Nichtparametrische Analyse von diagnostischen Tests

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von
Carola Werner
aus Wolfenbüttel

Göttingen, 2006

D7
Referent: Prof. Dr. Edgar Brunner
Korreferent: Prof. Dr. Manfred Denker
Tag der mündlichen Prüfung: 7. Juli 2006

Danksagung

Ich möchte mich als erstes bei Herrn Prof. Dr. Edgar Brunner bedanken, der das Thema meiner Arbeit vorgeschlagen hat und mich bei der Anfertigung engagiert unterstützt hat. Durch die Bereitstellung der Mittel und Möglichkeiten in der Abteilung für Medizinische Statistik und die Ermutigung zu Konferenzbeiträgen und Veröffentlichungen hat er meine wissenschaftliche Entwicklung bedeutend vorangetrieben.
Außerdem danke ich Herrn Prof. Dr. Manfred Denker für die Übernahme des Korreferates.
Weiterhin möchte ich mich bei Herrn Dr. Jörg Kaufmann für die fruchtbaren Diskussionen und Einblicke in die aktuelle Forschung in der Pharmaindustrie bedanken.
Außerdem möchte ich mich bei Herrn Prof. Dr. W. Röhl für den Kontakt zu Dr. Bernhardt bedanken, der mir mein erstes richtiges Beispiel für clustered data geliefert hat.
Schließlich gilt mein Dank all denen, die mich auf dem Weg hierher moralisch unterstützt und begleitet haben, allen voran Moritz Hiller. Ein großer Dank geht auch an Leif Boysen und Karin Neubert für die Hilfe auf den „letzten Metern“. Nicht zu vergessen natürlich auch die lieben Kollegen der Abteilungen Medizinische Statistik und Genetische Epidemiologie, alle netten Kommilitonen und Professoren vom Promotionsstudiengang „Angewandte Statistik und Empirische Methoden“ und meine Mama, die in der zweiten Klasse meine Liebe zur Mathematik geweckt hat.

Inhaltsverzeichnis

1 Einleitung

Die Biometrie beschäftigt sich mit der Anwendung statistischer Methoden in der Biologie, Psychologie, Land- und Forstwissenschaft und der Medizin. Im letztgenannten Fall stellt insbesondere die klinische Forschung ein wichtiges Teilgebiet dar. Hier steht der Vergleich verschiedener Behandlungs- und Therapiemethoden im Vordergrund. Grundlage fast jeder solchen klinischen Studie bilden jedoch diagnostische Verfahren, die Voraussetzung für das Erkennen einer Krankheit sind. Die Entwicklung neuer diagnostischer Verfahren in sogenannten Diagnosestudien muss deshalb nach genauso strengen wissenschaftlichen und regulatorischen Richtlinien erfolgen, wie sie auch für Arzneimittel und andere medizinische Produkte gelten. Eine der zentralen Aufgaben der Biometrie ist es somit, valide statistische Methoden für Diagnosestudien zu liefern.

Bei der Evalutation eines neuen diagnostischen Tests werden sogenannte Fälle („Kranke“) und Kontrollen („Gesunde“) mit gesichertem Gesundheitsstatus benötigt. Die Methode, mit der dieser Gesundheitsstatus erhoben wird, ist der sogenannte „Goldstandard“. Der Goldstandard ist ein anerkanntes Verfahren oder die Kombination mehrerer Verfahren, um mit sehr hoher Wahrscheinlichkeit die wahre Diagnose zu erhalten. Häufig besteht dennoch die Notwendigkeit, ein alternatives Verfahren zu entwickeln, was unter anderem daran liegen kann, dass der Goldstandard sehr aufwändig bzw. teuer ist, dass er zu invasiv ist oder sogar erst nach dem Tod des Patienten zu bestimmt werden kann.
In der Diagnosestudie werden die Kollektive der Gesunden und Kranken dann mit dem potentiellen neuen Diagnoseverfahren untersucht. Die dabei erhobenen Messwerte werden in beiden Kollektiven sicher nicht die gleiche Verteilung haben, insbesondere ist es sehr unrealistisch, gleiche Varianzen in beiden Gruppen anzunehmen. Aus statistischer Sicht liegt damit eine Situation vor, die dem sogenannten Behrens-Fisher-Problem sehr ähnlich ist. Dieses Problem besteht im parametrischen Fall darin, bei ungleichen und unbekannten Varianzen einen Unterschied in den Erwartungswerten aufzudecken. Sowohl parametrisch (Welch, 1938) als auch nichtparametrisch (Brunner & Munzel, 2000) gibt es für dieses Problem bereits zahlreiche Lösungsansätze zum Vergleich zweier Stichproben. In neueren Arbeiten werden außerdem nichtparametrische multivariate Verfahren präsentiert (Brunner *et al.* , 2002).
Zum Vergleich diagnostischer Tests wird die Theorie der ROC-Kurven verwendet. Ein Gütemaß für einen diagnostischen Test ist die Fläche unter der ROC-Kurve. Bereits Bamber (1975) und Hanley & McNeil (1982) bildeten die Verbindung zur Nicht-

parametrik durch die Feststellung, dass diese Fläche unter der ROC-Kurve genau der Größe entspricht, die von der nichtparametrischen Wilcoxon-Statistik geschätzt wird. Bisher gibt es in der Literatur keine einheitliche und zusammenhängende Theorie für die nichtparametrische Behandlung der Ergebnisse von Diagnosestudien für beliebige Designs. Die meisten Arbeiten beschränken sich auf den Vergleich von zwei diagnostischen Tests. Außerdem sind die Verfahren meist auf ein bestimmtes Skalenniveau der Messwerte limitiert.

Diese Beschränkungen sollen in der vorliegenden Arbeit behoben werden: es wird eine Theorie entwickelt, die mit leichten Modifikationen für alle Designs, die in Diagnosestudien üblich und sinnvoll sind, anwendbar ist. Diese Designs schließen Studien mit beliebig vielen Untersuchern und Methoden bzw. Modalitäten und Studien mit Mehrfachbeobachtungen an einem Patienten (sogenannte „clustered data“) ein. Die Art der erhobenen Messwerte ist außerdem nicht eingeschränkt, es sind dichotome, ordinale und stetige Werte zugelassen. Es wurde bei der Herleitung der Teststatistiken großer Wert darauf gelegt, dass alle Größen in Rängen bzw. Rangmittelwerten dargestellt werden. Dies ermöglicht auch Anwendern ohne profunde statistische Vorbildung die einfache Interpretation der Effekte.

Aufbau der Arbeit

Zunächst wird eine einführende Übersicht über das Anwendungsgebiet der Diagnosestudien gegeben. Hierbei wird zunächst auf allgemeine Designfragen eingegangen. Im folgenden Kapitel 3 wird die Theorie der Verteilung von Schätzern in drei verschiedenen Modellen hergeleitet und Test-Statistiken und Konfidenzintervalle konstruiert. Im anschließenden Exkurs in Kapitel 4 wird das Verhalten der Testverfahren bei dichotomen Daten beleuchtet. In Kapitel 5 werden die vorgestellten Verfahren auf praktische Beispiele angewendet und die verwendeten SAS-Makros vorgestellt. Im letzten Kapitel wird schließlich das Verhalten der Test-Statistiken bei kleinen und mittleren Stichprobenumfängen mithilfe von Simulationen untersucht. Den Abschluss bildet ein Ausblick auf offene Fragen im Gebiet der nichtparametrischen Auswertung von diagnostischen Studien.

2 Diagnostische Tests

Was ist Diagnostik überhaupt? Darauf soll hier keine Antwort gegeben werden, denn man müsste bei Erörterungen des Erkenntnisprozesses beginnen und mit den Grundsätzen ärztlichen Handelns fortfahren. Die Begriffe Krankheit, Diagnose, Diagnostik, Test, diagnostischer Prozess etc. wären zu definieren. Der Biometriker kann es sich aber einfach machen: Er betrachtet die diagnostische Maßnahme als Mittel, eine Wahrscheinlichkeit für die Richtigkeit der Vermutung, dass ein Patient an einer Krankheit leidet, in eine (möglichst) höhere Wahrscheinlichkeit zu transformieren.
Eine Einleitung findet man in einem Werk über Methoden der medizinischen Diagnostik (Köbberling *et al.* , 1991). Um die Diagnostik zu beschreiben, wird sie als Folge von binären Einzelentscheidungen aufgefasst. Bei diesen Einzelunterscheidungen werden diagnostische Tests eingesetzt, die zwischen zwei Zuständen entscheiden sollen: Krankheit vorhanden bzw. nicht vorhanden. Entsprechend ist auch das Testresultat eine Ja/Nein-Aussage: krank (=positiv) / nicht krank (=negativ). Bei Tests mit quantitativen Ergebnissen, wie z.B. bei Laborwerten, erfolgt die Überführung in eine solche binäre Aussage mit einem Trennwert („Cut-off“Wert). Hieraus lässt sich eine Vierfeldertafel erzeugen, die den Zustand des Patienten dem Testergebnis gegenüberstellt:

	Patient krank	nicht krank
Test positiv	richtig positiv (RP)	falsch positiv (FP)
Test negativ	falsch negativ (FN)	richtig negativ (RN)

Anhand dieser Tafel lassen sich spalten- und zeilenweise je die Verhältnisse der Einzelzellen zu den Summen bilden. Die Sensitivität ermittelt den Anteil der richtig positiv erkannten Patienten an allen Kranken (#RP/(#RP+#FN)), die Spezifität den Anteil der richtig negativ erkannten Patienten an den Nicht-Kranken (#RN/(#RN+#FP)). Sensitivität und Spezifität sind die Größen, die die Entwickler und Hersteller bei der Bewertung ihrer diagnostischen Tests verwenden können. Die Vorhersagewerte (zeilenweise Betrachtung) betrachten dagegen die Wahrscheinlichkeit, dass der Patient tatsächlich den Zustand aufweist, den der Test anzeigt (positiv prädiktiver Wert: #RP/(#RP+#FP), negativ prädiktiver Wert: #RN/(#RN+#FN)). Die Vorhersagewerte beschreiben damit die Wahrscheinlichkeiten aus der Sicht des Arztes, dem das Testergebnis vorliegt. Er kann mit diesen Werten das

Testergebnis hinsichtlich seiner Relevanz einschätzen. Allerdings sind diese Werte nicht unabhängig von der Prävalenz, also der Häufigkeit, mit der die Krankheit im untersuchten Kollektiv auftritt. Die Größen Sensitivität und Spezifität dagegen sind prävalenzunabhängig und deshalb für den Vergleich diagnostischer Verfahren besser geeignet.
Der wahre Gesundheitszustand des Patienten wird mit dem sogenannten Goldstandard bestimmt. Dies ist das beste zur Verfügung stehende Verfahren, um die Diagnose zu stellen. Manchmal muss man hierfür mehrere Verfahren kombinieren oder der Goldstandard ist erst nach dem Tod des Patienten oder nach einer Biopsie zu bestimmen. In dieser Arbeit wird davon ausgegangen, dass es immer möglich ist, den Goldstandard mit hoher Wahrscheinlichkeit richtig zu bestimmen. Um dies zu gewährleisten, ist es wichtig, systematische Fehler zu vermeiden, die zu Verzerrungen („Bias“) führen: der „selection bias“ und der „information bias“.
Der selection bias beschreibt die zu erwartende Verzerrung, wenn der Goldstandard ein invasives Verfahren ist. Dies wird dann nämlich nur bei den Testpositiven eingesetzt, während bei Testnegativen darauf verständlicherweise verzichtet wird. Es ist eine Überschätzung der Sensitivität zu erwarten.
Der information bias wird dadurch hervorgerufen, dass die Kenntnis des Goldstandards das Ergebnis des zu beurteilenden Tests beeinflusst. Dies ist insbesondere bei Verfahren zu erwarten, bei denen Befunde interpretiert werden müssen (zum Beispiel bei bildgebenden Verfahren).
Um diese systematischen Fehler zu vermeiden, ist es wichtig, sich ihrer bewusst zu sein und schon bei der Planung einer Studie Gegenmaßnahmen zu ergreifen.
In den folgenden Abschnitten werden wichtige Begriffe aus dem Bereich der Diagnosestudien vorgestellt und erläutert: ROC-Kurven, Indizes für die Treffsicherheit, Reader und Methoden, clustered data und die verwendeten Skalenniveaus. Außerdem wird eine kurze Übersicht über andere statistische Verfahren und Fragestellungen in der Diagnostik gegeben.

2.1 ROC-Kurven

In der diagnostischen Medizin unterliegt die Entwicklung neuer Verfahren und Prozeduren den gleichen strengen Richtlinien wie in der Entwicklung neuer Medikamente und Medizinprodukte. Die Studien müssen gründlich geplant und vorbereitet werden und auch die Auswertung bedarf spezieller statistischer Methoden. Das Kollektiv der Gesunden und der Kranken wird untersucht, anschließend muss beurteilt werden, wie gut das neue Verfahren in der Lage ist, die beiden Kollektive voneinander zu trennen. Diese Trennschärfe kann bei dichotomen Testergebnissen direkt über die Sensitivität und Spezifität berechnet werden. Sind die Testergebnisse allerdings stetig oder ordinal (haben sie also mehr als zwei Ausprägungen), bedarf es anderer Methoden der Analyse. Hier hat sich die Auswertung mithilfe von ROC-Kurven durchgesetzt, wel-

che unabhängig von einem gewählten Cut-off Wert die Trennschärfe (Accuracy) des Tests bewerten. Diese Theorie der ROC-Kurven stammt aus der Nachrichtentechnik (Signalerkennung) (Peterson *et al.* , 1954), wo sie den Zusammenhang zwischen richtig und falsch erkannten Signalen wiedergibt. Deshalb ist sie ein verbreitetes Verfahren, um solche Studien auszuwerten. Dabei wird jeder Messwert vom Minimum bis zum Maximum der bestimmten Messungen als Cut-off Wert gesetzt und jeweils die Sensitivität und Spezifität des Verfahrens geschätzt. Die so entstandenen Paare aus Sensitivität und Spezifität werden in einem Koordinatensystem mit x-Achse (1-Spezifität) und y-Achse (Sensitivität) aufgetragen. Durch Verbinden der Punkte entsteht die sogenannte „Receiver Operating Characteristic"(ROC)-Kurve.

In Abbildung 2.1 ist exemplarisch dargestellt, wie aus zwei Histogrammen (dunkelgrau: gesunde Patienten, hellgrau kranke Patienten) eine ROC-Kurve entsteht. Der erste Fall in Abbildung 2.1.a repräsentiert zwei Verteilungen, die sich überlappen. Daraus entsteht eine ROC-Kurve mit einem typischen, mittleren Verlauf. Um dieses Beispiel besser einordnen zu können sind außerdem die beiden Extremfälle für die Verläufe von empirischen ROC-Kurven dargestellt: besitzt das Verfahren eine Trennschärfe, die nicht besser als der Zufall ist, so überlagern sich die Histogramme fast vollständig und ie daraus resultierende ROC-Kurve wird sich kaum von der Winkelhalbierenden unterscheiden (vergleiche Abbildung 2.1.b). Ist das Verfahren dagegen perfekt, so wird die Kurve bis in die linke obere Ecke des Einheitssystems reichen (vergleiche Abbildung 2.1.c).

Der Verlauf dieser ROC-Kurve kann empirisch ermittelt werden, indem die Paare aus Sensitivität und Spezifität mit Hilfe einer Treppenfunktion verbunden werden. Eine andere Möglichkeit, den Kurvenverlauf zu bestimmen, basiert auf der Annahme, dass die ursprünglichen Verteilungen F_0 und F_1 der Gesunden und Kranken einer Normalverteilung folgen. In dem Fall werden zwei Parameter a (Differenz der Erwartungswerte) und b (Quotient der empirischen Varianzen) geschätzt. Der Plot der Punktepaare $[1-\Phi(c), 1-\Phi(\hat{b}c-\hat{a})]$ für $-\infty < c < \infty$ ergibt dann einen stetigen Schätzer für die ROC-Kurve.

Außerdem haben verschiedene Autoren die Kurve mit Hilfe von Kern-Dichte-Schätzern bestimmt (Lloyd, 1998; Lloyd & Yong, 1999). Hierbei konnte gezeigt werden, dass die Verzerrung des Schätzers bei optimaler Wahl der Bandbreite geringer ist, als die des empirischen Schätzers. Allerdings sind diese Verfahren von der optimalen Wahl der Bandbreite abhängig.

Die ersten Arbeiten zur Anwendung von ROC-Kurven bei der Auswertung diagnostischer Tests waren die von Hanley (1989) und Bamber (1975). Außerdem haben Ransohoff & Feinstein (1978) bereits auf Probleme bei der Planung und Auswertung diagnostischer Studien hingewiesen und die Aufmerksamkeit auf ROC-Kurven gelenkt. In einem Review-Artikel fassen Zweig & Campbell (1993) die Entwicklung der Verwendung der ROC-Kurven in der Diagnostik bis zum Jahre 1993 zusammen.

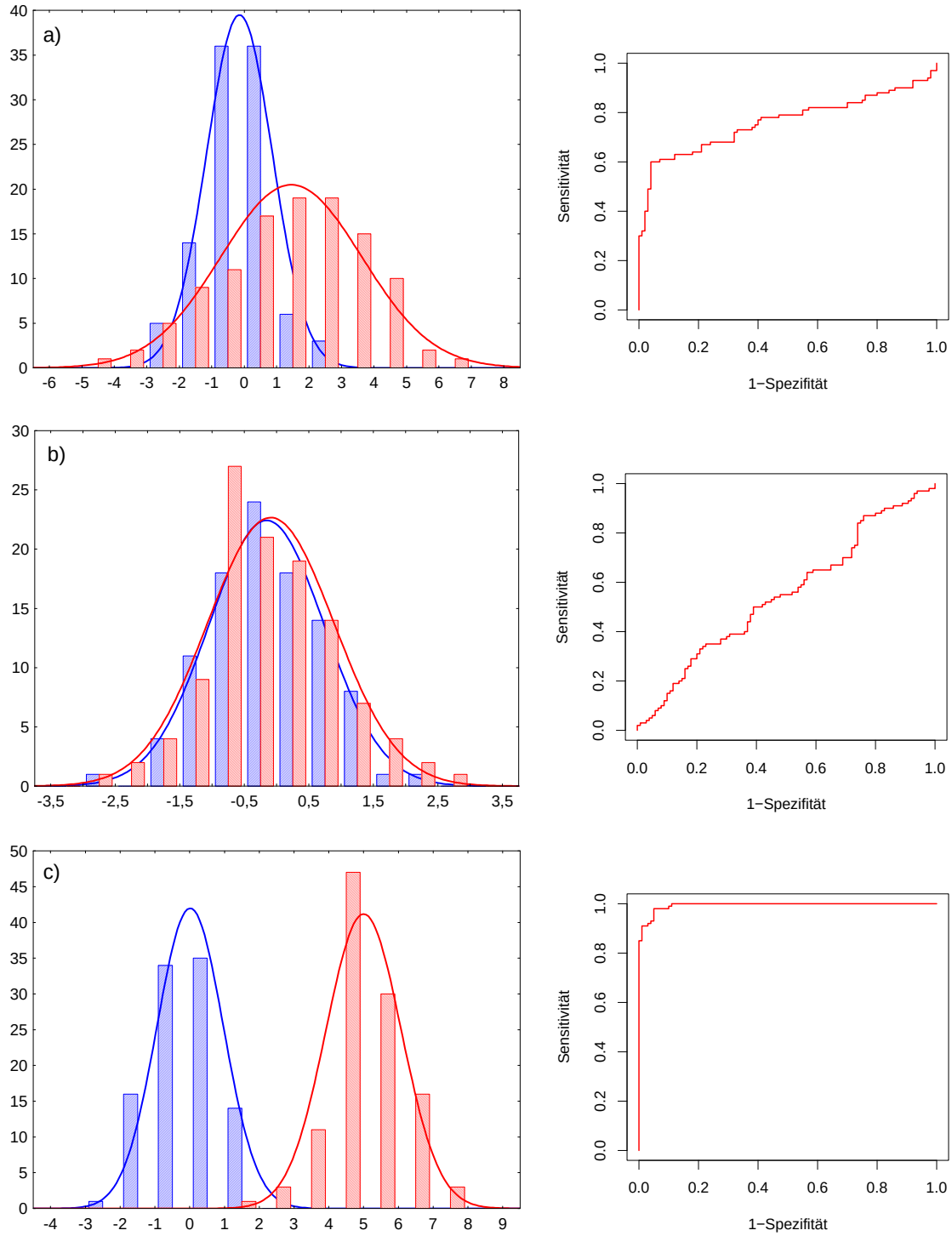

Abbildung 2.1: Empirische Dichten und ROC-Kurven: kranke Patienten (hellgraue Balken) und gesunde Patienten (dunkelgraue Balken). (a) Standardbeispiel, (b) Extremfall 1: nicht besser als der Zufall, (c) Extremfall 2: perfekte Trennung

2.2 Indizes für die Treffsicherheit

Der Vergleich diagnostischer Tests kann auf verschiedene Arten geschehen. Eine Möglichkeit ist die Reduzierung der ROC-Kurve auf einen einzigen Index, der die Trennschärfe des Verfahrens bzw. Tests treffend beschreiben soll. Dies hat natürlich den Verlust von Informationen zur Folge, deshalb sollte die Wahl des Indizes wohlüberlegt getroffen werden. Eine exemplarische Auswahl an Indizes wird im Folgenden vorgestellt.
Ein beliebter Index ist die Fläche unter der ROC-Kurve („Area under the curve"=AUC), weil sie erstens leicht zu beschreiben und zweitens leicht zu berechnen ist. Außerdem hat er auch eine sehr anschauliche Interpretation: er gibt die Wahrscheinlichkeit an, dass zwei zufällig gezogenen Patienten aus den beiden Kollektiven richtig zugeordnet werden würden (das heißt der kranke hat einen größeren Wert als der gesunde). Er gibt Auskunft über die Trennschärfe des Verfahrens auf dem gesamten Wertebereich. Wenn der Anwender keine speziellen Voraussetzungen an die Sensitivität oder Spezifität hat, ist dieser Index der beste. Man kann ihn parametrisch oder nichtparametrisch berechnen. Er wird die zentrale Größe in der vorliegenden Arbeit bilden.
Andere Möglichkeiten der Indexbildung wurden für den Fall vorgestellt, dass der Anwender eine bestimmte Mindest-Sensitivität oder -Spezifität haben möchte (McClish, 1989). Dies kann der Fall sein, wenn eine hohe Sensitivität oder Spezifität unerlässlich sind. In diesem Fall wird nicht die gesamte Fläche unter der Kurve geschätzt, sondern nur der Teil ab der vorgegebenen Grenze. Falls sich die beiden Kurven, die verglichen werden sollen, kreuzen, ist die Interpretation häufig leichter, wenn man nur Teilflächen statt der gesamten Fläche untersucht.
Weiterhin kann man den Youden-Index bestimmen, welcher als maximale Summe zwischen Sensitivität und Spezifität definiert ist. Er ist also der Punkt der Kurve, welcher am weitesten von der Winkelhalbierenden entfernt und somit am nächsten an der oberen linken Ecke des Systems ist. Zum Bestimmen des optimalen Cut-off Wertes ist dieser Index aufgrund seiner einfachen Berechnung sehr gut geeignet. Hier wird allerdings die Kurve auf nur einen einzigen Datenpunkt reduziert, weshalb sich dieser Index zur Beschreibung der Trennschärfe nicht durchgesetzt hat.
Außerdem gibt es die Möglichkeit, bei einer bestimmten, festgelegten Sensitivität die Spezifitäten der Verfahren zu vergleichen oder umgekehrt. Die Wahl dieser festen Sensitivität oder Spezifität ist allerdings eher willkürlich und es gilt hier das gleiche Argument wie beim Youden-Index: dieser Wert ist nicht gut geeignet, um das Verhalten des Tests im allgemeinen zu beschreiben.

