Ludwig-Maximilians-Universität München
Institut für Statistik

Masterarbeit

Testen und Schätzen individueller Behandlungseffekte bei hochdimensionalen Kandidaten-Biomarkern

---

Autorin: Katrin Hummrich
Betreuerin: Prof. Dr. Anne-Laure Boulesteix
Datum: 15. Juli 2016

Abstract

Hintergrund: In der personalisierten Medizin verfolgt man den Ansatz, dass nicht jede Behandlung für alle Patienten gleich wirkt. Eine statistische Methode zur individuellen Behandlungswahl, stellt die Entwicklung von prädiktiven Biomarkern dar.

Methoden: In dieser Arbeit wird eine Methode von Matsui et al. (2012) mit einem neuen Ansatz zur Entwicklung von hochdimensionalen Biomarkern anhand von omics-Daten verglichen. Der neue Ansatz stellt im Gegensatz zu der Methode von Matsui et al. (2012) eine multiple Herangehensweise mittels Lasso-Schätzer dar. Dabei wird sowohl ein einschrittiges als auch zweischrittige Verfahren betrachtet. Um den Treatmenteffekt zu testen werden Permutationstests vorgestellt. Wobei die Verfahren unterschiedliche Hypothesen betrachten.

Ergebnisse: Eine Simulationsstudie zeigt, dass die Vorhersagegenauigkeit durch den multiplen Ansatz verbessert werden kann. Der neue Ansatz erkennt mehr der wahren Effekte und lässt den fälschlicherweise geschätzten Effekten nicht so viel Einfluss zukommen wie die Methode von Matsui et al. (2012). Dabei hängen die Ergebnisse aller Methoden stark von den vorhandenen Effekten in den Daten ab.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die Medizin und ihre Behandlungsstrategien entwickeln sich stets weiter. In neuen Forschungsbereichen rückt der Patient mit seinen individuellen Eigenschaften in den Mittelpunkt. Hierbei spricht man von individualisierter bzw. personalisierter Medizin. Dieser Behandlungsansatz versucht zum einen genetische und klinische Eigenschaften von Patienten zusammen zu betrachten und nutzt zum anderen das aktuelle Wissen über die biologischen Vorgänge von Krankheiten, um darauf basierend eine patientenspezifische Behandlungsstrategie zu entwickeln (Ma et al., 2015). Chen et al. (2015, S. 1121) beschreiben das Ziel der personalisierten Medizin folgendermaßen: „Precision medicine will provide clinicians with new tools, knowledge and therapies to select which treatments will work best for which patients".

Ein Hilfsmittel um entscheiden zu können, welche Behandlung für welchen Patienten am besten geeignet ist, können sogenannte Biomarker sein. The National Institut of Health definiert Biomarker (biological Marker) als „A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" (Atkinson et al., 2001, S. 91). Dabei handelt es sich oft um Produkte von Organismen wie Enzyme, Hormone oder Ionen, die in Probenmaterial wie venösem Blut oder Urin festgestellt werden können (Bracht, 2009). In der modernen Krebstherapie werden Biomarker bereits genutzt, doch auch bei anderen Erkrankungen könnten sie wichtige Informationen für Diagnose, Prognose und Therapie liefern (Bracht, 2009). Je nach Verwendung in der Medizin unterscheidet man dabei verschiedene Arten. In dieser Arbeit besteht Interesse an sogenannten prognostischen und prädiktiven Biomarkern.

Prognostische Biomarker erlauben Aussagen über die voraussichtlichen Heilungschancen und/oder den Krankheitsverlauf ungeachtet jeglicher Behandlungen (Bracht, 2009; Chen et al., 2015).

Und prädiktive Biomarker geben entweder Auskunft über die Wahrscheinlichkeit zukünftig an einer Krankheit zu erkranken oder über das voraussichtliche Ansprechen auf eine bestimmte Behandlung (Bracht, 2009), wobei hier Letzteres von Interesse ist.

Bei der personalisierten Medizin können somit prädiktive Biomarker maßgebend für die Entscheidung der Therapiewahl sein. Die Erforschung und Validierung neuer Biomarker ist folglich für die Medizin von großer Bedeutung.

Dank der omics-Technologien stehen hierfür große Datenquellen zur Verfügung, die Informationen über Genexpressionen enthalten. Diese omics-Technologien stammen aus modernen Fachbereichen der Biologie, „[...] die sich mit der Analyse von Gesamtheiten ähnlicher Ein-

zelelemente [...] in einer lebenden Zelle beschäftigen“ (Guthke, 2010). Beispielsweise wird bei der Genomik ein sogenanntes Genom (= gesamte genetische Information eines Organismus) betrachtet oder bei der Proteomik alle Proteine eines Proteoms (Guthke, 2010). Diese Technologien erzeugen Datensätze mit mehreren Tausend Variablen, weshalb man es in der Regel mit hochdimensionalen Daten zu tun hat. Somit ermöglichen diese Daten auch die Entwicklung hochdimensionaler Biomarker im Gegensatz zu den bisher meist niedrigdimensionalen, wenn nicht sogar univariaten Biomarkern.
Dazu bedarf es jedoch neuer statistischer Methoden, weshalb sich diese Arbeit mit dem Schätzen und Testen individueller Behandlungseffekte bei hochdimensionalen Biomarkern befasst.
In Kapitel 2 werden dazu zunächst derzeitige statistische Methoden aus dem Bereich der personalisierten Medizin kurz betrachtet und der Nutzen dieser Arbeit eingeordnet. Der erste Teil von Kapitel 3 befasst sich anschließend mit dem Schätzen individueller Behandlungseffekte. Dabei wird zunächst ein Ansatz von Matsui et al. (2012) und danach ein neuer, davon inspirierter, Ansatz vorgestellt. Der zweite Teil von Kapitel 3 setzt sich mit dem Testen dieser Behandlungseffekte auseinander. Nach der Beschreibung des Permutationstests von Matsui et al. (2012), werden Möglichkeiten aus der Literatur aufgezeigt, wie man für den neuen Ansatz eine etwas andere Nullhypothese ebenfalls mit Permutationstests testen könnte. Kapitel 4 befasst sich mit einer Simulationsstudie zum Vergleich der unterschiedlichen Schätzmethoden. Kapitel 5 schließt dann die Arbeit mit einem Diskussionsteil ab.

2 Statistische Methoden der personalisierten Medizin

Wie in der Einleitung bereits beschrieben wurde, geht es bei der personalisierten Medizin darum, für jeden Patienten die beste Therapie zu finden. Dazu gibt es verschiedene statistische Ansätze, die helfen sollen für jeden Patienten die richtige Behandlungswahl zu treffen. Um den neuen Ansatz aus dieser Arbeit besser einordnen zu können, wird in diesem Kapitel ein kurzer Überblick über bisherige statistische Methoden im Bereich der personalisierten Medizin gegeben. Grundlage für diesen Überblick bieten vor allem zwei Reviews von Chen et al. (2015) und Ma et al. (2015).
Chen et al. (2015) geben in ihrer Arbeit einen statistischen Überblick bezüglich der Entwicklung von prädiktiven Biomarkern indem sie unterschiedliche Methoden vorstellen.
Bisher wurde meist angenommen, dass ein Medikament entweder für alle Patienten wirkt oder für keinen. Mit dem heutigen Wissen aus der molekularen Biologie zieht man nun auch in Betracht, dass lediglich eine Subgruppe der Patienten von einem Medikament profitieren könnte. Deshalb werden prädiktive Biomarker entwickelt, um bei der Behandlungswahl

Patienten zu identifizieren, die auf eine bestimmte Therapie ansprechen. In diesem Kontext werden mehrdimensionale Biomarker betrachtet.
Chen et al. (2015) gehen von einem klinischen Experiment mit einer Treatment- und einer Kontrollgruppe aus. Hierbei liegen für jeden Patienten omics-Daten vor, die vor der Behandlung gemessen wurden. Es wird nun angenommen, dass das Medikament nicht für alle Patienten gleich wirkt, sondern sich die Stichprobe in zwei Untergruppen unterteilen lässt. Es existiert folglich eine Gruppe von Patienten, die auf die Therapie ansprechen (g^+) und eine Gruppe von Patienten, die nicht auf die Therapie ansprechen (g^-). Um feststellen zu können, welcher Patient in welche Gruppe gehört, werden prädiktive Biomarker entwickelt. Nach Chen et al. (2015) setzt sich die Entwicklung von prädiktiven Biomarkern aus drei Schritten zusammen. Dabei gehen sie in jedem Schritt auf unterschiedliche Ansätze der konkreten Umsetzung ein.

1. Im ersten Schritt der Biomarker Identifikation geht es darum genetische Variablen zu erkennen, die sich zur Bildung eines mehrdimensionalen prädiktiven Biomarkers eignen könnten. Unabhängig davon, ob es sich bei der Zielvariable um eine binäre, stetige oder eine Überlebenszeit handelt, wird davon ausgegangen, dass innerhalb der Treatmentgruppe g^+ Patienten einen besseren Erwartungswert der Zielvariable haben als g^- Patienten. Denn g^+ Patienten profitieren von der Behandlung und somit verbessert sich ihr Gesundheitszustand durch die Behandlung. Das heißt sie leben beispielsweise länger oder ihr Tumor schrumpft. Bei g^- Patienten dagegen kann mit Hilfe der Therapie keine Verbesserung erzielt werden. Daher wird auch angenommen, dass g^+ Patienten in der Treatmentgruppe einen besseren Erwartungswert der Zielvariable haben als g^+ Patienten aus der Kontrollgruppe. Prädiktive Biomarker sollen nun unterschiedliche Behandlungseffekte bei den Patienten vorhersagen. Das heißt g^+ Patienten, die auf die Therapie ansprechen, unterscheiden sich in ihrem Messwert bezüglich des prädiktiven Biomarkers hinsichtlich der g^- Patienten, die nicht auf die Therapie ansprechen. Um feststellen zu können, welche genetischen Variablen sich als prädiktiver Biomarker eignen, wird für jede Genexpression ein eigenes generalisiertes lineares Modell geschätzt. Eine Möglichkeit ist es lediglich die Messwerte der Treatmentgruppe zu betrachten und jeweils zu überprüfen, ob der Haupteffekt der Genexpression einen signifikanten Einfluss auf die Zielvariable hat (Chen et al., 2015). Denn unter allen behandelten Patienten, ist bei denjenigen ein besserer Response zu beobachten, die auf die Behandlung ansprechen, da diese zur g^+ Gruppe gehören. Um diese schließlich zu identifizieren eignen sich die genetischen Variablen, die einen Einfluss auf die Zielvariable haben. Die Menge der genetischen Variablen mit einem signifikanten β-Koeffizienten bildet dann

die Menge U der potentiellen prädiktiven Biomarker.
Da klinische Studien häufig eine Treatment- und eine Kontrollgruppe besitzen, wird auch oft ein generalisiertes lineares Modell mit allen Beobachtungen geschätzt, das dann den Haupteffekt des Treatments und der genetischen Variable, sowie die Interaktion der beiden enthält (Chen et al., 2015). Ob eine Genexpression als prädiktiver Biomarker in Frage kommt, ist folglich nicht mehr vom Haupteffekt der Genexpression, sondern vom Interaktionseffekt zwischen Genexpression und Treatment abhängig. Dieses Vorgehen ist deutlich intuitiver, da es das Zusammenwirken von Treatment und genetischer Variable direkt in Form der Interaktion schätzt.
Freidlin und Simon (2005) schlagen dagegen vor, ein Modell zu fitten, das zwar die Interaktion enthält aber nicht den Haupteffekt der genetischen Variable. Um zu entscheiden welches Modell den Vorzug erhält, bedarf es laut Chen et al. (2015) noch weiterer Studien.

2. Im zweiten Schritt geht es darum die Patienten in die g^+ und g^- Gruppe einzuteilen, um eine Behandlungswahl treffen zu können. Da die wahren Label, g^+ und g^-, nicht bekannt sind, stellt eine klassische Subgroup Selection eine Herausforderung dar (Chen et al., 2015).
   Handelt es sich bei der Zielvariable um eine binäre Variable, werden oft die Ausprägungen der Zielvariable als Label genutzt. Das heißt man geht davon aus, dass zum Beispiel bei Eintreten einer Schrumpfung des Tumors, der Patient als g^+ Patient betrachtet werden kann. In diesem Fall werden die üblichen Methoden, wie logistische Regression, Klassifikationsbäume oder Random Forests für die Vorhersage der Klassenzughörigkeit genutzt. Dabei ist allerdings zu beachten, dass es sich hierbei nicht um wahre Labels handelt. Denn die beobachtete Outcomevariable stellt vielmehr eine binäre Zufallsvariable dar, mit den Erwartungswerten der Gruppen als Wahrscheinlichkeiten für g^+ bzw. g^-. Es könnte sich beispielsweise der Tumor aus einem anderen Grund verkleinert haben und nicht wegen der Behandlung, das heißt der Patient würde fälschlicherweise mit g^+ gelabelt werden. Somit können auch falsch gelabelte Beobachtungen vorliegen, die die Subgroup Selection behindern (Chen et al., 2015).
   Handelt es sich jedoch um eine stetige Outcomevariable oder um eine Überlebenszeit, müssen andere Methoden herangezogen werden. Dazu stellen Chen et al. (2015) verschiedene Methoden vor.
   Ein möglicher Ansatz ist es zunächst einen prädiktiven Score anhand der Menge U des ersten Schrittes, die alle Variablen mit einem signifikanten β-Koeffizienten enthält, zu bilden (Chen et al., 2015). Dieser prädiktive Score stellt folglich einen mehrdimen-

sionalen Biomarker basierend auf mehreren prädiktiven Genexpressionen dar. Ist U nicht zu groß, wird ein multiples Regressionsmodell mit allen Genexpressionen aus U gefittet. Sind es zu viele genetische Variablen in U, sollte eine dimensionsreduzierende Methode, wie eine Hauptkomponentenanalyse, vorgeschaltet werden. Die gewichtete Summe der Ausprägungen der Genexpressionen und der zugehörigen β-Koeffizienten aus der multiplen Regression, bildet dann den prädiktiven Score für jeden Patienten.
Eine Alternative stellt die Methode von Matsui et al. (2012) dar, die zur Score-Bildung mehrere einfache anstatt ein multiples Regressionsmodell benutzt. Dieser Ansatz wird in Kapitel 3.2.1 noch genauer vorgestellt.
Anschließend muss zur Gruppeneinteilung der stetige Score dichotomisiert werden, indem ein Cutoff-Point gesucht wird, der die Patienten in zwei Gruppen teilt (Chen et al., 2015). Dieser Cutoff-Point kann anhand von Percentilen des prädiktiven Scores oder durch vorher festgelegte Grenzwerte der Zielvariable definiert werden. Jiang et al. (2007) schlagen eine Methode für quantitative Biomarker vor, die einen Schwellenwert für die g^+ Gruppe entwickelt und validiert. Dabei entwickelt diese Methode nicht nur einen Cutoff-Point durch Maximieren der Log-Likelihood Teststatistik über alle möglichen Cutoff-Points, sondern testet gleichzeitig auch, ob es einen overall Treatmenteffekt für die gesamte Population gibt.
Andere Ansätze nutzen laut Chen et al. (2015) Klassifikations- und Regressionsbäume um die Patienten in homogene Gruppen bezüglich des Nutzens der Behandlung zu splitten. Jedoch werden hier in der Regel mehr als zwei Gruppen gebildet.
Und schließlich stellt die ASD Methode (adaptive signature design) bzw. die CVASD Methode (cross-validated adaptive signature design) noch eine weitere Alternative dar (Freidlin and Simon, 2005; Freidlin et al., 2010). Diese Methode basiert auf binären Zielvariablen, kann aber laut Freidlin et al. (2010) für Überlebenszeiten verallgemeinert werden (wie es im Prinzip die Methode von Matsui et al. (2012) macht). Hier werden die Patienten mit Hilfe von Odds Ratios den Gruppen zugeteilt. Dazu wird für jede genetische Variable aus U anhand der Regression ein Odds Ratio geschätzt. Jeder Patient von dem eine vorher definierte Mindestanzahl an genetischen Variablen ein Odds Ratio größer einem bestimmten Grenzwert hat, wird dann der Biomarker-positiven Gruppe zugeordnet. Der Unterschied zwischen der ASD Methode und der CVASD Methode liegt in der Einteilung in Trainings- und Testdaten. Bei der ASD Methode wird nur eine einmalige Unterteilung vorgenommen und bei der CVASD wird eine Kreuzvalidierung vorgenommen. Auf den Nutzen solcher Unterteilungen wird im 3. Schritt näher eingegangen.

3. Im dritten Schritt geht es darum den klinischen Nutzen des prädiktiven Biomarkers zu bewerten. Chen et al. (2015) schildern hierbei zwei Teile.
   Der erste Teil besteht darin die Vorhersagekraft des Klassifikators zu bestimmen. Chen et al. (2015) nennen hier unter anderem zwei gängige Vorgehensweisen für binäre Outcomevariablen und Survivaldaten. Bei binären Zielvariablen wird die Güte der Vorhersage meist über Anteile an richtig zugeordneten Beobachtungen definiert. In der Medizin werden hier in der Regel die Sensitivität und die Spezifität betrachtet, die in Kapitel 3.2.3 noch genauer definiert werden. Bei Überlebenszeiten wird betrachtet wie gut der prädiktive Score die Untergruppen trennt. Dies geschieht mit dem Logrank-Test, welcher überprüft, ob die Überlebenskurven der beiden Gruppen sich signifikant voneinander unterscheiden. Neben diesen zwei gängigen Methoden finden sich in der Literatur noch einige weitere Möglichkeiten, auf die hier nicht eingegangen wird. Unabhängig des Skalenniveaus der Zielvariable wird die Beurteilung der Vorhersage des prädiktiven Biomarkers meist mit Hilfe von Trainings- und Testdaten vorgenommen (Chen et al., 2015). Die Entwicklung des prädiktiven Biomarkers wird anhand der Trainingsdaten durchgeführt und danach wird dieser auf die Testdaten angewendet, um dessen Performance zu beurteilen. Bei der sogenannten Kreuzvalidierung, die bei der CVASD Methode bereits erwähnt wurde, werden die Daten mehrmals in Trainings- und Testdaten unterteilt. Wie zuvor beschrieben wurde, ist es eine methodische Herausforderung die Patienten den zwei Gruppen zuzuordnen. Genauso schwierig stellt sich die Evaluierung des Klassifikators aufgrund der fehlenden Labels dar. Weshalb Chen et al. (2015) diesen Teil der Entwicklung von prädiktiven Biomarkern als besonders schwierig bezeichnen.
   Der zweite Teil der Beurteilung des klinischen Nutzens besteht darin den geschätzten Treatmenteffekt zu testen. Dabei interessiert man sich oft sowohl für einen Effekt in der gesamten Population als auch für Effekte in der Untergruppe g^+. Dazu dienen die folgenden Hypothesen (der Subgroup Analyse) (Chen et al., 2015):

   - H00: es gibt allgemein keinen Treatmenteffekt in der gesamten Population.
   - H01: es gibt keinen Treatmenteffekt in der g^+ Gruppe.
   - H02: es gibt keinen Treatmenteffekt in der g^- Gruppe .

   Dazu wird entweder die gesamte Stichprobe betrachtet oder nur einzelne Untergruppen. Das hängt davon ab, welche Hypothese getestet werden soll. Je nachdem welche Ergebnisse die Tests liefern, können unterschiedliche Rückschlüsse gezogen werden. Erhält man beispielsweise für H01 ein signifikantes Ergebnis, aber für H00 keines, spricht das dafür, dass es nur in der g^+ Gruppe einen Treatmenteffekt gibt (Chen et al., 2015).

Es gibt allerdings noch andere Möglichkeiten den Treatmenteffekt zu testen, die hier nicht betrachtet wurden. In Kapitel 3.3 werden beispielsweise Permutationstests vorgestellt um die Interaktionen zwischen den genetischen Variablen und dem Treatment auf Signifikanz zu testen.

Während dem eben beschriebenen Prozess der Entwicklung von prädiktiven Biomarkern kann es an zwei Stellen zu multiplen Testproblemen kommen (Chen et al., 2015).
Zum einen im ersten Schritt bei der Identifikation von potentiellen Biomarkern. Denn hier wird für jede genetische Variable ein einzelnes Regressionsmodell geschätzt und anschließend der interessierende β-Koeffizient auf Signifikanz getestet. Chen et al. (2015) schlagen an dieser Stelle die Adjustierung der FDR (false discovery rate) vor, anstatt den globalen Fehler 1. Art zu adjustieren. Die FDR ist definiert als der Erwartungswert des Anteils der falsch positiven Testergebnisse an allen positiven Testergebnissen (Benjamini and Hochberg, 1995). Der Ansatz der FDR lässt vergleichsweise mehr signifikante Ergebnisse zu, was bei einer großen Menge von Variablen wie bei omics-Daten vorteilhaft ist (Chen et al., 2015). Dafür muss jedoch ein kleiner Anteil an falsch positiven Ergebnissen in Kauf genommen werden. Dies sollte bei der Entwicklung von prädiktiven Biomarkern allerdings verkraftbar sein, da hier laut Chen et al. (2015) das Weglassen von wichtigen Variablen einen größeren Einfluss auf die Performance des Klassifikators hat, als das Hinzunehmen von unwichtigen Variablen.
Zum anderen taucht das multiple Testproblem im dritten Schritt beim Testen des Treatmenteffekts auf, da hier mehrere Hypothesen getestet werden. Zwei mögliche Lösungsansätze stellen die Adjustierung der p-Werte zum Beispiel anhand der Bonferroni-Adjustierung oder das Testen der Hypothesen in fester Reihenfolge (fixed sequence testing) dar (Chen et al., 2015). Freidlin and Simon (2005) schlagen eine weitere Alternative vor, um den globalen α-Fehler des Designs kontrollieren zu können. Dieser setzt sich hierbei aus zwei Teilen zusammen $\alpha = \alpha_1 + \alpha_2$. Der Signifikanztest bezüglich des globalen Treatmenteffekts wird dabei zum Signifikanzniveau α_1 zu Beginn der Analyse mit allen Patienten durchgeführt. Und der Signifikanztest innerhalb der Patienten, die der g^+-Gruppe zugeteilt wurden, wird zum Niveau α_2 durchgeführt. Freidlin and Simon (2005) wählen beispielsweise $\alpha = 0.05$ und nehmen mit $\alpha_1 = 0.04$ und $\alpha_2 = 0.01$ eine 80%/20% Aufteilung vor.

