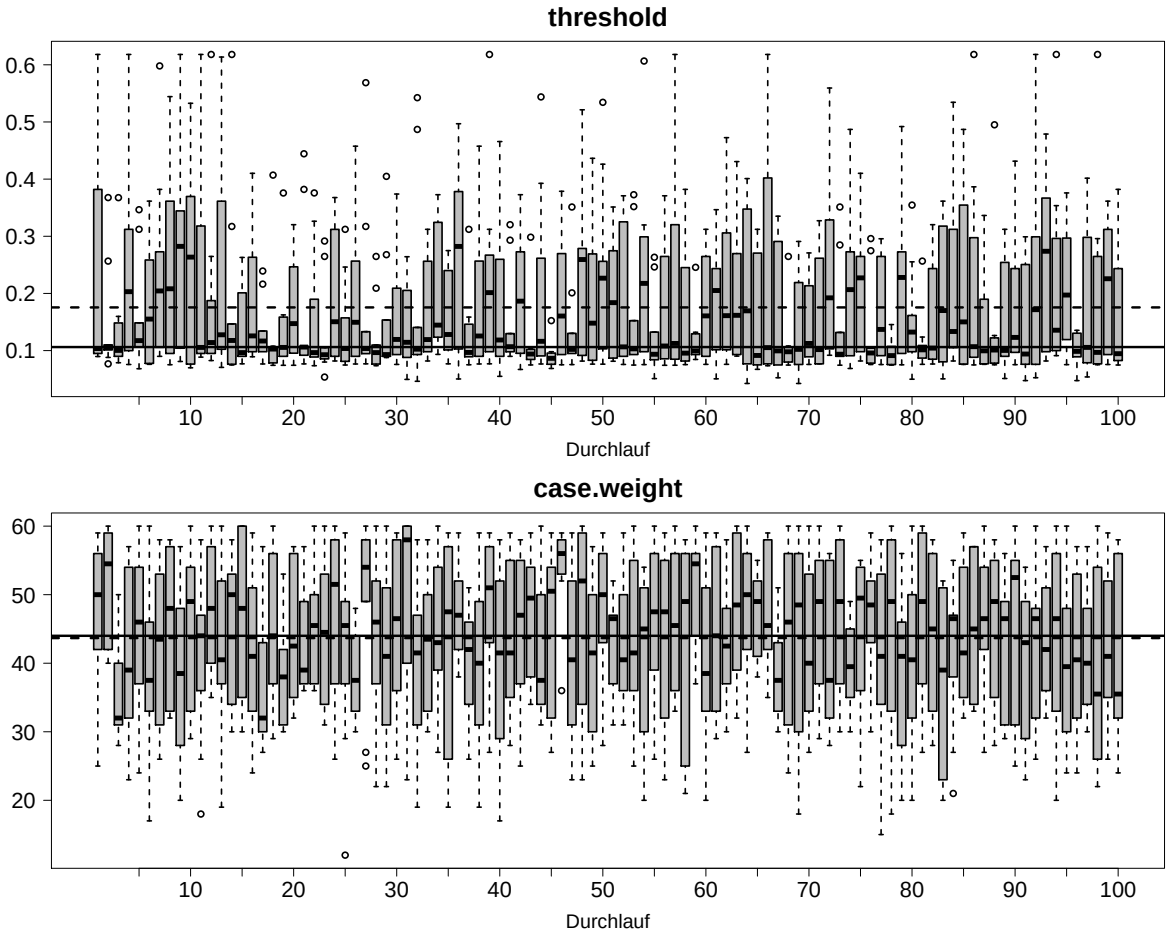

Abbildung G.5: Tuning-Ergebnisse der Random Forests ohne vorausgewählte Gene.