2.3 Reader und Methode

Eine diagnostische Studie dient der Evaluation neuer diagnostischer Verfahren oder zum Vergleich existierender Methoden. Einige Beispiele hierfür sind:

- eine Standardmethode und ein neues, nichtetabliertes Verfahren zur Entdeckung von Karzinomen,
- eine Röntgenaufnahme mit und eine Aufnahme ohne Kontrastmittel,
- eine Röntgenaufnahme mit einem etablierten und eine Aufnahme mit einem neuen Kontrastmittel,
- ein Gerät (CT oder MRT) mit verschiedenen Einstellungen bezüglich der Aufnahmegeschwindigkeit oder der Auflösung,
- die digitale Speicherung von Aufnahmen oder eine analoge Aufnahme.

Die Liste könnte lange weitergeführt werden. Außerdem sind auch Kombinationen der angeführten Beispiele möglich. Es muss natürlich darauf geachtet werden, dass ein Patient durch die Aufnahmen nicht über die Maße belastet wird. Es ist deshalb oft nicht möglich, verschiedene Röntgen- und außerdem mehrere MRT- oder CT-Aufnahmen an ein und demselben Patienten durchzuführen. Meistens interessieren den Anwender aber auch nur die Vergleiche zweier Methoden mit verschiedenen Nebenbedingungen. Deshalb wird für die Entwicklung der Theorie der Fall betrachtet, dass alle Methoden von allen Readern an jeder Person durchgeführt werden.
In Studien dieser Art zur Evaluation diagnostischer Tests werden von den Behörden in den Guidelines (European Medicines Agency (EMEA), 2001) mindestens drei Untersucher (Reader) gefordert. Schließlich soll ein Verfahren gute Trenneigenschaften unabhängig von der Person und den Fähigkeiten des Untersuchers aufweisen. Die Reader müssen verblindet agieren, das heißt sie dürfen die wahre Diagnose nicht kennen. Um die Kenntnis der wahren Diagnose zu vermeiden und außerdem zu verhindern, dass Informationen aus bereits befundeten Aufnahmen die Diagnose beeinflussen, muss die Untersuchung mit den verschiedenen Methoden an den verschiedenen Patienten randomisiert stattfinden. Nur dann kann sichergestellt werden, dass der Arzt die Diagnose mit jeder Methode unabhängig von der vorherigen treffen kann.
In dieser Arbeit werden nur vollständig verbundene Designs betrachtet. Hierbei wird jeder Patient von allen R Readern mit allen M Methoden beurteilt. Das resultiert dann in $d = MR$ wiederholten Messungen an einem Subjekt. Diese Messungen können umgeschrieben werden, sodass jede Reader-Methoden-Kombination mit einem Index l bezeichnet wird, wobei folgendes Schema (siehe Tabelle 2.1) verwendet wird: zunächst läuft der Index des Readers und die Methode wird festgehalten, dann

Methode m	1	1	1	...	M	M	M
Reader r	1	...	R	...	1	...	R
Index l	1	...	R		(M-1)R+1	...	d

Tabelle 2.1: Indizierung der Reader-Methoden-Kombinationen

läuft auch der Index der Methode, sodass man am Ende jede Reader-Methoden-Kombination abgedeckt hat.

Wenn ein Patient mehr als einmal befundet wird, muss die Abhängigkeit der Beobachtungen an einem Patienten berücksichtigt werden. Um die Abhängigkeiten zu modellieren, erweiterten Dorfman *et al.* (1992) in einem ersten Ansatz ihre eigene Methode mithilfe von Jackknife-Verfahren von unverbundenen auf verbundene Messungen. Anschließend evaluierten sie diese mithilfe von Simulationsstudien (Roe & Metz, 1997). Theoretische Begründungen für die Gültigkeit dieser Erweiterung lieferten sie allerdings nicht. DeLong *et al.* (1988) entwickelten nichtparametrische Methoden basierend auf der Theorie von U-Statistiken, die auf Sen (1960) zurückgeht. Von Hillis *et al.* (2005) wurden die verschiedenen Methoden verglichen, wobei die nichtparametrische Methode am besten abschnitt. Alle diese vorgestellten Verfahren berücksichtigen aber keine clustered data.

2.4 Clustered data

Clustered data im Sinne dieser Arbeit liegen immer dann vor, wenn ein Patient sowohl gesunde als auch kranke Beobachtungseinheiten liefert. In der Literatur gibt es verschiedene Definitionen des clusterings von Daten (siehe Beam (1998) für einen Überblick). Er spricht immer dann von clustering, wenn man mehr als eine Beobachtung an einem Subjekt durchführt und diese somit abhängig sind. Würde man dieses clustering ignorieren, ignoriert man immer die Korrelation zwischen den abhängigen Messwerten und wird falsche Schlüsse ziehen. Die Definition des clusterings von Beam lässt sich in zwei Stufen einteilen. Die erste Möglichkeit des clusterings bei diagnostischen Tests tritt auf, wenn an einem Patienten mehrere Beobachtungseinheiten des gleichen Gesundheitsstatus erhoben werden. Dieser Fall soll zur besseren Abgrenzung aber nur mit „Messwiederholung“ bezeichnet werden. Diese Abhängigkeiten müssen in der Schätzung der Kovarianz berücksichtigt werden.

Die zweite Möglichkeit - welche in dieser Arbeit als einzige mit clustered data bezeichnet werden soll - erhöht die Komplexität der Daten. Wenn die verschiedenen Beobachtungseinheiten an einem Patienten zusätzlich verschiedene Gesundheitszustände (gesund/krank) haben können, dann sind auch die gesunden und kranken Beobachtungen abhängig und dieser Abhängigkeit muss durch Einführung einer Kovarianz Rechnung getragen werden. Diese muss dann weitere mögliche Abhängig-

Tabelle 2.2: Strukur von clustered Daten, x steht für eine Beobachtung

	Methode 1						Methode 2					
	Reader 1		Reader 2		Reader 3		Reader 1		Reader 2		Reader 3	
Sub.	gesund	krank	gesund	krank	gesund	krank	gesund	krank	gesund	krank	gesund	krank
1	xxx	xx	xxx	xx	xxx	xx	xxx	xx	xxx	xx	xxx	xx
⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮	
n_c	xx	xxxx	xx	xxxx	xx	xxxx	xx	xxxx	xx	xxxx	xx	xxxx
1	xxx		xxx		xxx		xxx		xxx		xxx	
⋮	⋮		⋮		⋮		⋮		⋮		⋮	
n_0-n_c	xx		xx		xx		xx		xx		xx	
1		xxxx		xxxx		xxxx		xxxx		xxxx		xxxx
⋮		⋮		⋮		⋮		⋮		⋮		⋮
n_1-n_c		xxx		xxx		xxx		xxx		xxx		xxx

keiten berücksichtigen, wenn verschiedene Reader die verschiedenen Beobachtungen an einem Patienten bei verschiedenen Gesundheitszuständen diagnostizieren. In Tabelle 2.2 ist schematisch dargestellt, welche Daten an den n Patienten in einer Studie mit drei Readern und zwei Methoden erhoben werden können. Es gibt n_c Patienten, die sowohl gesunde als auch kranke Beobachtungen liefern. Außerdem gibt es insgesamt $n_0 - n_c$ zusätzliche Patienten, die gesunde Beobachtungen liefern und $n_1 - n_c$ Patienten, die kranke Beobachtungen liefern. Also hat man schließlich von n_0 Patienten gesunde Beobachtungen und von n_1 Patienten kranke Beobachtungen. Die drei Gruppengrößen n_0, n_1, n_c müssen nicht notwendigerweise identisch n sein. Wenn sie jedoch identisch sind, dann liegt ein sogenanntes „vollständiges Design“ vor, von jedem Patienten gibt es sowohl kranke, als auch gesunde Beobachtungen. Sind die Größen n_0, n_1, n_c nicht alle identisch, dann liegt dementsprechend ein unvollständiges Design bzw. ein Design mit „fehlenden Werten“ vor. Diese Werte fehlen, weil die Beobachtungseinheit nicht vorhanden ist (der entsprechende Patient hat keine kranken Beobachtungen), und nicht, weil sie nicht gemessen wurden. Jeder der Patienten liefert also eine Anzahl an gesunden und kranken Beobachtungseinheiten (in der Tabelle dargestellt durch „x“), die von jedem Reader mit jeder Methode beurteilt wird.
Außerdem kann die Anzahl der Beobachtungen pro Patient gleich und fest sein (Annahme der festen Messpunkte), wenn man zum Beispiel Gliedmaßen, Organe oder ähnliches betrachtet. Es ist aber auch denkbar, dass die Anzahl der Beobachtungen nicht von vornherein bekannt ist, sondern erst im Laufe der Untersuchung festgestellt wird und zwischen den Patienten auch variieren kann. Hier seien man zum Beispiel Untersuchungen an Tumoren (verschiedene kalte und heiße Herde) oder Leberflecken erwähnt, bei denen der Goldstandard und die Anzahl der Beobachtungseinheiten vor der Auswertung nicht bekannt ist.
In letzter Zeit haben sich mehrere Arbeiten mit dem allgemeinen Thema nichtparametrische Methoden für clustered data beschäftigt, unabhängig von der möglichen

Anwendung auf Diagnosestudien. Datta & Satten (2005) verwenden Permutationen im weiteren Sinne. Die Tests werden mehrfach durchgeführt, wobei jedes Mal eine andere Beobachtung innerhalb eines Clusters ausgewählt wird. Sie haben die Methode zunächst parametrisch für clustered data hergeleitet, in denen die Clustergröße informativ ist (Williamson *et al.* , 2003). In der neueren Arbeit haben sie die Theorie dann für Rangsummentests erweitert. Dieses Verfahren führen sie sowohl für unabhängige als auch für abhängige Gruppen durch. Sie betrachten allerdings nicht den Fall ungleicher Varianzen in den beiden Gruppen, was für die Betrachtung von ROC-Kurven aber notwendig ist.
In einer zweiten Arbeit (Rosner & Grove, 1999) wird ein stratifiziertes Verfahren vorgestellt. Hier wird für jede Clustergröße ein einzelner Test durchgeführt und diese dann geeignet kombiniert. Dieses Verfahren ist allerdings nur für unabhängige Gruppen anwendbar. Später wurde für das Verfahren auch noch eine asymptotische Theorie für große Stichproben hergeleitet (Rosner *et al.* , 2003).
(Benhin *et al.* , 2005) greifen das Problem der informativen Clustergröße auf. Die Arbeit befasst sich mit der parametrischen Auswertung cluster-korrelierter Daten, könnte aber auch als Grundlage nichtparametrischer Überlegungen dienen. Die Autoren stellen eine Methode vor, die die Resampling-Methoden von Datta und Satten umgeht und stattdessen "mean estimation equations" verwendet.
Die beschriebenen Verfahren für clustered data sind also nur begrenzt bei der Evaluation diagnostischer Tests einsetzbar. Clustered data in Kombination mit Diagnosestudien wurden bisher nur von Obuchowski (1997) betrachtet. Dort wurde aber nur der einfache Fall des Zweistichprobenvergleichs entwickelt.

2.5 Skalenniveau der Messwerte

Die Daten, die bei diagnostischen Tests erhoben werden, reichen von stetigen Messwerten wie Laborparametern, Messungen von Volumen und Umfängen über ordinale Daten wie Scores und graduelle Bewertungen von Krankheitsstadien bis hin zu dichotomen Daten (eine Krankheit oder ein Merkmal ist vorhanden oder nicht). Ohne Einschränkung der Allgemeinheit kann man annehmen, dass größere Werte für kranke Beobachtungen stehen (falls dies nicht der Fall ist, können die Werte durch eine geeignete Transformation modifiziert werden). Der Vorteil der nichtparametrischen Methoden liegt in diesem Bereich auf der Hand: man kann ein einziges statistisches Verfahren für alle Skalenniveaus anwenden. Es ist natürlich nicht immer das optimale Verfahren. Will man aber verschiedene diagnostische Tests mit verschiedenen Arten von Messwerten vergleichen, ist es zwingend notwendig, ein Verfahren zu verwenden, dass für alle Messwerte geeignet ist. Deshalb wird zunächst das allgemeine Verfahren vorgestellt werden und später noch auf seine Eigenschaften im Bereich der dichotomen Daten untersucht. Dichotome Ergebnisse spielen historisch eine zentrale Rolle bei Diagnosestudien und manche Ergebnisse treten auf natürliche Weise nur

binär auf.
Es ist also eine Theorie notwendig, die auf beliebige Verteilungen und Skalenniveaus anwendbar ist. Die Theorie des nichtparametrischen Behrens-Fisher-Problems wurde bereits von Brunner & Denker (1994) im Zusammenhang mit stetigen Messwerten erwähnt. Später wurde dieses Verfahren auf nichtstetige Verfahren erweitert (Brunner *et al.* , 2002). Die Theorie soll in dieser Arbeit auf ihre Anwendbarkeit in Diagnosestudien untersucht und ergänzt werden.

2.6 Andere Arbeiten über Diagnosestudien

Die Auswertung diagnostischer Tests mit statistischen Methoden ist in der letzten Zeit ein sehr großes Forschungsgebiet geworden. Da sind einerseits Methoden zur Evaluation sequentieller Tests (Su *et al.* , 2004; Thompson, 2003), das heißt, die Hintereinanderanwendung mehrerer diagnostischer Tests zur Erhöhung der Accuracy. Außerdem gibt es Techniken zur Einbindung von Kovariablen in die Auswertung der Studien (Schisterman *et al.* , 2004). Weiterhin wird untersucht, wie man diagnostische Studien mit fehlendem oder verzerrtem Goldstandard auswerten kann (Zhou & Castelluccio, 2003). Methoden zur Bestimmung einer drei- oder höherdimensionalen ROC-Kurve betrachten Nakas & Yiannoutsos (2004); Obuchowski (2005); Obuchowski *et al.* (2001). Hier wird dann von einer „Fläche unter der ROC-Oberfläche“ („Area under the surface“) gesprochen. Bei Obuchowski (2005) ist der Goldstandard nicht nur ordinal, sondern auf einer stetigen Skala gemessen. Es wird dann eine mittlere Accuracy angegeben. Ist man nicht an der gesamten Kurve, sondern nur an einem Teil interessiert, so gibt es Arbeiten von Dodd & Pepe (2003) und McClish (1989) zur „partial area under the curve“ (pAUC). Anwendungen von ROC-Kurven in Meta-Analysen haben zur Entwicklung der „Summary-ROC-Kurven“ geführt (Walter, 2003, 2005). Diese sogenannten SROC-Kurven helfen dabei, Ergebnisse aus verschiedenen Diagnosestudien, von denen nur die Falsch-Positiv-Raten und Richtig-Positiv-Raten zur Verfügung stehen. Es entsteht dann wieder eine ROC-Kurve; die Fläche darunter kann auch wieder mithilfe der Trapezmethode bestimmt werden. Die Fläche wird sowohl vollständig (Walter, 2003), als auch partiell (Walter, 2005) betrachtet.
Andere Autoren beschäftigen sich mit der Anwendung der ROC-Kurven in Äquivalenzstudien (Obuchowski, 2001; Lui & Zhou, 2004) bzw. in Non-Inferiority Studien (Lu *et al.* , 2003).
Obuchowski (2000) stellt Tafeln für die Stichprobenplanung von diagnostischen Studien vor. Außerdem gibt es von ihr auch Tafeln zur Bestimmung von Konfidenzintervallen (Obuchowski & Lieber, 1998), wenn die Accuracy eines Verfahrens sehr hoch (also sehr nah an 1) ist.

3 Theorie

Im Folgenden soll die Schätzung der Accuracy eines diagnostischen Tests mit Hilfe der Fläche unter der ROC-Kurve hergeleitet werden. Die dargestellte Theorie ist auf andere Definitionen der Accuracy (z.B. die partielle Fläche unter der Kurve) durchaus übertragbar, was an dieser Stelle aber nicht ausgeführt werden soll.
Die Herleitung der Effektschätzer wird zunächst am Beispiel der unabhängigen Beobachtungen wie in Kaufmann *et al.* (2005) erläutert, anschließend wird die Theorie für zwei verschiedene Formen von clustered data entwickelt (Werner & Brunner, 2006). Die statistischen Modelle, die betrachtet werden sollen, lassen sich in drei Kategorien einteilen:

- Modell 1 („Einfachmessung“): an jedem Patienten wird eine Beobachtung erhoben,
- Modell 2 („Mehrfachmessung“): an jedem Patienten werden mehrere Beobachtungen erhoben, diese haben jedoch alle denselben Gesundheitszustand,
- Modell 3 („clustered data“): an jedem Patienten werden mehrere Beobachtungen erhoben, diese können verschiedene Gesundheitszustände haben.

Die drei verschiedenen Modelle sind im Folgenden schematisch dargestellt. Beispielhaft werden die Datenstrukturen für eine Studie mit zwei Methoden und drei Readern gezeigt. In einer Doppelspalte gesund/krank steht jedes x für eine Beobachtungseinheit. Die wiederholten Messungen der Beobachtungseinheiten werden dadurch repräsentiert, dass die gleichen x in jeder Doppelspalte vorkommen. Die Vollständigkeit des Designs in Bezug auf fehlende Messungen ist daran zu erkennen, dass in jeder Zelle einer Zeile die Häufigkeit der x für gesund und krank immer dieselbe ist.

Betrachten wir zunächst das einfachste der drei Modelle mit nur einer Beobachtungseinheit pro Patient in Tabelle 3.1. Ein Beispiel hierfür ist die Untersuchung einer Blutprobe auf die Streptokokken-Dichte.
Im Modell 2 (Tabelle 3.2) ist es dann erlaubt, dass pro Subjekt mehr als eine Beobachtung erhoben wird. Dies kann zum Beispiel die Untersuchung mehrerer histologischer Schnitte an malignen und benignen Tumoren sein.
Das dritte Modell (Tabelle 3.3) weist die komplexeste Struktur der Daten auf. Es ist nun möglich, dass an einem Patienten sowohl kranke als auch gesunde Beobachtungseinheiten erhoben werden. Die Untersuchung aller wichtigen Arterien auf eine

Tabelle 3.1: Schematische Darstellung des Modells 1, x steht für eine Beobachtung

	Methode 1						Methode 2					
	Reader 1		Reader 2		Reader 3		Reader 1		Reader 2		Reader 3	
Sub.	gesund	krank	gesund	krank	gesund	krank	gesund	krank	gesund	krank	gesund	krank
1	x		x		x		x		x		x	
⋮	⋮		⋮		⋮		⋮		⋮		⋮	
n_0	x		x		x		x		x		x	
1		x		x		x		x		x		x
⋮		⋮		⋮		⋮		⋮		⋮		⋮
n_1		x		x		x		x		x		x

Tabelle 3.2: Schematische Darstellung des Modells 2, x steht für eine Beobachtung

	Methode 1						Methode 2					
	Reader 1		Reader 2		Reader 3		Reader 1		Reader 2		Reader 3	
Sub.	gesund	krank	gesund	krank	gesund	krank	gesund	krank	gesund	krank	gesund	krank
1	xxx		xxx		xxx		xxx		xxx		xxx	
⋮	⋮		⋮		⋮		⋮		⋮		⋮	
n_0	xx		xx		xx		xx		xx		xx	
1		xxxx		xxxx		xxxx		xxxx		xxxx		xxxx
⋮		⋮		⋮		⋮		⋮		⋮		⋮
n_1		xxx		xxx		xxx		xxx		xxx		xxx

Stenose stellt hier ein gutes Beispiel dar, denn nur sehr selten werden diese Arterien bei einem Patienten alle verschlossen sein. Wählt man außerdem als Studiengruppe Risikopatienten, so wird auch kaum einer der Patienten gar keinen Verschluss haben.
Allen Modellen gemeinsam ist die Abhängigkeit der wiederholten Messungen an einem Patienten durch verschiedene Reader oder Methoden. Dadurch entsteht der multivariate Charakter des Designs. Diese Abhängigkeit wird in der Literatur häufig mit „correlated“ oder bereits mit „clustered“ data bezeichnet. In dieser Arbeit soll das Wort clustered aber nur für das Modell 3 stehen.
Die Modelle unterscheiden sich vor allem in der Schätzung der Kovarianzstruktur. Sobald diese bestimmt wurde, können in allen drei Modellen nach einem ähnlichen Prinzip Hypothesen getestet werden oder Konfidenzintervalle für Effekte aufgestellt werden.
Die Herleitung der Theorie basiert vor allem auf Arbeiten zum multivariaten nichtparametrischen Behrens-Fisher Problem (Brunner *et al.* , 2002), dem multivariaten nichtparametrischen Modell für verschiedene Messwiederholungen und fehlende Werte (Brunner *et al.* , 1999) und dem allgemeinen nichtparametrischen Modell von Brunner & Denker (1994), in dem noch keine beliebigen Verteilungsfunktionen zugelassen waren. Die Theorie, die in den Arbeiten noch nicht abgedeckt ist, besteht

Tabelle 3.3: Schematische Darstellung des Modells 3, x steht für eine Beobachtung

	Methode 1						Methode 2					
	Reader 1		Reader 2		Reader 3		Reader 1		Reader 2		Reader 3	
Sub.	gesund	krank	gesund	krank	gesund	krank	gesund	krank	gesund	krank	gesund	krank
1	xxx	xx	xxx	xx	xxx	xx	xxx	xx	xxx	xx	xxx	xx
⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮	⋮	⋮	
n_c	xx	xxxx	xx	xxxx	xx	xxxx	xx	xxxx	xx	xxxx	xx	xxxx
1	xxx		xxx		xxx		xxx		xxx		xxx	
⋮	⋮		⋮		⋮		⋮		⋮		⋮	
n_0-n_c	xx		xx		xx		xx		xx		xx	
1		xxxx		xxxx		xxxx		xxxx		xxxx		xxxx
⋮		⋮		⋮		⋮		⋮		⋮		⋮
n_1-n_c		xxx		xxx		xxx		xxx		xxx		xxx

aus dem multivariaten verbundenen Behrens-Fisher Problem mit fehlenden Werten, wobei fehlende Werte sich hier wieder auf fehlende Beobachtungseinheiten und nicht fehlende Messungen bezieht.
Das Modell 2 und Modell 3 bilden die zentralen Modelle der vorliegenden Arbeit. Das Modell 1 ist - wie bereits erwähnt - als Spezialfall der anderen zu betrachten. Deswegen werden die Beweise nicht extra angegeben. Wenngleich man bemerken muss, dass die Beweise im Fall unabhängiger Gruppen von Kranken und Gesunden (also Modell 1) weitaus einfacher zu führen sind als im abhängigen Fall. Für die Beweistechniken, die in den ersten zwei Modellen zum Tragen kommen, sei der Leser auf die Arbeiten zum multivariaten Behrens-Fisher-Problem (Brunner *et al.* , 2002) und zum allgemeinen nichtparametrischen Modell (Brunner & Denker, 1994) hingewiesen.
Die ersten drei Abschnitte dieses Kapitels dienen der Herleitung der Schätzer und deren Verteilungen in den drei verschiedenen Modellen. Hier werden Gemeinsamkeiten und Unterschiede dargestellt. Bei der Beschreibung des Modells 1 geht es vor allem darum aufzuzeigen, wie man das nichtparametrische multivariate Behrens-Fisher-Problem mit der Begrifflichkeit der Diagnosestudien lesen kann. Die hier vorgestellte Vorgehensweise wird dann in den Modellen 2 und 3 weitergeführt. In den darauf folgenden Abschnitten werden Hypothesen (Abschnitt 3.4), Test-Statistiken (Abschnitt 3.5) und Konfidenzintervalle (Abschnitt 3.6) für alle drei Modelle beschrieben.

3.1 Modell 1: Einfachmessung

3.1.1 Modell und Notation

Wir betrachten n_0 gesunde und n_1 kranke Patienten. Bezeichne $X_{ik}^{(l)} \sim F_i^{(l)}$ die Beobachtung des k-ten ($k = 1, ..., n_i$) Patienten in Gruppe i ($i = 0, 1$), die mit Reader-Methoden-Kombination l erhoben wurde. Diese Beobachtungen sind unabhängig für verschiedene k, aber abhängig für verschiedene l bei gleichem k und i. Es gibt insgesamt $n_0 + n_1 = N$ unabhängige Beobachtungsvektoren. Die Verteilungsfunktionen können beliebig stetig oder unstetig sein, einzig Ein-Punkt-Verteilungen werden ausgeschlossen. Für die Herleitung der asymptotischen Ergebnisse sind folgende Annahmen zu machen.