Ma et al. (2015) stellen in ihrem Paper sogenannte Behandlungsregeln, auch ITR (individualized treatment rules) bezeichnet, vor. Hier ist es auch das Ziel anhand der Eigenschaften des Patienten zu entscheiden, welche Therapie die bessere für ihn ist. Jedoch ist die Herangehensweise etwas anders. Die Idee der ITR ist es die Behandlungsergebnisse von zwei zur Auswahl stehenden Therapien zu vergleichen. Dazu wird die Differenz der Erwartungswerte der Zielvariable gegeben die Therapieform und weiterer Kovariablen betrachtet. Diese Ko-

variablen können verschiedene Eigenschaften des Patienten sein, wie Alter, Geschlecht oder auch genetische Variablen. Kovariablen mit einem Einfluss auf die Zielvariable werden hier als Biomarker bezeichnet. Wobei der Einfluss der Variablen bei diesem Ansatz in einem gemeinsamen Modell gefittet wird und nicht anhand lauter einzelner Regressionen. Je nachdem ob es sich dabei um einen Haupt- oder Interakionseffekt handelt, werden sie als prognostisch oder prädiktiv angesehen. Dabei wird in diesem Fall kein stetiger Score aus diesen gebildet, sondern der Erwartungswert anhand eines Modells, das diese Variablen enthält geschätzt.
Ma et al. (2015) weisen darauf hin, dass die Modellwahl an das Skalenniveau der Zielvariable und die Dimensionalität der Daten angepasst werden muss. Bei einem binären Outcome wird beispielsweise häufig die logistische Regression verwendet und bei Survivaldaten das Cox-Modell oder das AFT Modell (accelerated failure time), das auf die proportional hazard Annahme verzichtet. Liegen hochdimensionale Daten vor wird meist auf Variablenselektion durch beispielsweise Lasso zurückgegriffen. Außerdem weisen Ma et al. (2015) noch auf fortgeschrittene Methoden hin, die zum einen eine robustere Inferenz bei Misspezifikation erzielen und zum anderen auf Techniken aus dem Bereich Machine Learning zurückgreifen. Dabei werden auch hier die Daten in Trainings- und Testdaten aufgeteilt, um die Koeffizienten der Kovariablen auf Basis der Trainingsdaten zu schätzen und dann anhand der geschätzten Koeffizienten die ITR auf die Testdaten anzuwenden und zu evaluieren.
Oftmals reicht es jedoch nicht, wenn die neue Therapie nur für einen Unterschied zwischen den Behandlungsergebnissen sorgt. Also wenn beispielsweise durch die neue Therapie die Überlebenszeit verlängert wird oder der Tumor geschrumpft wird. Sondern es wird ein Minimum an Verbesserung verlangt, um die Aussage treffen zu können, dass die neue Therapie für den Patienten besser ist als die herkömmliche. Das heißt der Unterschied der Erwartungswerte soll nicht nur ungleich Null sein, sondern einen vorher festgelegten Wert überschreiten, damit die Verbesserung als klinisch relevant angesehen wird.

Wie dieser kurze Methodenüberblick gezeigt hat, ist es eine komplexe Aufgabe für die Statistik, der personalisierten Medizin hilfreiche und valide Hilfsmittel an die Hand zu geben. Weshalb in diesem Bereich noch viel geforscht wird. Gerade die Entwicklung von mehrdimensionalen Biomarkern scheint ein vielversprechendes Forschungsfeld zu sein. Genau aus diesem Grund wird in dieser Arbeit zunächst eine interessante Methode dazu von Matsui et al. (2012) und anschließend ein neuer Ansatz vorgestellt. Dieser basiert auf der Idee von Matsui et al. (2012) und versucht diese durch eine abgeänderte Herangehensweise zu optimieren. Dazu werden ähnlich wie bei Chen et al. (2015) beschrieben, erst die hochdimensionalen Biomarker geschätzt, dann eine Vorhersage anhand dieser Biomarker gemacht und dessen Resultat evaluiert. Schließlich soll der Treatmenteffekt noch getestet werden.

3 Theorie

3.1 Notation

Bevor die im Folgenden betrachteten Theorien vorgestellt werden, wird zunächst die in dieser Arbeit geltende Notation eingeführt. Hierbei werden Vektoren mit fett gedruckten Kleinbuchstaben und Matrizen mit fett gedruckten Großbuchstaben dargestellt.

Es liegen Daten von insgesamt n Patienten vor mit dem Index $i = 1, .., n$. Die Zielvariable Y sei binär mit den unabhängigen Ausprägungen $\mathbf{y} = (y_1, ..., y_n)^T \in \{0, 1\}$. Es werden zwei Behandlungen betrachtet, eine neue Behandlung (Treatment) und eine herkömmliche oder eine Behandlung mit Placebo (Kontrolle). Die zugehörige Treatmentvariable T sei

$$T_i = \begin{cases} 1, \text{ Treatment} \\ 0, \text{ Kontrolle} \end{cases}$$

mit den Ausprägungen $\mathbf{t} = (t_1, ..., t_n)^T$. Die Kovariablenmatrix $\mathbf{X}_{(n \times p)}$ enthält die Werte der p omics-Variablen für alle n Patienten. Wobei $p > n$ gilt, dies bedeutet es liegen mehr Kovariablen (Genexpressionen) als Beobachtungen (Patienten) vor. In diesem Fall spricht man auch von hochdimensionalen Daten. Der Vektor $\mathbf{x}_j = (x_{1j}, .., x_{nj})^T$ mit $j = 1, ..., p$ enthält dabei die stetigen Ausprägungen der j-ten genetischen Variable X_j für alle n Beobachtungen.
Wie in Kapitel 2 angesprochen, sollten die Daten idealerweise in Trainings- und Testdaten unterteilt werden. Dies bringt den Vorteil, dass kein Modell geschätzt wird, das sich zu sehr an die beobachteten Daten anpasst. Denn um unverzerrte Schätzungen zu erhalten sollte die Stichprobe, die zur Modellbildung verwendet wird, unabhängig der Stichprobe, die zur Beurteilung der Performance verwendet wird, sein (Baek et al., 2009). Das heißt man könnte den Datensatz in zwei Teile unterteilen und dann einen zur Schätzung und den anderen zur Validierung nutzen, wie es beispielsweise bei der ASD Methode gemacht wird (Freidlin and Simon, 2005). Da es sich jedoch bei dem hier verwendeten Datentyp der omics-Daten um hochdimensionale Daten handelt, bedarf es für die Entwicklung der Biomarker einer möglichst großen Stichprobe (Freidlin et al., 2010). Daher bietet es sich an, statt den Datensatz nur einmal in Trainings- und Testdaten aufzuteilen, dies mehrmals zu machen, was der Kreuzvalidierung entspricht. Dadurch kann gewissermaßen die komplette Stichprobe für die Schätzung und für die Validierung genutzt werden (Freidlin et al., 2010).
Die Methode von Matsui et al. (2012) sowie der neue Ansatz nutzen daher die Kreuzvalidierung. Die Vorgehensweise der Kreuzvalidierung und die zugehörige Notation werden daher ebenfalls kurz beschrieben. Bei einer K-fachen Kreuzvalidierung wird folgendermaßen vor-

gegangen (Baek et al., 2009):

1. Splitte den Datensatz zufällig in K (etwa gleichgroße) Teile.
2. Nutze $K-1$ Teildatensätze (Trainingsdaten) zur Schätzung der Biomarker.
3. Berechne die Biomarker mittels dieser Schätzung für den übrigen Teildatensatz (Testdaten).
4. Wiederhole 2. und 3. für alle $k = 1, ..., K$ Teildatensätze.

3.2 Schätzen

In diesem Abschnitt wird zunächst beschrieben wie die prädiktiven und prognostischen Biomarker gebildet werden. Da sich die Biomarker im Folgenden aus mehreren genetischen Variablen zusammensetzen, werden sie auch häufig als Scores bezeichnet.
Dazu wird erst der Vorschlag von Matsui et al. (2012) gezeigt und dann der daraus resultierende neue Ansatz.

3.2.1 Methode von Matsui

Ziel der Studie von Matsui et al. (2012) ist es die Empfänglichkeit von Patienten für Krebsbehandlungen vorhersagen zu können, um für jeden Patienten die geeignete Behandlung zu finden. Dazu sollen genetische Scores oder auch (hochdimensionale) Biomarker entwickelt und validiert werden. Dabei werden zum einen die allgemeinen Risiken, repräsentiert durch einen prognostischen Score, und zum anderen die unterschiedlichen Empfänglichkeiten für die Behandlung, repräsentiert durch den prädiktiven Score, betrachtet.
In einer randomisierten Studie soll nun die Treatmentgruppe (T), die die Behandlung erhalten hat, mit der Kontrollgruppe (K), die ein Placebo erhalten hat, verglichen werden. Insgesamt werden p pretreatment (= vor der Behandlung gemessen) Genexpressionen, eine Treatmentvariable und ein binärer Outcome von n Patienten betrachtet. Wobei mit $p > n$ hochdimensionale Daten vorliegen. Ursprünglich haben Matsui et al. (2012) als Outcome Überlebenszeiten betrachtet. Um jedoch die Simulation in Kapitel 4 zu erleichtern wird hier nur eine binäre Responsevariable (Event eingetreten oder nicht eingetreten) angenommen. Das heißt die Methode von Matsui et al. (2012) wurde so angepasst, dass sie für binäre Zielvariablen anwendbar ist. Dabei wurde vor allem das Cox-Modell für Überlebensdaueranalysen durch eine logistische Regression für die Analyse von binären Outcomes ersetzt.

Logistische Regression

Wenn die Zielvariable nicht stetig sondern binär ist, verwendet man anstelle der normalen linearen Regression eine binäre Regression. Fahrmeir et al. (2009) leiten diese folgendermaßen her. Da es sich bei dem Erwartungswert einer binären Variable um eine Wahrscheinlichkeit handelt, möchte man bei binären Regressionsanalysen den Effekt der Kovariablen auf die (bedingte) Wahrscheinlichkeit

$$\pi_i = P(y_i = 1|x_{i1}, ..., x_{ip}) = E(y_i|x_{i1}, ..., x_{ip})$$

für $y_i = 1$ gegeben die Kovariablen modellieren. Eine lineares Modell

$$\pi_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} = \eta_i$$

bringt unter anderem den Nachteil mit sich, dass der Wertebereich nicht wie bei Wahrscheinlichkeiten zwischen 0 und 1 liegt. Wobei η_i den linearen Prädiktor bezeichnet. Daher wird bei binären Regressionsmodellen die Beziehung zwischen der Wahrscheinlichkeit π_i und dem linearen Prädiktor η_i üblicherweise mit Hilfe einer Verteilungsfunktion h dargestellt. Hierdurch gelten für den transformierten linearen Prädiktor dieselben Eigenschaften wie für Verteilungsfunktionen. Dies garantiert, dass der Wertebereich $[0, 1]$ eingehalten wird.
Wählt man für h die logistische Verteilungsfunktion, erhält man

$$\pi_i = h(\eta_i) = \frac{exp(\eta_i)}{1 + exp(\eta_i)} = \frac{exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})} \tag{1}$$

das sogenannte logistische Regressionsmodell. Zur besseren Interpretation formt man dieses meist um und erhält die logarithmierte Chance

$$log\left(\frac{\pi_i}{1 - \pi_i}\right) = log\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

bzw. die Chance

$$\frac{P(y_i = 1)}{1 - P(y_i = 1)} = exp(\beta_0)exp(\beta_1 x_{i1}) \cdot ... \cdot exp(\beta_p x_{ip}).$$

Die Chance setzt folglich die Wahrscheinlichkeit für $y_i = 1$ und für $y_i = 0$ ins Verhältnis. Dabei hat $exp(\beta_j)$ einen multiplikativen Einfluss. Erhöht sich beispielsweise x_{ij} um eine Einheit, so verändert sich die Chance multiplikativ um den Faktor $exp(\beta_j)$ (bei Konstanthaltung aller anderen Kovariablen).

Score-Bildung

Für die Bildung des prädiktiven Scores schlagen Matsui et al. (2012) folgendes vor:

1. Schätze für jede Genexpression X_j, $j = 1, ..., p$, ein eigenes Modell, hier eine logistische Regression wie in Gleichung (1),

$$\pi_i^{(j)} = \frac{exp(\eta_i^{(j)})}{1 + exp(\eta_i^{(j)})} = \frac{exp(\beta_0^{(j)} + \beta_T^{(j)} t_i + \beta_1^{(j)} x_{ij} + \beta_2^{(j)} t_i x_{ij})}{1 + exp(\beta_0^{(j)} + \beta_T^{(j)} t_i + \beta_1^{(j)} x_{ij} + \beta_2^{(j)} t_i x_{ij})}, \tag{2}$$

   wobei $\pi_i^{(j)}$ die Wahrscheinlichkeit $P(y_i^{(j)} = 1 | X_j = x_{ij}, T = t_i)$ darstellt. Also die Wahrscheinlichkeit, dass die Outcomevariable Y den Wert 1 annimmt gegeben die Genexpression und das Treatment. Außerdem bezeichnet $\eta_i^{(j)}$ den linearen Prädiktor im j-ten Modellmit $\beta_0^{(j)}$ dem Intercept, $\beta_T^{(j)}$ dem Haupteffekt des Treatments, $\beta_1^{(j)}$ dem Haupteffekt der betrachteten Genexpression und $\beta_2^{(j)}$ dem Interaktionseffekt zwischen Treatment und Genexpression. Das j in der Formel 2 ist erforderlich aus dem Grund, da für jede Genexpression (Kovariable) ein separates Modell gefittet wird.

2. Teste mit Hilfe des Wald-Tests für jedes Modell die Hypothese H_0: $\beta_2^{(j)} = 0$ gegen H_1: $\beta_2^{(j)} \neq 0$ mit $\alpha = 0.001$ (Matsui et al., 2012). Das heißt teste auf Signifikanz der Interaktion zum Niveau 0.001.
   Dieser Test basiert auf der Wald-Statistik $v^{(j)}$. Diese wird berechnet indem man den Schätzer quadriert und durch seine Varianz teilt. Asymptotisch ist die Wald-Statistik unter der Nullhypothese χ^2-verteilt (Fahrmeir et al., 2009)

$$v^{(j)} = \frac{(\hat{\beta}_2^{(j)})^2}{Var(\hat{\beta}_2^{(j)})} \overset{a}{\sim} \chi_1^2.$$

   Matsui et al. (2012) nutzen in ihrem Paper eine standardnormalverteilte Teststatistik als Gewicht. Lässt man den Schätzer unquadriert und teilt an Stelle der Varianz des Schätzers durch seine Standardabweichung erhält man die standardisierte Wald-Statistik $z^{(j)}$:

$$z^{(j)} = \frac{\hat{\beta}_2^{(j)}}{\sqrt{Var(\hat{\beta}_2^{(j)})}} \overset{a}{\sim} N(0, 1).$$

   Alle Genexpressionen X_j mit signifikantem Interaktionseffekt bilden die Menge

$$\Omega_1 = \{j \mid \text{p-Wert von } \beta_2^{(j)} \leq \alpha \ \forall j\}.$$

3. Bilde den prädiktiven Score U_i^M nach Matsui et al. (2012) für Patient i als gewichtete Summe der Ausprägungen der Genexpressionen aus der Menge Ω_1 mit der zugehörigen

standardisierten Wald-Statistik z des Interaktionseffekts:

$$U_i^M = \sum_{g \in \Omega_1} z_g x_{ig}.$$

4. Bilde den prognostischen Score W_i^M nach Matsui et al. (2012) für jeden Patienten i. Dabei ist das Vorgehen analog zu den ersten drei Schritten, jedoch ohne Betrachtung des Treatments (Matsui et al., 2012). Das heißt es werden ebenfalls logistische Einfachregressionen gefittet, die jedoch lediglich den Haupteffekt der jeweiligen Genexpression X_j schätzen:

$$\pi_i^{(j)} = \frac{exp(\eta_i^{(j)})}{1 + exp(\eta_i^{(j)})} = \frac{exp(\beta_0^{(j)} + \beta_1^{(j)} x_{ij})}{1 + exp(\beta_0^{(j)} + \beta_1^{(j)} x_{ij})}.$$

   Anschließend wird hier $\beta_1^{(j)}$, der Haupteffekt der Genexpression, mit Hilfe des Wald-Tests auf Signifikanz getestet. Anhand der standardisierten Wald-Statistik,

$$s^{(j)} = \frac{\hat{\beta}_1^{(j)}}{\sqrt{Var(\hat{\beta}_1^{(j)})}} \overset{a}{\sim} N(0,1)$$

   hier zur besseren Unterscheidung mit $s^{(j)}$ bezeichnet, wird folglich die Nullhypothese H_0: $\beta_1^{(j)} = 0$ auch mit $\alpha = 0.001$ getestet.

$$\Omega_2 = \{j \mid \text{p-Wert von } \beta_1^{(j)} \leq \alpha \ \forall j\}$$

   bildet hierbei die Menge der signifikanten Genexpressionen. Schließlich wird der prognostische Score für Patient i in Form einer gewichteten Summe der Genexpressionen aus Ω_2 und ihren zugehörigen standardisierten Wald-Statistiken s des Haupteffekts gebildet:

$$W_i^M = \sum_{g \in \Omega_2} s_g x_{i,g}.$$

Nachdem die beschriebenen Schritte für alle k Teildatensätze der Kreuzvalidierung durchgeführt wurden, hat nun jeder Patient i einen prädiktiven und einen prognostischen Score, die für weitere Analysen verwendet werden können.

3.2.2 Neuer Ansatz

Matsui et al. (2012) betrachten alle Genexpressionen einzeln, indem sie für jede Genexpression eine eigene Einfachregression schätzen, die den Einfluss lediglich dieser Genexpression auf die Responsevariable ohne Hinzunahme weiterer Kovariablen schätzen soll. Dabei entsteht

allerdings das Problem des multiplen Testens, wie es auch schon in Kapitel 2 kurz erläutert wurde. Matsui et al. (2012) scheinen dieses durch ein allgemein klein gewähltes Signifikanzniveau $\alpha = 0.001$ abschwächen zu wollen, anstatt eine explizite Korrektur der p-Werte vorzunehmen. Ma et al. (2015) kritisierten ebenfalls die bisher oftmals univariaten Herangehensweisen, da dadurch gemeinsame Effekte von potentiellen multiplen Biomarkern unberücksichtigt bleiben. Der neue Ansatz stellt daher eine multiple Alternative zur Schätzung der Biomarker dar.

Lasso-Regression

Wie zuvor angesprochen, handelt es sich bei omics-Daten um Datensätze mit mehreren Tausend Variablen. Bei dieser hohen Anzahl an Kovariablen liefern herkömmliche Parameterschätzungen in einem multiplen Regressionsmodell keine zufriedenstellenden Ergebnisse bezüglich Vorhersagegenauigkeit und Interpretierbarkeit (Tibshirani, 1996). Übersteigt sogar die Anzahl der Kovariablen die der Beobachtungen ist ein lineares Modell gar nicht mehr schätzbar. Deshalb wird hier die sogenannte Lasso-Regression (least absolute shrinkage and selection operator) von Tibshirani (1996) angewendet. Diese Methode schrumpft die Koeffizienten und setzt dabei einige auf Null. Dadurch wird die Interpretation vereinfacht, da die wichtigsten Effekte, ähnlich wie bei der Subset Selection, selektiert und nicht mehr durch viele kleine Effekte verschleiert werden. Vergleichbar mit der Ridge Regression sorgt das Schrumpfen der Koeffizienten für eine stabilere Schätzung, was die Vorhersagegenauigkeit verbessert. Damit vereint Lasso die Vorteile der Subset Selection und der Ridge Regression. Die Outcomevariable sei hier zunächst $Y \in \mathbb{R}$. Ansonsten werden weiterhin p Kovariablen und n Beobachtungen betrachtet. Die übliche Annahme von entweder unabhängigen Beobachtungen oder bedingt unabhängigen y_i's gegeben den x_{ij}'s mit $i = 1, ..., n$ und $j = 1, ..., p$ sei ebenfalls getroffen. Zusätzlich wird hier angenommen, dass alle Kovariablen standardisiert sind, so dass $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$ und $\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$ gilt. Bei der Lasso-Regression , wie bei der normalen Regression (OLS), die Residuenquadratsumme $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2$ minimiert (Fahrmeir et al., 2009), jedoch unter einer Nebenbedingung. Die Definition von Lasso sieht dann folgendermaßen aus (Tibshirani, 1996; Hastie et al., 2009):

$$\begin{aligned}(\hat{\beta}_0, \hat{\beta})_{Lasso} = \underset{\beta_0, \beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \right) \\ \text{unter der Nebenbedingung} \sum_{j=1}^{p} |\beta_j| \leq t,\end{aligned} \tag{3}$$

wobei t ein nichtnegativer Skalar ist. Die Nebenbedingung von Lasso sorgt folglich dafür, dass die Summe der Beträge der β-Koeffizienten einen bestimmten Wert t nicht überschreitet.

Dadurch müssen einige β-Koeffizienten entweder geschrumpft oder sogar auf Null gesetzt werden, was eine Reduzierung auf die relevanten Effekte bewirkt. Die vorherige Standardisierung der Kovariablen garantiert dabei, dass alle β-Koeffizienten gleichermaßen durch die Nebenbedingung bestraft werden.

Der Lasso-Schätzer (3) lässt sich mit Hilfe der Lagrange-Methode auch in penalisierter Form mit dem Penalisierungsparameter λ darstellen (Hastie et al., 2009):

$$(\hat{\beta}_0, \hat{\beta})_{Lasso} = \underset{\beta_0, \beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left(\frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right),$$

mit $\lambda \geq 0$. Umso größer man λ (bzw. umso kleiner man t) wählt, desto stärker ist die Bestrafung oder Schrumpfung. Zur optimalen Wahl von λ wird oft auf ein Kreuzvalidierungsverfahren zurückgegriffen (Hastie and Qian, 2014).