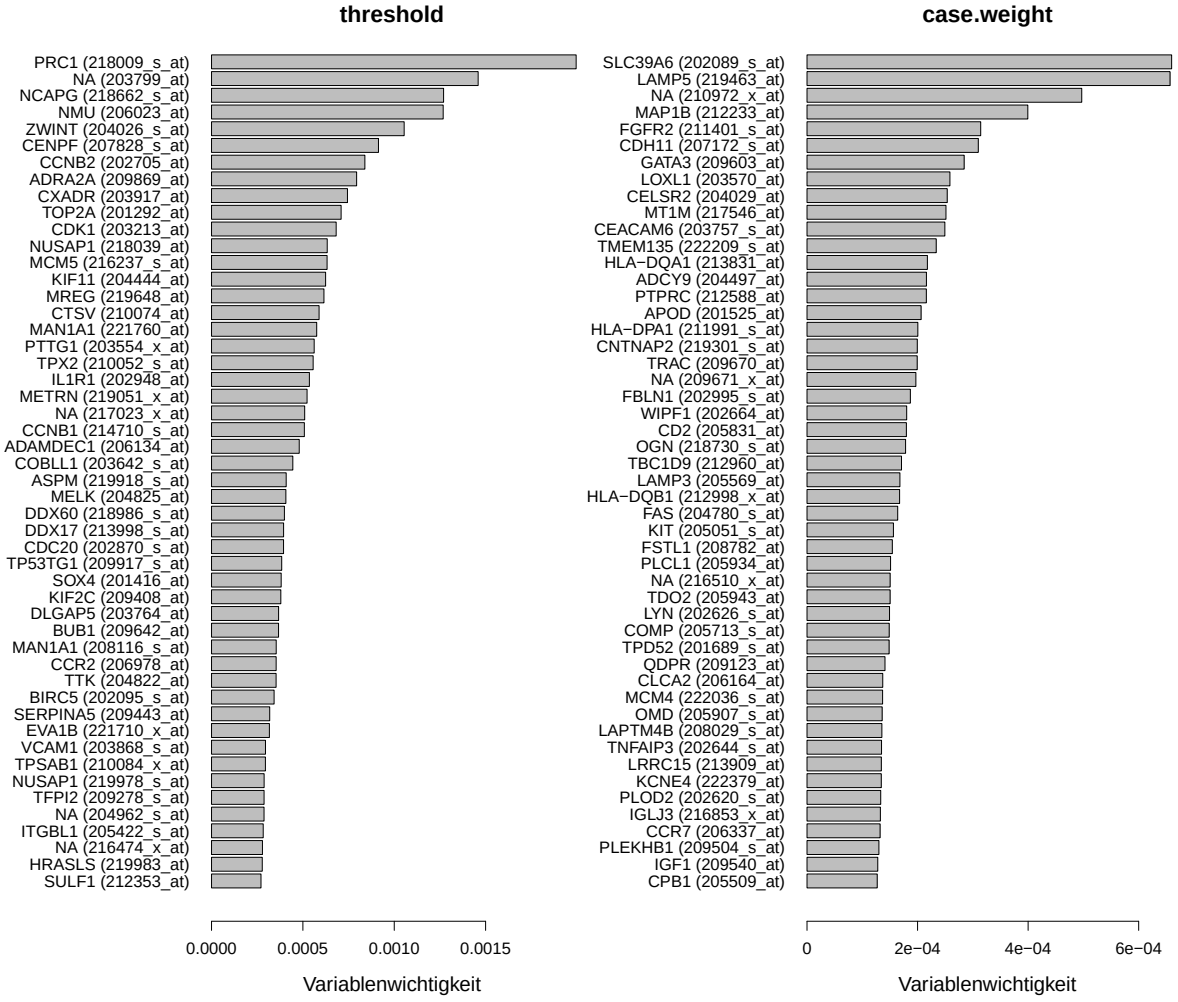

Abbildung G.6: Plot der Variablenwichtigkeit für die Random Forests ohne vorausgewählte Gene. Für jedes der getunten Modelle sind die 50 Gene mit der größten Variablenwichtigkeit enthalten.