(V1) Für $l, j = 1, ..., d$ muss die bivariate Verteilung von $(X_{ik}^{(l)}, X_{ik}^{(j)})$ für alle $k = 1, ..., n_i, i = 0, 1$ identisch sein.

(A1) $N \to \infty$, so dass $N/n_i \leq N_0 < \infty$, $i = 0, 1$ d.h. der Quotient der Anzahl der Beobachtungen und der Anzahl der Patienten in einer Gruppe muss gleichmäßig beschränkt sein.

Mit anderen Worten heißt (A1), dass die Stichprobenumfänge der beiden Gruppen nicht zu stark unbalanciert sein dürfen.

3.1.2 Ein Schätzer für die Accuracy

Betrachten wir zunächst nur eine einzige Reader-Methoden Kombination. Die Fläche unter der ROC-Kurve kann man mithilfe der Verteilungsfunktionen der Gesunden und Kranken F_0 und F_1 folgendermaßen definieren:

$$\text{AUC} = p = \int F_0 dF_1.$$

Dies ist genau der Effekt, der von der Mann-Whitney Statistik geschätzt wird. Dieser wird in der Nichtparametrik „relativer Effekt“ genannt wird, da er die beiden Verteilungsfunktionen in Relation zueinander setzt. Als Schätzer für die F_i ($i = 0, 1$) betrachten wir die normalisierte Version der empirischen Verteilungsfunktionen, die mithilfe der Zählfunktion $c(x)$ (s. S. 83) aufgestellt wird:

$$\widehat{F}_i(x) = \frac{1}{n_i} \sum_{k=1}^{n_i} c(x - X_{ik}).$$

Die Verwendung dieser Verteilungsfunktionen führt zur Definition der Ränge als

$$R_{ik} = n_0 \widehat{F}_0(X_{ik}) + n_1 \widehat{F}_1(X_{ik}) + \frac{1}{2}.$$

Die Verwendung dieser sogenannten Mittelränge wird notwendig, wenn Bindungen in den Daten auftreten. Das ist zum Beispiel bei ordinalen Daten der Fall. Setzt man die empirischen Verteilungsfunktionen $\widehat{F}_i$ in das Integral ein, so erhält man den Schätzer für den relativen Effekt, der sich auch über die Ränge R_{ik} der Beobachtungen X_{ik} darstellen lässt:

$$\widehat{p} = \int \widehat{F}_0 d\widehat{F}_1 = \frac{1}{n_0}(\overline{R}_{1\cdot} - \frac{n_1+1}{2}) = \frac{1}{N}(\overline{R}_{1\cdot} - \overline{R}_{0\cdot}) + \frac{1}{2}.$$

Diese Herleitung ist natürlich für jede Reader-Methoden-Kombination l gültig. Damit sichergestellt wird, dass man für jede dieser Kombinationen die Accuracy unabhängig davon erhält, was in den anderen Kombinationen beobachtet wurde, werden jedes Mal neue Ränge vergeben. Die empirischen Verteilungsfunktionen der $F_i^{(l)}$ $(i = 0, 1)$ sind also immer innerhalb einer Reader-Methoden-Kombination l definiert:

$$\widehat{F}_i^{(l)}(x) = \frac{1}{n_i}\sum_{k=1}^{n_i} c(x - X_{ik}^{(l)}).$$

Die Verwendung dieser Verteilungsfunktionen führt dann zur Definition der Ränge innerhalb einer Reader-Methoden-Kombination l als

$$R_{ik}^{(l)} = n_0\widehat{F}_0^{(l)}(X_{ik}^{(l)}) + n_1\widehat{F}_1^{(l)}(X_{ik}^{(l)}) + \frac{1}{2}.$$

Man erhält also für jede Kombination einen Schätzer für die AUC:

$$\widehat{p}_{mr} = \widehat{p}_l = \frac{1}{N}(\overline{R}_{1\cdot}^{(l)} - \overline{R}_{0\cdot}^{(l)}) + \frac{1}{2}.$$

So bleibt die anschauliche Interpretation der Accuracy erhalten. Die $d = MR$ Accuracies werden in einem Vektor $\widehat{\mathbf{p}}$ zusammengefasst:

$$\widehat{\mathbf{p}} = (\widehat{p}_{11}, ..., \widehat{p}_{1R}, ..., \widehat{p}_{M1}, ..., \widehat{p}_{MR})' = (\widehat{p}_1, ..., \widehat{p}_d)'.$$

Die Reihenfolge der Indizes wurde bereits auf Seite 9 vorgestellt. Dieser Schätzer für die Accuracy ist erwartungstreu und konsistent. Der Beweis hierfür wird in Brunner *et al.* (2002) ausführlich dargestellt. Entgegen dieser Theorie findet man in der Literatur immer wieder die Behauptung, der Schätzer, der nach der Trapezregel die Fläche unter der ROC-Kurve angibt, unterschätzt bei ordinalen Daten die wahre Fläche systematisch (Zhou *et al.* , 2002) und deshalb sollte besser ein Schätzer verwendet werden, der Parameter einer Normalverteilung schätzt. Diese Behauptung bedarf einer genaueren Untersuchung.
Die Behauptung stammt ursprünglich aus der Arbeit von Hanley & McNeil (1982). Dort unterstellte man den ordinalen Beobachtungen eine unterliegende stetige „wahre“ Verteilung und folgerte, dass dann auch die ROC-Kurve stetig sein sollte. Diese

Kurve würde natürlich mit den fünf Paaren an Sensitivität und Spezifität, die man bei Verwendung eines 5-Punkte-Scores erhielte, nur sehr ungenau beschrieben. In diesem Fall kann man sagen, dass die Fläche unter der Treppenfunktion die Fläche unter der „wahren“ stetigen Kurve unterschätzt. Wenn man allerdings davon ausgeht, dass die Scores tatsächlich einer diskreten, nichtstetigen Verteilung entstammen, ist es nicht sinnvoll, stetige Verteilungen anzupassen, um den Schätzer für die Fläche unter der Kurve zu erhalten.
Newcombe (2006) unterteilt die möglichen Verteilungen, die den Beobachtungen unterliegen können, in drei Gruppen:

1. stetige Verteilungen, deren Daten nahezu stetig gemessen werden,
2. stetige Verteilungen, deren Daten mit sehr vielen Bindungen, also diskret gemessen werden, sowie
3. diskrete Verteilungen.

In den Fällen 1 und 3 wird der nichtparametrische Schätzer empfohlen, lediglich im Fall 2 werden auch parametrische Methoden in Erwägung gezogen.
Insgesamt ist zu sagen, dass das Missverständnis, der nichtparametrische Schätzer unterschätze die wahre Fläche systematisch, daher rührt, dass die ersten Arbeiten zur Accuracy immer annahmen, dass die latente, nicht zu beobachtende Variable, stetig ist. Sobald diese Annahme nicht gemacht wird, kann davon ausgegangen werden, dass der Schätzer erwartungstreu ist. Unglücklicherweise werden die alten Arbeiten (Hanley & McNeil, 1982) auch heutzutage immer noch zitiert, wenn behauptet wird, der nichtparametrische Schätzer sei in allen Fällen schlechter als ein angepasster stetiger Schätzer (Lloyd, 1998; Lloyd & Yong, 1999).

3.1.3 Verteilung des Schätzers

Es werden keine Annahmen an die Gleichheit der Varianzen der beiden Verteilungen F_0 und F_1 gemacht. Deshalb liegt das Schätzproblem vor, das aus dem Behrens-Fisher-Problem bekannt ist. Aus Brunner *et al.* (2002) kann man die Verteilung von $\sqrt{N}(\widehat{\mathbf{p}} - \mathbf{p})$ ableiten. Es kann gezeigt werden, dass der Vektor der Accuracies asymptotisch äquivalent zu einer Summe $\sqrt{N}\mathbf{B}$ von stochastisch unabhängigen Zufallsvektoren ist. Es wird vorausgesetzt, dass der kleinste Eigenwert der Kovarianzmatrix von $\sqrt{N}\mathbf{B}$ größer als 0 ist. Seien γ_i die Eigenwerte von $\sqrt{N}\mathbf{B}$.

(V2) Es existiert eine Konstante g_0, sodass $\gamma_{min} = \min_i \gamma_i > g_0 > 0$.

Damit folgt unter dem Nachweis der Lindeberg-Bedingung dann die asymptotische multivariate Normalverteilung des Vektors nach dem Zentralen Grenzwertsatz.
Der Schätzer für die Kovarianzmatrix $\mathbf{V}_N = \mathrm{Var}(\sqrt{N}(\widehat{\mathbf{p}} - \mathbf{p}))$ wird in Brunner *et al.* (2002) mithilfe der Asymptotischen Rangtransformation hergeleitet. Hier sollen nur

die Resultate und deren Übertragung auf diagnostische Studien dargestellt werden. Dafür werden zusätzlich zu den Rängen $R_{ik}^{(l)}$ der Beobachtungen innerhalb einer Reader-Methoden-Kombination auch die Intern-Ränge $R_{ik}^{(i,l)}$ der Beobachtungen $X_{ik}^{(l)}$ innerhalb der einzelnen Gesundheitszustände (Gruppen) und Reader-Methoden-Kombinationen benötigt. Bezeichne im Folgenden $Z_{ik}^{(l)} = R_{ik}^{(l)} - R_{ik}^{(i,l)}$ die Differenz der Ränge und Internränge für jeden Patienten, welche zu einem Vektor $\mathbf{Z}_{ik} = (Z_{ik}^{(1)}, ..., Z_{ik}^{(MR)})'$ zusammengefasst werden können. Der Vektor der arithmetischen Mittelwerte wird mit $\overline{\mathbf{Z}}_{i\cdot} = 1/n_i \sum_{k=1}^{n_i} \mathbf{Z}_{ik}$ bezeichnet. Einen konsistenten Schätzer $\widehat{\mathbf{V}}_N$ für $\mathbf{V}_N$ erhält man mit $\widehat{\mathbf{V}}_N = \widehat{\mathbf{V}}_{N,0} + \widehat{\mathbf{V}}_{N,1}$, wobei gilt

$$\widehat{\mathbf{V}}_{N,i} = \frac{N}{(N-n_i)^2 n_i} \widehat{\mathbf{S}}_i, \ i = 0, 1$$

und

$$\widehat{\mathbf{S}}_i = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (\mathbf{Z}_{ik} - \overline{\mathbf{Z}}_{i\cdot})(\mathbf{Z}_{ik} - \overline{\mathbf{Z}}_{i\cdot})'$$

die empirische Kovarianzmatrix der $\mathbf{Z}_{i1}, ..., \mathbf{Z}_{in_i}$ bezeichnet. Die empirische Varianz von $\sqrt{N}(\widehat{\mathbf{p}} - \mathbf{p})$ ist eine Linearkombination der Einzelvarianzen der Gesunden und Kranken. Da die Annahmen über die Verteilungsfunktionen nicht die Gleichheit der Varianzen in den einzelnen Gruppen beinhalteten, können diese Varianzen - auch unter Hypothese - verschieden sein. Das wird durch die Struktur von $\widehat{\mathbf{V}}_N$ berücksichtigt.

3.2 Modell 2: Mehrfachmessungen

3.2.1 Modell und Notation

Im Modell 2 betrachten wir den ersten Schritt zur Clusterbildung der Daten. Die Patienten liefern nun nicht mehr nur einen Beobachtungspunkt, sondern mehrere. Allerdings wird nach wie vor angenommen, dass diese Beobachtungseinheiten alle den gleichen Gesundheitszustand haben. Es gibt also Beobachtungsvektoren $X_{iks}^{(l)} \sim F_i^{(l)}$ mit $i = 0, 1$ und $k = 1, ..., n_i$ und $s = 1, ..., m_{ik}$. Insgesamt erhält man somit

$$N = m_0 + m_1 = \sum_{k=1}^{n_0} m_{0k} + \sum_{k=1}^{n_1} m_{1k}$$

Beobachtungseinheiten an insgesamt $n = n_0 + n_1$ Patienten. Die Größen m_0 und m_1 sind dementsprechend die Gesamtanzahlen an gesunden bzw. kranken Beobachtungseinheiten. Die Beobachtungen sind unabhängig für verschiedene k, aber abhängig für verschiedene s bei gleichem k. Jede dieser N Beobachtungseinheiten wird insgesamt d-mal gemessen.

Der Index l soll auch hier anzeigen, unter welcher Reader-Methoden-Kombination die Beobachtung erhoben wurde. Wir nehmen an, dass jeder Reader mit jeder Methode jeden Patienten untersucht hat, deshalb ist m_{ik} identisch für alle l und es bedarf deshalb auch keines weiteren Indexes.
Damit die asymptotischen Resultate hergeleitet werden können, muss zusätzlich zu Annahme (A1) auch folgende Annahme (A2) gemacht werden.

(V3) Die bivariate Verteilung der $(X_{iks}^{(l)}, X_{iks'}^{(j)})$ hängt nicht von s, s' und k ab.

(A2) $m_{ik} \leq M_0 < \infty,\ i = 0, 1, k = 1, ..., n_i$ für alle N d.h. die Anzahl der Beobachtungen pro Person muss gleichmäßig beschränkt sein.

Diese Annahmen sind für die Praxis nicht restriktiv, weil sie nur sicherstellen, dass die Anzahl der Gesunden und Kranken nicht zu weit voneinander entfernt ist. Einzige weitere Annahme an die Verteilungsfunktionen ist der Ausschluss von Ein-Punkt-Verteilungen, um positive Varianzschätzer zu erhalten.

3.2.2 Schätzer für die Accuracy

Die Accuracy wird auch im Modell 2 über die Fläche unter der ROC-Kurve geschätzt. Dieses kann wiederum über das Integral der Verteilungsfunktionen definiert werden:

$$\text{AUC} = p = \int F_0 dF_1.$$

Der Unterschied zum Modell 1 wird bei den empirischen Versionen der Verteilungsfunktionen deutlich werden. Die empirischen Verteilungsfunktionen $\widehat{F}_i^{(l)}(x)$ können für jede Reader-Methoden-Kombination l aufgestellt werden.

$$\widehat{F}_i(x)^{(l)} = \frac{1}{n_i} \sum_{k=1}^{n_i} \frac{1}{m_{ik}} \sum_{s=1}^{m_{ik}} c(x - X_{iks}^{(l)}).$$

Hier muss berücksichtigt werden, dass mehrere Beobachtungen an einem Patienten erhoben worden sind. Dies geschieht durch einen ungewichteten Mittelwert der Patientenmittelwerte. Diese Art der Wichtung gibt also jedem Patienten das gleiche Gewicht, unabhängig davon, wieviele Beobachtungen er liefert. In der Literatur findet man verschiedene Ansätze zur Gewichtung, aber Datta & Satten (2005) schreiben, dass die hier verwendete besser ist, wenn die Einzelmessungen an einem Patienten korreliert sind. Das führt dazu, dass einzelne Patienten mit sehr vielen Beobachtungen das Ergebnis nicht dominieren können.
Um auch in diesem Modell die übliche Rangdarstellung des Schätzers für $\mathbf{p}$ zu erhalten, bedarf es allerdings einiger Vorüberlegungen. Dafür muss zunächst die mittlere Verteilungsfunktion definiert werden:

$$H(x) = \frac{1}{N} \left(m_0 F_0(x) + m_1 F_1(x) \right). \tag{3.1}$$

Die Umstellung ist in folgendem Lemma festgehalten.

Lemma 3.1. *Es gilt* $p = q_1 - q_0 + \frac{1}{2}$ *mit* $q_i = \int H dF_i$ *und* $p = \int F_0 dF_1$.

Beweis:

$$\begin{aligned} p = \int F_0 dF_1 &= \frac{m_0}{N} \int F_0 dF_1 + \frac{m_1}{N} \int F_0 dF_1 \\ &= \frac{m_0}{N} \int F_0 dF_1 + \frac{m_1}{N} - \frac{m_1}{N} \int F_1 dF_0 \\ &= \int H dF_1 - \int H dF_0 + \frac{m_1}{N} - \frac{m_1}{N} \int F_1 dF_1 + \frac{m_0}{N} \int F_0 dF_0 \\ &= \int H dF_1 - \int H dF_0 + \frac{N}{2N} = q_1 - q_0 + \frac{1}{2} \quad \square \end{aligned}$$

Hierbei bezeichnen q_1, q_0 die relativen Effekte, wenn man F_1 bzw. F_0 relativ zu der mittleren Verteilungsfunktion H betrachtet, im Gegensatz zu den relativen Effekten p, die die Verteilung der Kranken nur zur Verteilung der Gesunden in Relation setzen. Die Verbindung zu den Rängen liefert dann die empirische Version der mittleren Verteilungsfunktion $H(x)$:

$$\widehat{H}(x) = \frac{1}{N} \sum_{i=0}^{1} \sum_{k=1}^{n_i} \sum_{s=1}^{m_{ik}} c(x - X_{iks}), \tag{3.2}$$

wobei $c(x)$ die mittlere Zählfunktion bezeichnet. Die Auswertung von $\widehat{H}$ für eine beliebige Zufallsvariable X_{iks} liefert schon fast den Rang dieser Beobachtung unter allen N Beobachtungen. Die bisherigen Überlegungen gelten wieder für alle Reader-Methoden-Kombinationen, deshalb wird nun der Index l wieder verwendet. Bezeichne $R_{iks}^{(l)}$ den Rang der Beobachtung $X_{iks}^{(l)}$ unter allen N Beobachtungen mit der Reader-Methoden-Kombination l. Dann gilt:

$$\widehat{H}^{(l)}(X_{iks}^{(l)}) = \frac{1}{N}(R_{iks}^{(l)} - \frac{1}{2}).$$

Für die Schätzung der Accuracy mithilfe der Ränge muss nun auch der Rangmittelwert analog zu den empirischen Verteilungsfunktionen neu definiert werden. Aus der Definition der empirischen Verteilung folgt, dass der ungewichtete Mittelwert der Rangmittelwerte $\overline{R}_{ik\cdot}^{(l)}$ pro Person und Reader-Methoden-Kombination verwendet werden muss. Definiere also:

$$\overline{R}_{i\cdot\cdot}^{(l)} = \frac{1}{n_i} \sum_{k=1}^{n_i} \overline{R}_{ik\cdot}^{(l)} = \frac{1}{n_i} \sum_{k=1}^{n_i} \frac{1}{m_{ik}} \sum_{s=1}^{m_{ik}} R_{iks}^{(l)}.$$

Wenn man diese Überlegungen verwendet und in (3.1) die Verteilungsfunktionen durch ihre empirischen Pendants ersetzt, so erhält man für $i = 0, 1$ und für jede Reader-Methoden-Kombination l

$$\begin{aligned}\widehat{q}_i^{(l)} = \int \widehat{H}^{(l)}(x) d\widehat{F}_i^{(l)}(x) &= \frac{1}{n_i}\sum_{k=1}^{n_i}\frac{1}{m_{ik}}\sum_{s=1}^{m_{ik}}\widehat{H}^{(l)}(X_{iks}^{(l)}) \\ &= \frac{1}{n_i}\sum_{k=1}^{n_i}\frac{1}{m_{ik}}\sum_{s=1}^{m_{ik}}\frac{1}{N}(R_{iks}^{(l)} - \frac{1}{2}) = \frac{1}{N}\overline{R}_{i\cdot\cdot}^{(l)} - \frac{1}{2N}\end{aligned}$$

Zusammengesetzt ergibt das für den Schätzer der Accuracy für eine Kombination l von Reader und Methode:

$$\widehat{p}_l = \frac{1}{N}(\overline{R}_{1\cdot\cdot}^{(l)} - \overline{R}_{0\cdot\cdot}^{(l)}) + \frac{1}{2}.$$

An dieser Stelle muss angemerkt werden, dass der naive Schätzer, den man erhält, wenn man in $\int F_0 dF_1$ direkt die empirischen Verteilungsfunktionen einsetzt (wie im Modell 1), nicht zur Differenz der ungewichteten Rangmittelwerte führt. Betrachtet man den folgenden Ausdruck (der Index l für die Reader-Methoden-Kombination wurde hier weggelassen)

$$\begin{aligned}\int \widehat{F}_0 d\widehat{F}_1 &= \frac{1}{n_1}\sum_{k=1}^{n_1}\frac{1}{m_{1k}}\sum_{s=1}^{m_{1k}}\widehat{F}_0(X_{1ks}) \\ &= \frac{1}{n_1}\sum_{k=1}^{n_1}\frac{1}{m_{1k}}\sum_{s=1}^{m_{1k}}\frac{1}{n_0}\sum_{l=1}^{n_0}\frac{1}{m_{0l}}\sum_{t=1}^{m_{0l}} c(X_{1ks} - X_{0lt}),\end{aligned}$$

so ist zu sehen, dass dieser Mittelwert nicht in die gewohnten Ränge transformierbar ist.
Die Schätzer für alle Reader und Methoden können, wie vorher, in einem Vektor angeordnet werden:

$$\widehat{\mathbf{p}} = (\widehat{p}_{11}, ..., \widehat{p}_{1R}, ..., \widehat{p}_{M1}, ..., \widehat{p}_{MR})' = (\widehat{p}_1, ..., \widehat{p}_d)'.$$

Dieser Schätzer ist konsistent, aber im Gegensatz zum Schätzer im Modell 1 nur asymptotisch erwartungstreu. Das liegt daran, dass verschiedene m_{ik} bei den einzelnen Patienten auftreten können. Für den Beweis der Konsistenz wird zunächst gezeigt, dass die empirischen Verteilungsfunktionen beschränkte zweite Momente haben.

Lemma 3.2. *Unter Voraussetzung (A1) und (A2) gilt für alle N und für alle $i = 0, 1$ und $k = 1, ..., n_i$ und $s = 1, ..., m_{ik}$*

1. $E(\widehat{H}(x) - H(x))^2 \leq \frac{M_0}{N}$.

2. $E(\widehat{H}(X_{iks}) - H(X_{iks}))^2 \leq \frac{M_0}{N}$.

Beweis:
Der zweite Teil der Aussage folgt aus dem ersten mithilfe von Fubinis Theorem. Deshalb wird hier nur erste Teil gezeigt.

$$\begin{aligned}
&E(\widehat{H}(x) - H(x))^2 = \\
&= \frac{1}{N^2} \sum_{i=0}^{1} \sum_{i'=0}^{1} \sum_{k=1}^{n_i} \sum_{k'=1}^{n_{i'}} \sum_{s=1}^{m_{ik}} \sum_{s'=1}^{m_{i'k'}} E\left[(c(x - X_{iks}) - F_i(x))(c(x - X_{i'k's'}) - F_{i'}(x))\right] \\
&\leq \frac{1}{N^2} \sum_{i=0}^{1} \sum_{k=1}^{n_i} \sum_{s,s'=1}^{m_{ik}} E\left[(c(x - X_{iks}) - F_i(x))(c(x - X_{iks'}) - F_i(x))\right] \\
&\leq \frac{1}{N^2} \sum_{i=0}^{1} \sum_{k=1}^{n_i} \sum_{s,s'=1}^{m_{ik}} 1 \leq \frac{1}{N^2} \sum_{i=0}^{1} \sum_{k=1}^{n_i} m_{ik}^2 \\
&\leq \frac{1}{N^2} \sum_{i=0}^{1} \sum_{k=1}^{n_i} m_{ik} M_0 = \frac{M_0}{N},
\end{aligned}$$

wobei im ersten Schritt verwendet wird, dass die beiden Terme für verschiedene i, k unabhängig sind und der Erwartungswert der einzelnen Terme $= 0$ ist. Damit ist die Behauptung gezeigt. □
Mithilfe des Lemmas kann nun die Konsistenz des hergeleiteten Schätzers $\widehat{\mathbf{p}}$ für die Accuracy gezeigt werden.