Im Falle einer binären Outcomevariable $Y = \{0, 1\}$ wird, wie bereits erwähnt, häufig die logistische Regression verwendet. Der Lasso-Schätzer maximiert dann die Log-Likelihood anstatt die Residuenquadratsumme zu minimieren. Wie bereits definiert, ist $\pi = P(Y = 1|X = \mathbf{x}) = \frac{exp(\beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{x}_j)}{1 + exp(\beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{x}_j)}$ und analog $1 - \pi = P(Y = 0|X = \mathbf{x}) = \frac{1}{1 + exp(\beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{x}_j)}$. Die Dichte von einer Realisation y_i der Zufallsvariable Y sieht dann folgendermaßen aus (Fahrmeir et al., 2007):

$$f(y_i|\pi) = P(Y = y_i|\pi) = \pi^{y_i}(1 - \pi)^{1 - y_i}.$$

Die Likelihoodfunktion ist dann die Dichte unabhängiger und identischer Wiederholungen (Fahrmeir et al., 2007):

$$L(\pi) = f(y_1, ..., y_n|\pi) = f(y_1|\pi) \cdot ... \cdot f(y_n|\pi) = \prod_{i=1}^{n} \pi^{y_i}(1 - \pi)^{1 - y_i}.$$

Logarithmiert man diese Likelihoodfunktion und setzt $\frac{exp(\beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{x}_j)}{1 + exp(\beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{x}_j)}$ ein, erhält man folgende Log-Likelihood:

$$\begin{aligned} l(\pi) = log(L(\pi)) &= \sum_{i=1}^{n} y_i \cdot log(\pi) + (1 - y_i) \cdot log(1 - \pi) \\ &= \sum_{i=1}^{n} y_i \cdot (\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j) - log(1 + exp(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)). \end{aligned}$$

Schließlich ergibt sich mit dieser Log-Likelihood die folgende penalisierte Form des Lasso-Schätzers für logistische Regressionen (Hastie et al., 2009):

$$(\hat{\beta}_0, \hat{\beta})_{Lasso} = \underset{\beta_0, \beta \in \mathbb{R}^{p+1}}{\operatorname{argmax}} \left(\sum_{i=1}^{n} \left[y_i \cdot (\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}) - log(1 + exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right). \tag{4}$$

Es ist noch anzumerken, dass es sich bei dem Lasso-Schätzer um keine geschlossene Form handelt (und durch die L_1 Penalisierung die Lösung nicht linear in y_i ist)(Hastie et al., 2009). Um die quadratische Gleichung zu lösen gibt es verschiedene Algorithmen. Das Paket glmnet nutzt beispielsweise einen Algorithmus mit CCD (cyclical coordinate descent) (Hastie and Qian, 2014).

Idee des neuen Ansatzes

Bevor der neue Ansatz formal definiert wird, soll zunächst die Idee kurz beschrieben werden. Statt vieler Einfachregressionen wie bei Matsui et al. (2012), soll hier eine Lasso-Regression geschätzt werden und dann auf Basis dieser Variablenselektion die Scores gebildet werden. Das heißt alle Genexpressionen kommen als Kovariablen mit ins Modell, was den Vorteil hat, dass auch die Zusammenhangsstruktur der Kovariablen berücksichtigt werden kann. Ein weiterer Vorteil ist, dass hier kein multiples Testproblem mehr entsteht, da die Variablenselektion nicht mehr unter Anwendung von statistischen Signifikanztests geschieht. Alle Genexpressionen mit einem Haupteffekt ungleich 0 werden Teil des prognostischen Scores. Dieser stellt eine gewichtete Summe der Ausprägungen dieser Genexpressionen mit ihren zugehörigen β-Werten dar. Analog setzt sich der prädiktive Score aus der Summe der Ausprägungen der Genexpressionen mit ihren zugehörigen Interaktionseffekten zusammen. Die β-Koeffizienten können hier als Gewichte verwendet werden, da die Lasso-Regression auf standardisierten Werten basiert und somit auch die daraus resultierenden Koeffizienten standardisiert sind. Das heißt die Größe der Koeffizienten hängt nicht von der Skala, auf der die beobachteten Werte gemessen wurden, ab. Diese Standardisierung erfolgt bei der Anwendung mit dem R-Paket glmnet intern (Hastie and Qian, 2014). Das heißt die Werte werden zur Schätzung standardisiert und anschließend wieder zurück transformiert. Die resultierenden Koeffizienten sind somit ebenfalls standardisiert, werden aber anschließend auch auf die ursprüngliche Skala zurück transformiert, so dass sie als Gewicht für die ursprünglichen Werte geeignet sind.

Die erste intuitive Variante dieses Ansatzes erfolgt einschrittig. Einschrittig bedeutet dabei, dass gleich im ersten Schritt sowohl alle Haupteffekte der Genexpressionen als auch ihre Interaktionen mit dem Treatment mit in die Lasso-Regression aufgenommen werden. Die-

se Vorgehensweise bringt den Nachteil mit sich, dass durch das einschrittige Vorgehen die Möglichkeit besteht, dass die Interaktionen zu viel Bedeutung in der Schätzung bekommen. Daher werden zusätzlich zweischrittige Ansätze vorgeschlagen.
Zweischrittig bedeutet dabei, dass Haupteffekte und Interaktionen in zwei aufeinander folgenden Schritten betrachtet werden. Im ersten Schritt wird eine Lasso-Regression geschätzt, die nur die Haupteffekte der Genexpressionen enthält. Damit soll sichergestellt werden, dass zunächst möglichst viel Streuung anhand der Haupteffekte erklärt wird und erst im zweiten Schritt die Reststreuung mit Hilfe der Interaktionseffekte erklärt wird. Dadurch wird verhindert, dass den Interaktionen zu viel Bedeutung zukommt. Damit die Interaktionen nur die Reststreuung erklären, wird der lineare Prädiktor des Haupteffektmodells des ersten Schrittes als Offset für das Interaktionsmodell des zweiten Schrittes verwendet. Im zweiten Schritt stellt sich dabei die Frage welche Interaktionen ins Modell aufgenommen werden sollen. Häufig wird die Meinung vertreten, dass nur Interaktionsterme ins Regressionsmodell aufgenommen werden sollen, wenn auch die zugehörigen Haupteffekte im Modell sind. Das bedeutet in diesem Fall, dass lediglich Interaktionsterme von Genexpressionen mit einem β-Koeffizienten ungleich 0 aus dem ersten Schritt mit in das Modell aus dem zweiten Schritt aufgenommen werden dürfen. Nachteil dieses Vorgehens ist es, dass durchaus nicht alle Genexpressionen, die einen Interaktionseffekt mit dem Treatment haben, auch einen Haupteffekt auf die Zielvariable haben müssen. Inhaltlich bedeutet das, dass nicht nur genetische Variablen, die für den prognostischen Score von Bedeutung sind für den prädiktiven Score relevant sein können. Bei diesem Ansatz könnte somit die Gefahr bestehen, dass wichtige Genexpressionen für den prädiktiven Score verloren gehen. Weshalb eine weitere Variante des zweischrittigen Ansatzes betrachtet wird. Diese lässt im zweiten Schritt alle Interaktionsterme zu. Dadurch erscheint diese Variante inhaltlich gesehen sinnvoller.
Welches Verfahren zu bevorzugen ist und in welchen Situationen wird schließlich mit Hilfe der Simulationsstudie in Kapitel 4 untersucht.

Genaues Vorgehen

Wie oben bereits beschrieben wurde, werden die Kovariablen vor der Schätzung der Lasso-Regression standardisiert. Da es sich bei der Treatmentvariable um eine binäre Variable handelt, kann diese nicht sinnvoll standardisiert werden. Deshalb wurde hier der Haupteffekt des Treatments vorab anhand einer normalen logistischen Regression ohne Penalisierung wie in Gleichung (2) geschätzt

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_T t_i = \eta_i^t \tag{5}$$

und der daraus resultierende lineare Prädiktor η_i^t für die darauf folgenden Lasso-Regressionen als Offset verwendet. Inhaltlich lässt sich dieses Vorgehen folgendermaßen interpretieren. Es wird erst betrachtet welchen Anteil der Streuung das Treatment ungeachtet anderer Einflussfaktoren erklären kann und dann wird versucht die Reststreuung anhand der genetischen Variablen zu erklären. In den folgenden Modellen findet sich daher kein Parameter für das Treatment, da dessen Haupteffekt bereits im Offset enthalten ist.

Das einschrittige Verfahren

Wie bereits beschrieben, werden beim einschrittigen Verfahren alle Haupt- und Interaktionseffekte auf einmal ins Modell aufgenommen. Damit ergibt sich nach Formel (4) der Lasso-Schätzer

$$\begin{aligned}(\hat{\beta}_0, \hat{\beta})_{Lasso} = \underset{\beta_0, \beta \in \mathbb{R}^{2p+1}}{\operatorname{argmax}} \Bigg(& \sum_{i=1}^{n} \Big[y_i \cdot (\beta_0 + \sum_{j=1}^{p} \beta_j^H x_{ij} + \sum_{j=1}^{p} \beta_j^I (t_i x_{ij}) + \eta_i^t) \\ & - log(1 + exp(\beta_0 + \sum_{j=1}^{p} \beta_j^H x_{ij} + \sum_{j=1}^{p} \beta_j^I (t_i x_{ij}) + \eta_i^t)) \Big] \\ & - \lambda \sum_{j=1}^{p} (|\beta_j^H| + |\beta_j^I|) \Bigg),\end{aligned} \tag{6}$$

mit $\boldsymbol{\beta}^H = (\beta_1^H, ..., \beta_p^H)^T$ dem Parametervektor der Haupteffekte und $\boldsymbol{\beta}^I = (\beta_1^I, ..., \beta_p^I)^T$ dem Parametervektor der Interaktionseffekte.

Die zweischrittigen Verfahren

Hier wird zuerst das Haupteffektmodell

$$\begin{aligned}(\hat{\beta}_0, \hat{\beta})_{Lasso} = \underset{\beta_0, \beta \in \mathbb{R}^{p+1}}{\operatorname{argmax}} \Bigg(& \sum_{i=1}^{n} \Big[y_i \cdot (\beta_0 + \sum_{j=1}^{p} \beta_j^H x_{ij} + \eta_i^t) \\ & - log(1 + exp(\beta_0 + \sum_{j=1}^{p} \beta_j^H x_{ij} + \eta_i^t)) \Big] \\ & - \lambda \sum_{j=1}^{p} |\beta_j^H| \Bigg),\end{aligned} \tag{7}$$

mit dem Offset η^t aus Gleichung (5) geschätzt. Im zweiten Schritt wird anschließend das Interaktionsmodell mit dem resultierenden linearen Prädiktor $\eta_i^h = \beta_0 + \sum_{j=1}^{p} (\beta_j^H x_{ij}) + \eta_i^t$ aus dem Haupteffektmodell als Offset gefittet.

Version I

Sei $S = \{j|\beta_j^H \neq 0 \ \forall j\}$ die Menge der Kovariablen mit einem β-Koeffizienten ungleich Null. Dann sieht bei der ersten zweischrittigen Version das Interaktionsmodell des zweiten Schrittes folgendermaßen aus

$$\begin{aligned}(\hat{\beta}_0, \hat{\beta})_{Lasso} = \underset{\beta_0, \beta \in \mathbb{R}^{p+1}}{\operatorname{argmax}} \Bigg(& \sum_{i=1}^{n} \Big[y_i \cdot (\beta_0 + \sum_{s \in S} \beta_s^I (t_i x_{is}) + \eta_i^h) \\ & - log(1 + exp(\beta_0 + \sum_{s \in S} \beta_s^I (t_i x_{is}) + \eta_i^h)) \Big] \\ & - \lambda \sum_{s \in S} |\beta_s^I| \Bigg),\end{aligned} \tag{8}$$

dabei gilt $\beta_j^I = 0$ für $j \notin S$. Dieses lässt also nur Interaktionseffekte von Genexpressionen mit einem Haupteffekt ungleich Null zu.

Version II

Die andere zweischrittige Version lässt im zweiten Schritt alle Interaktionen zu. Somit ergibt sich dieses Interaktionsmodell:

$$\begin{aligned}(\hat{\beta}_0, \hat{\beta})_{Lasso} = \underset{\beta_0, \beta \in \mathbb{R}^{p+1}}{\operatorname{argmax}} \Bigg(& \sum_{i=1}^{n} \Big[y_i \cdot (\beta_0 + \sum_{j=1}^{p} \beta_j^I (t_i x_{ij}) + \eta_i^h) \\ & - log(1 + exp(\beta_0 + \sum_{j=1}^{p} \beta_j^I (t_i x_{ij}) + \eta_i^h)) \Big] \\ & - \lambda \sum_{j=1}^{p} |\beta_j^I| \Bigg).\end{aligned} \tag{9}$$

Score-Bildung

Wie bei der Idee des neuen Ansatzes bereits beschrieben, bilden die Scores die gewichteten Summen der Ausprägungen der Genexpressionen mit den zugehörigen Haupt- bzw. Interaktionseffekten.

Der prädiktive Score für Person i berechnet sich somit nach dem neuen Ansatz folgendermaßen und wird mit U_i^* bezeichnet:

$$U_i^* = \sum_{j=1}^{p} \hat{\beta}_j^H x_{ij}.$$

Der prognostische Score für Person i berechnet sich analog:

$$W_i^* = \sum_{j=1}^{p} \hat{\beta}_j^I x_{ij}.$$

Wobei je nach Schätzung der Lasso-Regressionen einige β-Koeffizienten gleich Null sind und somit die zugehörigen Genexpressionen nicht in die gewichtete Summe eingehen.

3.2.3 Prädiktionsmodell

In diesem Abschnitt werden zwei Prädiktionsmodelle vorgestellt, die im Simulationsteil genutzt werden um die Ergebnisse der unterschiedlichen Methoden zur Score-Bildung zu vergleichen.
Ähnlich wie bei Chen et al. (2015) beschrieben sollen anhand der in Kapitel 3.2 geschätzten Scores die Ausprägungen der Zielvariable vorhergesagt werden. Die bereits beobachteten Werte dieser Outcomevariable dienen dann als Label, um überprüfen zu können wie gut die Vorhersage ist.

Methode von Matsui
Bei dem Ansatz von Matsui et al. (2012) werden dazu, wie bei der Score-Bildung, die aus der Kreuzvalidierung gebildeten Trainings- und Testdaten genutzt. Anhand der Trainingsdaten wird ein Prädiktionsmodell geschätzt und anschließend die Prädiktionsgenauigkeit des Modells auf den Testdaten evaluiert.
Das Prädiktionsmodell ist somit wieder ein logistisches Regressionsmodell und sieht folgendermaßen aus:

$$\pi_i = P(y_i = 1|T = t_i, W^M = w_i^M, U^M = u_i^M) = \frac{exp(\hat{\beta}_0 + \hat{\beta}_T t_i + \hat{\beta}_1 w_i^M + \hat{\beta}_2 u_i^M + \hat{\beta}_3 t_i u_i^M)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_T t_i + \hat{\beta}_1 w_i^M + \hat{\beta}_2 u_i^M + \hat{\beta}_3 t_i u_i^M)}, \tag{10}$$

wobei π_i die Wahrscheinlichkeit darstellt, dass y_i den Wert 1 annimmt, gegeben die Werte der Kovariablen T, W^M und U^M. Die geschätzten Koeffizienten $\hat{\beta}_0$ für den Intercept, $\hat{\beta}_T$ für den Haupteffekt des Treatments, $\hat{\beta}_1$ für den Haupteffekt des prognostischen Scores, $\hat{\beta}_2$ für den Haupteffekt des prädiktiven Scores und $\hat{\beta}_3$ für den Interaktionseffekt zwischen Treatment und prädiktivem Scores stammen dabei aus einer Schätzung auf den Trainingsdaten.
Man könnte sich an dieser Stelle auch überlegen das Prädiktionsmodell ohne den Haupteffekt des prädiktiven Scores zu schätzen, da es bei diesem eigentlich nur um die Interaktion geht. Abbildung 12 im Anhang vergleicht die beiden Optionen und zeigt, dass kaum Unterschiede bestehen.

Neuer Ansatz

Da der neue Ansatz zur Score-Bildung bereits die aus den Trainingsdaten geschätzten β-Koeffizienten als Gewichte nutzt, muss hier das Prädiktionsmodell nicht mehr explizit geschätzt werden. Die Vorhersage kann dann folgendermaßen für jede Beobachtung aus dem Testdatensatz berechnet werden:

$$P(y_i = 1|\mathbf{x}_i, t_i) = \frac{exp(\eta_i^t + \hat{\beta}_1^H x_{i1} + ... + \hat{\beta}_p^H x_{ip} + \hat{\beta}_1^I x_{i1} + ... + \hat{\beta}_p^I x_{ip})}{1 + exp(\eta_i^t + \hat{\beta}_1^H x_{i1} + ... + \hat{\beta}_p^H x_{ip} + \hat{\beta}_1^I x_{i1} + ... + \hat{\beta}_p^I x_{ip})},$$

wobei man die geschätzten Koeffizienten $\hat{\boldsymbol{\beta}}^H = (\hat{\beta}_1^H, ..., \hat{\beta}_p^H)^T$ und $\hat{\boldsymbol{\beta}}^I = (\hat{\beta}_1^I, ..., \hat{\beta}_p^I)^T$ aus der Lasso-Regression des einschrittigen Verfahrens (6) bzw. der zweischrittigen Verfahren ((7), (8), (9)) erhält. Die Schätzung des Intercepts und des Haupteffekts des Treatments stecken in dem linearen Prädiktor η^t, der aus dem vorab gefitteten logistischen Regressionsmodell (5) resultiert.

Eine Möglichkeit die Performance des Prädiktionsmodells darzustellen ist die Konfusionsmatrix, wie sie in Tabelle 1 zu sehen ist (Pepe, 2003; Swets, 1988).

Tabelle 1: Konfusionsmatrix

	Label 0	Label 1	
Vorhersage 0	richtig negativ (TN)	falsch negativ (FN)	TN + FN
Vorhersage 1	falsch positiv (FP)	richtig positiv (TP)	FP + TP
	TN + FP	FN + TP	n

Diese zeigt die richtigen (TP und TN) und die falschen (FP und FN) Vorhersagen. Wurde beispielsweise bei Patient i Genesung beobachtet, es gilt also $y_i = 1$ und somit hat Patient i das Label 1, und die Vorhersage besagt 0 also keine Genesung, so handelt es sich um eine falsche Vorhersage in Form eines falsch negativen Ergebnisses. Betrachtet man den Anteil falscher Prädiktionen ($\frac{FP+FN}{N}$), so erhält man den Missklassifikationsfehler.
Aus dem oben definierten Prädiktionsmodell erhält man jedoch zunächst nur Wahrscheinlichkeiten dafür, dass die Zielvariable Y den Wert 1 annimmt und noch keine Entscheidung über die Ausprägung 1 oder 0. Dazu muss vorher noch ein sogenannter Schwellenwert festgelegt werden, der angibt ab welcher Wahrscheinlichkeit die Ausprägung der Zielvariable auf 1 gesetzt wird (Pepe, 2003). In der Praxis wird dieser Schwellenwert häufig intuitiv bei 0.5 oder nahe 0.5 gewählt. Je nach gewähltem Schwellenwert können sehr unterschiedliche Ergebnisse herauskommen und sich folgendermaßen auch unterschiedliche Missklassifikationsfehler ergeben. Eine Möglichkeit der grafischen Darstellung der Vorhersagekraft für verschiede-

ne Schwellenwerte, stellt die sogenannte ROC (receiver operating characteristic) dar (Pepe, 2003; Swets, 1988). Die ROC Kurve bildet dabei den Anteil der richtig Positiven gegenüber dem Anteil der falsch Positiven ab.
Zwei Begriffe, die man in der Medizin häufig in diesem Zusammenhang hört und in Kapitel 2 genannt wurden, sind die Sensitivität und die Spezifität (Pepe, 2003).

Sensitivität ist die bedingte Wahrscheinlichkeit, dass 1 vorhergesagt wird, gegeben das wahre Label ist 1 $\widehat{=}$ $P(Vorhersage = 1|Label = 1)$. Als Schätzer dient der Anteil richtig Positiver (TPF): $TPF = \frac{TP}{TP+FN}$.

Spezifität ist die bedingte Wahrscheinlichkeit, dass 0 vorhergesagt wird, gegeben das wahre Label ist 0 $\widehat{=}$ $P(Vorhersage = 0|Label = 0)$. Als Schätzer dient der Anteil richtig Negativer (TNF): $TNF = \frac{TN}{TN+FP}$.

Die ROC Kurve bildet hierbei die Sensitivität auf der y-Achse und (1 - Spezifität) auf der x-Achse für alle möglichen Schwellenwerte (vom kleinsten bis zum größten) ab. (1 - Spezifität) ist dann $1 - TNF = \frac{FP}{TN+FP} = FPF$, also der Anteil der falsch Positiven (FPF). Bei einem guten Vorhersagemodell verläuft die ROC Kurve in einem großen Bogen über der Winkelhalbierenden. Ein schlechtes Vorhersagemodell hat eine ROC Kurve eng an der Winkelhalbierenden liegend. Abbildung 1 verdeutlicht dies anhand zwei fiktiver Beispiele.
Zum Vergleich von ROC Kurven wie in Abbildung 1 wird auch häufig die Fläche unter den Kurven herangezogen. Das sogenannte AUC (Area under the curve) fasst somit die ROC Kurve in einem Wert zusammen indem es die Fläche unter ihr angibt(Pepe, 2003). Damit stellt das AUC ein Maß, das ebenfalls zum Vergleich von verschiedenen Prädiktionsmodellen herangezogen werden kann, aber unabhängig der Wahl des Schwellenwertes ist, dar.
L_1 bezeichne die Menge der Beobachtungen mit Label 1, also $y_j = 1$ und L_0 die Menge der Beobachtungen mit Label 0, also $y_i = 0$. Angenommen man zieht aus jeder Gruppe zufällig eine Beobachtung, dann gibt das AUC die Wahrscheinlichkeit an, dass π_j von der Beobachtung aus L_1 größer ist als π_i der Beobachtung aus L_0 (LeDell et al., 2015). Somit lässt sich das AUC folgendermaßen empirisch berechnen(LeDell et al., 2015):

$$\widehat{AUC} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I(\hat{\pi}_j > \hat{\pi}_i),$$

mit n_0 der Anzahl der Elemente aus L_0 und n_1 der Anzahl der Elemente aus L_1 und I der Indikatorfunktion, die zählt wie oft $\hat{\pi}_j > \hat{\pi}_i$ gilt. Das AUC nimmt im allgemeinen Werte zwischen 0.5 (unbrauchbares Vorhersagemodell) und 1.0 (perfektes Vorhersagemodell) an.