Satz 3.3. *Unter Voraussetzung (A1) und (A2) gilt:* $\widehat{\mathbf{p}}$ *konvergiert in* L_2*-Norm gegen* $\mathbf{p}$ *mit der Rate* $\frac{1}{\sqrt{N}}$.

Beweis:
Es reicht, das Resultat getrennt für jedes Element $p^{(l)}$ von $\mathbf{p}$ zu zeigen. Für die bessere Lesbarkeit wird der Index l im Folgenden weggelassen.
Aus der L_2-Konvergenz von $\hat{q}_1, \hat{q}_0$ folgt direkt die L_2-Konvergenz von $\widehat{p}$:

$$\begin{aligned}
E(\widehat{p} - p)^2 &= E(\hat{q}_1 - \hat{q}_0 - (q_1 - q_0))^2 \\
&\leq 2E(\hat{q}_1 - q_1)^2 + 2E(\hat{q}_0 - q_0)^2.
\end{aligned}$$

Dementsprechend reicht es aus, die Behauptung für $q_i, i = 0, 1$ zu zeigen. Mit der Jensen- und der c_r-Ungleichung folgt für $i = 0, 1$

$$\begin{aligned}
&E(\widehat{q}_i - q_i)^2 \\
&= \left(\int \widehat{H} d\widehat{F}_i - \int H dF_i \right)^2 \\
&= E\left(\int (\widehat{H} - H) d\widehat{F}_i + \int H d(\widehat{F}_i - F_i) \right)^2
\end{aligned}$$

$$
\begin{aligned}
&\leq 2E\left(\int(\widehat{H}-H)d\widehat{F}_i\right)^2 + 2E\left(\int H d(\widehat{F}_i - F_i)\right)^2 \\
&\leq 2E\int(\widehat{H}-H)^2 d\widehat{F}_i \\
&\quad + \frac{2}{n_i^2}\sum_{k,k'=1}^{n_1}\frac{1}{m_{ik}m_{ik'}}\sum_{s,s'=1}^{m_{ik},m_{ik'}} E(H(X_{iks})-q_i)(H(X_{ik's'})-q_i) \\
&= \frac{2}{n_i}\sum_{k=1}^{n_i}\frac{1}{m_{ik}}\sum_{s=1}^{m_{ik}} E(\widehat{H}(X_{iks})-H(X_{iks}))^2 \\
&\quad + \frac{2}{n_i^2}\sum_{k=1}^{n_i}\frac{1}{m_{ik}^2}\sum_{s,s'=1}^{m_{ik}} E(H(X_{iks})-q_i)(H(X_{iks'})-q_i) \\
&\leq \frac{2}{n_i}\sum_{k=1}^{n_i}\frac{M_=}{N} + \frac{2}{n_i^2}\sum_{k=1}^{n_i}1 = 2\left(\frac{M_0}{N}+\frac{1}{n_i}\right) \leq \frac{2(M_0+N_0)}{N} \longrightarrow 0,
\end{aligned}
$$

wobei Lemma 3.2 im letzten Schritt verwendet wird. □

Der voranstehende Satz zeigt, dass der Schätzer $\widehat{\mathbf{p}}$ zumindest asymptotisch unverzerrt ist. Alternativ wäre die Schätzung der AUC mithilfe der gewichteten Rangmittelwerte denkbar. Die Diskussion um die Unterschiede und Vorzüge von gewichteten und ungewichteten Schätzern wird aber ausführlich für das nächsthöhere Modell 3 geführt und soll an dieser Stelle deshalb nur kurz erwähnt werden.

3.2.3 Verteilung des Schätzers

Im folgenden Abschnitt soll die asymptotische Äquivalenz des Schätzers für die Accuracy zu einer Summe von unabhängigen Zufallsvariablen gezeigt werden. Auch hier kann man dann die asymptotische multivariate Normalverteilung des Vektors nachweisen.

Theorem 3.4. *Der Vektor* $\sqrt{N}(\widehat{\mathbf{p}}-\mathbf{p})$ *ist asymptotisch äquivalent zu einem Vektor* $\sqrt{N}\mathbf{B}_N = \sqrt{N}(B_N^{(1)},...,B_N^{(d)})$*, dessen Komponenten*

$$
\begin{aligned}
B_N^{(l)} &= \int F_0^{(l)} d\widehat{F}_1^{(l)} - \int F_1^{(l)} d\widehat{F}_0^{(l)} + 1 - 2p^{(l)} \qquad (3.3)\\
&= \frac{1}{n_1}\sum_{k=1}^{n_1}\frac{1}{m_{1k}}\sum_{s=1}^{m_{1k}} F_0^{(l)}(X_{1ks}^{(l)}) - \frac{1}{n_0}\sum_{l=1}^{n_0}\frac{1}{m_{0l}}\sum_{t=1}^{m_{0l}} F_1^{(l)}(X_{0lt}^{(l)}) + 1 - 2p^{(l)}
\end{aligned}
$$

Summen von unabhängigen Zufallsvariablen sind.

Beweis: Die Darstellung als Summe von unabhängigen Zufallsvariablen sieht man durch einfache Umstellung:

$$B_N^{(l)} = \left(\sum_{i=0}^{1} \sum_{k=1}^{n_i} \zeta_{ik} \right) + 1 - 2p^{(l)}$$

wobei

$$\zeta_{ik} = \frac{2i-1}{n_i} \frac{1}{m_{ik}} \sum_{s=1}^{m_{ik}} F_{i^*}^{(l)}(X_{iks}^{(l)}).$$

Hierbei steht i^* für 0 (bzw. 1), wenn $i = 1$ (bzw. $i = 0$) ist. Da der Beweis der asymptotischen Äquivalenz sehr technisch ist, wurde er in den Anhang verschoben (siehe S.84). □

Aus diesem Theorem folgt auch im Modell 2 die asymptotische multivariate Normalverteilung des Vektors $\sqrt{N}(\widehat{\mathbf{p}} - \mathbf{p})$ unter der Lindeberg-Bedingung, wenn (V1) vorausgesetzt wird. Die Schätzung der Kovarianzmatrix $\mathbf{V}_N = \mathrm{Var}(\sqrt{N}(\widehat{\mathbf{p}} - \mathbf{p}))$ erfolgt ähnlich wie im Modell 1. Zunächst werden dafür die asymptotischen Rangtransformationen (ART) definiert:

$$\begin{aligned} Y_{0ks}^{(l)} &= F_1^{(l)}(X_{0ks}^{(l)}) \\ Y_{1ks}^{(l)} &= F_0^{(l)}(X_{1ks}^{(l)}) \end{aligned}$$

Diese entsprechen genau den Zufallsvariablen aus der asymptotischen Äquivalenz. Wären diese Zufallsvariablen beobachtbar, so könnte man mit deren empirischen Kovarianzmatrizen die Matrix $\mathbf{V}_N$ approximieren. Da sie aber nicht beobachtbar sind, müssen sie durch Schätzer ersetzt werden. Im Modell 1 konnte man eine Darstellung mithilfe von Rängen und Internrängen erhalten. Die Darstellung der Ränge im Modell 2 mit der empirischen Verteilungsfunktion $\widehat{H}^{(l)}$ wurde bereits gezeigt, es fehlen also noch die Internränge. Hierfür kann die ungewichtete empirische Verteilungsfunktion $\widehat{F}_i^{(l)}$ nicht verwendet werden. Die Ränge innerhalb eines Gesundheitszustandes können hingegen mit der folgenden gewichteten empirischen Verteilungsfunktion

$$\widetilde{F}_i^{(l)}(x) = \frac{1}{m_i} \sum_{k=1}^{n_i} \sum_{s=1}^{m_{ik}} c(x - X_{iks}^{(l)}) \tag{3.4}$$

berechnet werden. Man sieht leicht, dass sich diese gewichteten Mittelwerte genau zu $\widehat{H}$ addieren:

$$\widehat{H}^{(l)}(x) = \frac{1}{N}(m_0 \widetilde{F}_0^{(l)}(x) + m_1 \widetilde{F}_1^{(l)}(x)).$$

Dadurch verändert sich die Notation in den folgenden Definitionen. Die Elemente der Vektoren $\mathbf{Z}$, die die Differenzen der Ränge und Internränge enthalten, müssen

folgendermaßen umdefiniert werden:

$$\begin{aligned} Z_{0ks}^{(l)} &= m_1 \widetilde{F}_1^{(l)}(X_{0ks}) = R_{0ks}^{(l)} - R_{0ks}^{(0,l)} \\ Z_{1ks}^{(l)} &= m_0 \widetilde{F}_0^{(l)}(X_{1ks}) = R_{1ks}^{(l)} - R_{1ks}^{(1,l)}. \end{aligned}$$

Außerdem werden die folgenden Mittelwerte definiert:

$$\begin{aligned} \overline{Z}_{ik\cdot}^{(l)} &= \overline{R}_{ik\cdot}^{(l)} - \overline{R}_{ik\cdot}^{(i,l)} = \frac{1}{m_{ik}} \sum_{s=1}^{m_{ik}} (R_{iks}^{(l)} - R_{iks}^{(i,l)}) \\ \overline{Z}_{i\cdot\cdot}^{(l)} &= \frac{1}{n_i} \sum_{k=1}^{n_i} \overline{Z}_{ik\cdot}^{(l)} = \frac{1}{n_i} \sum_{k=1}^{n_i} \frac{1}{m_{ik}} \sum_{s=1}^{m_{ik}} (R_{iks}^{(l)} - R_{iks}^{(i,l)}). \end{aligned}$$

Dann erhält man auch im Modell 2 den Schätzer $\widehat{\mathbf{V}}_N = \widehat{\mathbf{V}}_{N,0} + \widehat{\mathbf{V}}_{N,1}$ durch Addition der beiden Einzelschätzer innerhalb der Kranken und der Gesunden. Es muss allerdings beachtet werden, dass $N = \sum_{i,k} m_{ik}$ die Summe aller Beobachtungseinheiten an den n_0 Gesunden und n_1 Kranken Patienten ist und nicht mehr nur die Anzahl der Patienten. Der Schätzer setzt sich wie im vorherigen Kapitel aus den empirischen Kovarianzmatrizen der $\overline{\mathbf{Z}}_{i1\cdot}, ..., \overline{\mathbf{Z}}_{in_i\cdot}$ zusammen:

$$\widehat{\mathbf{V}}_{N,i} = \frac{N}{(N - m_i)^2 n_i} \widehat{\mathbf{S}}_i, \ i = 0, 1,$$

wobei gilt:

$$\widehat{\mathbf{S}}_i = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (\overline{\mathbf{Z}}_{ik\cdot} - \overline{\mathbf{Z}}_{i\cdot\cdot})(\overline{\mathbf{Z}}_{ik\cdot} - \overline{\mathbf{Z}}_{i\cdot\cdot})'.$$

Theorem 3.5. *Der Schätzer $\widehat{\mathbf{V}}_N$ ist konsistent für $\mathbf{V}_N = \mathrm{Var}(\sqrt{N}(\hat{\mathbf{p}} - \mathbf{p}))$ in dem Sinne, dass*

$$E(\widehat{v}(l, l') - v(l, l'))^2 \to 0 \text{ für } N \to \infty \ (l, l' = 1, ..., d),$$

wobei $\widehat{v}(l, l')$ und $v(l, l')$ die Einträge der Matrizen $\widehat{\mathbf{V}}_N$ und $\mathbf{V}_N$ darstellen.

Beweis: siehe Appendix (Seite 86).

Somit sind auch für Modell 2 alle notwendigen Schätzer und deren Verteilungen hergeleitet.

3.3 Modell 3: Clustered Data

3.3.1 Modell und Notation

In diesem Kapitel wird nun schließlich ein Verfahren für clustered data hergeleitet, also für den Fall, dass an einem Patienten sowohl gesunde als auch erkrankte Beobachtungseinheiten vorhanden sind. Sei n die Anzahl der beobachteten Patienten bzw.

Cluster. Wir nehmen an, dass an jedem dieser Cluster k jeweils m_{0k} gesunde und m_{1k} erkrankte Beobachtungeinheiten vorliegen. Jede einzelne Beobachtungseinheit wird auch manchmal mit *ROI=Region of Interest* abgekürzt. Dies ist die kleinste Einheit im Cluster. Insgesamt liefert also jeder Cluster $m_{0k} + m_{1k}$ Beobachtungseinheiten.
Wir bezeichnen mit $X_{iks}^{(l)} \sim F_i^{(l)}$ die s-te ($s = 1, ..., m_{ik}$) ROI von Patient k ($k = 1, ..., n$), welche entweder gesund ($i = 0$) oder erkrankt ($i = 1$) ist und mit Reader-Methoden-Kombination $l = 1, ..., d$ gemessen wurde. Die schematische Darstellung dieser Daten befindet sich auf Seite 15. Zwei Beobachtungseinheiten an verschiedenen Patienten können als unabhängig betrachtet werden, wohingegen bei zwei Beobachtungseinheiten an einem Patienten Abhängigkeiten zugelassen werden müssen, unabhängig vom wahren Gesundheitszustand der Beobachtungseinheit. Außerdem sind auch die mehrfachen Beurteilungen einer ROI möglicherweise abhängig.
Insgesamt gibt es also

$$N = m_0 + m_1 = \sum_{k=1}^{n} (m_{0k} + m_{1k})$$

Beobachtungseinheiten, die von n Patienten stammen. Die Größen m_0 und m_1 bezeichnen damit die Gesamtanzahlen an gesunden und kranken Beobachtungseinheiten. Im Folgenden sind n_0 und n_1 die Anzahlen der Patienten von den insgesamt n beobachteten, die mindestens eine gesunde ROI bzw. mindestens eine erkrankte ROI haben. Außerdem bezeichnet n_c die Anzahl der Patienten, die sowohl mindestens eine erkrankte als auch eine nichterkrankte ROI liefern. Weiterhin wird ein Index λ_{ik} eingeführt, der anzeigt, ob ein Patient k eine Beobachtung vom Gesundheitszustand i hat oder nicht:

$$\lambda_{ik} = \begin{cases} 1, & \text{Patient } k \text{ hat eine Beobachtung vom Zustand } i \\ 0, & \text{sonst} \end{cases}$$

Das Produkt jeweils zweier Indizes pro Patient $\lambda_{0k}\lambda_{1k}$ liefert die Information, ob dieser Patient Beobachtungen mit beiden Gesundheitszuständen hat. Der Unterschied zum Modell 2 besteht darin, dass zusätzlich die gesunden und kranken Beobachtungen abhängig sein können. Um aus diesem Modell das Modell 2 zu erhalten, müssen alle diese Produkte $= 0$ sein.

Annahmen

An die Marginal-Verteilungen F_i ($i = 0, 1$) der Daten wird keine spezielle Annahme gemacht. Ausgenommen sind lediglich Ein-Punkt-Verteilungen. Für die bivariate Verteilungen muss allerdings wieder folgende Annahme getroffen werden:

(V4) Die bivariate Verteilung der $(X_{iks}^{(l)}, X_{i'ks'}^{(j)})$ ist unabhängig von i, i', s, s' und k.

An die Stichprobenumfänge und hier speziell an die Clustergrößen müssen einige Anforderungen der Balanciertheit gestellt werden. Es müssen wiederum (A1) und (A2) gelten, wobei die Bedeutung von (A1) im Kontext der clustered data umformuliert wird:

(A1) $N \to \infty$, so dass $N/n_i \leq N_0 < \infty$, $i = 0, 1$ d.h. der Quotient der Anzahl der Beobachtungen und der Anzahl der Patienten mit Beobachtungen in einem Gesundheitszustand muss gleichmäßig beschränkt sein.

(A3) $N \to \infty$, so dass $N/n_c \leq N_C < \infty$, d.h. die Anzahl der Cluster mit nur einem Gesundheitszustand muss gleichmäßig beschränkt sein.

Die letzte Annahme ist in der Praxis sicherlich nicht immer gegeben. Wenn die Anzahl der verbundenen Messungen, also der Patienten mit gesunden und kranken ROIs verschwindend gering ist, ist es nicht mehr möglich, sinnvoll Kovarianzen zu schätzen. Es sollte hier bereits bei der Planung der Studie darauf geachtet werden, dass gar keine verbundenen Messungen auftreten.
Die Schätzung der Accuracy im Modell 3 wird auf zwei verschiedene Arten erfolgen: zunächst wird ein ungewichteter Schätzer vorgestellt, anschließend ein gewichteter. Die Begriffe „ungewichtet“ und „gewichtet“ beziehen sich auf das Gewicht, welches der einzelne Patient in der Auswertung erhält. Beim ungewichteten Schätzer bekommt jeder Patient das gleiche Gewicht, beim gewichteten Schätzer werden die Patienten mit der Anzahl ihrer Beobachtungen gewichtet.

3.3.2 Ein ungewichteter Schätzer

Der Schätzer für die Accuracy eines einzelnen Readers wird analog zu den Schätzern in den vorherigen Modellen hergeleitet. Es wird ein Schätzer für die Verteilungsfunktionen aufgestellt. Hier bildet man die empirische Verteilungsfunktion so, dass zunächst die Zählfunktionen innerhalb eines Patienten und eines Gesundheitsstatus gemittelt werden, danach wird über alle Patienten hinweg gemittelt. Bezeichne mit $\widehat{F}_i^{(l)}(x)$ die empirische Verteilungsfunktion von $X_{i11}^{(l)}, ..., X_{inm_{in}}^{(l)}$ $(i = 0, 1;\ l = 1, ..., d)$, die folgendermaßen definiert ist:

$$\widehat{F}_i^{(l)}(x) = \frac{1}{n_i} \sum_{k=1}^{n} \frac{\lambda_{ik}}{m_{ik}} \sum_{s=1}^{m_{ik}} c(x - X_{iks}^{(l)}).$$

Der Schätzer für die Fläche unter der ROC-Kurve lässt sich in diesem Modell nicht direkt über $\widehat{F}_0^{(l)}$ und $\widehat{F}_1^{(l)}$ darstellen, man benötigt außerdem die empirische Verteilungsfunktion $\widehat{H}^{(l)}$,

$$\widehat{H}^{(l)}(x) = \frac{1}{N} \sum_{i=0}^{1} \sum_{k=1}^{n_i} \sum_{s=1}^{m_{ik}} c(x - X_{iks}^{(l)}).$$

Damit erhält man eine Formulierung für der Ränge der Beobachtungen:

$$\widehat{H}^{(l)}(X_{iks}^{(l)}) = \frac{1}{N}(R_{iks}^{(l)} - \frac{1}{2}).$$

Den Schätzer für die Accuracy erhält man mithilfe dieser Rangdarstellung über die Differenz der relativen Effekte q_0, q_1 in den beiden Gruppen (siehe Lemma 3.1). Dann lautet der Schätzer in Rangdarstellung:

$$\widehat{p}_l \;=\; \frac{1}{N}\left[\overline{R}_{1..}^{(l)} - \overline{R}_{0..}^{(l)}\right] + \frac{1}{2}.$$

Hierbei ist

$$\overline{R}_{i..}^{(l)} = \frac{1}{n_i}\sum_{k=1}^{n} \lambda_{ik}\overline{R}_{ik.}^{(l)}$$

der ungewichtete Mittelwert der Ränge $R_{iks}^{(l)}$ der Beobachtungen $X_{iks}^{(l)}$. Diesen erhält man durch die Verwendung der ungewichteten empirischen Verteilungsfunktionen $\widehat{F}_i^{(l)}$. Der Vorteil des ungewichteten Schätzers liegt darin, dass er Patienten mit vielen Beobachtungen nicht stärker gewichtet als solche mit wenigen Beobachtungen. Für jede Reader-Methoden-Kombination l kann die AUC geschätzt werden und die Schätzer werden wie gewohnt in einem Vektor $\widehat{\mathbf{p}}$ zusammengefasst:

$$\widehat{\mathbf{p}} = (\widehat{p}_{11}, ..., \widehat{p}_{1R}, ..., \widehat{p}_{M1}, ..., \widehat{p}_{MR})'. \tag{3.5}$$

Satz 3.6. *Unter den Annahmen (A1)-(A3) gilt: Der in 3.5 definierte Schätzer $\widehat{\mathbf{p}}$ ist asymptotisch erwartungstreu und konsistent für $\mathbf{p}$.*

Beweis: Der Beweis kann ähnlich geführt werden, wie im Modell 2 (siehe Satz 3.3). Im eigentlichen Satz wurde die Unabhängigkeit von X_{0ks} und X_{1ks} für gleiche k nicht ausgenutzt. Allerdings muss die Abschätzung des Lemmas neu gezeigt werden. Es gilt aber auch hier:

$$\begin{aligned}
&E(\widehat{H}(x) - H(x))^2 \\
&= \frac{1}{N^2}\sum_{i=0}^{1}\sum_{i'=0}^{1}\sum_{k=1}^{n}\sum_{k'=1}^{n}\sum_{s=1}^{m_{ik}}\sum_{s'=1}^{m_{i'k'}} \\
&\qquad \lambda_{ik}\lambda_{i'k'}E\left[(c(x - X_{iks}) - F_i(x))(c(x - X_{i'k's'}) - F_{i'}(x))\right] \\
&\leq \frac{1}{N^2}\sum_{i=0}^{1}\sum_{i'=0}^{1}\sum_{k=1}^{n}\sum_{s,s'=1}^{m_{ik}} \\
&\qquad \lambda_{ik}\lambda_{i'k}E\left[(c(x - X_{iks}) - F_i(x))(c(x - X_{i'ks'}) - F_{i'}(x))\right] \\
&\leq \frac{1}{N^2}\sum_{i,i'=0}^{1}\sum_{k=1}^{n}\sum_{s=1}^{m_{ik}}\sum_{s'=1}^{m_{i'k}}\lambda_{ik}\lambda_{i'k} \leq \frac{1}{N^2}\sum_{i,i'=0}^{1}\sum_{k=1}^{n} m_{ik}m_{i'k} \\
&= \frac{1}{N^2}\sum_{k=1}^{n} m_{\cdot k}m_{\cdot k} \leq \frac{1}{N^2}\sum_{k=1}^{n} 2M_0 m_{\cdot k} = \frac{2M_0}{N},
\end{aligned}$$

wobei im ersten Schritt diesmal nur verwendet wird, dass die beiden Terme für verschiedene k unabhängig sind. □
Wie zuvor ist es auch in diesem Modell möglich, die asymptotische Normalität des Schätzvektors nachzuweisen. Dafür wird zunächst die asymptotische Äquivalenz zu einem Vektor unabhängiger Summanden bewiesen.

Theorem 3.7. *Der Vektor* $\sqrt{N}(\widehat{\mathbf{p}}-\mathbf{p})$ *ist asymptotisch äquivalent zu einem Vektor* $\sqrt{N}\mathbf{B}_N = \sqrt{N}(B_N^{(1)},...,B_N^{(d)})$*, dessen Komponenten*

$$\begin{aligned} B_N^{(l)} &= \int F_0^{(l)} d\widehat{F}_1^{(l)} - \int F_1^{(l)} d\widehat{F}_0^{(l)} + 1 - 2p^{(l)} \qquad (3.6) \\ &= \frac{1}{n_1}\sum_{k=1}^{n} \frac{\lambda_{1k}}{m_{1k}} \sum_{s=1}^{m_{1k}} F_0^{(l)}(X_{1ks}^{(l)}) - \frac{1}{n_0}\sum_{l=1}^{n} \frac{\lambda_{0l}}{m_{0l}} \sum_{t=1}^{m_{0l}} F_1^{(l)}(X_{0lt}^{(l)}) + 1 - 2p^{(l)} \end{aligned}$$

Summen von unabhängigen Zufallsvariablen sind.