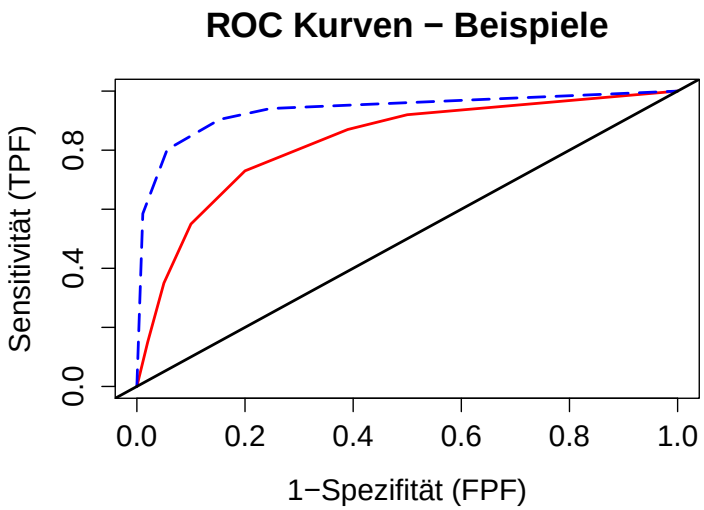

Abbildung 1: Die blaue ROC Kurve (gestrichelte Linie) stellt eine bessere Vorhersage als die rote Kurve (durchgezogene Linie) dar, da sie die rote Kurve dominiert. Das heißt die blaue Kurve ist immer über der roten Kurve. (Quelle: eigene Berechnungen)

Da die Analyse jedoch auf Kreuzvalidierung basiert, soll hier nicht das AUC über alle Beobachtungen berechnet werden, sondern das cvAUC (cross validated AUC). Das heißt ein AUC, das die K-fache Aufteilung in Trainings- und Testdaten berücksichtigt. Bezeichne dazu $B_n^1, ..., B_n^K$ wie bei LeDell et al. (2015) die K Unterteilungen in Trainings- und Testdaten, wobei $B_n^k \in \{0,1\}^n$. Die Beobachtungen der k-ten Testdaten sind dann gekennzeichnet durch $\{i : B_n^k(i) = 1\}$ und die Beobachtungen der zugehörigen Trainingsdaten durch $\{i : B_n^k(i) = 0\}$. $n_0^k = \sum_{i=1}^n I(y_i = 0)I(B_n^k(i) = 1)$ bezeichnet dann die Anzahl Beobachtungen aus dem k-ten Testdatensatz mit $y_i = 0$ und $n_1^k = \sum_{i=1}^n I(y_i = 1)I(B_n^k(i) = 1)$ die Anzahl Beobachtungen aus dem k-ten Testdatensatz mit $y_i = 1$.
Für einen einzelnen Testdatensatz $\{i : B_n^k(i) = 1\}$ der Kreuzvalidierung sieht das AUC nach LeDell et al. (2015), folgendermaßen aus:

$$\widehat{AUC}_k = \frac{1}{n_0^k n_1^k} \sum_{i=1}^{n_0^k} \sum_{j=1}^{n_1^k} I(\hat{\pi}_j > \hat{\pi}_i).$$

Das cvAUC über alle Testdaten der K-fachen Kreuzvalidierung ist dann als Mittelwert aller AUCs der einzelnen K Testdaten definiert (LeDell et al., 2015):

$$\widehat{cvAUC} = \frac{1}{K} \sum_{k=1}^{K} \widehat{AUC}_k = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_0^k n_1^k} \sum_{i=1}^{n_0^k} \sum_{j=1}^{n_1^k} I(\hat{\pi}_j > \hat{\pi}_i). \tag{11}$$

3.3 Testen

Gemäß Chen et al. (2015) gehört es zur Entwicklung von prädiktiven Biomarkern dazu, dass man den Treatmenteffekt nicht nur schätzt sondern auch testet. Dabei können verschiedene Hypothesen interessant sein. Dieses Kapitel beschreibt dazu erst den Permutationstest von Matsui et al. (2012) und anschließend welchen Test der neue Ansatz anstreben würde. Dabei werden zwei Permutationstest vorgestellt, die für die Umsetzung des Tests von Nutzen sein könnten.

3.3.1 Methode von Matsui

Matsui et al. (2012) testen in ihrem Paper die Nullhypothese,

$$H_0^M : \text{es gibt für die gesamte Population keinen Treatmenteffekt} \mathrel{\widehat{=}} H_0^M : \beta_T = \beta_3 = 0.$$

Dazu stellen sie den Treatmenteffekt durch eine Funktion $\Psi(u^M)$ vom prädiktiven Score U^M dar. Da es sich in diesem Unterkapitel bei allen Scores um die Scores nach Matsui et al. (2012) handelt, wird zur besseren Lesbarkeit im Folgenden auf den Index M verzichtet. Die an das binäre Setting angepasste Funktion für einen Patienten i sieht dabei folgendermaßen aus:

$$\begin{aligned}\Psi(u_i) &= log\left(\frac{P(y_i = 1|t_i = 1)}{1 - P(y_i = 1|t_i = 1)}\right) - log\left(\frac{P(y_i = 1|t_i = 0)}{1 - P(y_i = 1|t_i = 0)}\right) \\ &= (\beta_0 + \beta_T \cdot 1 + \beta_1 w_i + \beta_2 u_i + \beta_3 u_i \cdot 1) - (\beta_0 + \beta_t \cdot 0 + \beta_1 w_i + \beta_2 u_i + \beta_3 u_i \cdot 0) \\ &= \beta_t + \beta_3 u_i.\end{aligned}$$

$\Psi(u_i)$ steht für die Differenz der logarithmierten Chancen. Eine Chance, auch odds genannt, stellt dabei die Wahrscheinlichkeit für $y_i = 1$ ins Verhältnis zur Wahrscheinlichkeit für $y_i = 0$ (Fahrmeir et al., 2009). Eine Chance größer 1 bzw. eine logarithmierte Chance größer 0 bedeutet, dass bei gegebenen Kovariablenwerten die Wahrscheinlichkeit für $y_i = 1$ größer ist als für $y_i = 0$. Steht beispielsweise $y_i = 1$ für Genesung, so würde $\Psi(u_i) > 0$ bedeuten, dass mit Treatment die logarithmierte Chance auf Genesung größer ist als ohne. Es gibt folglich einen Behandlungseffekt oder -nutzen für Patient i mit dem prädiktiven Score u_i. Die Schätzungen der β-Koeffizienten erhält man dabei aus dem vorher beschriebenen Prädiktionsmodell (10). Um die Nullhypothese 2-seitig zu testen, schlagen Matsui et al. (2012) einen Permutationstest mit der Teststatistik

$$T = \int |\hat{\Psi}(u)| \, \mathrm{d}u$$

vor. Die Schätzung von T unter H_0 wird dabei durch Permutation des Treatments erzielt. Das heißt die Einträge der Spalte mit den Treatmentwerten werden zufällig umsortiert. Dann werden wie zuvor die Scores für jeden Patienten und die zugehörigen β-Koeffizienten geschätzt und anschließend die $\hat{\Psi}(u_i)$ berechnet. Dieses Vorgehen wird mehrmals wiederholt und schließlich kann ein p-Wert über die Anzahl der Permutationen mit $T \geq T_{obs}$ gebildet werden. Wobei T_{obs} für die Teststatistik der ursprünglich beobachteten Daten steht. Tritt $T \geq T_{obs}$ zu häufig auf weist dies auf keinen Treatmenteffekt hin.

3.3.2 Idee des neuen Ansatzes

Liegt das Interesse nicht am globalen Treatmenteffekt wie in H_0^M, sondern ausschließlich am Interaktionseffekt zwischen dem Treatment und den genetischen Variablen, sieht die interessierende Nullhypothese folgendermaßen aus:

$$H_0^* : \text{es gibt keine Interaktionen} \mathrel{\widehat{=}} H_0^* : \beta_3 = 0, \text{ für alle genetischen Variablen.}$$

Um wirklich nur den Interaktionseffekt ohne den Haupteffekt auf Signifikanz zu testen, kann der eben beschriebene Permutationstest nicht ohne weiteres verwendet werden. Denn ein einfaches Permutieren des Treatments zerstört nicht nur die Struktur der Interaktionen, sondern gleichzeitig auch die Struktur des Haupteffekts des Treatments. Daher werden andere Methoden des Permutationstests benötigt, um vergleichbar wie bei Matsui et al. (2012) die neue Nullhypothese testen zu können.
Dazu werden im Folgenden zwei Permutationstests für Interaktionen vorgestellt, die für die neue Nullhypothese adäquater erscheinen.

Permutationstest von Werft et al.
Werft et al. (2012) schlagen einen PRR Test (permutation of regerssor residuals) zur Identifikation von prädiktiven Biomarkern vor. Dabei verfolgen sie einen genweisen Ansatz. Das heißt sie betrachten für jede genetische Variable ein einzelnes generalisiertes lineares Modell und testen auch für jede genetische Variable einzeln die Nullhypothese

$$H_0^{Werft} : \text{es gibt keine Interaktion} \mathrel{\widehat{=}} H_0^{Werft} : \beta_2^{(j)} = 0, \ j = 1, ..., p,$$

wobei der Interaktionseffekt $\beta_2^{(j)}$ für die j-te Genexpression aus dem folgenden Modell stammt,

$$E(Y) = f^{-1}[\beta_0^{(j)} + \beta_T^{(j)} T + \beta_1^{(j)} X_j + \beta_2^{(j)} T X_j + \sum_{s=1}^{q} \beta_{s+2}^{(j)} O_s],$$

das hier in ganz allgemeiner Form dargestellt wird, so dass je nach Daten die passende Linkfunktion f einfügt werden kann. Zusätzlich zu den genetischen Variablen X_j, $j = 1, ..., p$, und der Treatmentvariable T, werden hier noch mögliche weitere Kovariablen O_s, $s = 1, ..., q$, wie klinische Eigenschaften der Patienten beobachtet.
Da dieser Ansatz jede genetische Variable einzeln betrachtet, tritt hier erneut das multiple Testproblem auf. Werft et al. (2012) greifen dabei ebenfalls auf die Adjustierung der FDR, die in Kapitel 2 kurz beschrieben wurde, zurück.
Die Idee des Tests ist es die interessierende, zu testende Größe durch die Residuen aus einem Modell von allen anderen Variablen auf diese interessierende Größe zu ersetzen. Das bedeutet es wird erst ein Modell

$$E(X_j T) = \gamma_0^{(j)} + \gamma_T^{(j)} T + \gamma_1^{(j)} X_j + \sum_{s=1}^{q} \gamma_{s+1}^{(j)} O_s$$

mit dem (stetigen) Interaktionsterm der genetischen Variable und der Treatmentvariable als abhängige Variable gefittet. Die Koeffizienten dieses Modells sind dabei zur besseren Unterscheidung mit γ's dargestellt. Die daraus resultierenden Residuen

$$r = X_j T - (\hat{\gamma}_0^{(j)} + \hat{\gamma}_T^{(j)} T + \hat{\gamma}_1^{(j)} X_j + \sum_{s=1}^{q} \hat{\gamma}_{s+1}^{(j)} O_s)$$

sind gemäß ihrer Definition unkorreliert mit den Kovariablen, aber korreliert mit der abhängigen Variable, in diesem Fall also dem Interaktionsterm. Da das Maximum der Likelihoodfunktion das gleiche ist, egal ob man für das generalisierte lineare Modell die Residuen r als erklärende Variable einsetzt oder die eigentlich beobachtete Kovariable, in diesem Fall den Interaktionsterm, hat das zur Folge, dass auch der Likelihood-Ratio-Test (LR Test) auf das selbe Ergebnis kommt, wenn man statt dem eigentlichen Interaktionsterm die Residuen verwendet (Werft et al., 2012).
Werft et al. (2012) berechnen also zunächst den p-Wert $\tilde{p}$ für die ursprünglichen Daten anhand des LR Tests

$$LR(X_j) = -2log(\frac{L(\beta_2^{0(j)})}{L(\hat{\beta}_2^{(j)})}). \tag{12}$$

Wobei $\beta_2^{0(j)}$ den Interaktionsterm unter der Nullhypothese (also in diesem Fall gleich Null) bezeichnet und $\hat{\beta}_2^{(j)}$ den geschätzten Interaktionseffekt des Modells mit den Residuen anstelle des Interaktionsterms.
Anschließend werden die Residuen immer wieder randomisiert r_b^*, $b = 1, ..., B$ und die p-Werte p_b^* mittels dem LR Test berechnet.
Schließlich ist der p-Wert p_j für die j-te genetische Variable des PRR Tests nach Werft et al.

(2012) folgendermaßen definiert:

$$p_j = \frac{I(p_b^* \leq \tilde{p})}{B}.$$

Dabei zählt die Indikatorfunktion I wie viele p-Werte nach Randomisierung der Residuen kleiner gleich dem p-Wert der beobachteten Daten sind. Je seltener dies auftritt, desto wahrscheinlicher ist es, dass es sich bei dem beobachteten Interaktionseffekt um einen tatsächlich signifikanten Effekt handelt. Denn werden nach der Zerstörung der Interaktionsstruktur überwiegend größere p-Werte beobachtet, spricht das dafür, dass der kleine p-Wert der ursprünglich beobachteten Daten nicht zufällig beobachtet wurde.

Aufgrund der univariaten Herangehensweise ist diese Methode jedoch nicht optimal für die Testidee des neuen Ansatzes, der versucht aus der univariaten Methode von Matsui et al. (2012) einen multiplen Ansatz zu machen. Vielleicht wäre es aber eine Möglichkeit, um mit der Methode von Matsui et al. (2012) die Nullhypothese H_0^* zu testen, was auch interessant sein könnte.

Permutationstest von Wang et al.

Wang et al. (2015) schlagen einen Permutationstest vor, der die Nullhypothese testen soll, welche besagt, dass es keine Interaktionen gibt. Das heißt hier wird nicht für jede Kovariable ein eigener Test durchgeführt, sondern für alle gleichzeitig. Dadurch werden im Vergleich zur vorher betrachteten Methode von Werft et al. (2012) alle verfügbaren Informationen genutzt und die Wahrscheinlichkeit auf ein falsch positives Ergebnis durch die vielen Kovariablen nicht erhöht (Wang et al., 2015).

Der Test ist für randomisierte Studien mit Z (≥ 2) verschiedenen Behandlungen anwendbar. Von den insgesamt n beobachteten Patienten haben jeweils n_z, $z = 1, .., Z$, die Behandlung z erhalten. Dabei handelt es sich bei den Beobachtungen um unabhängige und identisch verteilte Zufallsvariablen des Zufallsvektors $(Y, T, X_1, ..., X_p)$. Die hier betrachtete Zielvariable Y ist stetig und die Treatmentvariable T ist dummykodiert mit der Referenzkategorie Z. Dabei hat die Matrix $\mathbf{T}$ mit der dummykodierten Treatmentvariable die Dimension $(Z-1) \times n$ und die Kovariablenmatrix $\mathbf{X}$ hat die Dimension $p \times n$. Wang et al. (2015) stellen folgendes lineares Modell auf:

$$\mathbf{y} = \boldsymbol{\alpha}^T\mathbf{T} + \boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\gamma}^T\mathbf{T} \otimes \mathbf{X} + \boldsymbol{\epsilon}_0,\ \boldsymbol{\epsilon}_0 \perp (\mathbf{T}^T, \mathbf{X}^T), \tag{13}$$

mit den Parametervektoren $\boldsymbol{\alpha}$ für die Behandlungseffekte, $\boldsymbol{\beta}$ für die Haupteffekte der Kovariablen und $\boldsymbol{\gamma}$ für die Interaktionseffekte zwischen Behandlung und Kovariable. $\boldsymbol{\epsilon}_0 \perp (\mathbf{T}^T, \mathbf{X}^T)$ bedeutet, dass der zufällige Fehlerterm unabhängig vom Treatment und den Kovariablen ist.

Mit $\otimes$ wird das Kronecker Produkt dargestellt. Dieses Modell stellt eine verallgemeinerte Form des bisher betrachteten Modells beim Schätzen der Biomarker dar, da es mehr als zwei Behandlungsformen zulässt und kann somit für mehrere Biomarker verwendet werden. Die interessierende Nullhypothese von Wang et al. (2015) lässt sich dann folgendermaßen darstellen:

$$H_0^{Wang} : \text{es gibt keine Interaktionen} \mathrel{\widehat{=}} H_0^{Wang} : \boldsymbol{\gamma} = \mathbf{0}_{(Z-1)p\times 1}.$$

Wang et al. (2015) betrachten zunächst für jede Behandlung z ein einzelnes lineares Modell

$$\mathbf{y} = \boldsymbol{\beta}^{(z)^T}\mathbf{X} + \boldsymbol{\epsilon}_z,\ \boldsymbol{\epsilon}_z \perp \mathbf{X}, \tag{14}$$

wobei die oben definierte Nullhypothese sich dann äquivalent darstellen lässt durch:

$$H_0^{Wang} : \boldsymbol{\beta}^{(1)} = ... = \boldsymbol{\beta}^{(Z)}.$$

Das bedeutet, wenn die Haupteffekte der Kovariablen in den Modellen getrennt nach den Behandlungen gleich sind, gibt es keine Interaktionseffekte. Denn in diesem Fall scheint der Effekt der Kovariablen über alle Behandlungsformen gleich zu sein. Dies verdeutlichen Wang et al. (2015) nochmal durch folgende Umformung:

$$\begin{aligned}
&\boldsymbol{\beta}^{(z)} = \boldsymbol{\beta} + \boldsymbol{\gamma}^{(z)},\ \text{für } z = 1, ..., Z-1\\
&\boldsymbol{\beta}^{(Z)} = \boldsymbol{\beta}.
\end{aligned}$$

Wobei $\boldsymbol{\gamma}^{(z)}$ ein Vektor der Interaktionseffekte der Behandlung z mit den p Kovariablen ist. Für die Referenzgruppe Z wird dabei kein Interaktionseffekt geschätzt.
Wäre die Anzahl der Kovariablen im Verhältnis zu den Beobachtungen klein ($n > p$), dann wäre es eine Option das lineare Modell (13) und das Modell

$$\mathbf{y} = \boldsymbol{\alpha}^T\mathbf{T} + \boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\epsilon}_0 \tag{15}$$

unter der Nullhypothese, also ohne Ineraktionen, zu schätzen und anschließend einen Likelihood-Ratio-Test vergleichbar wie in (12) durchzuführen. Ist jedoch die Anzahl der Kovariablen groß im Verhältnis zu den Beobachtungen, sodass in den Modellen (13) und (15) mehr Parameter geschätzt werden müssen wie es Beobachtungen gibt oder es gilt ohnehin $n < p$, dann sind die Modelle (13) und (15) nicht schätzbar und somit ist der Likelihood-Ratio-Test nicht mehr anwendbar (Wang et al., 2015). Daher schlagen Wang et al. (2015) einen Permutationstest vor, bei dem es auch möglich ist eine Variablenselektion vorzunehmen.
Dabei wird die Permutation folgendermaßen durchgeführt (Wang et al., 2015). Das Modell

(15) unter der Nullhypothese kann derart umgeformt werden

$$\mathbf{y} = \boldsymbol{\alpha}^T\mathbf{T} + \boldsymbol{\epsilon}, \tag{16}$$

wobei hier $\boldsymbol{\epsilon} = \boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\epsilon}_0$ gilt, da sich die Reststreuung durch das Weglassen der Haupteffekte der Kovariablen um die erklärte Streuung dieser Effekte erhöht. Betrachtet man dazu die übliche Definition des Fehlerterms ergibt sich $\boldsymbol{\epsilon} = \mathbf{y} - \boldsymbol{\alpha}^T\mathbf{T} = \boldsymbol{\beta}^T\mathbf{X} + \boldsymbol{\epsilon}_0 \perp \mathbf{T}|\mathbf{X}$. Das heißt, der Fehlerterm aus (16) ist unabhängig vom Treatment gegeben die Kovariablen. Da es sich um eine randomisierte Studie handelt, sind $\mathbf{T}$ und $\mathbf{X}$ ebenfalls unabhängig voneinander. Es gilt somit $\mathbf{T} \perp \mathbf{X}$. Kombiniert man diese Unabhängigkeitsannahmen, erhält man, dass das Treatment sowohl vom Fehlerterm, als auch von den Kovariablen unabhängig ist. Damit ergibt sich, wenn man $\boldsymbol{\epsilon} = \mathbf{y} - \boldsymbol{\alpha}^T\mathbf{T}$ einsetzt,

$$\mathbf{T} \perp (\mathbf{y} - \boldsymbol{\alpha}^T\mathbf{T}, \mathbf{X}). \tag{17}$$

Im ersten Schritt ersetzen Wang et al. (2015) in der ursprünglichen Datenmatrix $\mathbf{D} = (\mathbf{Y}, \mathbf{T}, \mathbf{X})$ die Werte der Zielvariable durch die Residuen aus Modell (16). Dadurch erhalten sie die transformierte Datenmatrix $\tilde{\mathbf{D}} = (\mathbf{y} - \boldsymbol{\alpha}^T\mathbf{T}, \mathbf{T}, \mathbf{X})$. Aufgrund der Unabhängigkeiten aus (17) kann nun die Spalte mit dem Treatment permutiert werden, während die anderen Spalten mit den Residuen und den Kovariablen gleich bleiben, was die Datenmatrix $\tilde{\mathbf{D}}^\ell = (\mathbf{y} - \boldsymbol{\alpha}^T\mathbf{T}, \mathbf{T}^\ell, \mathbf{X})$ ergibt. Die beiden Datensätze $\tilde{\mathbf{D}}$ und $\tilde{\mathbf{D}}^\ell$ werden als gleich wahrscheinlich angesehen (Wang et al., 2015). Im letzten Schritt werden dann die randomisierten Residuen wieder zurücktransformiert in die beobachteten y-Werte und man erhält den Datensatz $\tilde{\mathbf{D}}^\ell_{(-1)} = (\mathbf{Y}, \mathbf{T}^\ell, \mathbf{X})$. Wobei dieser wieder gleich wahrscheinlich ist, wie der ursprünglich beobachtete Datensatz $\mathbf{D}$. Sei T eine beliebige Teststatistik, so kann die Verteilung unter H_0 durch die Permutationsverteilung, die mit Hilfe der $T(\tilde{\mathbf{D}}^\ell_{(-1)})$ gebildet wird, dargestellt werden. $T(\mathbf{D})$ kann schließlich als Zufallsstichprobe dieser Permutationsverteilung betrachtet werden (Wang et al., 2015).