Beweis: Die Darstellung als Summe von unabhängigen Zufallsvariablen sieht man durch einfache Umstellung:

$$B_N^{(l)} = \left(\sum_{k=1}^{n} \zeta_k\right) + 1 - 2p^{(l)}$$

wobei

$$\zeta_k = \frac{\lambda_{1k}}{n_1 m_{1k}} \sum_{s=1}^{m_{1k}} F_0^{(l)}(X_{1ks}^{(l)}) - \frac{\lambda_{0k}}{n_0 m_{0k}} \sum_{s=1}^{m_{0k}} F_1^{(l)}(X_{0ks}^{(l)}).$$

Hauptunterschied zum Beweis von Theorem 3.4 ist die Einführung des Indexes λ_{ik}, der anzeigt, welcher Patient mindestens eine Beobachtung in Gruppe i hat. Dadurch wird der Beweis nicht schwieriger, nur komplizierter aufzuschreiben. Berücksichtigt man die Voraussetzung, dass nicht zu viele der Patienten nur einen Gesundheitszustand haben dürfen (A3), dann wird klar, dass der Beweis völlig analog zum Beweis von Theorem 3.4 im Modell 2 zu führen ist. □
Um die Verteilung der Schätzer zu bestimmen, ist die Annahme (V1) an die Kovarianzmatrix $\mathbf{V}_N$ bzw. deren Eigenwerte γ_i notwendig.

Korollar 3.8. *Unter der zusätzlichen Annahme (V1) gilt: Der Vektor* $\sqrt{N}(\widehat{\mathbf{p}}-\mathbf{p})$ *folgt asymptotisch einer multivariaten Standardnormalverteilung mit Kovarianzmatrix* $\mathbf{V}_N$*.*

Beweis: Die Behauptung folgt aus der asymptotischen Äquivalenz von $\sqrt{N}(\widehat{\mathbf{p}}-\mathbf{p})$ zu einer Summe von unabhängigen Zufallsvariablen und unter Nachweis der Lindeberg-Bedingung aus dem zentralen Grenzwertsatz. □

Die Kovarianzmatrix $\mathbf{V}_N = \text{Var}(\sqrt{N}(\widehat{\mathbf{p}} - \mathbf{p}))$ ist asymptotisch die gleiche wie die Kovarianzmatrix $\mathbf{V}_B = \text{Var}(\sqrt{N}\mathbf{B}_N)$. Wenn man einen konsistenten Schätzer für $\mathbf{V}_B$ herleitet, so dient dieser auch als Schätzer für $\mathbf{V}_N$. Dafür werden zunächst die asymptotischen Rangtransformationen (ART) eingeführt. Diese sind allerdings nichtbeobachtbare Zufallsvariablen. Sie müssen durch Rangschätzer ersetzt werden. Mithilfe dieser Schätzer kann dann der Schätzer für die Kovarianzmatrix $\mathbf{V}_B$ bestimmt werden. Dieses zweistufige Verfahren hat den Vorteil, dass alle Schätzer mithilfe von Rängen berechnet werden können.
Zunächst werden die asymptotischen Rangtransformationen bestimmt:

$$\begin{aligned} Y_{0ks}^{(l)} &= F_1^{(l)}(X_{0ks}^{(l)}) \\ Y_{1ks}^{(l)} &= F_0^{(l)}(X_{1ks}^{(l)}). \end{aligned}$$

Die Mittelwerte der ART innerhalb eines Patienten werden mit $\overline{Y}_{ik\cdot}^{(l)}$ bezeichnet und die ungewichteten Mittelwerte über alle Patienten mit $\overline{Y}_{i\cdot\cdot}^{(l)}$. Als nächstes betrachten wir zunächst zwei Komponenten l und j des Vektors $\mathbf{B}_N$ und berechnen deren Kovarianz:

$$\begin{aligned} & Cov(\sqrt{N}B_N^{(l)}, \sqrt{N}B_N^{(j)}) \\ &= NCov(\overline{Y}_{1\cdot\cdot}^{(l)} - \overline{Y}_{0\cdot\cdot}^{(l)}, \overline{Y}_{1\cdot\cdot}^{(j)} - \overline{Y}_{0\cdot\cdot}^{(j)}) \\ &= N\Big(Cov(\overline{Y}_{1\cdot\cdot}^{(l)}, \overline{Y}_{1\cdot\cdot}^{(j)}) + Cov(\overline{Y}_{0\cdot\cdot}^{(l)}, \overline{Y}_{0\cdot\cdot}^{(j)}) \\ &\quad -Cov(\overline{Y}_{1\cdot\cdot}^{(l)}, \overline{Y}_{0\cdot\cdot}^{(j)}) - Cov(\overline{Y}_{0\cdot\cdot}^{(l)}, \overline{Y}_{1\cdot\cdot}^{(j)})\Big) \\ &= \frac{N}{n_1^2}\sum_{k=1}^{n}\lambda_{1k}Cov(\overline{Y}_{1k\cdot}^{(l)}, \overline{Y}_{1k\cdot}^{(j)}) + \frac{N}{n_0^2}\sum_{k=1}^{n}\lambda_{0k}Cov(\overline{Y}_{0k\cdot}^{(l)}, \overline{Y}_{0k\cdot}^{(j)}) \\ &\quad -\frac{N}{n_0 n_1}\sum_{k=1}^{n}\lambda_{0k}\lambda_{1k}Cov(\overline{Y}_{1k\cdot}^{(l)}, \overline{Y}_{0k\cdot}^{(j)}) - \frac{N}{n_0 n_1}\sum_{k=1}^{n}\lambda_{0k}\lambda_{1k}Cov(\overline{Y}_{0k\cdot}^{(l)}, \overline{Y}_{1k\cdot}^{(j)}) \\ &= \frac{N}{n_1}s_{1,1}^{(l,j)} + \frac{N}{n_0}s_{0,0}^{(l,j)} - \frac{Nn_c}{n_0 n_1}c_{1,0}^{(l,j)} - \frac{Nn_c}{n_0 n_1}c_{0,1}^{(l,j)}. \end{aligned}$$

Die einzelnen $s_{i,i}^{(l,j)}$ und $c_{0,1}^{(l,j)}$ können in Matrizenform als $\mathbf{S}_0$,$\mathbf{S}_1$,$\mathbf{C}_0$ und $\mathbf{C}_1$ zusammengestellt werden. Wenn die Zufallsvariablen $Y_{iks}^{(l)}$ beobachtbar wären, wären die empirischen Kovarianzen von $\overline{Y}_{ik\cdot}^{(l)}$ und $\overline{Y}_{ik\cdot}^{(j)}$ natürliche Schätzer für $s_{i,i}^{(l,j)}$ $(i = 0, 1)$ und die empirischen Kovarianzen von $\overline{Y}_{0k\cdot}^{(l)}$ und $\overline{Y}_{1k\cdot}^{(j)}$ wären natürliche Schätzer für $c_{0,1}^{(l,j)}$ und $c_{1,0}^{(l,j)}$. Bezeichne die entsprechenden empirischen Größen mit $\tilde{s}_{i,i}^{(l,j)}$, $\tilde{c}_{0,1}^{(l,j)}$ und $\tilde{c}_{1,0}^{(l,j)}$:

$$\tilde{s}_{i,i}^{(l,j)} = \frac{1}{n_i - 1}\sum_{k=1}^{n}\lambda_{ik}(\overline{Y}_{ik\cdot}^{(l)} - \overline{Y}_{i\cdot\cdot}^{(l)})(\overline{Y}_{ik\cdot}^{(j)} - \overline{Y}_{i\cdot\cdot}^{(j)})$$

$$\begin{aligned}\tilde{c}_{0,1}^{(l,j)} &= \frac{1}{n_c-1}\sum_{k=1}^{n}\lambda_{0k}\lambda_{1k}(\overline{Y}_{0k\cdot}^{(l)}-\overline{Y}_{0\cdot\cdot}^{(l)})(\overline{Y}_{1k\cdot}^{(j)}-\overline{Y}_{1\cdot\cdot}^{(j)})\\ \tilde{c}_{1,0}^{(l,j)} &= \tilde{c}_{0,1}^{(j,l)}\end{aligned}$$

Die $\tilde{c}_{i,i'}^{(l,j)}$ stellen sicher, dass die Korrelationen zwischen gesunden und kranken Beobachtungen innerhalb eines Patienten berücksichtigt werden. $\tilde{c}_{1,0}^{(l,j)}$ schätzt zum Beispiel die Korrelation zwischen den Beurteilungen von Untersucher l für die kranken und den Beurteilungen von Untersucher j für die gesunden Beobachtungseinheiten innerhalb eines Clusters. Diese Schätzer kann man in vier Matrizen anordnen, entsprechend ihrer Reihenfolge im Vektor. Damit erhalten wir $\tilde{\mathbf{S}}_0$ aus $\tilde{s}_{0,0}^{(l,j)}$, $\tilde{\mathbf{S}}_1$ aus $\tilde{s}_{1,1}^{(l,j)}$, $\tilde{\mathbf{C}}_0$ aus $\tilde{c}_{0,1}^{(l,j)}$ und $\tilde{\mathbf{C}}_1 = \tilde{\mathbf{C}}_0'$:

$$\tilde{\mathbf{S}}_0 = \begin{pmatrix} \tilde{s}_{0,0}^{(1,1)} & \cdots & \tilde{s}_{0,0}^{(d,1)} \\ \vdots & \ddots & \vdots \\ \tilde{s}_{0,0}^{(1,d)} & \cdots & \tilde{s}_{0,0}^{(d,d)} \end{pmatrix} \tag{3.7}$$

Die anderen Matrizen entstehen analog.
Die $Y_{iks}^{(l)}$ sind jedoch, wie oben bereits erwähnt, nicht beobachtbar. Zur Schätzung wird eine weitere empirische Verteilungsfunktion

$$\widetilde{F}_i^{(l)}(x) = \frac{1}{m_i}\sum_{k=1}^{n}\lambda_{ik}\sum_{s=1}^{m_{ik}} c(x - X_{iks}^{(l)}) \tag{3.8}$$

benötigt. Die Auswertung dieser Funktion an einer Zufallsvariablen $X_{iks}^{(l)}$ liefert die sogenannten Internränge $R_{iks}^{(i,l)}$:

$$\widetilde{F}_i^{(l)}(X_{iks}^{(l)}) = \frac{1}{m_i}(R_{iks}^{(i,l)} - \frac{1}{2}).$$

Durch Ersetzen der unbekannten Verteilungen mit den empirischen Verteilungsfunktionen $\widetilde{F}_i^{(l)}$ erhalten wir Rangschätzer, die mit $Z_{iks}^{(l)}$ bezeichnet werden:

$$Z_{0ks}^{(l)} = \widetilde{F}_1^{(l)}(X_{0ks}^{(l)}) = \frac{1}{m_1}(R_{0ks}^{(l)} - R_{0ks}^{(0,l)}), \tag{3.9}$$

$$Z_{1ks}^{(l)} = \widetilde{F}_0^{(l)}(X_{1ks}^{(l)}) = \frac{1}{m_0}(R_{1ks}^{(l)} - R_{1ks}^{(1,l)}). \tag{3.10}$$

Hier bezeichnet $R_{iks}^{(l)}$ den Rang von Beobachtung $X_{iks}^{(l)}$ unter allen Beurteilungen in der Reader-Methoden-Kombination l. $R_{iks}^{(i,l)}$ ist der Internrang von Beobachtung $X_{iks}^{(l)}$

in Gruppe i *und* unter allen Beurteilungen in der Reader-Methoden-Kombination l. Die Mittelwerte innerhalb eines Patienten und Gesundheitszustands

$$\overline{Z}_{0k.}^{(l)} = \frac{1}{m_{0k}} \sum_{s=1}^{m_{0k}} \widetilde{F}_1^{(l)}(X_{0ks}^{(l)}) \quad = \quad \frac{1}{m_1}(\overline{R}_{0k.}^{(l)} - \overline{R}_{0k.}^{(0,l)}), \tag{3.11}$$

$$\overline{Z}_{1k.}^{(l)} = \frac{1}{m_{1k}} \sum_{s=1}^{m_{1k}} \widetilde{F}_0^{(l)}(X_{1ks}^{(l)}) \quad = \quad \frac{1}{m_0}(\overline{R}_{1k.}^{(l)} - \overline{R}_{1k.}^{(1,l)}), \tag{3.12}$$

werden in Vektoren $\overline{\mathbf{Z}}_{ik.} = (\overline{Z}_{ik.}^{(1)}, ..., \overline{Z}_{ik.}^{(d)})'$ zusammengestellt. Diese existieren immer dann, wenn Patient k eine Beobachtung vom Gesundheitszustand i hat. Sei außerdem $\overline{\mathbf{Z}}_{i..} = (\overline{Z}_{i..}^{(1)}, ..., \overline{Z}_{i..}^{(d)})'$ der Vektor der arithmetischen Mittelwerte der $\overline{Z}_{ik.}^{(l)}$. Diese Rangschätzer sind nun beobachtbare Größen, die anstelle der ARTs in den Schätzern verwendet werden können. Zur Notation ist zu bemerken, dass diese zum Modell 2 leicht verändert ist, die Vektoren Z_{iks} waren im vorherigen Modell mit dem jeweils anderen Stichprobenumfang normiert. Diese Notation würde hier aber zu umständlichen Ausdrücken führen.
Weiterhin kann man Schätzer für die jeweiligen Varianzkomponenten $\mathbf{S}_0$,$\mathbf{S}_1$,$\mathbf{C}_0$ und $\mathbf{C}_1$ angeben, indem die nichtbeobachtbaren Zufallsvariablen durch die Rangschätzer ersetzt werden:

$$\begin{aligned}
\widehat{\mathbf{S}}_0 &= \frac{1}{n_0 - 1} \sum_{k=1}^{n} \lambda_{0k}(\overline{\mathbf{Z}}_{0k.} - \overline{\mathbf{Z}}_{0..})(\overline{\mathbf{Z}}_{0k.} - \overline{\mathbf{Z}}_{0..})' \\
\widehat{\mathbf{S}}_1 &= \frac{1}{n_1 - 1} \sum_{k=1}^{n} \lambda_{1k}(\overline{\mathbf{Z}}_{1k.} - \overline{\mathbf{Z}}_{1..})(\overline{\mathbf{Z}}_{1k.} - \overline{\mathbf{Z}}_{1..})' \\
\widehat{\mathbf{C}}_0 &= \frac{1}{n_c - 1} \sum_{k=1}^{n} \lambda_{0k}\lambda_{1k}(\overline{\mathbf{Z}}_{0k.} - \overline{\mathbf{Z}}_{0..})(\overline{\mathbf{Z}}_{1k.} - \overline{\mathbf{Z}}_{1..})'.
\end{aligned}$$

Aus (3.7) folgt dann, dass die Summe dieser Matrizen ein Schätzer für die Kovarianzmatrix $\mathbf{V}_B$ und somit auch für $\mathbf{V}_N$ ist:

$$\widehat{\mathbf{V}}_N = \frac{N\widehat{\mathbf{S}}_0}{n_0} + \frac{N\widehat{\mathbf{S}}_1}{n_1} - \frac{Nn_c(\widehat{\mathbf{C}}_0 + \widehat{\mathbf{C}}_0')}{n_0 n_1}. \tag{3.13}$$

Dass diese Definition einen konsistenten Schätzer liefert, ist Aussage des folgenden Satzes.

Satz 3.9. *Der Schätzer $\widehat{\mathbf{V}}_N$ ist konsistenter Schätzer für $\mathbf{V}_N = Cov(\sqrt{N}(\hat{\mathbf{p}} - \mathbf{p}))$ in dem Sinne, dass*

$$E(\widehat{v}(l,l') - v(l,l'))^2 \to 0 \text{ für } N \to \infty \ (l,l' = 1, ..., d),$$

wobei $\widehat{v}(l,l')$ und $v(l,l')$ die Einträge der Matrizen $\widehat{\mathbf{V}}_N$ und $\mathbf{V}_N$ darstellen.

Der Beweis kann komponentenweise geführt werden, denn die Summe konsistenter Schätzer bildet wieder einen konsistenten Schätzer. Für Details sei an dieser Stelle auf den Appendix (Seite 90) verwiesen, aber die Konsistenz der Schätzer für die Kovarianzmatrizen innerhalb einer Gruppe i kann direkt aus Modell 2 übernommen werden.

Korrekturmöglichkeiten für die Kovarianzmatrix

Paarige Beobachtungen liegen immer dann vor, wenn $\lambda_{0k}\lambda_{1k} \neq 0$ ist. Die Qualität der Schätzung der Kovarianz zwischen gesunden und erkrankten Beobachtungen hängt davon ab, wieviele paarige Beobachtungen zur Schätzung verwendet werden können. Es kann somit vorkommen, dass die geschätzte Matrix nicht positiv semidefinit ist. Dies ist genau dann der Fall, wenn die Schätzung der Teilmatrizen $\mathbf{C}$ so schlecht wird, dass das Bilden der Differenz in Gleichung 3.13 zu einer Matrix mit negativen Eigenwerten führt. Das Fehlen eines Gesundheitszustandes bei einem Patienten wird im Folgenden als „fehlender Wert“ bezeichnet werden. Auf echte missing values, also fehlende Bewertungen einzelner Reader oder mit einzelnen Methoden soll an dieser Stelle nicht näher eingegangen werden. In der Darstellung von Seite 15 entsprechen die fehlenden Werte den leeren Zellen in der Tabelle in den unteren Zeilen, echte missing values würden dargestellt durch verschieden viele x in den unterschiedlichen Wiederholungen der Zellen pro Reader oder Methode.
Stanish *et al.* (1978) haben ein Verfahren zur Korrektur der Analyse longitudinaler Daten mit fehlenden Werten vorgeschlagen, das unter bestimmten Voraussetzungen anwendbar ist. Hierfür muss der Mechanismus, der bestimmt, ob ein Wert fehlt, stochastisch unabhängig von den tatsächlich beobachteten Werten sein. Anschaulich heißt das, dass die Höhe der Messwerte nicht mit dem Fehlen der Werte zusammenhängen darf. Es ist häufig schwierig, diese Voraussetzung rechnerisch nachzuweisen. Deshalb muss man diskutieren, ob es sinnvoll ist, diese Annahme zu machen. Wenn zum Beispiel bei einer Untersuchung der Zähne die kranken Beobachtungen nicht mehr gemacht werden können, weil die meisten kranken Zähne bereits ausgefallen sind, so ist der „Fehl-Mechanismus“ sicherlich nicht unabhängig vom Messwert. Hier kann das Verfahren also nicht angewendet werden. Wenn dagegen alle Herzkranzgefäße auf einen Verschluss untersucht werden, so ist das Fehlen gesunder oder kranker Gefäße bei einem Patienten unabhängig vom Grad des Verschlusses.
Eine weitere Methode, das Problem nicht positiv semidefiniter Kovarianzmatrizen zu lösen, bieten Rousseeuw & Molenberghs (1993) in ihrer Arbeit über die Transformation einer Korrelationsmatrix. Sie bestimmen mit Optimierungsprozessen eine Matrix, die positiv semidefinit und elementweise so nah wie möglich an der Originalmatrix ist.

3.3.3 Ein gewichteter Schätzer

Eine andere Möglichkeit, die Accuracy zu schätzen, ist die Verwendung gewichteter Schätzer. Hierbei wird der Rangmittelwert innerhalb eines Clusters - also an einem Patienten - mit der Anzahl der Beobachtungen an diesem Cluster gewichtet. Das resultiert in einem Schätzer für die Accuracy, den man auch erhielte, wenn alle Beobachtungen unabhängig wären. Es bekommt also jede Beobachtung das gleiche Gewicht und jede Person wird dementsprechend mit der Anzahl ihrer Beobachtungen gewichtet. Hierbei muss man sicherstellen, dass jede weitere Beobachtung auch wirklich weitere Informationen liefert und nicht einfach nur die Präzision der Messung erhöht.
Es gibt auch andere Methoden, den Schätzer zu gewichten. Donner & Klar (2000) schreiben allerdings: *„Optimal weighting in sense of smaller variation of the estimator could be achieved by weighting with the inverse of the cluster-variance. But since this variance is unknown, weighting by sample-size is a well-known stable and robust alternative“*. Der ungewichtete Schätzer ist ein Spezialfall einer Gewichtung: hier sind die Gewichte für alle Cluster identisch.
Der gewichtete Schätzer lautet:

$$\widetilde{\mathrm{AUC}} = \widetilde{p} = \frac{1}{N}\left(\widetilde{R}_{1\cdot\cdot} - \widetilde{R}_{0\cdot\cdot}\right) + \frac{1}{2} \quad \text{mit} \quad \widetilde{R}_{i\cdot\cdot} = \frac{1}{m_i}\sum_{k=1}^{n}\sum_{s=1}^{m_{ik}} R_{iks}$$

Für diesen Schätzer muss auch eine andere Gewichtung der geschätzten Kovarianzmatrix verwendet werden. Wenn man keine Informationen über die Struktur der Kovarianzmatrix einfließen lässt, so erhält man einen Schätzer analog zum Schätzer im ungewichteten Fall. Der einzige Unterschied sind die Gewichtungen mit den Stichprobenumfängen.
Die asymptotischen Rangtransformationen sind beim gewichteten Schätzer genauso definiert wie beim ungewichteten. Die Unterschiede treten erst durch die Bildung der Mittelwerte über die Ränge auf. Die empirischen Verteilungsfunktionen $\widehat{F}_i$ aus der Herleitung der ungewichteten Schätzer werden allerdings nicht verwendet, hier werden die $\widetilde{F}_i$ (siehe (3.4) S. 25) benötigt. Deshalb werden die empirischen Vektoren, die man nun erhält, mit einem Index w für den gewichteten Schätzer bezeichnet. Man definiert mithilfe der Ränge und Internränge wie vorher:

$$\begin{aligned} Z^{(l)}_{0ks} &= \widetilde{F}^{(l)}_1(X^{(l)}_{0ks}) = \frac{1}{m_1}(R^{(l)}_{0ks} - R^{(0,l)}_{0ks}) \\ Z^{(l)}_{1ks} &= \widetilde{F}^{(l)}_0(X^{(l)}_{1ks}) = \frac{1}{m_0}(R^{(l)}_{1ks} - R^{(1,l)}_{1ks}) \end{aligned}$$

Der Unterschied zum vorher betrachteten ungewichteten Fall tritt dann im nächsten Schritt auf: hier wird ein Vektor der Rangsummen und nicht der Rangmittelwerte

pro Patient betrachtet und diese werden mit einem Index w gekennzeichnet ($Z_{ik\cdot}^{(l)w}$):

$$\begin{aligned} Z_{0k\cdot}^{(l)w} &= \frac{1}{m_1}(R_{0k\cdot}^{(l)} - R_{0k\cdot}^{(0,l)}) = \frac{m_{0k}}{m_1}(\overline{R}_{0k\cdot}^{(l)} - \overline{R}_{0k\cdot}^{(0,l)}) \\ Z_{1k\cdot}^{(l)w} &= \frac{1}{m_0}(R_{1k\cdot}^{(l)} - R_{1k\cdot}^{(1,l)}) = \frac{m_{1k}}{m_0}(\overline{R}_{1k\cdot}^{(l)} - \overline{R}_{1k\cdot}^{(1,l)}). \end{aligned}$$

Die Schätzer $Z_{ik\cdot}^{(l)w}$ werden in einem Vektor $\mathbf{Z}_{ik\cdot}^{w} = (Z_{ik\cdot}^{(1)w}, ..., Z_{ik\cdot}^{(d)w})'$ angeordnet. Der Mittelwertsvektor über alle Patienten sei mit $\mathbf{Z}_{i\cdot\cdot}^{w} = (Z_{i\cdot\cdot}^{(1)w}, ..., Z_{i\cdot\cdot}^{(d)w})'$ bezeichnet. Als Schätzer für die Elemente der Kovarianzmatrix lassen sich dann die folgenden Matrizen aufstellen:

$$\begin{aligned} \widehat{\mathbf{S}}_0^w &= \frac{1}{n_0 - 1}\sum_{k=1}^{n} \lambda_{0k}(\overline{\mathbf{Z}}_{0k\cdot}^{w} - \overline{\mathbf{Z}}_{0\cdot\cdot}^{w})(\overline{\mathbf{Z}}_{0k\cdot}^{w} - \overline{\mathbf{Z}}_{0\cdot\cdot}^{w})' \\ \widehat{\mathbf{S}}_1^w &= \frac{1}{n_1 - 1}\sum_{k=1}^{n} \lambda_{1k}(\overline{\mathbf{Z}}_{1k\cdot}^{w} - \overline{\mathbf{Z}}_{1\cdot\cdot}^{w})(\overline{\mathbf{Z}}_{1k\cdot}^{w} - \overline{\mathbf{Z}}_{1\cdot\cdot}^{w})' \\ \widehat{\mathbf{C}}_0^w &= \frac{1}{n_c - 1}\sum_{k=1}^{n} \lambda_{0k}\lambda_{1k}(\overline{\mathbf{Z}}_{0k\cdot}^{w} - \overline{\mathbf{Z}}_{0\cdot\cdot}^{w})(\overline{\mathbf{Z}}_{1k\cdot}^{w} - \overline{\mathbf{Z}}_{1\cdot\cdot}^{w})' \\ \widehat{\mathbf{C}}_1^w &= \frac{1}{n_c - 1}\sum_{k=1}^{n} \lambda_{0k}\lambda_{1k}(\overline{\mathbf{Z}}_{1k\cdot}^{w} - \overline{\mathbf{Z}}_{1\cdot\cdot}^{w})(\overline{\mathbf{Z}}_{0k\cdot}^{w} - \overline{\mathbf{Z}}_{0\cdot\cdot}^{w})'. \end{aligned}$$

Abschließend kann man die Komponenten auch wieder zu einem konsistenten Schätzer zusammenfügen.

$$\widehat{\mathbf{V}}_N^w = \frac{Nn_0\widehat{\mathbf{S}}_0^w}{m_0^2} + \frac{Nn_1\widehat{\mathbf{S}}_1^w}{m_1^2} - \frac{Nn_c\widehat{\mathbf{C}}_0^w}{m_0m_1} - \frac{Nn_c\widehat{\mathbf{C}}_1^w}{m_0m_1} \tag{3.14}$$

Alle Beweise für den gewichteten Schätzer können analog zu den Beweisen für den ungewichteten Schätzer hergeleitet werden. Da der Schwerpunkt dieser Arbeit aber auf den ungewichteten Schätzern liegt, soll das hier nicht ausgeführt werden.

3.3.4 Vergleich der beiden Schätzer

Es ist offensichtlich, dass der gewichtete und der ungewichtete Schätzer im Fall gleicher Clustergrößen identische Ergebnisse liefern. Für unterschiedliche Clustergrößen soll im Folgenden die Güte der verschiedenen Schätzer verglichen werden.
Die Effizienz (im Sinne geringerer Varianzen) der beiden Schätzer kann nicht ausgerechnet werden. Man kann aber einen Anhaltspunkt für deren Verhältnis zueinander bekommen, wenn man ein simples Modell annimmt: man betrachte zwei unabhängige Vektoren $\mathbf{X}_1$ und $\mathbf{X}_2$, deren Komponenten jeweils der gleichen compound

symmetry Struktur folgen. Bestimmt man nun die Varianzen des gewichteten und ungewichteten Mittelwerts dieser beiden Vektoren, so erhält man folgendes Resultat. Die Varianz des gewichteten Mittelwerts kann nur dann kleiner sein als die des ungewichteten, wenn die Korrelation innerhalb eines Vektors kleiner ist als $(m_1+m_2)/(m_1+m_2+2m_1m_2)$. Diese Behauptung ist im folgenden Satz zusammengefasst.

Proposition 3.10. *Sei $m_1 \neq m_2$. Betrachte zwei unabhängige Vektoren $\mathbf{X}_1 = (X_{11}, ..., X_{1m_1})'$ und $\mathbf{X}_2 = (X_{21}, ..., X_{2m_2})'$. Die Komponenten sind paarweise korreliert. Die Kovarianzmatrix weist also folgende compound symmetry Struktur auf : $(\sigma - c)\mathbf{I}_{m_i} + c\mathbf{J}_{m_i}$. Damit ergibt sich die Korrelation zwischen zwei Elementen eines Vektors als $\rho = \frac{\sigma}{c}$.*
Die Varianz des gewichteten Mittelwerts der beiden Vektoren $\mathbf{X}_1$ und $\mathbf{X}_2$

$$G = \frac{1}{m_1 + m_2}(m_1\overline{X}_{1\cdot} + m_2\overline{X}_{2\cdot})$$

ist genau dann kleiner als die des ungewichteten Mittelwerts

$$U = \frac{1}{2}(\overline{X}_{1\cdot} + \overline{X}_{2\cdot}),$$

wenn gilt:

$$\rho < \frac{m_1 + m_2}{m_1 + m_2 + 2m_1m_2}.$$

Beweis: siehe Appendix (Seite 91).

Für praktische Zwecke heißt das, dass der ungewichtete Schätzer dem gewichteten vorzuziehen ist, sobald von einer Korrelation zwischen den einzelnen Komponenten auszugehen ist. Außerdem sieht man, dass mit steigender Clustergröße diese Beziehung noch strenger wird, d.h. je größer der Cluster, desto geringer darf die Korrelation sein, um einen gewichteten Schätzer mit einer geringeren Varianz zu erhalten.
In Abbildung 3.1 wurde für verschiedene AUC, Clustergesamtgrößen (von oben nach unten von 2 bis 4) und Korrelationen simuliert, wie groß die Varianz der beiden Schätzer ist. Unter Clustergesamtgröße hat man hier die Summe aus gesunden und kranken ROIs an einem Patienten zu verstehen. Die tatsächliche Anzahl der Beobachtungen an einem Patienten zu einem Gesundheitszustand ist hier natürlich immer noch verschieden, sie wird für jeden Patienten zufällig bestimmt. Wie man sieht, ist die Variabilität des ungewichteten Schätzers (durchgezogene Linie) bereits ab einer moderaten Korrelation unterhalb der Variabilität des gewichteten Schätzers (gestrichelte Linie). Außerdem ist zu beobachten, dass die Schwankung mit der Clustergröße geringer wird und die Variabilität des ungewichteten Schätzers im Mittel kleiner ist, als die des gewichteten.

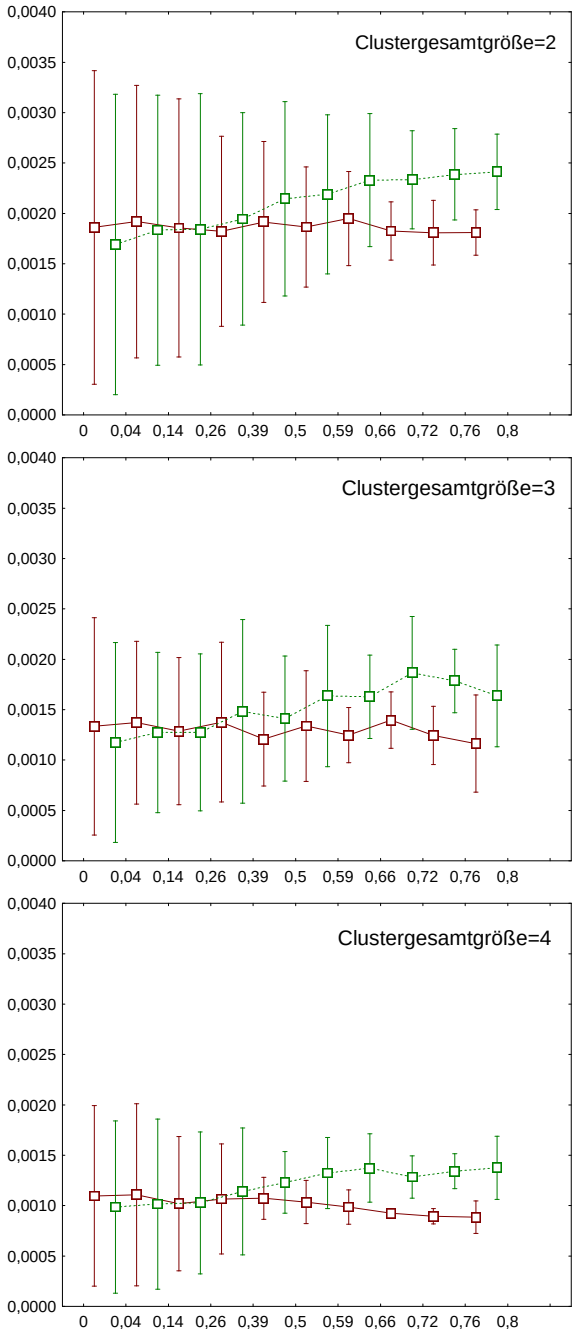

Abbildung 3.1: Empirische Standardabweichung der AUC-Schätzer (Mittelwert ± Standardabweichung) bei Stichprobenumfang $n = 50$ und variierender AUC (die ganze Bandbreite, Verschiebungen von 0.5 bis 2) in Abhängigkeit von der Anzahl der Wiederholungen (Clustergesamtgröße $= m_{0k}+m_{1k}$) und der Clusterkorrelation für den ungewichteten (durchgezogene, rote Linie) und den gewichteten (gestrichelte, grüne Linie) Schätzer.

Darstellung der ROC-Kurve

Die graphische Darstellung der ROC-Kurve von abhängigen Daten (also im Modell 2 und Modell 3) ist nicht klar definiert. Es bieten sich zwei Möglichkeiten:

1. die Erstellung der ROC-Kurve aus allen Einzelbeobachtungen,

2. die Verwendung patientenbezogener Daten, d.h. der Rangmittelwerte pro Patient, zur Erstellung einer ROC-Kurve.

Diese beiden Möglichkeiten haben unterschiedliche Eigenschaften und Vorzüge. Verwendet man die erste Darstellung, so ist es möglich, aus dem Graphen einen optimalen Cut-off mithilfe des Youden-Index zu bestimmen. In dieser Graphik entspricht jeder Punkt einem tatsächlich gemessenen und messbaren Wert. Diese Eigenschaft ist besonders bei ordinalen Messwerten wichtig.
Allerdings entspricht die Fläche unter der ersten Kurve genau der Fläche unter der Kurve, wie sie definiert ist, wenn man den gewichteten Schätzer für die Accuracy verwendet: jede Beobachtung erhält hier die gleiche Gewichtung, d.h. jeder Patient wird mit der Anzahl der Beobachtungen gewichtet. Will man also die Kurve zeichnen, die dem Schätzer für die Accuracy mit einem ungewichteten Mittelwert entspricht, so ist es besser, die Rangmittelwerte der einzelnen Patienten aufzuzeichnen. Das Bilden der Rangmittelwerte speziell bei ordinalen Messwerten führt allerdings dazu, dass die zu zeichnende Kurve mehr Punkte enthält, als die erste. Hier entspricht dann nicht mehr jeder Punkt einem ursprünglichen Score.
Wie man sich leicht vorstellen kann, sind die Unterschiede zwischen den beiden Graphen umso größer, je größer die einzelnen Cluster sind und je mehr die Beobachtungen innerhalb eines Clusters schwanken. Exemplarisch sind hier die beiden verschiedenen Kurven für einen Score von 1-5 bei Clustergrößen zwischen 1 und 3 angegeben. Man sieht deutlich, dass die zweite Kurve mehr Sprungstellen hat.

3.4 Hypothesen

Die Hypothesen in diagnostischen Studien können entsprechend der Hypothesen im linearen Modell aufgestellt werden. Im Standarddesign mit mehreren Readern und mehreren Methoden sind die Haupteffekte für Reader und Methode und die Wechselwirkung zwischen Reader und Methode von Interesse. Diese Hypothesen können mit Hilfe von Kontrastmatrizen formuliert werden, wenn die einzelnen Accuracies als Vektor dargestellt werden

$$\mathbf{p} = (p_{11}, ..., p_{MR}).$$

Die einzelnen Hypothesen werden dann wie folgt gebildet:

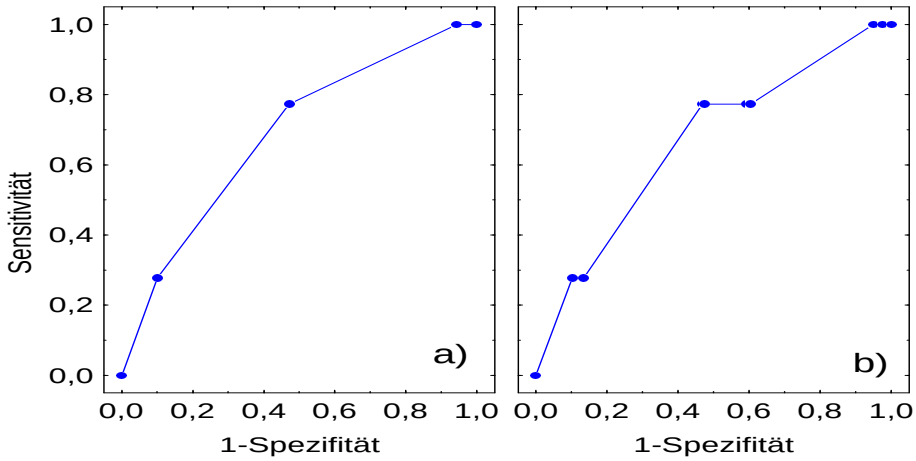

Abbildung 3.2: ROC-Kurven für (a) die Originalwerte und (b) die Rangmittelwerte

- **Methoden-Effekt**: Unterscheiden sich die Methoden in der Accuracy? Diese Frage wird mit der Hypothese $H_0^M : \overline{p}_{1\cdot} = ... = \overline{p}_{M\cdot}$ dargestellt. Die entsprechende Kontrastmatrix lautet dann $\mathbf{C}_M = \mathbf{P}_M \otimes (1/R)\mathbf{1}'_R$, wobei $\otimes$ das Kroneckerprodukt (s.S. 83) und $\mathbf{P}_M = \mathbf{I}_M - (1/M)\mathbf{J}_M$ bezeichnet. Die Hypothese kann dann also in Matrixform folgendermaßen umgeschrieben werden: $H_0^M : \mathbf{C}_M\mathbf{p} = \mathbf{0}$.

- **Reader-Effekt**: Sind die diagnostischen Fähigkeiten der einzelnen Reader gleich? Diese Frage entspricht der Hypothese: $H_0^R : \overline{p}_{\cdot 1} = ... = \overline{p}_{\cdot R}$, welche mithilfe der Matrix $\mathbf{C}_R = (1/M)\mathbf{1}'_M \otimes \mathbf{P}_R$, mit $\mathbf{P}_R = \mathbf{I}_R - (1/R)\mathbf{J}_R$, umgeschrieben werden kann als $H_0^R : \mathbf{C}_R\mathbf{p} = \mathbf{0}$.

- **Wechselwirkung**: Sind die Beurteilungen der einzelnen Reader homogen über die Modalitäten hinweg? Die entsprechende Hypothese lautet dann $H_0^{MR} : p_{mr} = \overline{p}_{m\cdot} + \overline{p}_{\cdot r} - \overline{p}_{\cdot\cdot}$. Sie wird in Matrixnotation dargestellt als $H_0^{MR} : \mathbf{C}_{MR}\mathbf{p} = \mathbf{0}$, wobei $\mathbf{C}_{MR} = \mathbf{P}_M \otimes \mathbf{P}_R$.

Falls die einzelnen Faktoren, Reader und Modalität, noch eine weitere Struktur tragen (also z.B. die Reader in zwei Gruppen nach Berufserfahrung eingeteilt werden können oder die Modalitäten in zwei verschiedene Gerätegruppen aufgeteilt sind), so kann dies modelliert werden, indem man die Kontrastmatrizen entsprechend anpasst. Angenommen, Reader 1-3 haben eine lange Berufserfahrung und Reader 4-8 sind noch Anfänger im Bereich der bildgebenden Diagnostik. Dies entspricht in der Theorie der linearen Modelle einem hierarchischen Modell: die Reader sind unter dem Faktor Berufserfahrung verschachtelt. Interessant ist hier also nur der Vergleich der Berufserfahrung. Eine entsprechende Kontrastmatrix $\mathbf{C}_G = (1/M)\mathbf{1}'_M \otimes \mathbf{e}_G$ mit

$\mathbf{e}_G = (1/3)\mathbf{1}_3 \oplus (-1/5)\mathbf{1}_5$ kann aufgestellt werden, mithilfe derer die Hypothese „kein Unterschied in den Gruppen“ getestet werden kann.

3.5 Test-Statistik

Um Hypothesen der Form $\mathbf{Cp} = \mathbf{0}$, wobei $\mathbf{C}$ eine geeignete Kontrastmatrix ist, im faktoriellen Modell zu testen, gibt es zwei verschiedene Teststatistiken, basierend auf quadratischen Formen: die Wald-Typ Statistik (WTS) und die ANOVA-Typ Statistik (ATS). Die WTS wurde von DeLong *et al.* (1988) für unabhängige (also nicht in Clustern vorliegende) Beobachtungen vorgeschlagen und wäre die logische Erweiterung des Zweistichprobenvergleichs, der von Obuchowski (1997) für clustered data vorgeschlagen wurde.
Unter der Voraussetzung, dass $\mathbf{V}_N \to \mathbf{V} \neq \mathbf{0}$ für $N \to \infty$, sodass $r(\mathbf{CV}_N) = r(\mathbf{CV})$, hat die WTS

$$Q_N = N\widehat{\mathbf{p}}'\mathbf{C}'[\mathbf{C}\widehat{\mathbf{V}}_N\mathbf{C}']^+\mathbf{C}\widehat{\mathbf{p}}$$

unter Hypothese asymptotisch eine $\chi^2_{r(\mathbf{CV})}$-Verteilung, wobei das $^+$ für die Moore-Penrose-Inverse der Matrix steht. Da diese Statistik bei kleinen Stichproben und insbesondere für eine hohe Zahl an Faktorstufen zu liberalen Entscheidungen führt, betrachten wir außerdem die ATS, die für Modell 1 bereits in Kaufmann *et al.* (2005) vorgeschlagen wurde. Diese erhält man aus der WTS durch Ersetzen des Varianzschätzers mit einer bekannten Matrix (in diesem Fall der Einheitsmatrix). Die Statistik wird mithilfe der Projektormatrix $\mathbf{T} = \mathbf{C}'[\mathbf{CC}']^+\mathbf{C}$ berechnet, welche aus den entsprechenden Kontrastmatrizen $\mathbf{C}$ bestimmt wird. Unter H_0 ist die ATS

$$F_N = \frac{N}{tr(\mathbf{T}\widehat{\mathbf{V}}_N)}\ \widehat{\mathbf{p}}'\ \mathbf{T}\ \widehat{\mathbf{p}}$$

approximativ χ^2_f/f-verteilt, wobei der Freiheitsgrad f durch

$$\widehat{f} = \frac{\left[tr(\mathbf{T}\hat{\mathbf{V}}_N)\right]^2}{tr(\mathbf{T}\hat{\mathbf{V}}_N\mathbf{T}\hat{\mathbf{V}}_N)}$$

geschätzt wird. Die Approximation der Verteilung geht auf Box (1954) zurück, der die Idee, die Verteilung durch Gleichsetzen der ersten zwei Momente der jeweiligen Verteilungen zu bestimmen, ausgearbeitet und formalisiert hat. Die Verwendung dieser Teststatistik in der Nichtparametrik geht auf Brunner *et al.* (1997) zurück, speziell im Bereich faktorieller Designs mit hohen Faktorstufenanzahlen. Denn gerade in dieser Situation tendiert die WTS zu sehr liberalen Entscheidungen. Zahlreiche Studien haben bereits die Überlegenheit der ATS gegenüber der WTS bei kleinen Stichproben und vielen Faktorstufen gezeigt. Bei vergleichbarer (und teilweise sogar besserer) Power hält sie das Niveau besser ein. In der bisherigen Literatur wurden

die Simulationen nur unter den Hypothesen $F_0 = F_1$ oder $p = 1/2$ durchgeführt. Bei Diagnosestudien interessieren aber viel mehr die Hypothesen der Form $p_1 = p_2$, und das in einem Bereich, in dem p_i erheblich größer ist als $1/2$, denn nur dann repräsentiert das p_i einen sinnvollen diagnostischen Test.

3.6 Konfidenzintervalle

Zur graphischen Darstellung und Einordnung der Relevanz statistisch signifikanter Ergebnisse sind Konfidenzintervalle ein unerlässliches Hilfsmittel. Deshalb soll im Folgenden gezeigt werden, welche Möglichkeiten der Darstellung sich in Diagnosestudien anbieten.

Individuelle Intervalle

Zunächst einmal kann man Konfidenzintervalle für jede einzelne Accuracy bestimmen. Diese folgen direkt aus der asymptotischen Normalverteilung des Vektors $\sqrt{N}(\widehat{\mathbf{p}} - \mathbf{p})$. Mithilfe der Verteilung kann ein $(1-\alpha)$-Konfidenzintervall folgendermaßen formuliert werden:

$$\widehat{p}_U^{(l)} = \widehat{p}_l - \widehat{\sigma}_l \sqrt{N} u_{1-\alpha/2} \quad \text{und} \quad \widehat{p}_O^{(l)} = \widehat{p}_l + \widehat{\sigma}_l \sqrt{N} u_{1-\alpha/2}. \tag{3.15}$$

Hier bezeichnet $\widehat{\sigma}_l^2$ das l-te Diagonalelement der geschätzten Kovarianzmatrix $\widehat{\mathbf{V}}_N$ und $u_{1-\alpha/2}$ das $(1-\alpha/2)$-Quantil der Normalverteilung. Mit Hilfe dieser Konfidenzintervalle lässt sich eine mögliche Reader-Methoden-Interaktion grafisch darstellen. Außerdem ist man mit den Konfidenzintervallen in der Lage, die Größe und somit die medizinische Relevanz der einzelnen Effekte zu beurteilen und zu interpretieren.

Mittlere Intervalle für die Methoden

Wenn keine Interaktion vorliegt, ist es sinnvoll, gemittelt über die einzelnen Reader Konfidenzintervalle für jede Methode zu bestimmen. Die Herleitung dieser Intervalle ist mit entsprechenden Kontrastvektoren möglich. Wir können aus der asymptotischen Äquivalenz für einen beliebigen Kontrastvektor $\mathbf{c}$ die Verteilung von $\sqrt{N}\mathbf{c}'(\widehat{\mathbf{p}} - \mathbf{p})$ herleiten. Diese Statistik ist asymptotisch normalverteilt mit Erwartungswert 0 und Varianz $\mathbf{c}'\mathbf{V}_N\mathbf{c}$. Wenn man nun $\mathbf{c}$ so wählt, dass es über die einzelnen Reader innerhalb einer Modalität mittelt, dann erhält man über die gleiche Konstruktion wie oben m verschiedene Konfidenzintervalle für die einzelnen Modalitäten. Diese entsprechen dem *mean reader*, der noch nicht von allen Zulassungsbehörden bei der Auswertung neuer Diagnostika akzeptiert wird. Wenn man mit $\mathbf{e}_m = (0, .., 1, .., 0)'$ einen Vektor bezeichnet, der an der m-ten Stelle eine 1 hat und ansonsten gleich 0 ist, so liefert die Anwendung des Kontrastvektors

$\mathbf{c}_m = \mathbf{e}_m \otimes (1/R)\mathbf{1}_R$ ein Konfidenzintervall für Modalität m, gemittelt über die R Untersucher:

$$\mathbf{c}_m'\widehat{\mathbf{p}} \pm (\mathbf{c}_m'\widehat{\mathbf{V}}_N\mathbf{c}_m \cdot u_{1-\alpha/2})/\sqrt{N}.$$

Intervalle für die Differenzen der Methoden

Zum Vergleich der Methoden untereinander ist es darüber hinaus möglich, Konfidenzintervalle für die paarweisen Differenzen - entweder die Differenzen der über die Reader gemittelten Accuracies oder für jeden Reader getrennt - anzugeben. Hier sollte man allerdings beachten, dass die paarweisen Intervalle, würde man sie mit den gleichen Quantilen aufstellen wie oben, insgesamt nur noch eine viel geringe Überdeckungswahrscheinlichkeit aufweisen als 95%. Deshalb ist es hier sinnvoll, eine Adjustierung der Konfidenzintervalle z.B. nach Bonferroni anzuwenden. Einen entsprechenden Kontrastvektor für die paarweisen Differenzen zweier Methoden i und j erhält man mithilfe eines Vektors $\mathbf{d}_{ij} = \mathbf{e}_i - \mathbf{e}_j$, der einen Vektor bezeichnet, der eine 1 an Position i, eine -1 an Position j und sonst den Wert 0 hat. Durch Multiplikation mit der entsprechenden zentrierenden Matrix erhält man den Vektor $\mathbf{c}_{ij} = \mathbf{d}_{ij} \otimes (1/R)\mathbf{1}_R$, mit dessen Hilfe man das gewünschte Konfidenzintervall erhält:

$$\mathbf{c}_{ij}'\widehat{\mathbf{p}} \pm (\mathbf{c}_{ij}'\widehat{\mathbf{V}}_N\mathbf{c}_{ij} \cdot u_{1-\alpha/2})/\sqrt{N}.$$

Es ist natürlich auch möglich, für beliebige andere Linearkombinationen Konfidenzintervalle aufzustellen. Es kann zum Beispiel von Interesse sein, zwei Untergruppen der Reder miteinander zu vergleichen. Hier kommt wieder der Kontrastvektor zur Anwendung, der die Differenz zweier Untergruppen von Readern bildet (siehe Abschnitt 3.4).