Die Teststatistik von Wang et al. (2015) wird dann anhand der folgenden fünf Schritte geschätzt:

1. Schätze für jede Behandlung z ein Modell wie in (14), welches gegebenfalls eine Variablenselektion enthalten kann.

2. Berechne für jede Beobachtung i einen Vorhersagewert $\hat{y}_i^{(z)}$ basierend auf dem Modell aus Schritt 1. Dann berechne den Vorhersagefehler über alle Modelle $Err_1 = \frac{1}{n}\sum_{z=1}^{Z}\sum_{i=1}^{n}(y_i - \hat{y}_i^{(z)})^2$.

3. Schätze nun ein gemeinsames Modell für alle Behandlungen wie in Modell (15). Gegebenenfalls kann hier wieder eine Variablenselektion integriert werden.

4. Berechne nun wieder für jede Beobachtung einen Vorhersagewert $\hat{y}_i$ mit dem Modell aus Schritt 3 und berechne anschließend den Vorhersagefehler des Modells $Err_2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

5. Bilde die Teststatistik $\Delta = Err_1 - Err_2$ und gib den p-Wert über alle m Permutationen an. Der von Wang et al. (2015)vorgeschlagene p-Wert sieht dabei folgendermaßen aus:

$$\frac{1 + \sum_{i=1}^{m} I(|\Delta_i| \geq |\Delta_{obs}|)}{1 + m},$$

   wobei I eine Indikatorfunktion darstellt und Δ_{obs} die Teststatistik des ursprünglich beobachteten Datensatzes ist.

Der p-Wert wird folglich groß, wenn viele Teststatistiken nach Permutation betragsmäßig größer gleich der Teststatistik des ursprünglich beobachteten Datensatzes sind. Ein großer p-Wert spricht somit für die Nullhypothese und dafür, dass es keine Interaktionen gibt. Umgekehrt wird die Nullhypothese abgelehnt, wenn es nur wenige Teststatistiken nach Permutation gibt, die größer gleich der Teststatistik des ursprünglich beobachteten Datensatzes sind, weil sich dann ein kleiner p-Wert ergibt. Denn $|\Delta_i| \geq |\Delta_{obs}|$ bedeutet, dass die Zerstörung der Interaktionsstruktur nichts am Ergebnis der Vorhersagefehler geändert hat. Die Vorhersagefehler 1 und 2 unterscheiden sich im gleichen Maße wie beim ursprünglich beobachteten Datensatz. Es scheint somit keine Interaktionseffekte zu geben.

Diese Methode erscheint für die Testidee des neuen Ansatzes geeigneter zu sein, da hier dieselbe Nullhypothese getestet wird. Ein weiterer Vorteil bezüglich der Hochdimesionalität der omics-Daten liefert die Möglichkeit eine Variablenselektion in die Testmethode zu integrieren. Jedoch ist die Tatsache, dass in den Schritten 2 und 4 der Trainingsfehler betrachtet wird nicht optimal, da auf denselben Daten das Prädiktionsmodell gefittet und angewendet wird. Das kann zu Over-Fitting führen, vor allem bei hochdimensionalen Daten, weshalb dieses Vorgehen (für die Ziele dieser Arbeit) nicht empfehlenswert ist. Jedoch wäre es überlegenswert die Methode in Kombination mit Kreuzvalidierung durchzuführen. Demzufolge würden in den Schritten 1 und 3 die Modelle jeweils anhand der Trainingsdaten geschätzt werden und danach in den Schritten 2 und 4 auf die zugehörigen Testdaten angewendet werden. Nachdem man dies für alle K Testdatensätze wiederholt hat, könnte man anschließend den mittleren Vorhersagefehler über alle Testdatensätze angeben. Abschließend

könnte der von Wang et al. (2015) vorgeschlagene p-Wert berechnet werden.
Außerdem ist die Testmethode von Wang et al. (2015) für stetige Zielvariablen konzipiert. Ändert man jedoch die Definition des Vorhersagefehlers, kann diese Methode auch für binäre Outcomevariablen angewendet werden. Da bei einer binären Zielvariable die quadrierte Differenz von Vorhersage und beobachtetem Wert ein ungeeignetes Maß ist, sollte man stattdessen diesen Vorhersagefehler

$$Err = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

mit der Indikatorfunktion I betrachten. Für $\hat{y}_i$ gilt dabei

$$\hat{y}_i = \begin{cases} 1, & \pi_i \geq c \\ 0, & \text{sonst.} \end{cases}$$

Wobei c ein vorher zu definierender Schwellenwert ist, der festlegt ab welcher Wahrscheinlichkeit π_i die Vorhersage $\hat{y}_i$ auf 1 gesetzt wird. Folglich zählt die Indikatorfunktion wie oft Vorhersage und Beobachtung nicht übereinstimmen. Dividiert man diese absolute Häufigkeit durch die Anzahl der insgesamt betrachteten Beobachtungen erhält man den Anteil falscher Vorhersagen. Dieser kann dann als Vorhersagefehler für die binäre Outcomevariable verwendet werden.
Eine angepasste Form der Methode von Wang et al. (2015) könnte folglich eine geeignete Testmethode für die Idee des neuen Ansatzes liefern. Die genaue Ausarbeitung und Simulation dieser Testmethode konnte jedoch im Rahmen dieser Arbeit nicht mehr durchgeführt werden.

4 Simulation

In diesem Kapitel wird eine Simulationsstudie durchgeführt, um die zuvor beschriebenen Schätzmethoden zu vergleichen. Hierzu wird betrachtet, ob die Verfahren die generierten Effekte zum einen erkennen und zum anderen richtig schätzen und schließlich eine gute Vorhersage liefern. Im ersten Teil dieses Kapitels wird der Aufbau der Simulationsstudie beschrieben und darauf folgend die Ergebnisse dargestellt.

4.1 Aufbau der Simulation

Da bei der Simulation die Trainings- und Testdaten so generiert werden, dass beide aus derselben Verteilung stammen, kann auf Kreuzvalidierung verzichtet werden. Bei den Simulationen werden folgende fixe Parameter gewählt:

- Es wird ein Testdatensatz mit $n_{Test} = 10000$ Beobachtungen generiert.
- Dazu werden 100 Trainingsdatensätze mit je $n_{Training} = 200$ Beobachtungen generiert. Anhand des Testdatensatzes können dann die Schätzungen, basierend auf den 100 Trainingsdatensätzen, evaluiert werden. Dazu kann hier das normale AUC herangezogen werden anstelle des cvAUC.
- Die omics-Daten werden durch standard normalverteilte Variablen X_j, $j = 1, ..., p$ mit $p = 1000$, dargestellt. Es gilt somit $X_j \sim N(0, 1)\ \forall j$
- Die binäre Variable T zeigt an, ob die Beobachtung in die Treatment- oder in die Kontrollgruppe gehört. Dabei gilt wie bei einer randomisierten Studie $\pi = 0.5$.
- Die binäre Zielvariable Y nimmt den Wert 1 an, wenn ein positives Ereignis wie beispielsweise Genesung oder Schrumpfung eines Tumors eintritt.

Die Abhängigkeit der Zielvariable von den Kovariablen wird dabei per backward Simulation hergestellt. Das heißt es wird vorgegeben welchen Einfluss die Kovariablen haben sollen und dann gemäß dieser Zusammenhangsstrukturen die Ausprägungen der Zielvariable gebildet. Die unabhängige Generierung der Kovariablen stellt dabei eine eher unrealistische Situation dar. Jedoch sind die wahren Strukturen von omics-Daten sehr komplex, sodass eine realistische Darstellung im Rahmen dieser Arbeit nicht möglich war.
Außerdem werden folgende Annahmen getroffen. Für den Großteil der Haupt- und Interaktionseffekte der Kovariablen wird angenommen, dass sie keinen Einfluss haben. Das heißt, der zugehörige β-Koeffizient ist Null. Einflussreiche Genexpressionen sollen dabei einen negativen Haupteffekt und/oder einen positiven Interaktionseffekt haben. Der Haupteffekt des Treatments soll ebenfalls positiv sein.
Des weiteren gibt es folgende variable Parameter, die zur Gestaltung unterschiedlicher Datensituationen gewählt werden können:

- die Stärke des Haupteffekts des Treatments, sowie der Haupt- und Interaktionseffekte der einflussreichen genetischen Variablen
- die Anzahl der einflussreichen Haupt- und Interaktionseffekte
- die Aufteilung der einflussreichen Interaktionseffekte in Interaktionseffekte mit oder ohne zugehörigem Haupteffekt

Tabelle 2 zeigt je Setting die gewählten Parameter. Da bei der logistischen Regression üblicherweise $exp(\beta)$ interpretiert wird, wird hier auch bei der Stärke der Effekte $exp(\beta)$ angegeben. $exp(\beta)$ entspricht dabei dem Chancenverhältnis, auch Odds Ratio genannt. Wird

beispielsweise x_{i1} um eine Einheit erhöht, gilt für das Chancenverhältnis:

$$\frac{P(y_i = 1|x_{i1}, ...)/P(y_i = 0|x_{i1}, ...)}{P(y_i = 1|x_{i1} + 1, ...)/P(y_i = 0|x_{i1} + 1, ...)} = exp(\beta_1),$$

wobei β_1 der zugehörige Haupteffekt ist (Fahrmeir et al., 2009). Ist $exp(\beta_1) > 1$, bedeutet das, dass sich die Chance auf $y_i = 1$ für $x_{i1} + 1$ im Vergleich zu x_{i1} um den Faktor $exp(\beta_1)$ erhöht (bei Konstanhaltung aller anderen Kovariablen). Analog reduziert sich die Chance auf $y_i = 1$ für $x_{i1} + 1$ im Vergleich zu x_{i1} um den Faktor $exp(\beta_1)$, wenn $exp(\beta_1) < 1$ gilt. $exp(\beta_1) = 1$ bedeutet, dass die Variable X_1 keinen Einfluss hat, was gleichbedeutend mit $\beta_1 = 0$ ist.
Da nur eine begrenzte Anzahl an Settings simuliert werden kann, wird versucht mit den 12 gewählten Settings möglichst viele denkbare Varianten abzudecken. So gibt es einige Settings mit wenigen, aber dafür starken Haupteffekten und einige mit vielen schwachen oder mittleren Haupteffekten. Diese werden mit unterschiedlich starken Treatment- und Interaktionseffekten kombiniert. Bei den Interaktionseffekten wird auch variiert wie viele Interaktionen mit einflussreichen Haupteffekten zusammenhängen und wie viele unabhängig davon auftreten. Außerdem wird betrachtet wie die Methoden darauf reagieren, wenn entweder keine Haupt- oder Interaktionseffekte der Kovariablen vorhanden sind oder kein Haupteffekt des Treatments existiert.

Im Anhang befinden sich die Abbildungen 13 und 14, die die Verteilungen der Treatment- und der Zielvariable der Trainingsdatensätze zeigen. Entsprechend der Randomisierung liegt der Anteil an Behandlungen bei etwa 50%. Bei der Zielvariable ist der Anteil an Events etwas größer, die Daten sind aber nicht zu unbalanciert für brauchbare Analysen.

4.2 Gütemaße für die Verfahren

Bei der Auswertung der Simulationsergebnisse werden zwei Kriterien betrachtet. Zum einen geht es darum wie gut die Vorhersage auf Basis der unterschiedlichen Methoden je Setting gelingt. Dazu wird das AUC, wie es bereits im Theorieteil beschrieben wurde, betrachtet.
Zum anderen soll beurteilt werden, wie gut die Methoden die wahre Zusammenhangsstruktur erfassen. Hierbei werden die geschätzten β-Koeffizienten mit den β-Koeffizienten, die der Simulation zu Grunde liegen also den „wahren“ β-Koeffizienten, verglichen. Dieser Vergleich erfolgt zuerst quantitativ und schließlich auch qualitativ. Mit quantitativ ist hier gemeint, dass zunächst gezählt wird, wie viele der wahren Effekte erkannt und wie viele fälschlicherweise als solche ausgewiesen werden. Tabelle 3 verdeutlicht diese Herangehens-

Tabelle 2: Übersicht der Simulationssettings

Setting	Anzahl β_s^H	Stärke $exp(\beta_s^H)$	Stärke $exp(\beta_T)$	Anzahl β_s^I	Stärke $exp(\beta_s^I)$
1	wenige (5)	stark (0.2)	mittel (2)	wenige (4/1)	stark (5)
2	wenige (5)	stark (0.2)	mittel (2)	wenige (1/4)	stark (5)
3	wenige (5)	stark (0.2)	mittel (2)	viele (5/95)	schwach (1.4)
4	wenige (5)	stark (0.2)	mittel (2)	keine	-
5	keine	-	mittel	wenige (0/5)	stark (5)
6	viele (100)	schwach (0.71)	stark (5)	viele (80/20)	mittel (2)
7	viele (100)	schwach (0.71)	stark (5)	viele (20/80)	mittel (2)
8	viele (100)	schwach (0.71)	keiner	viele (80/20)	mittel (2)
9	viele (100)	schwach (0.71)	keiner	wenige (4/1)	stark (5)
10	viele (100)	mittel (0.5)	mittel (2)	viele (80/20)	schwach (1.4)
11	viele (100)	mittel (0.5)	mittel (2)	viele (20/80)	schwach (1.4)
12	viele (100)	mittel (0.5)	mittel (2)	keine	-

β_s^H stellt die Haupteffekte ungleich Null dar und β_s^I stellt analog die Interaktionseffekte ungleich Null dar, mit $s = 1, ..., S$. Wobei hier $S \in \{0, 5, 100\}$. Bei der Anzahl der β_s^I gibt die erste Zahl an wie viele Interaktionen von genetischen Variablen kommen, die auch einen von Null verschiedenen Haupteffekt haben und die zweite Zahl zeigt an wie viele genetische Variablen ohne Haupteffekt einen Interaktionseffekt haben.

weise noch einmal. Ist beispielsweise das wahre β_1 der genetischen Variable X_1 ungleich Null und das geschätzte $\hat{\beta}_1$ ist dagegen gleich Null, so ist das ein falsch negatives Ergebnis. Denn die Variable hätte in Wahrheit einen Einfluss gehabt, dieser wurde aber nicht erkannt. Beim

Tabelle 3: Betrachtung der richtig erkannten Effekte und der fälschlicherweise geschätzten Effekte.

	wahres $\beta = 0$	wahres $\beta \neq 0$	
$\hat{\beta} = 0$	richtig negativ (TN)	falsch negativ (FN)	Anzahl Kovariablen ohne geschätzten Effekt
$\hat{\beta} \neq 0$	falsch positiv (FP)	richtig positiv (TP)	Anzahl geschätzter Effekte
	Anzahl Kovariablen ohne Effekt	Anzahl wahrer Effekte	Gesamtanzahl Kovariablen

qualitativen Vergleich geht es dagegen darum zu betrachten, ob die richtig erkannten Effekte auch in ihrer Effektstärke mit den wahren Effekten übereinstimmen und wie groß falsche Effekte geschätzt wurden. Dazu werden wieder die Odds Ratios, also $exp(\beta)$, betrachtet.

4.3 Ergebnisse der Simulation

Bei den folgenden Grafiken sei darauf hingewiesen, dass auf die Skalierung der y-Achsen zu achten ist, da diese nicht immer gleich ist. Sowohl zwischen verschiedenen Abbildungen als auch zwischen einzelnen Plots auf einer Abbildung können Unterschiede vorliegen.

Vorhersagegenauigkeit

Um die Vorhersagegenauigkeiten der unterschiedlichen Schätzmethoden zu bewerten, wird das AUC herangezogen. Abbildung 2 zeigt für jedes Setting je Methode einen Boxplot. Ein Boxplot beinhaltet somit alle AUC-Werte der 100 Trainingsdatensätze. Die Grafik gibt folglich einen Überblick in welchem Bereich sich die AUCs befinden und wie groß die Streuung ist. Je breiter die Box ist, desto größer ist die Streuung und desto instabiler ist die jeweilige Methode einzuschätzen. Die schwarz gestrichelte Linie in Abbildung 2 und 3 ist bei 0.5 eingezeichnet, um schneller erfassen zu können wann es sich um eine unbrauchbare Vorhersage handelt.

In Abbildung 2 ist deutlich zu erkennen, dass die Vorhersagegenauigkeiten bei Settings mit wenigen starken Haupteffekten (1-4) im allgemeinen besser sind als bei Settings mit vielen schwachen (6-9) oder mittleren Haupteffekten (10-12). Dabei sind die Ergebnisse bei vielen schwachen Haupteffekten tendenziell besser als bei vielen mittleren. Die höchsten AUC-Werte werden in Setting 4 erzielt. Dieses Setting beinhaltet keine Interaktionseffekte und wenige starke Haupteffekte. Bei dem anderen Setting ohne Interaktionseffekte (Setting 12), das viele schwache Haupteffekte hat, ergeben sich jedoch deutlich niedrigere AUC-Werte. Existiert kein Haupteffekt des Treatments fallen die AUCs deutlich ab. Es ist allerdings zu beachten, dass es sich hier in beiden Fällen um Settings mit vielen schwachen Haupteffekten der Kovariablen handelt, die allgemein schlechter abgeschnitten haben. Wenn es keine Haupteffekte der genetischen Variablen gibt, sorgt dies ebenfalls für eine Verschlechterung der Vorhersagegenauigkeiten. Dabei schneiden die erste Version der zweischrittigen Verfahren und die Methode von Matsui et al. (2012) schlechter ab als die anderen beiden Methoden. Bei der ersten zweischrittigen Variante dürfte das daran liegen, dass dieses Verfahren per Konstruktion im zweiten Schritt keine Interaktionen mehr schätzen kann, wenn es keine Haupteffekte gibt und dadurch nicht alle Informationen der Daten nutzen kann. Die Konstruktion dieses Ansatzes sorgt ebenfalls dafür, dass die AUCs sinken, wenn es überwiegend Interaktionen gibt, die ohne einen zugehörigen Haupteffekt der genetischen Variable auftreten.

Insgesamt betrachtet schneiden die AUC-Werte des einschrittigen Verfahrens und der zweiten Version der zweischrittigen Verfahren am besten ab.

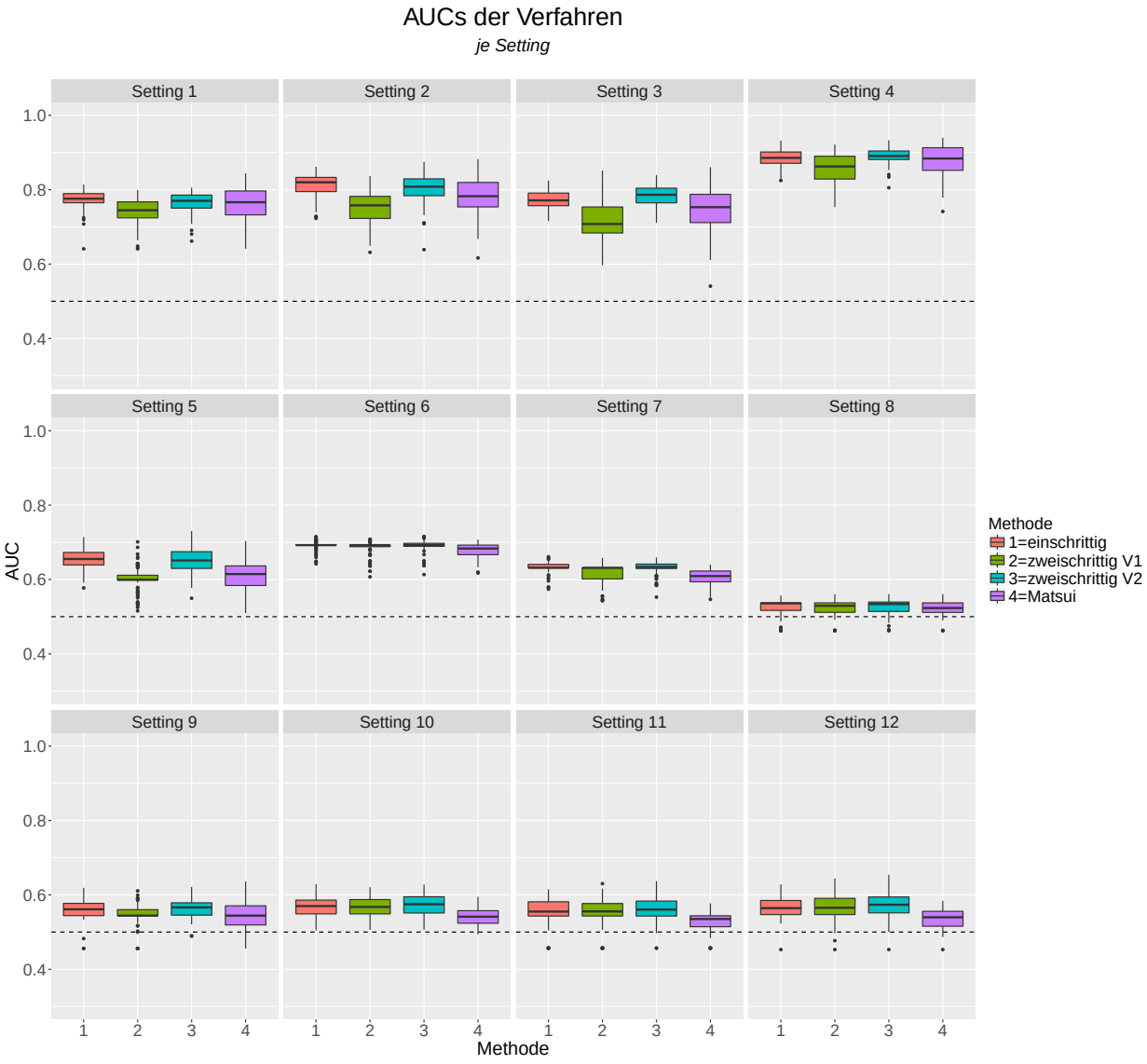

Abbildung 2: Für jedes Setting wird je Methode ein Boxplot mit den AUCs der 100 Trainingsdatensätze abgebildet. Die schwarz gestrichelte Linie bei 0.5. zeigt unbrauchbare Vorhersagen an.

Diese Tendenz bestätigt sich, wenn man alle Settings zusammen betrachtet. So ist in Abbildung 3 zu sehen, dass die Mediane der AUC-Werte des einschrittigen Verfahrens ($x_{med} = 0.64$) und der zweiten Version der zweischrittigen Ansätze ($x_{med} = 0.64$) etwas größer ausfallen, als bei der ersten Version der zweischrittigen Ansätze ($x_{med} = 0.62$) und der Methode von Matsui et al. (2012) ($x_{med} = 0.61$). Bei der Interpretation der AUC-Werte ist zu beachten, dass unter den 12 Settings mehr Settings mit vielen schwachen oder mittleren Haupteffekten vorliegen, die die Mediane der AUC-Werte nach unten ziehen. Deshalb sind die Boxen auch nicht ganz symmetrisch, sondern zeigen eine leicht linkssteile Verteilung an. An der Breite der Boxen kann jedoch erkannt werden, dass die Ergebnisse aller Methoden stark von der vorliegenden Datensituation abhängen. Da diese in der Praxis in der Regel unbekannt ist, muss bei allen Methoden auch mit schlechteren Vorhersagegenauigkeiten gerechnet werden.

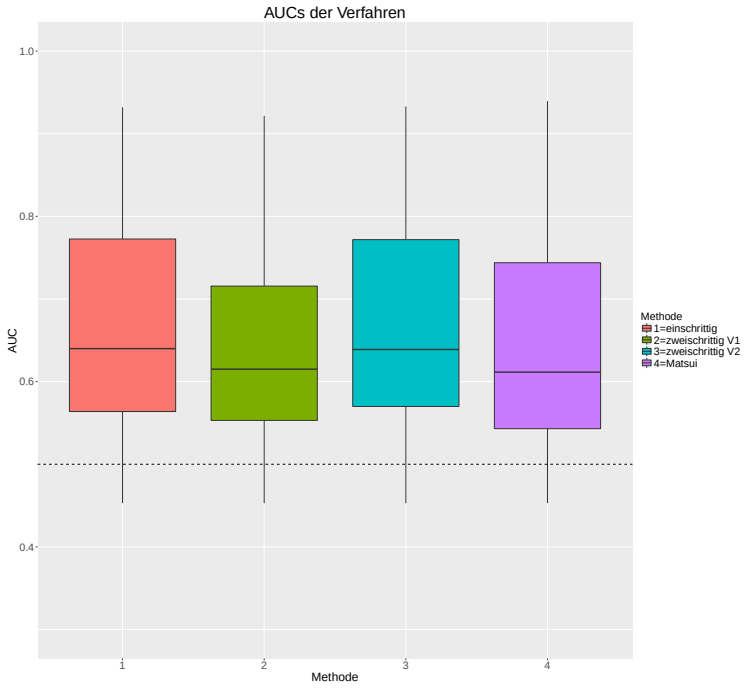

Abbildung 3: Für jede Methode wird ein Boxplot mit den AUCs über alle 12 Settings hinweg abgebildet. Die schwarz gestrichelte Linie bei 0.5. zeigt unbrauchbare Vorhersagen an.

Anzahl erkannter Effekte

Nun soll betrachtet werden, ob auch die richtigen Effekte erkannt werden und wie viele zusätzliche Effekte die Methoden fälschlicherweise schätzen. Dazu werden die richtig und falsch positiven Haupt- bzw. Interaktionseffekte der Kovariablen betrachtet.
Abbildung 4 und 5 zeigen die Anzahlen richtig und falsch positiver Haupteffekte der 100 Trainingsdatensätze je Setting und Methode mittels Boxplots. Dabei ergeben sich bei den

zweischrittigen Verfahren identische Ergebnisse bei den Haupteffekten, da sie auf demselben Haupteffektmodell basieren. Bei den richtig positiven Haupteffekten zeigt die rote Linie an wie viele wahre Haupteffekte in dem jeweiligen Setting generiert wurden.

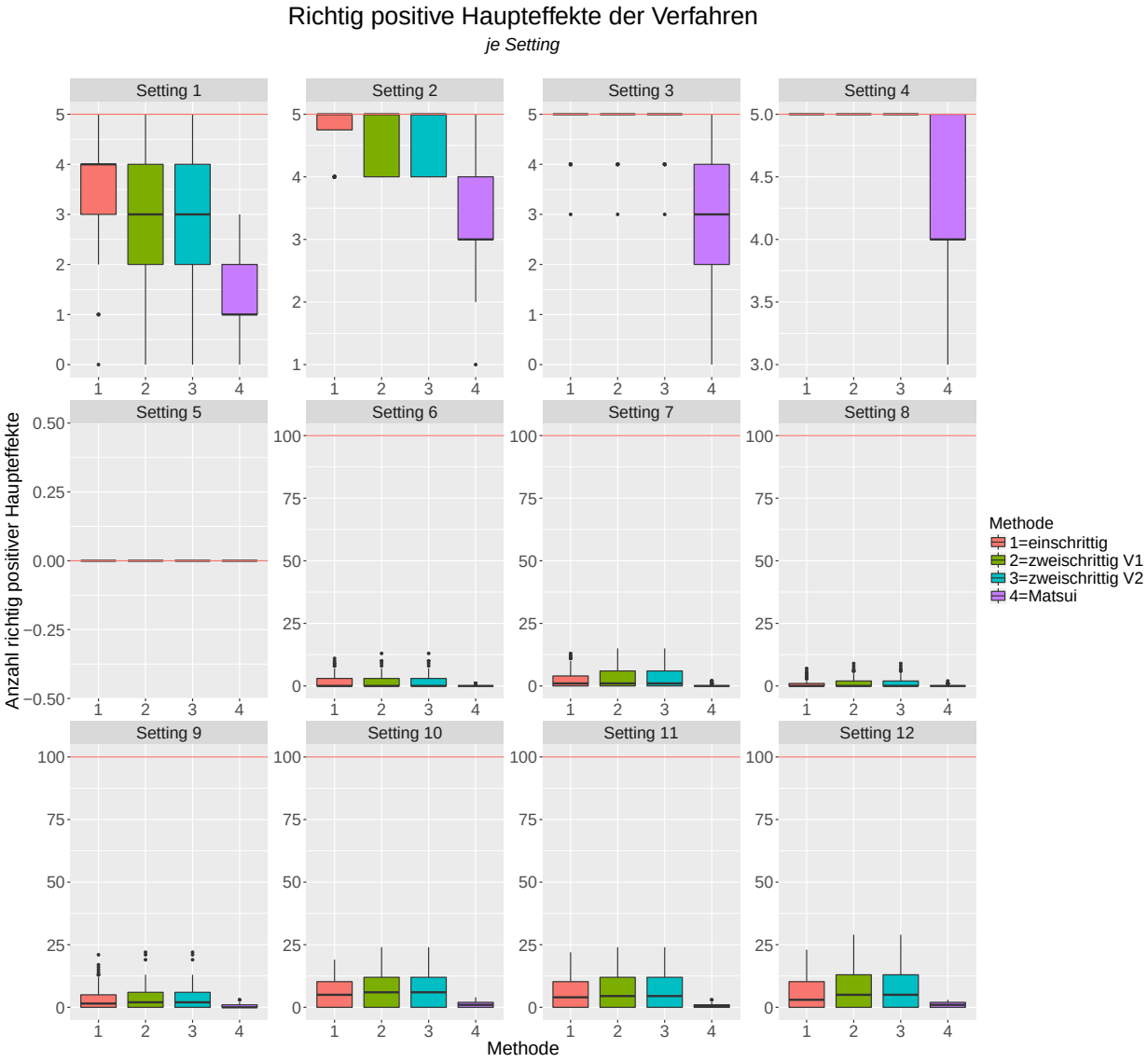

Abbildung 4: Für jedes Setting und jede Methode wird ein Boxplot mit den Anzahlen der richtig positiven Haupteffekte abgebildet. Die rote Linie zeigt die Anzahl wahrer Haupteffekte in dem jeweiligen Setting an.

Abbildung 4 zeigt, dass die Methode von Matsui et al. (2012) im Allgemeinen weniger wahre Haupteffekte erkennt als die drei Varianten des neuen Ansatzes. Die drei Versionen des neuen Ansatzes erkennen in den Settings mit wenigen starken Haupteffekten (1-4), alle oder beinahe alle Effekte. Gibt es viele Haupteffekte, bleiben bei allen Methoden die meisten wahren

Effekte unerkannt. Der Median der erkannten Effekte liegt hier durchgehend im einstelligen Bereich, obwohl es insgesamt 100 wahre Effekte gegeben hätte. Dabei ist zu beobachten, dass etwas mehr Effekte erkannt werden bei den Settings mit mittleren Haupteffekten (10-12), als bei den Settings mit schwachen Haupteffekten (6-9). Jedoch wird im besten Fall in diesen Settings ein Viertel der wahren Haupteffekte erkannt. In Setting 5 gibt es keine Haupteffekte, weshalb auch keine richtig positiven Haupteffekte existieren.

In Abbildung 5 wird ersichtlich, dass die Methode von Matsui et al. (2012) kaum falsch Posi-

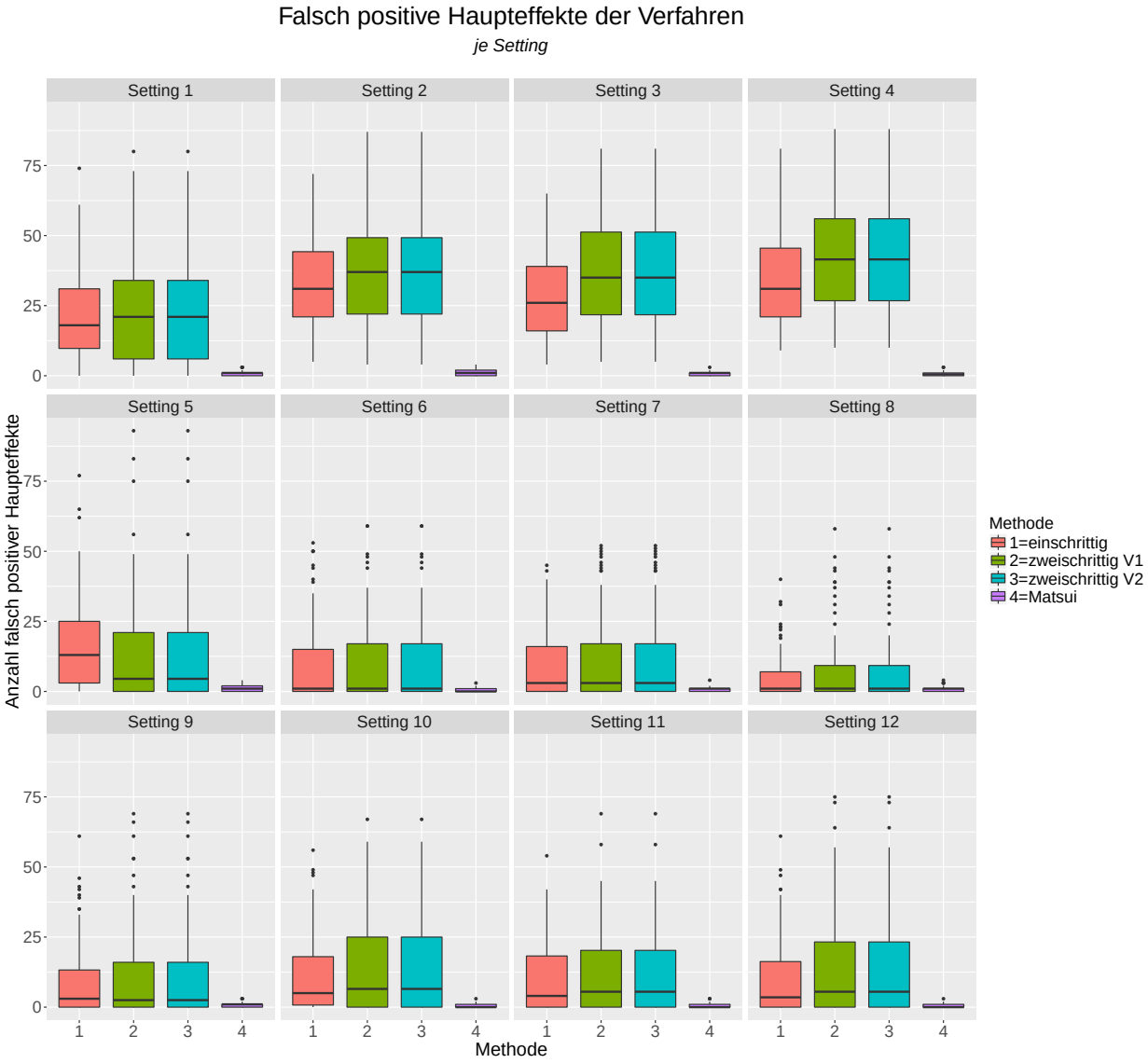

Abbildung 5: Für jedes Setting und jede Methode wird ein Boxplot mit den Anzahlen der falsch positiven Haupteffekte abgebildet.

tive erzeugt, ganz im Gegensatz zum neuen Ansatz, der in allen drei Varianten, einige Effekte schätzt, die nicht generiert wurden. Hierbei fällt auf, dass die Anzahl der falsch positiven

Haupteffekte mit der Effektstärke tendenziell ansteigt. Vergleicht man den einschrittigen Ansatz mit dem zweischrittigen, erkennt man, dass die zweischrittigen Verfahren in allen Settings (außer Setting 5 ohne Haupteffekte) mehr falsch positive Haupteffekte aufweisen als die einschrittige Variante. Beim neuen Ansatz befinden sich die Mediane der richtig und falsch Positiven bei Settings mit vielen Haupteffekten in einem ähnlichen Bereich. Allerdings ist die Verteilung der falsch Positiven deutlich links steil, das heißt es gibt viele kleine Anzahlen und wenige sehr große Anzahlen. Bei den Settings mit wenigen wahren Haupteffekten übersteigt die Anzahl der falsch positiven die der wahren Effekte deutlich.

Abbildung 6 und 7 zeigen richtig und falsch positive Interaktionseffekte. Die rote Linie bei den richtig Positiven zeigt hierbei wieder die Anzahl der wahren Interaktionseffekte an. Bei Setting 4 und 12 existieren wiederum keine richtig Positiven, da in diesen Fällen keine wahren Interaktionseffekte generiert wurden.
In Abbildung 6 ist zu erkennen, dass die Methoden die wahren Interaktionseffekte nicht so gut erfassen wie die wahren Haupteffekte. Auch hier werden im allgemeinen wenige starke Interaktionseffekte (3, 10, 11) besser erkannt als viele schwache (1, 2, 5, 9) oder mittlere (6, 7, 8). Die erste Variante der zweischrittigen Verfahren hat dabei die wenigsten richtig positiven Interaktionseffekte. Außerdem bringt diese Methode auffallend viele falsch Positive hervor, wie Abbildung 7 zeigt. Insgesamt übersteigt oft, bei allen Versionen des neuen Ansatzes, die Anzahl der falsch positiven Interaktionseffekte die der richtig positiven deutlich. Das heißt beim neuen Ansatz handelt es sich bei der Mehrheit der geschätzten Interaktionseffekte oftmals um falsch Positive. Die zweite Version der zweischrittigen Verfahren hat bei Setting 5, das keine Haupteffekte hat, besonders viele falsch positive Interaktionen. Ansonsten schneidet dieses Verfahren tendenziell besser ab als das einschrittige Verfahren. Die Methode von Matsui et al. (2012) hat erneut die wenigsten falsch Positiven.

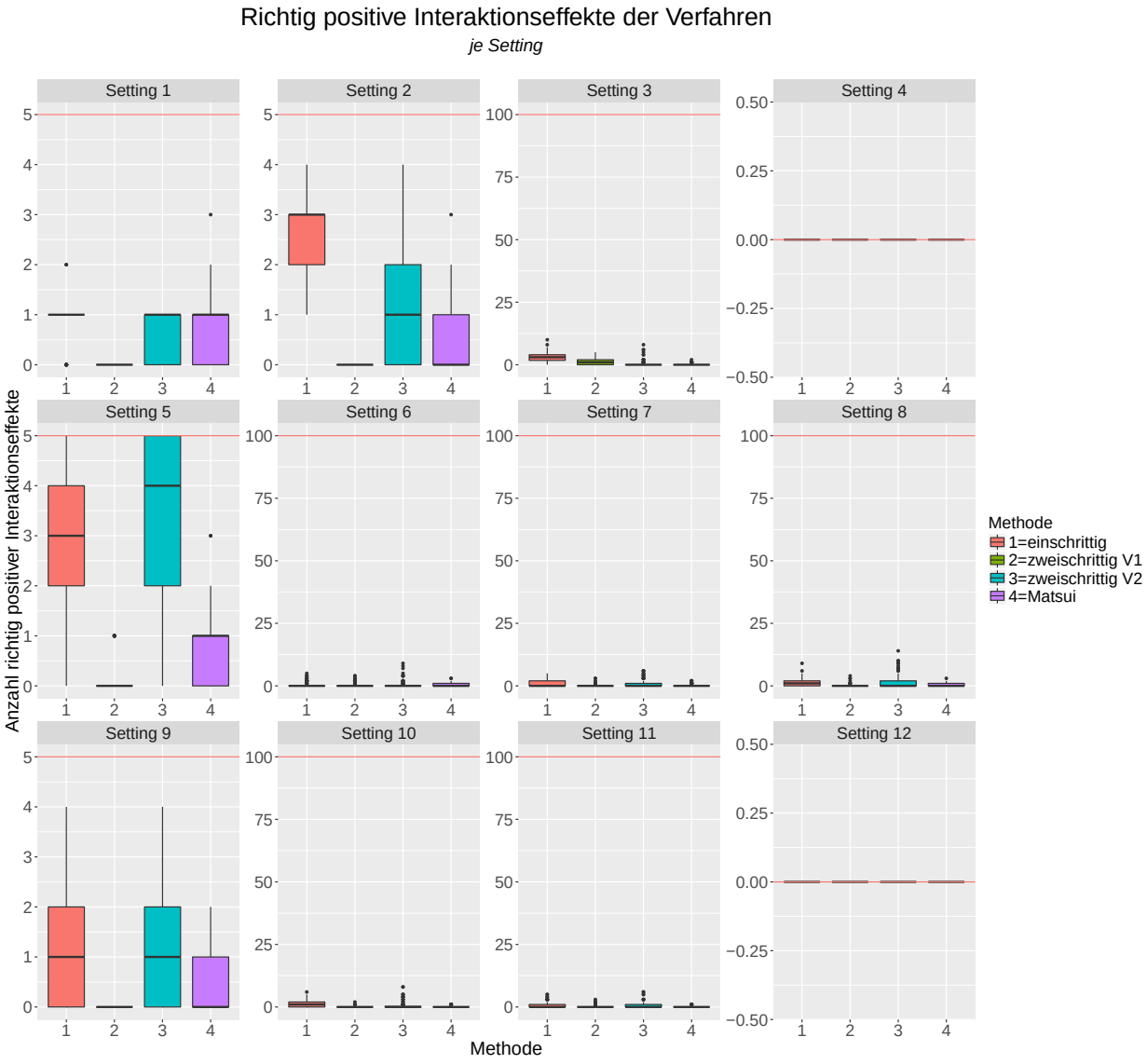

Abbildung 6: Für jedes Setting und jede Methode wird ein Boxplot mit den Anzahlen der richtig positiven Interaktionseffekte abgebildet. Die rote Linie zeigt die Anzahl wahrer Interaktionseffekte in dem jeweiligen Setting an.

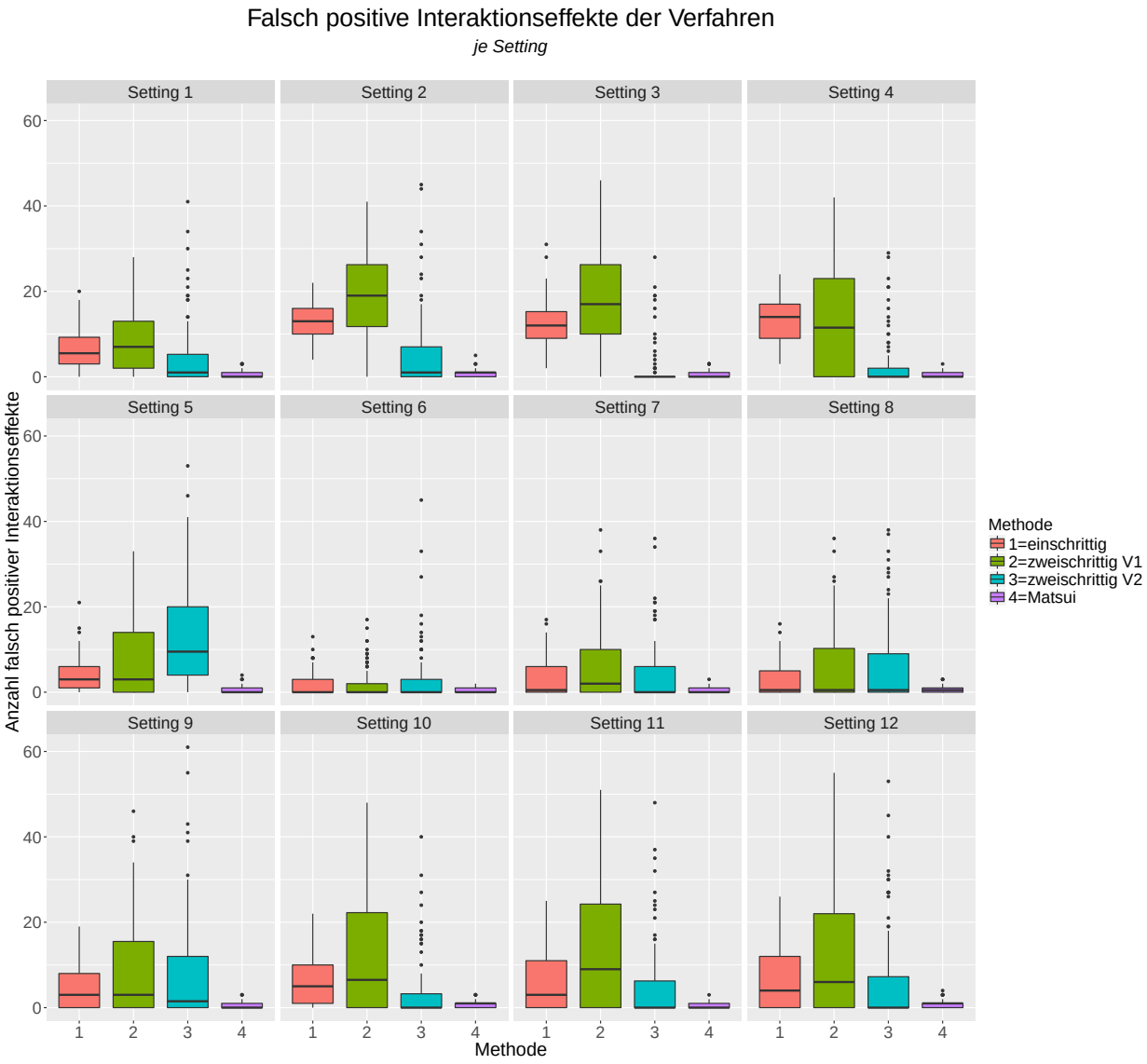

Abbildung 7: Für jedes Setting und jede Methode wird ein Boxplot mit den Anzahlen der falsch positiven Interaktionseffekte abgebildet.