Probleme und Lösungen

Wenn die AUC sehr groß ist, der Test also eine sehr hohe Accuracy besitzt, so ist die Approximation dieser Konfidenzintervalle, die in ihrer Konstruktion auf Wald zurückgehen, sehr ungenau. Besonders dramatisch wird es, wenn die Kollektive durch den Test komplett getrennt werden, das heißt, wenn ein perfekter Test vorliegt. Dann brechen die Berechnungen zusammen. Studien haben gezeigt (Obuchowski & Lieber, 2002; Tsimikas *et al.* , 2002), dass die Überdeckungswahrscheinlichkeit in diesen Situationen schon beim einfachen Reader-Methoden-Vergleich des Modells 1 - insbesondere bei kleinen Stichprobenumfängen - extrem gering ist. Außerdem kann es vorkommen, dass die obere Grenze des Konfidenzintervalles die 1 übersteigt und damit über den Definitionsbereich des relativen Effekts bzw. der Accuracy hinausgeht. Eine Möglichkeit, um zumindest bereichserhaltende Konfidenzintervalle zu berechnen, bietet die sogenannte Delta-Methode, die auf Cramér (1928) zurückgeht und in van der Vaart (1998) ausführlich erläutert wird. Bei dieser Methode werden die

beschränkten Schätzer für die relativen Effekte zunächst mit einer geeigneten Transformation auf die gesamte reelle Achse $\mathbb{R}$ umgewandelt. Dann wird das asymptotische Konfidenzintervall in $\mathbb{R}$ aufgestellt und die so bestimmten Grenzen mit dem Inversen der Transformation wieder in den Definitionsbereich der relativen Effekte zurückgebracht. Eine in der Statistik häufig verwendete Transformation ist die *logit*-Funktion

$$logit(x) = log\left(\frac{x}{1-x}\right),$$

welche gute Eigenschaften hat. Wendet man die beschriebene Delta-Methode auf die Schätzer für die AUC an, so erhält man als transformierte Effekte

$$\widetilde{p}^{(l)} = \log[\widehat{p}^{(l)}/(1-\widehat{p}^{(l)})].$$

Die Konfidenzintervalle werden zunächst für diese transformierten Effekte aufgestellt:

$$\widehat{p}_U^{(l)} = \widetilde{p}^{(l)} - \frac{\widehat{\sigma}_l \cdot u_{1-\alpha/2}}{\widehat{p}^{(l)}(1-\widehat{p}^{(l)})\sqrt{N}} \tag{3.16}$$

$$\widehat{p}_O^{(l)} = \widetilde{p}^{(l)} + \frac{\widehat{\sigma}_l \cdot u_{1-\alpha/2}}{\widehat{p}^{(l)}(1-\widehat{p}^{(l)})\sqrt{N}}. \tag{3.17}$$

Mithilfe der inversen *logit*-Transformation erhält man dann wieder die Konfidenzintervalle für den ursprünglichen relativen Effekt:

$$\widehat{p}_{u*}^{(l)} = \frac{\exp(\widehat{p}_U^{(l)})}{1+\exp(\widehat{p}_U^{(l)})} \quad \text{und} \quad \widehat{p}_{o*}^{(l)} = \frac{\exp(\widehat{p}_O^{(l)})}{1+\exp(\widehat{p}_O^{(l)})}. \tag{3.18}$$

Die schlechte Approximation für hohe Accuracies kann mit der hier beschriebenen Delta-Methode aber nur teilweise und auch nur dann behoben werden, wenn die empirische Accuracy kleiner als 1 ist. Die *logit*-Funktion ist nämlich für $x \in \{0,1\}$ nicht definiert. Dies ist aber der Fall, wenn man einen perfekten Test vorliegen hat, d.h. wenn die empirischen Verteilungen der Gesunden und Kranken vollständig disjunkt sind. In dem Fall gibt es verschiedene Auswege: entweder addiert man eine kleine Konstante zum Wert dazu, sodass die Transformation funktioniert, oder man verwendet die „Rule of Three“, die auf Hanley & Lippman-Hand (1983) zurückgeht. Hier wurde gezeigt, dass die obere Grenze eines Konfidenzintervalles für eine Häufigkeit, deren Schätzer 0 ist, genau $3/n$ ist. Diese Methode ließe sich umkehren und würde dann in einer unteren Grenze $p_U^{HL} = 1 - \frac{3}{n}$ für den Wert 1 resultieren.
Vor allem im Falle eines perfekten Tests, der eine Accuracy von 1 hat, versagen die meisten Approximationsmethoden, da sie darauf basieren, die Verteilung der Accuracy zu bestimmen. Obuchowski & Lieber (2002) schlagen als Lösung des Problems die Verwendung von tabellierten Werten vor, in denen für kleine Stichprobenumfänge ($n < 30$) und verschiedene Situationen untere Konfidenzgrenzen angegeben sind. Sie

erwähnen aber auch, dass die unteren Grenzen stark von der Form der Kurve und dem Skalenniveau der erhobenen Daten abhängen.
Tsimikas *et al.* (2002) schlagen vor, sogenannte „profile likelihood“ Methoden anzuwenden. Hierbei wird eine Gridsearch verwendet, die rechnerisch sehr aufwändig ist. Die Autoren schlagen außerdem eine Anwendung für stratifizierte und für korrelierte Diagnostik-Studien vor, die Problematik der clustered data bleibt aber unerwähnt.
Bei kleinen Stichproben haben Obuchowski & Lieber (1998) festgestellt, dass es keine optimale Methode für die Konstruktion von Konfidenzintervallen gibt. In Abhängigkeit von der Skala der Messwerte und der Größe der AUC geben sie jeweils eine Empfehlung für das beste Verfahren an, sagen aber auch, dass für sehr große Flächen keines der Verfahren geeignet ist, die sie untersucht haben.
Abschließend ist also zur Problematik der Konfidenzintervalle festzustellen, dass es keine einheitliche Lösung gibt. Dennoch bieten aber bereits die vorgestellten Konfidenzintervalle bzw. die nach der Delta-Methode transformierten Intervalle durchaus die Möglichkeit, die Daten graphisch darzustellen.

4 Exkurs: Dichotome Testergebnisse

Einen besonderen Fall stellen dichotome Testergebnisse dar, das heißt, wenn der dem Test unterliegende Cut-off Mechaninismus bereits etabliert ist oder gar nicht existiert. Auch in diesem Fall ist die Anwendung der vorher beschriebenen Theorie möglich. In diesem Gebiet gibt es allerdings auch eine Vielzahl anderer Auswertungsmöglichkeiten, die im folgenden Kapitel kurz beschrieben und mit der hier vorgestellten Methode verglichen werden sollen.
Besonders in der Zahnmedizin tauchen sehr oft dichotome und clustered data auf, wenn mehrere oder alle Zähne eines Patienten untersucht werden. Die Diagnose in der Zahnmedizin ist dann häufig nur dichotom: kariös oder nicht kariös, paradontös oder nicht paradontös. Hier wurden diverse Verfahren zum Vergleich von Sensitivität und Spezifität entwickelt, unter anderem von Ahn (1997); Davies *et al.* (1997); Hujoel *et al.* (1990); Obuchowski (1998).
Zunächst werden im folgenden Kapitel häufig verwendete Schätzer für das Modell 1 im dichotomen Fall vorgestellt und mit dem Verfahren in der vorliegenden Arbeit verglichen. Es folgt eine Beschreibung der aktuellen Diskussion in der Literatur zur Kombination von Sensitivität und Spezifität. Für das Modell 3, also die clustered data, gibt es bisher nur wenige Ansätze, wie die Sensitivität und Spezifität gemeinsam ausgewertet werden können.

4.1 Schätzer für die Accuracy

4.1.1 Schätzer im Modell 1

Wir betrachten zunächst das Modell 1, d.h. den Fall, dass an jedem der N Patienten nur eine Messung vorgenommen wurde. Es gibt n_0 gesunde $(K-)$ und n_1 kranke $(K+)$ Patienten. Es können dann richtig positive (RP), richtig negative (RN), falsch positive (FP) und falsch negative (FN) Diagnosen getroffen werden. Die Vierfeldertafel für einen Reader oder eine Methode ist in Tabelle 4.1 dargestellt. Die Prävalenz der Krankheit im Kollektiv wird durch $\hat{P} = n_1/N$ geschätzt. Die einfachste Vorgehensweise bei dichotomen Daten ist der getrennte Vergleich von Sensitivität und Spezifität. Diese werden über die relativen Häufigkeiten geschätzt :

$$\widehat{\text{Sens}} = \frac{\#RP}{n_1} \text{ und } \widehat{\text{Spez}} = \frac{\#RN}{n_0}.$$

	$K+$	$K-$	
$T+$	$\#RP$	$\#FP$	
$T-$	$\#FN$	$\#RN$	
	n_1	n_0	N

Tabelle 4.1: Vierfeldertafel für ein dichotomes Testergebnis

Anschließend können auch Schätzer $\hat{V}_{Se}, \hat{V}_{Sp}$ für die Varianzen bestimmt werden:

$$\hat{V}_{Se} = \frac{\widehat{\text{Sens}}(1 - \widehat{\text{Sens}})}{n_1} = \frac{\#RP \cdot \#FN}{n_1}$$

und analog für die Spezifität. Mithilfe eines einfachen Wald-Tests werden Unterschiede getestet.
Häufig ist man aber an der gemeinsamen Analyse von Sensitivität und Spezifität, also eher einem globalen als einem geschichteten Verfahren, interessiert. Eine zunächst einfache Methode der Schätzung der Accuracy ist der Anteil der richtig diagnostizierten Personen am Gesamtkollektiv, die sogenannte Trefferquote (TQ). Sie wurde erstmals von Metz (1978) erwähnt. Der Schätzer lautet:

$$\widehat{TQ} = \frac{\#RP + \#RN}{N} = \frac{n_1\widehat{\text{Sens}} + n_0\widehat{\text{Spez}}}{n_0 + n_1} = \hat{P} \cdot \widehat{\text{Sens}} + (1 - \hat{P}) \cdot \widehat{\text{Spez}}. \tag{4.1}$$

Der Schätzer ist also das gewichtete Mittel aus geschätzter Sensitivität und Spezifität, wobei die Wichtung mit den relativen Stichprobenumfängen der einzelnen Gruppen erfolgt. Man sieht leicht, dass dieser Schätzer direkt von der Prävalenz der Krankheit im untersuchten Kollektiv abhängt. Es ist somit sehr einfach, diesen Wert zu manipulieren, indem man entweder sehr viele oder sehr wenige Kranke bzw. Gesunde rekrutiert.
Deshalb ist es sinnvoller, einen prävalenzunabhängigen Schätzer zu definieren. Die Fläche unter der ROC-Kurve erfüllt diese Eigenschaft, die sie auch im Fall stetiger oder ordinaler Testergebnisse zu einem guten analytischen Werkzeug macht. Betrachtet man die algebraische Herleitung des Schätzers für die Accuracy über die Trapezregel (siehe Abbildung 4.1), so kann man feststellen, dass diese AUC genau dem ungewichteten Mittelwert zwischen Sensitivität und Spezifität entspricht.

$$\widehat{AUC} = \underbrace{\widehat{\text{Sens}} \cdot \widehat{\text{Spez}}}_{1} + \underbrace{\frac{\widehat{\text{Sens}}(1 - \widehat{\text{Spez}})}{2}}_{2} + \underbrace{\frac{\widehat{\text{Spez}}(1 - \widehat{\text{Sens}})}{2}}_{3} = \frac{\widehat{\text{Sens}} + \widehat{\text{Spez}}}{2}. \tag{4.2}$$

Die Ziffern 1,2,3 stehen für die einzelnen Flächen in der Abbildung.
Es hängt natürlich vor allem von der Fragestellung des Anwenders ab, ob die beiden Größen gleiches Gewicht (AUC) haben sollen oder nicht (Trefferquote). Die Unterschiede zwischen den beiden Schätzern unterliegen den gleichen Diskussionen und

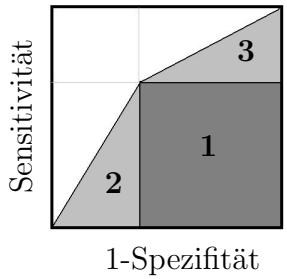

Abbildung 4.1: Algebraische Bestimmung der AUC über die Trapezregel, zur Erklärung der Ziffern siehe Formel (4.2)

Angriffspunkten, mit denen man immer konfrontiert ist, wenn man unbalancierte Gruppen hat. Youden (1950) schlug für die Wahl des optimalen Cut-Off Wertes die direkte Summe aus Sensitivität und Spezifität vor, also einen Schätzer, der dem ungewichteten Mittel entspricht. Allerdings verwendete er zur Normierung des Wertes die Subtraktion von 1. Dieser sogenannte „Youden-Index“ ist eine etablierte Größe in der Biometrie und liefert somit ein Argument für die AUC im Gegensatz zur Trefferquote.
Ein Blick in die Literatur zeigt, dass die Diskussion über diese beiden unterschiedlichen Schätzer schon lange geführt wird. Bereits Feinstein (1975) diskutiert die Unterschiede der beiden Schätzer und kommt zu dem Schluss, dass die Prävalenzabhängigkeit der Trefferquote problematisch ist und zu Manipulationen führen kann. Er zieht hier Youdens Index der Trefferquote eindeutig vor, wendet allerdings ein, dass bei der Zusammenfassung von Sensitivität und Spezifität in einem Index immer die einzelnen Werte und somit deren Dualität verloren gehen.
Shapiro (1999) erwähnt die Trefferquote, gibt aber zu bedenken, dass sie irreführend sein kann. Er illustriert dies mit folgendem Beispiel: würde man alle schwangeren Frauen in den USA auf eine HIV-Infektion testen wollen, hätte ein Test, der einfach jede dieser Frauen als HIV-negativ klassifiziert eine Trefferquote von vermutlich 99%, wobei die Sensitivität 0% betrüge.
Parker & Davis (1999) definieren die Accuracy eines dichotomen Tests als Summe aus Sensitivität und Spezifität und geben hierfür auch exakte Konfidenzbereiche (statt *„crude approximations“*) an.
In Zhou *et al.* (2002) wird die Trefferquote als *„häufig verwendetes Maß wegen seiner Einfachheit“* erwähnt, aber die Autoren sagen auch, dass sie einige weitere Summenmaße vorstellen werden, die der Trefferquote überlegen sind. An dieser Stelle findet sich dann auch der Hinweis, dass die überlegenen Maße mit der ROC-Kurve assoziiert sind.
Bei Lu *et al.* (2003) wird beschrieben, dass die Trefferquote nur in Kohortenstu-

dien Sinn macht, denn nur hier ist die Prävalenz nicht frei wählbar. Für eine Non-Inferiority Fragestellung schlagen sie als alternative Hypothese vor, dass beide Größen, also Sensitivität *und* Spezifität, des neuen Tests denen des alten Tests nicht unterlegen sein dürfen.
In den Points to Consider der European Medicines Agency (EMEA) (2001) wird zunächst gefordert: *„In a suitable experiment the probability of a correct test result is estimated as the proportion of cases for which the test result is correct.“* Weiter hinten im Dokument wird dann allerdings verlangt, dass es prävalenzunabhängige Maße geben soll, die die Accuracy der Testverfahren vergleichen.
Zusammenfassend hat dieser Literaturüberblick gezeigt, dass die Diskussion immer noch aktuell ist, da vermutlich die meisten Anwender die Trefferquote wegen ihrer Einfachheit vorziehen. Gerade deshalb soll die Empfehlung in dieser Arbeit aber sein, dass man auch bei dichotomen Daten sehr wohl und sehr gut die AUC als Schätzer für die Accuracy bei dichotomen Testergebnissen verwenden kann.

4.1.2 Clustered data - Modell 3

Für clustered data bei dichotomen Daten gibt es in der Literatur verschiedene Ansätze, um Sensitivität und Spezifität getrennt voneinander zu vergleichen. Die Vorgehensweise bei Ahn (1997) beruht auf der Schätzung der Sensitivität $\widehat{\text{Sens}}_k$ pro Person und dem anschließenden gewichteten Mitteln über die Personen, d.h. der Schätzer für die Sensitivität ignoriert die Zugehörigkeit der Daten zu einer bestimmten Person.

$$\widehat{\text{Sens}} = \frac{\sum_{k=1}^{n} m_{ik} \widehat{\text{Sens}}_k}{\sum_{k=1}^{n} m_{ik}} \tag{4.3}$$

Diese Abhängigkeit innerhalb der Personen wird dann allerdings beim Varianzschätzer berücksichtigt:

$$\hat{V}_{Se} = \frac{1}{n(n-1)} \sum_{k=1}^{n} \left(\frac{m_{ik}}{\overline{m}_{i\cdot}} (\text{Sens}_k - \widehat{\text{Sens}})^2 \right)$$

Die Varianzschätzung geht auf Arbeiten von Rao & Scott (1992) zurück, die einen Schätzer für die Varianz eines Quotienten von relativen Häufigkeiten vorgeschlagen hat. Für den Vergleich von zwei Sensitivitäten wird dann natürlich auch noch die Kovarianz zwischen den einzelnen Beobachtung benötigt. Diese wird analog zur Varianz mit der empirischen Kovarianz geschätzt.
Für die getrennte Bewertung von Sensitivität und Spezifität wurde außerdem ein allgemein gewichteter Schätzer (Lee & Dubin, 1994; Lee, 1996) vorgeschlagen, welcher durch Wichtung mit dem Inversen der Clustergröße genau zum ungewichteten Schätzer wird. Der Vergleich zwischen den Gewichtungen (Ahn, 1997) ergibt auch hier wieder das gleiche Ergebnis wie im nicht-dichotomen Fall: je größer die Korrelation innerhalb des Clusters desto besser schneidet der ungewichtete Schätzer ab.
Will man nun jedoch Sensitivität und Spezifität gemeinsam auswerten, braucht man

ein Analogon zur Trefferquote oder der Accuracy für clustered data. Hierzu finden sich in der Literatur verschiedene Ansätze. In dem Buch von Zhou *et al.* (2002) werden einige vorgestellt, welche jeweils die in (4.3) definierten Schätzer für Sensitivität und Spezifität bei clustered data verwenden. Eine Möglichkeit, Sensitivität und Spezifität gemeinsam auszuwerten, bietet der Likelihood-Quotient

$$\widehat{LR} = \frac{\widehat{\text{Sens}}}{1 - \widehat{\text{Spez}}}.$$

Die Varianzschätzung erfolgt hierbei durch

$$\widehat{Var(LR)} = \frac{1 - \widehat{\text{Sens}}}{\#RP} + \frac{\widehat{\text{Spez}}}{\#RN}.$$

Eine weitere Alternative ist das Odds Ratio

$$\widehat{OR} = \frac{\widehat{\text{Sens}}/(1 - \widehat{\text{Sens}})}{(1 - \widehat{\text{Spez}})/\widehat{\text{Spez}}},$$

also die relative Wahrscheinlichkeit für einen positiven Test bei Vorhandensein der Krankheit im Vergleich zum Nichtvorhandensein der Krankheit. Die Varianzschätzung

$$\widehat{Var(OR)} = \widehat{OR}\left(\frac{1}{\#RP} + \frac{1}{\#RN} + \frac{1}{\#FP} + \frac{1}{\#FN}\right)$$

ist allerdings nur dann möglich, wenn alle Zellen der Vierfeldertafel besetzt sind.
Unerwähnt bleibt in dem Buch allerdings die Anwendung von AUCs für dichotome Daten. Die in dieser Arbeit vorgestellte Theorie zur Berechnung und zum Vergleich von AUCs lässt sich aber direkt auf dichotome Daten übertragen, schließlich wurden in den Herleitungen beliebige Verteilungsfunktionen zugelassen. Dies entspricht dann wieder dem ungewichteten Mittelwert zwischen Sensitivität und Spezifität, der wegen der Analogie zum Youden-Index auf jeden Fall vorzuziehen ist. Damit steht eine konsistente Theorie für alle Skalenniveaus von Beobachtungsdaten zur Verfügung.
Die Güte der Approximation der vorgestellten Verfahren bei dichotomen Daten wird mihilfe von Simulationen in Kapitel 6 untersucht.

5 Beispiele

In diesem Kapitel soll die Anwendung der Theorie an realen Datensätzen illustriert werden. Verschiedene Studiendesigns werden untersucht und Daten hierzu ausgewertet. Zunächst wird der allgemeine Fall ordinaler Daten beschrieben, hier werden drei Beispiele vorgestellt: ein Vergleich zweier Reader im MRT, eine Diagnosestudie mit und ohne CAD und ein Vergleich hoher und niedriger Röhrenspannungen im Flachscanner. Anschließend wird ein Beispiel mit dichotomen Daten vorgestellt. Außerdem werden am Ende des Kapitels SAS-Makros beschrieben, die im Zuge dieser Arbeit zur Anwendung der Methoden entstanden sind.

5.1 Clustered ordinale Daten

5.1.1 Reader-Vergleich im MRT

In der Arbeit über nichtparametrische Verfahren zum Vergleich von ROC-Kurven bei clustered data von Obuchowski (1997) wird ein Beispiel ausgewertet, dessen zugehörige Daten in Tabelle 5.1 dargestellt sind. Sie stammen aus einer Studie, in der mithilfe der dreidimensionalen Magnet-Resonanz-Angiographie (MRA) die arterielle atherosklerotische Stenose quantifiziert wurde. Zwei Radiologen untersuchten 65 Halsschlagadern (links und rechts) an 36 Patienten mithilfe der MRA. Außerdem wurden die Patienten mithilfe von Digitaler-Selenium-Angiographie (DSA) untersucht, um den Goldstandard zu erhalten. Es gab Patienten, bei denen beide Arterien einen Verschluss hatten (k), bei denen nur die rechte oder nur die linke Arterie verschlossen waren und solche Patienten, bei denen beide Arterien ohne Verschluss waren (g). Bei 7 von den 36 untersuchten Patienten wurde nur eine Arterie untersucht, was zu ungleichen Clustergrößen führte. Es ist davon auszugehen, dass es zwischen rechter und linker Halsschlagader keinen systematischen Unterschied gibt.