Geschätzte Effektstärken

Da nicht nur entscheidend ist wie viele der wahren Effekte erkannt und wie viele Effekte fälschlicherweise als solche deklariert werden, sondern auch wie groß diese Effekte geschätzt werden, folgt abschließend eine Betrachtung der Odds Ratios, also $exp(\beta)$, der richtig und falsch Positiven. Die schwarz gestrichelte Linie bei der eins ist dabei zur besseren Orientierung eingezeichnet, da so schneller erfasst werden kann in welche Richtung die Odds Ratios gehen, also ob die Chance sich erhöht oder verringert. Bei den richtig erkannten Effekten zeigt die rote Linie $exp(\beta)$ der wahren Effekte.

Als erstes werden die Odds Ratios der Haupteffekte betrachtet. Abbildung 8 zeigt die Odds Ratios der richtig erkannten Haupteffekte. Dabei ist zu beachten, dass es keine Boxplots für Setting 5 gibt, weil es bei diesem Setting keine wahren Haupteffekte gibt und somit keine richtig Positiven existieren. Es ist zu erkennen, dass die Richtung der Haupteffekte in der Regel durch alle Methoden richtig geschätzt wird. Da bis auf wenige Ausnahmen keine Odds Ratios größer als eins zu beobachten sind. Alle drei Versionen des neuen Ansatzes unterschätzen $exp(\beta)$ und weisen allgemein sehr ähnliche Ergebnisse auf. Die Methode von Matsui et al. (2012) ist bei den Settings mit wenigen starken Haupteffekten (1-4) oder vielen mittleren (10-12) am nächsten an den wahren Odds Ratios dran. Bei den Settings mit vielen schwachen Haupteffekten (6-9) überschätzt die Methode von Matsui et al. (2012) jedoch die Odds Ratios etwas. Außerdem fällt auf, dass die Schätzungen bei den Settings mit wenigen starken Haupteffekten (1-4) mehr Varianz aufweisen, was an den breiteren Boxen zu erkennen ist.

Abbildung 9 zeigt ergänzend dazu wie groß $exp(\beta)$ für die falsch positiven Haupteffekte ausfallen. Die schwarz gestrichelte Linie bei der eins zeigt dabei wo die geschätzten Odds Ratios idealerweise liegen sollten. Denn ist das Chancenverhältnis nahe eins hat die zugehörige Kovariable kaum Einfluss. Hier ist zu erkennen, dass alle Versionen des neuen Ansatzes die Odds Ratios der falsch Positiven in der Regel nahe eins schätzen. Es sind jedoch auch einige Ausreißer nach oben und unten zu beobachten. Die Methode von Matsui et al. (2012) liegt dagegen mit seinen geschätzten Odds Ratios für die falsch Positiven deutlich weiter von der eins entfernt. Dabei sind sehr breite Boxen zu beobachten, was für eine instabile Schätzung spricht. Dieses Resultat ergibt sich möglicherweise durch die univariate Herangehensweise von Matsui et al. (2012), die die Effekte der genetischen Variablen zu isoliert voneinander betrachtet.

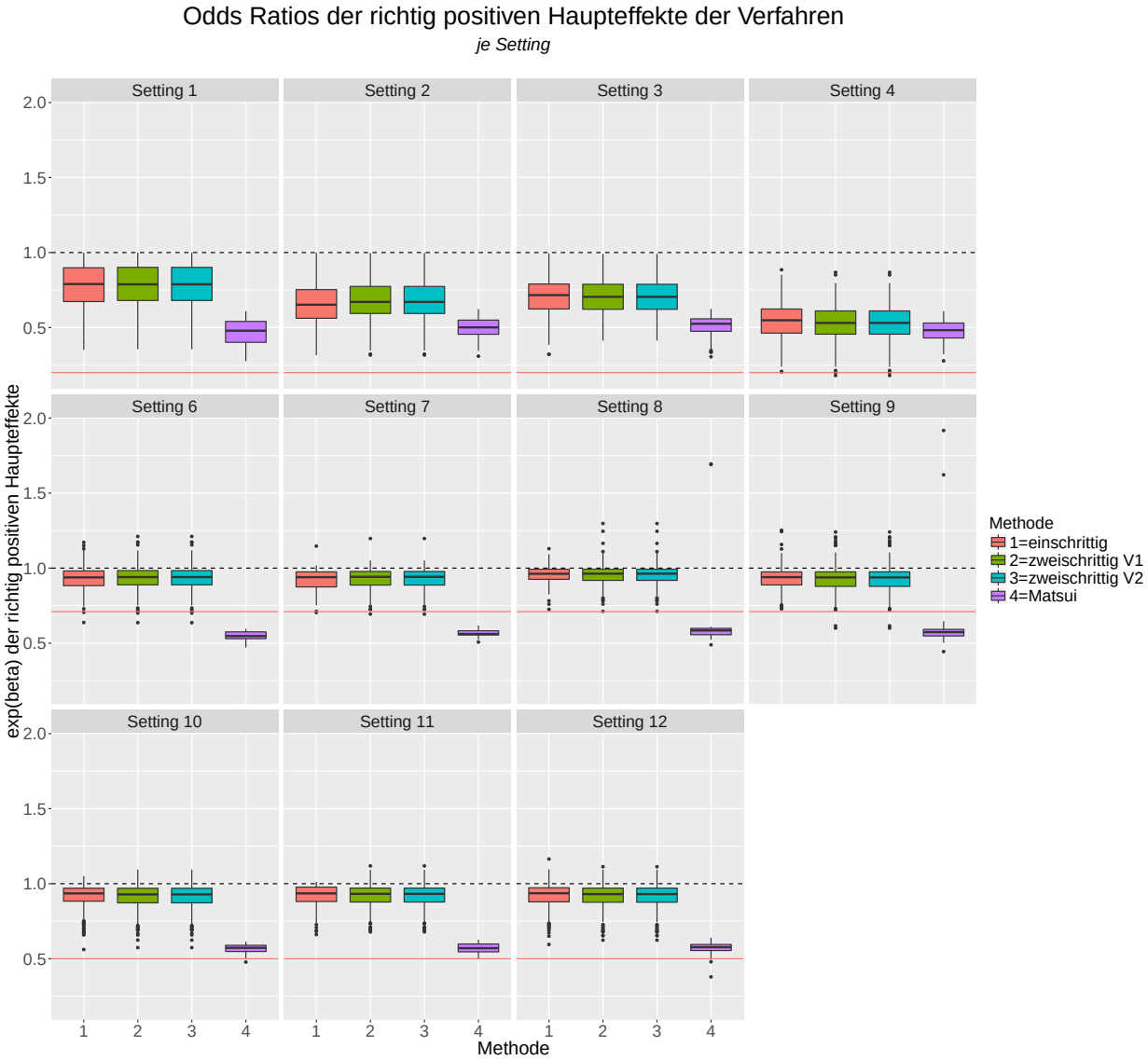

Abbildung 8: Für jedes Setting und jede Methode wird ein Boxplot mit den Odds Ratios der richtig erkannten Haupteffekte abgebildet. Die rote Linie zeigt die Odds Ratio der wahren Haupteffekte in dem jeweiligen Setting an. Die schwarz gestrichelte Linie bei 1.0 hilft die Richtung der Odds Ratios schneller zu erfassen.

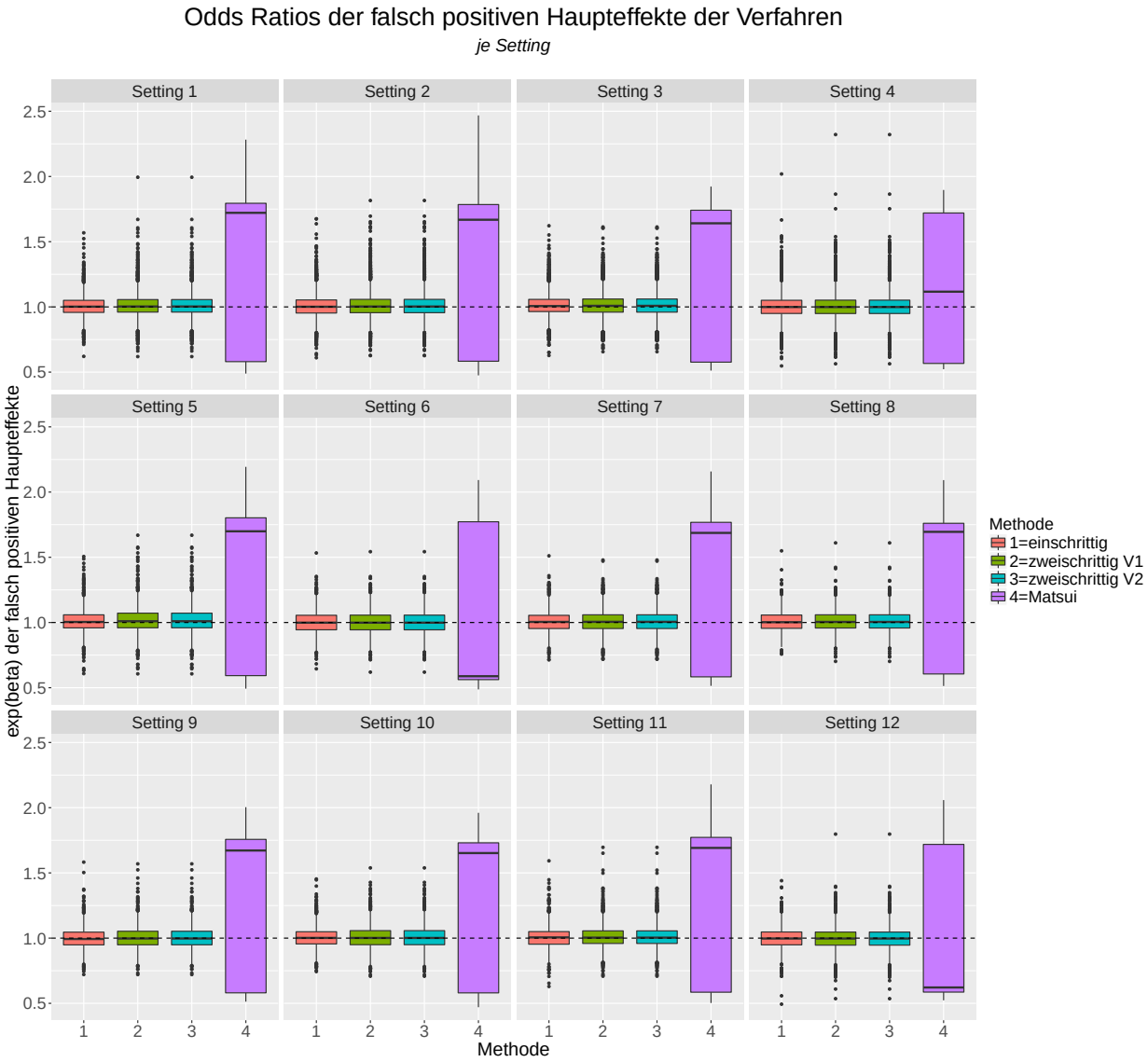

Abbildung 9: Für jedes Setting und jede Methode wird ein Boxplot mit den Odds Ratios der falsch positiven Haupteffekte abgebildet. Die schwarz gestrichelte Linie bei 1.0 zeigt an wo die Odds Ratios der falsch Positiven idealerweise liegen.

Für die Interaktionseffekte zeigt sich ein ähnliches Bild. Dabei fehlen hier Setting 4 und 12, weil diese keine wahren Interaktionseffekte haben und somit keine richtig Positiven existieren. Für die erste Version der zweischrittigen Verfahren sind bei Setting 1, 2 und 9 keine Boxen abgebildet, da diese Methode in diesen Fällen keine richtig positiven Interaktionseffekte hat. Wie Abbildung 10 erkennen lässt, liegen die Odds Ratios der Methode von Matsui et al. (2012) immer größer eins, die geschätzten Effekte gehen somit in die richtige Richtung. Allerdings überschätzt diese die Odds Ratios teilweise, vor allem bei Settings mit vielen schwachen oder mittleren Interaktionseffekten (Setting 3, 6, 7, 8, 10, 11). Der neue Ansatz unterschätzt die Odds Ratios durchweg und kommt dabei teilweise auch unter eins. Das heißt hier zeigen geschätzte Effekte auch in die falsche Richtung. Besonders negativ fällt dabei die erste Version der zweischrittigen Verfahren auf, die vor allem bei Setting 11 mit der Schätzung der Interaktionseffekte überwiegend falsch liegt.

Abbildung 11 zeigt, dass auch bei den falsch positiven Interaktionseffekten die Verfahren des neuen Ansatzes überwiegend Odds Ratios nahe eins schätzen. Allerdings streuen die Schätzungen der ersten Variante der zweischrittigen Verfahren hier im Vergleich zu den anderen beiden Verfahren des neuen Ansatzes mehr und weichen teilweise etwas deutlicher von der eins ab. Die Methode von Matsui et al. (2012) weist bei den falsch positiven Interaktionseffekten ebenfalls Odds Ratios auf, die deutlich von eins abweichen. Diese Abweichung ist dabei tendenziell noch größer als bei den falsch positiven Haupteffekten. Die Boxen dieser Methode sind abermals vergleichsweise breit.

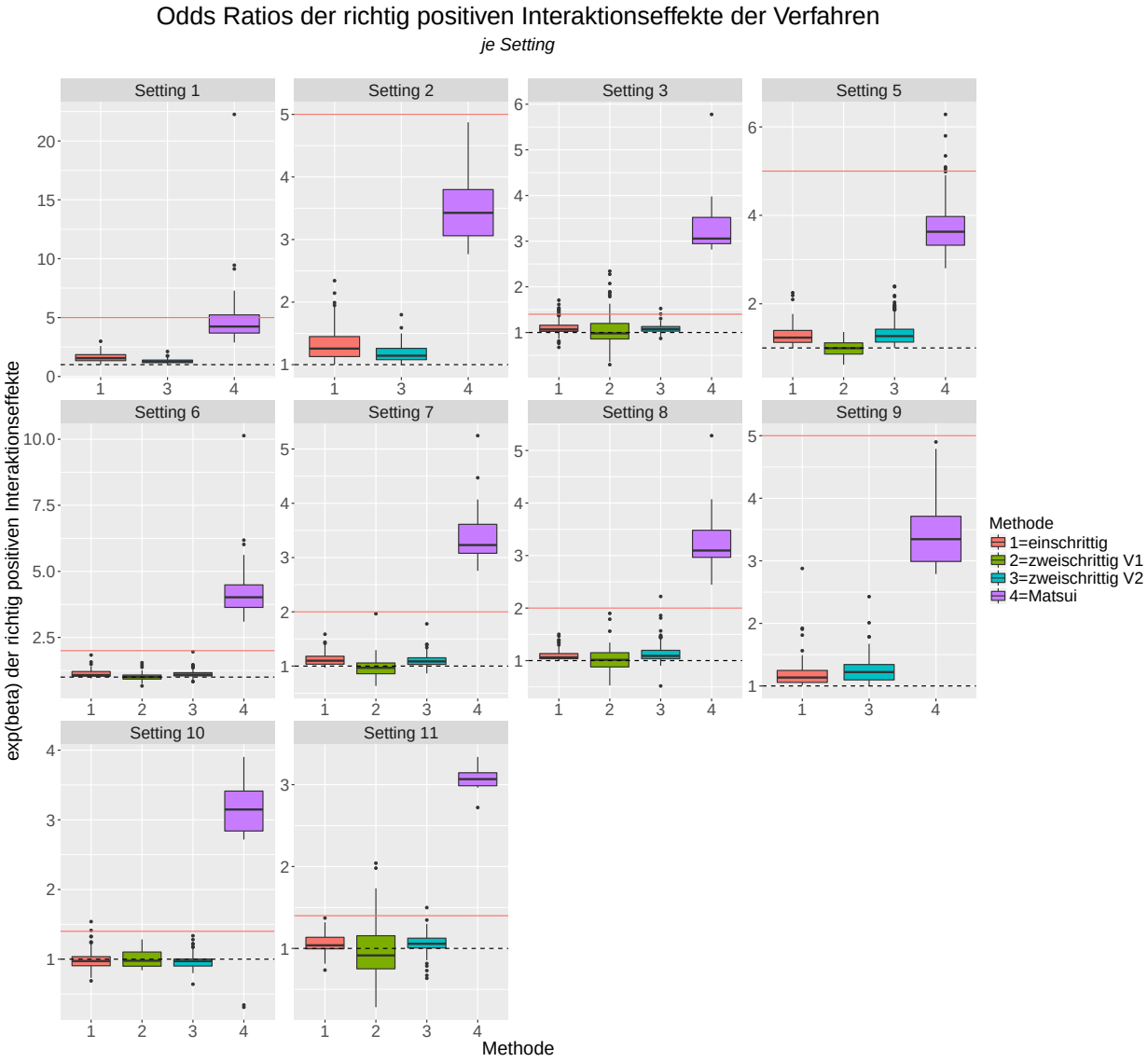

Abbildung 10: Für jedes Setting und jede Methode wird ein Boxplot mit den Odds Ratios der richtig erkannten Interaktionseffekte abgebildet. Die rote Linie zeigt die Odds Ratio der wahren Interaktionseffekte in dem jeweiligen Setting an. Die schwarz gestrichelte Linie bei 1.0 hilft die Richtung der Odds Ratios schneller zu erfassen.

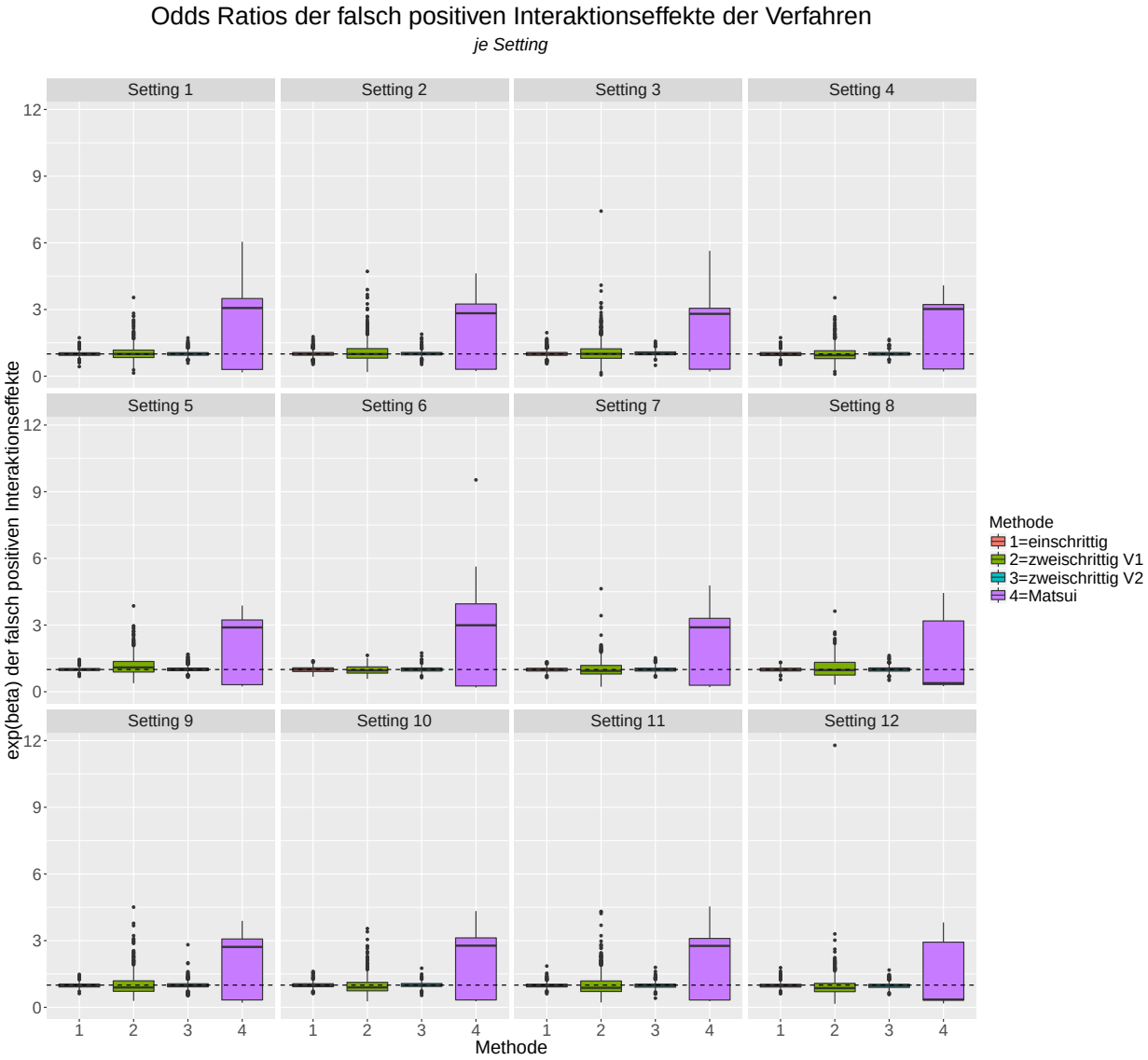

Abbildung 11: Für jedes Setting und jede Methode wird ein Boxplot mit den Odds Ratios der falsch positiven Interaktionseffekte abgebildet. Die schwarz gestrichelte Linie bei 1.0 zeigt wo die Odds Ratios der falsch Positiven idealerweise liegen.

Fazit der Ergebnisse

Keine der Methoden kann bei dieser Simulation über alle Settings hinweg voll überzeugen. Denn gibt es beispielsweise viele mittlere Haupteffekte, liegen die AUC-Werte aller Methoden zwischen 0.5 und 0.6, was keiner besonders guten Vorhersage entspricht. Bei diesen Settings bleibt auch der Großteil der wahren Haupteffekte unerkannt. Es können jedoch auch sehr gute Vorhersagen (mit AUC-Werten zwischen 0.7 und 0.9) getroffen werden, wenn wenige starke Haupteffekte vorhanden sind, die von den Methoden auch größtenteils erfasst werden. Der erste zweischrittige Ansatz macht bei dieser Simulation keinen empfehlenswerten Eindruck. Durch die Einschränkung im zweiten Schritt auf Interaktionen mit Haupteffekt werden die Interaktionseffekte nicht zufriedenstellend geschätzt, da einige wahre Effekte nicht erkannt werden. Zugleich werden einige Interaktionen fälschlicherweise geschätzt. Auch die Odds Ratios der Interaktionen können durch den ersten zweischrittigen Ansatz nicht akkurat geschätzt werden.

Die Methode von Matsui et al. (2012) bringt nicht so viele falsch Positive hervor, weder Haupt- noch Interaktionseffekte. Jedoch schätzt sie die Einflüsse dieser falsch Positiven teilweise sehr groß, was ebenfalls nicht optimal ist. Dies schlägt sich auch in den AUC-Werten nieder, die im Vergleich zum neuen Ansatz niedriger ausfallen.

Alles in allem überzeugen das einschrittige Verfahren und die zweite Version der zweischrittigen Ansätze am meisten. Diese weisen durchweg die höchsten AUCs auf und erkennen die meisten der wahren Effekte. Wobei die AUC-Werte teilweise nur geringfügig besser sind als die der anderen beiden Methoden. Die Odds Ratios der richtig positiven Haupt- und Interaktionseffekte werden zwar unterschätzt, jedoch ist die Richtung der Effekte immer richtig erfasst. Andererseits ergeben sich bei diesen Methoden auch einige falsch positive Haupt- und Interaktionseffekte. Allerdings liegen die geschätzten Odds Ratios hierfür nahe eins, so dass sie keinen zu großen Einfluss haben. Die beiden Methoden unterscheiden sich hinsichtlich der Anzahlen der falsch positiven Haupt- bzw. Interaktionseffekte. Beim einschrittigen Verfahren ergeben sich etwas mehr falsch positive Interaktionseffekte. Das könnte daran liegen, dass durch die gleichzeitige Schätzung der Haupt- und Interaktionseffekte den Interaktionen mehr Bedeutung zukommt. Dagegen sind bei der zweischrittigen Variante mehr falsch positive Haupteffekte zu beobachten, die möglicherweise durch das vorab geschätzte Haupteffektmodell zustande kommen.

5 Diskussion

In dieser Arbeit wurden Methoden zur Schätzung von prädiktiven Biomarkern anhand von omics-Daten vorgestellt. Diese sollen dabei helfen im Sinne der personalisierten Medizin individuelle Behandlungseffekte zu erkennen, um für jeden Patienten die richtige Behandlungswahl zu treffen. Die Methode von Matsui et al. (2012) geht hierbei univariat bei der Entwicklung von hochdimensionalen Biomarkern vor. Im Gegensatz hierzu benutzt der neue Ansatz den Lasso-Schätzer, um eine multiple Herangehensweise umzusetzen. Dabei wurden sowohl einschrittige, als auch zweischrittige Verfahren betrachtet. Die Simulationsstudie zeigt, dass der neue Ansatz durchaus Vorteile gegenüber der Methode von Matsui et al. (2012) mit sich bringt und sich weitere Analysen und Verbesserungen dieses Ansatzes lohnen könnten. Denn der neue Ansatz liefert insgesamt bessere Vorhersagegenauigkeiten als die Methode von Matsui et al. (2012) und erkennt mehr der wahren Effekte. Dabei scheinen angesichts der Ergebnisse aus dieser Arbeit die einschrittige Variante und die zweite Version der zweischrittigen Verfahren vielversprechender als die erste Version der zweischrittigen Verfahren zu sein. Bei Betrachtung der unterschiedlichen Settingergebnisse fällt auf, dass die Ergebnisse aller Methoden stark davon abhängen welche Effekte die vorliegenden Daten aufweisen.

Es gibt einige Aspekte, die bei der Einordnung der Ergebnisse der Schätzmethoden dieser Arbeit berücksichtigt werden sollten. Die Simulation bringt einige Einschränkungen mit sich. Bei der Kovariablenstruktur wurden keine Abhängigkeiten berücksichtigt, wie sie in der Regel in Genomikdaten vorhanden sind. Es wäre somit durchaus interessant zu sehen welche Ergebnisse die Methoden liefern, wenn man sie auf Daten anwendet, die eine komplexere Datenstruktur aufweisen. Dazu bedarf es jedoch mehr Wissen über den Aufbau von omics-Daten, um diese wirklich realistisch generieren zu können. Die in dieser Arbeit beschriebenen Ergebnisse geben folglich eher einen Hinweis darauf wie gut die Methoden funktionieren. Es wäre beispielsweise denkbar, dass der Unterschied zwischen dem neuen Ansatz und der Methode von Matsui et al. (2012) deutlicher wird, wenn eine komplexere Datenstruktur vorliegt. Denn in diesem Fall wird die multiple Herangehensweise des neuen Ansatzes vermutlich größere Vorteile mit sich bringen. Außerdem konnte nur eine begrenzte Anzahl an Settings simuliert werden. Das heißt es gibt sicher noch viele weitere interessante Kombinationen von Effekten.
Des weiteren wäre es interessant die Methoden auf reale Daten aus der Praxis anzuwenden und zu betrachten wie unterschiedlich die resultierenden Ergebnisse sind.
Die Methode von Matsui et al. (2012) nimmt keine explizite Adjustierung für das multiple

Testen vor, sondern verwendet einfach ein von vornherein sehr klein gewähltes α-Niveau für die einzelnen Signifikanztests der β-Koeffizienten. Hier könnte man auch überlegen anders vorzugehen und beispielsweise die FDR zu kontrollieren. Wobei das Problem der Methode weniger darin liegt, dass sie eine hohe Anzahl falsch Positiver liefert. Problematischer ist, dass die Methode von Matsui et al. (2012) den Einfluss dieser falsch Positiven zu groß schätzt.
In dieser Arbeit wurde der Lasso-Schätzer für den neuen Ansatz verwendet. Es wird jedoch auch die Meinung vertreten, dass in Fällen mit $p >> n$ oder mit vielen korrelierten Kovariablen das sogenannte Elastic-Net eine bessere Alternative darstellt (Friedman et al., 2010). Da beides bei omics-Daten auftreten kann, könnte man den neuen Ansatz auch mit Elastic-Net durchführen und die Ergebnisse vergleichen.
Laut Ma et al. (2015) handelt es sich in der Onkologie, wo prädiktive Biomarker immer mehr Anwendung finden, bei der Zielvariable meist um Überlebenszeiten und nicht um binäre Größen. Es wäre vielleicht lohnenswert den neuen Ansatz auch für Survivaldaten umzusetzen, indem man anstelle der logistischen Regression das Cox-Modell oder das AFT-Modell verwendet.
Außerdem wurden in dieser Arbeit die Testmethoden nur theoretisch vorgestellt, aber nicht praktisch umgesetzt. Eine praktische Umsetzung wäre folglich auch noch ein interessanter Punkt für weitere Analysen.

A Anhang

Vergleich der zwei Prädiktionsmodelle für die Methode von Matsui

Beim Prädikitonsmodell für die Methode von Matsui et al. (2012) stellte sich die Frage, ob der Haupteffekt des prädiktiven Scores mit aufgenommen werden soll oder nicht. Abbildung 12 zeigt dazu die AUC-Werte des Prädiktionsmodells mit Haupteffekt des prädiktiven Scores im Vergleich zu den AUC-Werten des Präditktionsmodells ohne diesen Haupteffekt. Wie zu erkennen ist, scheint es keinen Unterschied zu machen, welches Prädiktionsmodell man verwendet. Das liegt vermutlich an der Konstruktion des prädiktiven Scores, der aus genetischen Variablen gebildet wird, die hauptsächlich prädiktiven und kaum prognostischen Charakter haben, weshalb der Haupteffekt des Scores nur sehr geringen Einfluss aufweisen dürfte.

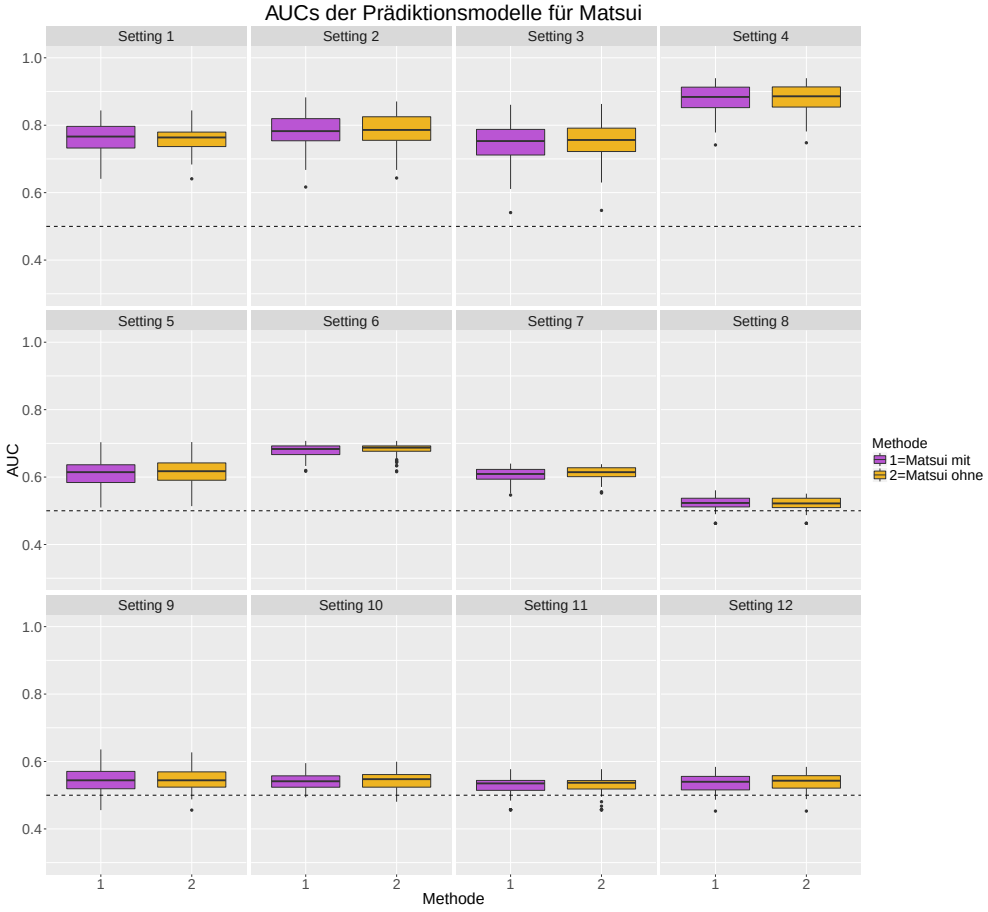

Abbildung 12: Diese Abbildung gilt dem Vergleich der beiden potentiellen Prädiktionsmodelle für die Methode von Matsui. Dazu wird für jedes Setting je Prädiktionsmodell ein Boxplot mit den AUCs der 100 Trainingsdatensätze abgebildet. Die schwarz gestrichelte Linie bei 0.5. zeigt unbrauchbare Vorhersagen an.

Verteilungen der Treatmentvariable und der Zielvariable bei der Simulation
Abbildungen 13 und 14 zeigen den Anteil Behandlungen bzw. Events in den Trainingsdatensätzen je Setting. Die roten Punkte bilden dabei den Anteil im zugehörigen Testdatensatz ab. Gemäß $\pi = 0.5$ um randomisierte Daten zu simulieren, fällt etwa die Hälfte in die Behandlungsgruppe und die andere Hälfte in die Kontrollgruppe. Bei der Zielvariable liegen etwas mehr Events wie nicht-Events vor. Die Daten sind jedoch nicht zu unbalanciert. Die roten Punkte liegen immer in der Box, nahe des Medians, das heißt die Testdaten sind von den Verteilungen her wie die Trainingsdatensätze und stellen keine Extreme dar.

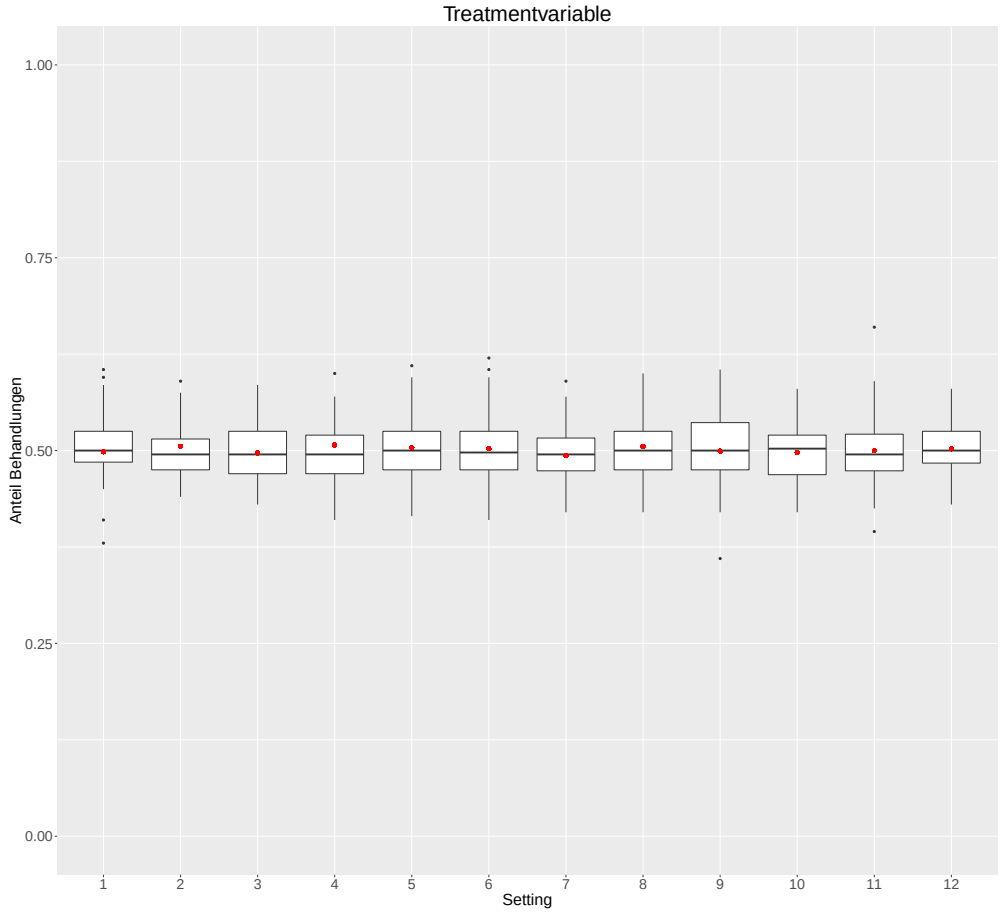

Abbildung 13: Je Setting wird mittels Boxplot der Anteil an Behandlungen in den Trainingsdatensätze dargestellt. Die roten Punkte zeigen dazu die Anteile in den jeweiligen Testdatensätzen.

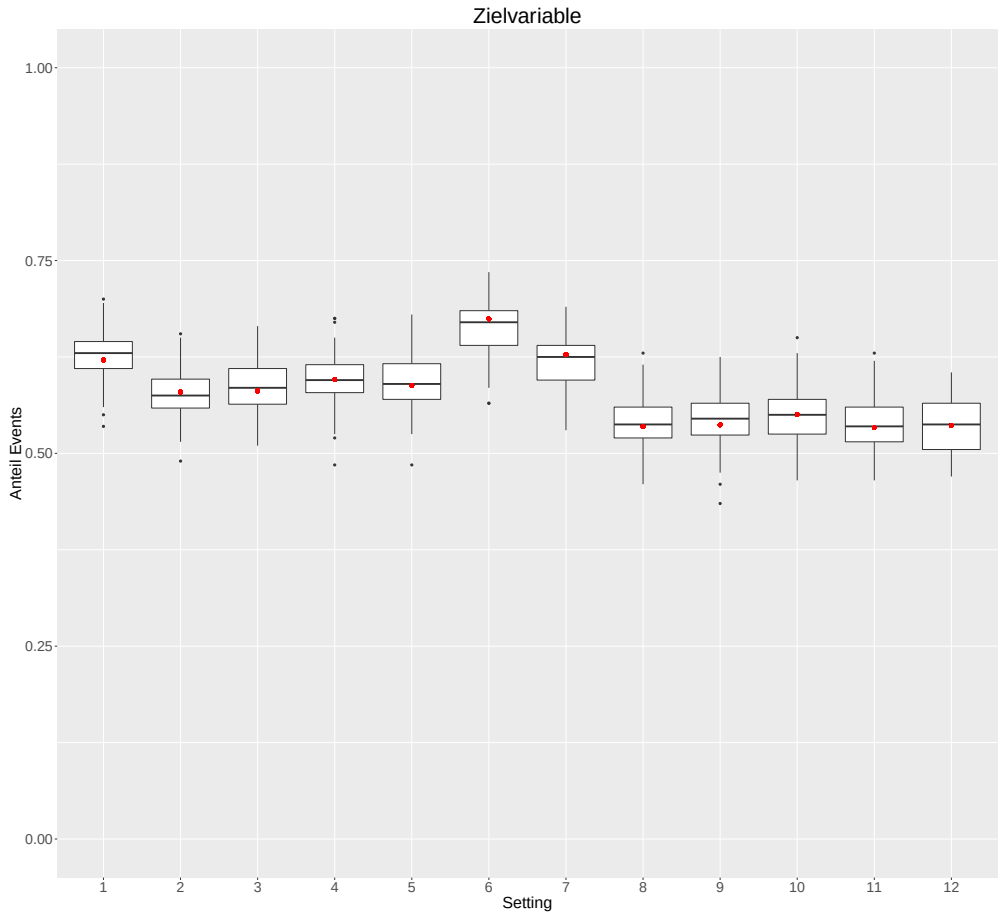

Abbildung 14: Je Setting wird mittels Boxplot der Anteil an Events in den Trainingsdatensätze dargestellt. Die roten Punkte zeigen dazu die Anteile in den jeweiligen Testdatensätzen.

Elektronischer Anhang

Im elektronischen Anhang befinden sich neben dieser Arbeit als pdf-Datei noch folgende weitere Dateien:

- Im Ordner Grafiken sind alle Grafiken aus dieser Arbeit als pdf-Dateien abgespeichert, sowie die R-Files zur Erstellung der Grafiken.
- Der Ordner Funktionen enthält alle programmierten Funktionen zur Umsetzung der Schätzmethoden und der Simulationen, sowie Funktionen zur Auswertung der Simulationsergebnisse.
- Im Ordner Simulation befinden sich schließlich die rda-Dateien mit den abgespeicherten Ergebnissen der Simulation und alle R-Files zur Durchführung der Simulationen.

Literatur

A. J. Atkinson, W. A. Colburn, V. G. DeGruttola, D. L. DeMets, G. J. Downing, D. F. Hoth, and et al. Biomarkers and surrogate endpoints: Preferres definitions and conceptual framework. *Clin. Pharmacol. Ther.*, 69(3):89–95, 2001.

S. Baek, C.-A. Tsai, and J. J. Chen. Development of biomarker classifiers from high-dimensional data. *Briefings in Bioinformatics*, 10(5):537–546, 2009.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

K. Bracht. Biomarker: Indikatoren für diagnose und therapie. *Pharmazeutische Zeitung online*, 12, 2009.

J. J. Chen, T.-P. Lu, Y.-C. Chen, and W.-J. Lin. Predictive biomarkers for treatment selection: statistical considerations. *Biomarkers in Medicine*, 9(11):1121–1135, 2015.

L. Fahrmeir, R. Künstler, I. Pigeot, and G. Tutz. *Statistik: Der Weg zur Datenanalyse*. Berlin, Heidelberg: Springer-Verlag, 7 edition, 2007.

L. Fahrmeir, T. Kneib, and S. Lang. *Regression: Modelle, Methoden und Anwendungen*. Berlin, Heidelberg: Springer-Verlag, 2 edition, 2009.

B. Freidlin and R. Simon. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research*, 11(21):7872–7878, 2005.

B. Freidlin, W. Jiang, and R. Simon. The cross-validated adaptive signature design. *Clinical Cancer Research*, 16(2):691–698, 2010.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

R. Guthke. Zur bedeutung der bioinformatik im kontext der -"omics„-technologien, 2010. URL `http://spectronet.de/story_docs/vortraege_2010/101108_technologietag_jett/101108_12_guthke_hki.pdf`.

T. Hastie and J. Qian. Glmnet vignette. Technical report, Stanford, 2014.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction.* Berlin: Springer Series in Statistics, 2 edition, 2009.

W. Jiang, B. Freidlin, and R. Simon. Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *Journal of the National Cancer Institute*, 99(13):1036–1043, 2007.

E. LeDell, M. Petersen, and M. van der Laan. Computationally efficient confidence intervals for cross-validated area under the roc curve estimates. *Electronic Journal of Statistics*, 9 (1):1583–1607, 2015.

J. Ma, B. P. Hobbs, and F. C. Stingo. Statistical methods for establishing personalized treatment rules in oncology. *BioMed Research International*, 2015:1–13, 2015.

S. Matsui, R. Simon, P. Qu, J. D. Shaughnessy Jr, B. Barlogie, and J. Crowly. Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine. *Clinical Cancer Research*, 18(21):6065–6073, 2012.

M. S. Pepe. *The statistical evaluation of medical tests for classification and prediction.* Oxford University Press, USA, 2003.

J. A. Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, 1988.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

R. Wang, D. A. Schoenfeld, B. Hoeppner, and A. E. Evins. Detecting treatment-covariate interactions using permutation methods. *Statistics in Medicine*, 34(12):2035–2047, 2015.

A. Werft, A. Benner, and A. Kopp-Schneider. On the identification of predictive biomarkers: Detecting treatment-by-gene interaction in high-dimensioanl data. *Computational Statistics and Data Analysis*, 56(5):1275–1286, 2012.

Eigenständigkeitserklärung:

Ich versichere, dass ich die vorgelegte Masterarbeit eigenständig und ohne fremde Hilfe verfasst, keine anderen als die angegebenen Quellen verwendet und die den benutzten Quellen entnommenen Passagen als solche kenntlich gemacht habe. Diese Masterarbeit ist in dieser oder einer ähnlichen Form in keinem anderen Kurs vorgelegt worden.

München, den