Auswertung

Die Ergebnisse der Analysen nach Obuchowski (1997) und mit der neuen Methode (Modell 3) sind in Tabelle 5.2 zusammengefasst. Man sieht, dass sich die Auswertungsmethoden nicht besonders stark unterscheiden. Da hier ein Zweistichprobenvergleich durchgeführt wurde, ist der Unterschied lediglich auf die unterschiedliche Definition der Schätzer (gewichtet und ungewichtet) zurückzuführen. Die Accuracies der beiden Reader unterscheiden sich nur minimal, was sich in den hohen p-Werten

Tabelle 5.1: Originaldatensatz aus Obuchowski (1997)

Fall	Reader 1		Reader 2		Goldstandard	
	links	rechts	links	rechts	links	rechts
1	87	79	87	83	k	k
2	88	95	94	93	k	k
3	100	68	100	79	k	k
4	65	89	61	91	k	k
5	97	100	97	100	k	k
6	100	89	99	88	k	k
7	77		77		k	g
8	94	45	89	44	k	g
9	89	-11	92	-17	k	g
10	95	30	91	14	k	g
11	95	10	95	-122	k	g
12	100	19	100	-3	k	g
13		75		75	g	k
14	70	70	73	81	g	k
15	26	86	27	93	g	k
16	0	65	8	70	g	k
17	6	100	23	100	g	k
18	76	96	79	95	g	k
19	44	100	55	100	g	k
20	73	97	67	94	g	k
21		85		86	g	k
22	58	100	60	99	g	k
23	-2	100	-2	100	g	k
24	-3	0	-14	-11	g	g
25	47	-94	-3	-90	g	g
26	-16	75	17	85	g	g
27	5	-1	-22	-14	g	g
28	-6		-52		g	g
29	37	47	38	35	g	g
30	42	4	40	6	g	g
31	55	55	23	45	g	g
32	4	-13	-87	-65	g	g
33	49	35	24	9	g	g
34		53		66	g	g
35	2		-15		g	g
36	-48		-78		g	g

Tabelle 5.2: Ergebnisse Halsschlagader aus Obuchowski (1997)

Methode	Reader 1		Reader 2		p-Wert
	AUC	Konf.intervall	AUC	Konf.intervall	Differenz
Obuchowski	0.984	[0.963;1.00]	0.985	[0.966;1.00]	0.89
ATS	0.982	[0.923;0.996]	0.979	[0.952;0.991]	0.80

mit beiden Teststatistiken widerspiegelt. Ein Blick auf die Konfidenzintervalle bei Obuchowski zeigt, dass diese als obere Grenze jeweils 1 aufweisen. Da sie nicht symmetrisch sind, die Formel zur Berechnung in Obuchowski (1997) aber symmetrische Konfidenzintervalle liefern müsste, sind die Intervalle wohl einfach nur bei 1 abgeschnitten worden. Da hier extrem hohe Accuracies vorliegen, wurde bei der Berechnung der Konfidenzintervalle nach der neuen Methode die Delta-Transformation angewandt, damit die Konfidenzintervalle nicht größer als 1 werden können. Wie man sieht, resultierte dies in unsymmetrischen Intervallen.

5.1.2 Diagnose mit und ohne CAD

In der Abteilung Diagnostische Radiologie der Universität Göttingen wurde eine prospektive Studie zur Untersuchung der computerunterstützten Diagnose (CAD) von Mammakarzinomen durchgeführt. Die Ergebnisse wurden von freundlicherweise von Frau PD Dr. med. Silvia Obenauer zur Verfügung gestellt und mithilfe von cand.med. Besim Angic erhoben. Drei Untersucher haben jeweils 100 maligne Fälle (histologisch gesichert) und 100 Kontrollen (aus einer Screening-Population) beurteilt. Hierbei erfolgte die Beurteilung wie in der täglichen Routine, nämlich nach der Betrachtung aus zwei Perspektiven (cranio-caudal, CC und medio-lateral-schräg, MLO) und für die rechte und linke Brust nacheinander. Die Befundung wurde in BIRADS [1] Kategorien von 1 bis 5 durchgeführt:

1 Negativ: Unauffälliges Mammogramm ohne Hinweis auf fokale Befunde.

2 Benigne Veränderungen: Mammographischer Nachweis eindeutig benigner Veränderungen z.B. kalzifizierte Fibroadenome, Talgdrüsenverkalkungen, Ölzysten, Lipome, intramammäre Lymphknoten, etc. Insgesamt keinerlei Hinweis auf Malignität.

3 Wahrscheinlich benigne Veränderungen - eine kurzfristige Kontrolle wird empfohlen: Mammographische Veränderungen mit hoher Wahrscheinlichkeit für Benignität. Zur Befunderhärtung bei vermuteter Gutartigkeit wird dennoch eine Kontrolle empfohlen.

[1] Breast Imaging Reporting and Data System

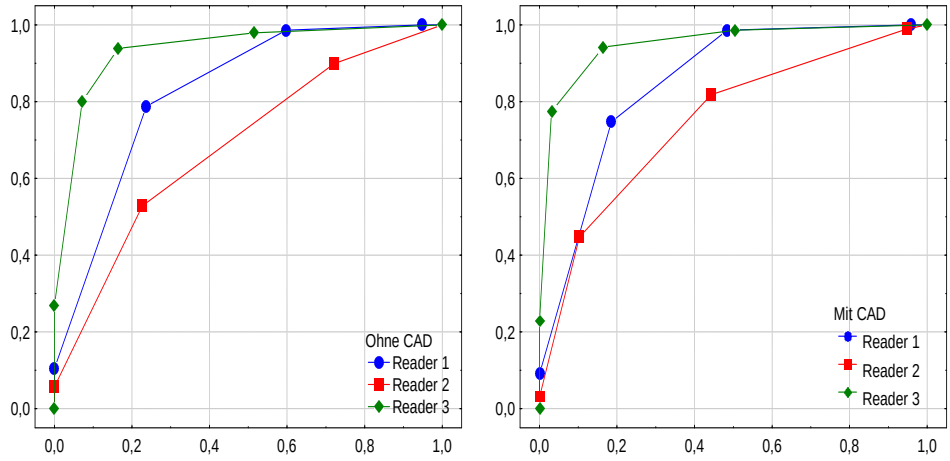

Abbildung 5.1: ROC-Kurven für die Analyse ohne (links) und mit (rechts) CAD

4 Malignomverdächtige Veränderungen - unklare Veränderungen, die bildmorphologisch keine eindeutigen Benignitätskriterien aufweisen. Zur weiteren Abklärung sollte eine Biopsie in Erwägung gezogen werden.

5 Hochmalignomverdächtige Veränderungen: Läsionen mit hoher Wahrscheinlichkeit der malignen Entartung.

Die Reader hatten unterschiedliche Erfahrungen (Student ohne Mammographie-Erfahrung, Assistenzarzt mit ca. sechsmonatiger Mammographie-Erfahrung, Facharzt mit mehrjähriger Mammographie-Erfahrung). Ziel der Studie war es zu überprüfen, ob die unterschiedlichen Erfahrungen zu unterschiedlichen Ergebnissen mit und ohne CAD führten. Da von fast jeder Patientin die Diagnosen von beiden Brüsten vorlagen, handelt es sich um clustered data des Modells 3. Bei den malignen Fällen gab es in den meisten Fällen auch Aufnahmen von der gesunden Brust, bei den Kontrollen gab es meist zwei gesunde Beobachtungen.

Auswertung

Abbildung 5.1 zeigt die empirischen ROC-Kurven der drei Reader mit den zwei Methoden. Es gab insgesamt 88 verbundene Beobachtungen, bei den Gesunden waren 194 Befundungen erfolgt, bei den Kranken 97 Befunde. Durch diese Unbalanciertheit kann es zu Unterschieden zwischen dem gewichteten und ungewichteten Schätzer für die AUC kommen.
Betrachtet man die ROC-Kurven, so ist deutlich zu sehen, dass der erfahrene Radiologe (Reader 3) am besten abschneidet, während der Assistenzarzt (Reader 1) immer noch besser befundet als der Student (Reader 2). Wie sehr jedoch das CAD

Tabelle 5.3: Ergebnisse der CAD-Studie

Methode	Reader	$\hat{p}$	Konfidenzintervall	$\tilde{p}$
ohne CAD	1	0.809	[0.770; 0.849]	0.826
ohne CAD	2	0.649	[0.608; 0.691]	0.684
ohne CAD	3	0.894	[0.854; 0.934]	0.935
mit CAD	1	0.834	[0.789; 0.878]	0.849
mit CAD	2	0.712	[0.669; 0.754]	0.748
mit CAD	3	0.911	[0.881; 0.940]	0.950

die Diagnose unterstützt, ist aus den reinen Kurven nicht sehr gut zu erkennen. Dafür ist es einfacher, die AUCs zu vergleichen. Die Ergebnisse der Studie sind in Tabelle 5.3 zusammengefasst.

Hier sieht man, dass der Gewinn an Accuracy beim Studenten (Reader 2) am höchsten ist, jedoch auch die Diagnosen der beiden Ärzte durch das CAD noch an Accuracy gewinnen. Sogar die Konfidenzintervalle bei jeder Methode sind für die drei Reader voneinander getrennt. Der gewichtete Schätzer $\tilde{p}$ für die Accuracy ist in diesem Beispiel immer größer als der ungewichtete, was ausschließlich daraus resultiert, dass die Messwiederholungen, die hier auftreten, lediglich bei den gesunden Patientinnen vorkommen: es gibt keine einzige Patientin mit zwei malignen Fällen. Vergleicht man die AUCs nun mithilfe eines statistischen Tests, erhält man folgende Ergebnisse (siehe Tabelle 5.4). Mit keiner der vier Teststatistiken ist die Wechselwirkung signifikant. Da es sich hier bei den Stichprobenumfängen um eine große Studie handelt, waren auch keine großen Unterschiede zwischen ATS und WTS zu erwarten. Weiterhin ist zu sehen, dass der Unterschied zwischen den Readern und den Methoden signifikant ist, d.h. die unterschiedliche Erfahrung führt zu unterschiedlichen Accuracies und das CAD führt im Mittel zu einer Erhöhung der Accuracy.

Tabelle 5.4: p-Werte und Teststatistiken für CAD-Studie

	ungewichteter Schätzer			gewichteter Schätzer		
Effekt	Statistik	FG	p-Wert	Statistik	FG	p-Wert
Methode ATS	10.040	1	0.0015	10.053	1	0.0015
Methode WTS	10.040	1	0.0015	10.053	1	0.0015
Reader ATS	51.166	1.9	<0.001	41.160	1.9	<0.001
Reader WTS	121.62	2	<0.001	89.644	2	<0.001
Wechselwirkung ATS	2.507	1.87	0.0855	2.939	1.72	0.0611
Wechselwirkung WTS	4.091	2	0.1293	4.573	2	0.1016

5.1.3 Vergleich hoher und niedriger Röhrenspannungen

Zum Vergleich verschiedener Röhrenspannungen bei digitaler Selenium-Radiographie wurde von Bernhardt *et al.* (2004) eine großangelegte Phantomstudie durchgeführt. Es sollte überprüft werden, ob es möglich ist, die Scans ohne einen zu großen Verlust der Bildqualität bei niedrigeren Spannungen durchzuführen. Die Wahl der Spannung wird hierbei auch noch durch die Dauer des Scans bestimmt: ist die Spannung zu gering, so dauert die Aufnahme zu lange. Der Patient wäre dann der Strahlung zu lange ausgesetzt, was nicht zumutbar wäre.
Es wurden insgesamt 54000 Bewertungen von Scans erhoben. Hierbei wurden fünf Untersucher mit verschiedener Berufserfahrung eingesetzt und drei verschiedene Ausgangsspannungen verwendet. Es wurden verschiedene künstliche Strukturen (u.a. Glasperlen, Netze, Katheter), die bestimmte Lungenverschlusskrankheiten bzw. deren Symptome simulieren, auf insgesamt 50 Plexiglasfolien appliziert. Diese Folien wurden dafür in 12 Segmente eingeteilt und jede Struktur insgesamt 60 mal zufällig auf den Folien verteilt. Außerdem wurde zu jeder Struktur ein leeres Segment, d.h. eines ohne diese entsprechende Struktur, als Kontrolle zugeordnet. Somit sind für jede Struktur 60 „Gesunde“ und 60 „Kranke“, also Segmente ohne bzw. mit dieser Struktur, vorhanden. Dieser Versuchsaufbau führte dazu, dass in jedem Segment entweder keine, eine oder mehrere Strukturen vorkommen konnten. Dann wurden die 50 Folien nacheinander auf einem Brustphantom plaziert, welches dann bei den drei verschiedenen Spannungen gescannt wurde. Anschließend hat jeder Untersucher jedes Bild in einer zufälligen Reihenfolge befundet. Hierbei musste er für jedes Segment und jede Struktur auf einer Skala von 1-5 diagnostizieren, ob die Struktur hier vorhanden ist oder nicht. Ein Auszug der Daten wird in Tabelle 5.5 präsentiert. Man beachte, dass aus Platzgründen nur die Bewertungen der ersten drei Reader dargestellt werden.
Sämtliche Strukturen, die gemeinsam auf einer Folie appliziert sind, sind als abhängig zu betrachten, da sie mit einem Scan abgebildet wurden. Die Auswertung des Versuchs durch die Autoren in der ursprünglichen Arbeit (Bernhardt *et al.* , 2004) wurde in Ermangelung einer passenden Methodik mit dem Programm Rockit von Berbaum, Dorfman, Metz (Dorfman *et al.* , 1992) durchgeführt. Hier müssen alle Beobachtungen an verschiedenen Personen durchgeführt werden, die einzigen Abhängigkeiten, die erlaubt sind, sind wiederholte Messungen von verschiedenen Readern oder mit verschiedenen Geräten. In diesem Beispiel gibt es dann also gleich zwei Verletzungen dieser Annahme:

1. Die Strukturen treten mit Wahrscheinlichkeit 1 mehr als einmal auf mindestens einer Folie auf (Modell 2, abhängige Messungen mit gleichem Goldstandard).

2. Es gibt auf fast jeder Folie ein zugeordnetes leeres Feld für jede Struktur (Modell 3, abhängige Messungen mit unterschiedlichem Goldstandard).

Tabelle 5.5: Auszug der Daten von Folie No.26-31 und Reader 1 bis 3 (von 5) für die drei Röhrenspannungen (V1-V3) und die Struktur „Netz“

			Reader 1			Reader 2			Reader 3		
Folie	Segment	Gold	V1	V2	V3	V1	V2	V3	V1	V2	V3
26	1	0	5	5	5	4	4	5	5	4	5
26	3	1	1	1	1	1	1	1	1	1	1
26	9	1	5	5	5	5	5	5	5	5	5
27	3	0	5	5	5	5	5	4	5	5	4
27	9	0	5	5	5	5	5	5	5	5	5
27	7	1	1	1	1	1	1	1	2	1	1
28	7	1	2	1	1	1	1	1	1	1	1
28	9	1	3	1	1	2	2	1	2	2	1
29	7	0	5	5	5	4	5	5	4	5	5
29	9	0	5	5	5	5	5	5	5	5	5
29	6	1	5	2	1	2	1	4	3	2	2
30	6	1	3	2	2	4	3	1	5	3	2
30	7	1	3	2	2	4	3	3	3	2	3
31	6	0	5	5	5	4	5	4	4	5	4
31	1	1	5	2	1	4	4	2	5	3	2
31	7	1	3	1	1	4	1	1	4	1	1

Wir haben den Versuch deshalb mit den hier entwickelten Methoden noch einmal neu ausgewertet. Insgesamt wurden in der Studie sechs verschiedene Strukturen untersucht. Da jede für sich eine andere Krankheit bzw. medizinisches Symptom simuliert, wurden die Strukturen getrennt voneinander (sozusagen alle als Primärvariablen) und nicht in einem multivariaten Design untersucht. Exemplarisch wird die Struktur „Netz“ vorgestellt, bei der Gazeteile auf der Folie verteilt wurden.

Auswertung

Zunächst zeigt Abbildung 5.2 die Accuracies und die dazugehörigen Konfidenzintervalle für alle 15 diagnostischen Tests (also für die drei Spannungen mal fünf Reader). Offensichtlich ist Methode 1 (150 kVp) schlechter als Methode 2 (90 kVp) und Methode 3 (70 kVp). Weiterhin sieht man, dass die Fähigkeiten der einzelnen Reader nicht homogen über die drei Methoden hinweg sind. Speziell Reader 5 fällt aus dem allgemeinen Muster heraus. Also sollten die Daten zunächst einmal in einem zweifaktoriellen Design mit Wechselwirkung ausgewertet werden. Da die WTS in einem so hochdimensionalen faktoriellen Setting zu extrem liberalen Entscheidungen führt, werden nur die Ergebnisse der ATS betrachtet. Die erste Spalte von Tabelle 5.6 zeigt die p-Werte und Teststatistiken der globalen Analyse, außerdem die Ergebnis-

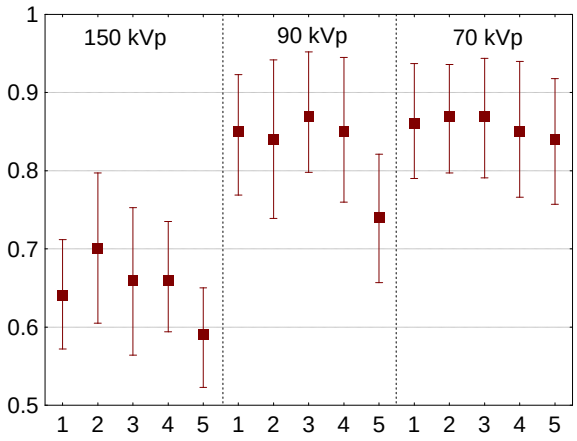

Abbildung 5.2: Empirische AUCs mit 95%-Konfidenzintervallen für die fünf Reader und die drei verschiedenen Methoden

se aus der Originalarbeit von Bernhardt *et al.* (2004), bezeichnet mit DBM. Wie man sieht, gibt es einen signifikanten Methoden- und Readereffekt mit der ATS. Die Wechselwirkung dagegen ist nicht signifikant. Die Ergebnisse aus der Originalarbeit zeigen auch einen signifikanten Effekt, leider ist dort nicht eindeutig dargestellt, welche Hypothese mit dem F-Test genau getestet wird. Die Ergebnisse sind der Vollständigkeit halber aber in der Tabelle mit angegeben. Eindeutig lässt sich nur sagen, dass auf eine mögliche Wechselwirkung nicht getestet wurde, da schließlich nur ein Test und ein p-Wert angegeben sind. Die weitere Analyse der Daten zeigt in den nebenstehenden Spalten die Paarvergleiche der drei Methoden und man sieht, dass die Methoden 2 und 3 beide signifikant von Methode 1 verschieden sind, untereinander dagegen nicht. Auch in der Originalarbeit wurden paarweise Vergleiche durchgeführt, auch hier wiederum mit einer Methode, die keine Wechselwirkung

Tabelle 5.6: ATS und p-Werte für den ungewichteten Schätzer, Teststatistiken und p-Werte nach Dorfman-Berbaum-Metz (aus Bernhardt *et al.* , 2004)

	Global		2 vs. 3		1 vs. 3		1 vs. 2	
Effekt	ATS	p-Wert	ATS	p-Wert	ATS	p-Wert	ATS	p-Wert
Spannung	32.5	<0.0001	1.356	0.2442	38.012	<0.0001	48.261	<0.0001
Reader	3.4	0.0245	2.278	0.0790	1.738	0.1613	4.434	0.0074
S x R	1.094	0.3565	1.384	0.2462	0.872	0.4278	0.999	0.3755
	F-Test	p-Wert	F-Test	p-Wert	F-Test	p-Wert	F-Test	p-Wert
DBM	5.86	0.02	3.73	0.12	3.70	0.09	8.25	0.04

Tabelle 5.7: Mittlere Konfidenzintervalle für die einzelnen Methoden

150 kVp:	$0.651 \in [0.584; 0.717]$
90 kVp:	$0.831 \in [0.760; 0.902]$
70 kVp:	$0.858 \in [0.797; 0.919]$

berücksichtigt. Die p-Werte und deren Interpretation gehen in eine ähnliche Richtung wie die Ergebnisse der ATS, allerdings ist der Vergleich zwischen Methode 1 und 3 kanpp nicht signifikant.
Die Konfidenzintervalle für die Methoden, jeweils gemittelt über die fünf Reader, sind in Tabelle 5.7 dargestellt. Man kann hier analog zu den Testergebnissen ablesen, dass die niedrigeren Spannungen signifikant besser abschneiden als die hohe Spannung. Dieses Ergebnis ist konträr zur allgemeinen Meinung, dass die Spannung nicht hoch genug gewählt werden kann. Allerdings müssen bei der abschließenden Bewertung der Ergebnisse natürlich alle Strukturen berücksichtigt werden. Außerdem ist es wichtig, den Trade-Off zwischen Spannung und Dauer der Aufnahme (und deshalb Strahlenbelastung eines Patienten) zu bedenken.

5.2 Dichotome Daten

5.2.1 Richtiger Zeitpunkt bei Kontrastmittel-unterstützter Diagnose

Die Daten der folgenden Studie wurden freundlicherweise von Dr. Kaufmann von der Firma Schering zur Verfügung gestellt. In der Studie mit 22 Patienten ohne Verschluss und 105 Patienten mit Verschluss der Oberschenkelarterie (*femoralis superficialis*) sollte untersucht werden, zu welchem Zeitpunkt der mit Kontrastmittel unterstützten Magnetresonanz-Angiographie (MRA) der Verschluss besser diagnostiziert werden kann: direkt nach dem Einspritzen des Kontrastmittels (first pass, Zeitpunkt 1) oder nach Erreichen des Gleichgewichts (equilibrium, Zeitpunkt 2) nach ca. 30 Minuten. Da dieser Zeitpunkt des Gleichgewichts schwierig zu bestimmen ist, sollte überprüft werden, ob die gleiche Diagnose auch direkt nach der Kontrastmittelgabe gestellt werden kann. Hierzu wurden die Patienten zu beiden Zeitpunkten von drei erfahrenen Blinded-Readern untersucht. Die Diagnose wurde dichotom erhoben, also 1 für Stenose und 0 für keine Stenose. Der Goldstandard wurde mithilfe von DSA (digital substraction angiography) bestimmt. Es liegt also ein zweifaktorieller Versuch vom Modell 1 vor, mit den Faktoren Zeitpunkt und Reader.

Tabelle 5.8: Trefferquote und AUC im Vergleich

Reader	Zeitpunkt	Spezifität	Sensitivität	Trefferquote	AUC
1	1	95.454	75.238	78.74	85.35
1	2	95.454	78.095	81.10	86.77
2	1	68.181	86.666	83.46	77.42
2	2	31.818	92.381	81.89	62.10
3	1	63.636	90.476	85.83	77.06
3	2	72.727	92.381	88.98	82.55
Mittelwert	1	75.757	84.127	82.68	79.94
Mittelwert	2	66.666	87.619	83.99	77.14

Auswertung

Folgende Größen (siehe Tabelle 5.8) wurden in der Studie berechnet: Sensitivität und Spezifität für jeden Reader und Zeitpunkt, außerdem die Mittelwerte (MW) über die Reader pro Zeitpunkt. Weiterhin wurde sowohl die Trefferquote als auch die AUC berechnet. Da in diesem Beispiel die Prävalenz der Krankheit nicht 50% ist, unterscheiden sich die beiden Indizes. Speziell bei Reader 2, der in der kleinen Gruppe der Gesunden enorme Probleme mit der richtigen Diagnose hatte (Spezifität=31.8%) sind die Unterschiede groß: eine Trefferquote von 81.9% steht einer AUC von 62.1% gegenüber. Wenn man nun eine Pauschalaussage über das Verfahren im Allgemeinen machen möchte, ist die Trefferquote irreführend hoch, denn die schlechte Performance bei der Diagnose gesunder Patienten wird durch den gewichteten Mittelwert verdeckt. Will man dagegen Sensitivität und Spezifität getrennt beurteilen, so ist keiner der beiden aggregierten Indizes notwendig.
In Abbildung 5.3 sind die 95%-Konfidenzintervalle für die AUC und die Trefferquote dargestellt. Hier ist noch einmal deutlich der Unterschied zwischen den beiden Größen zu sehen. Außerdem ist die Wechselwirkung zwischen Reader und Zeitpunkt bei der AUC viel deutlicher zu sehen. Die p-Werte der entsprechenden ATS zeigen dann auch die bereits in den Grafiken deutlichen Effekte: sowohl der Reader-Effekt (p=0.0018) als auch die Wechselwirkung (p=0.0081) sind statistisch signifikant, der Zeit-Effekt (p=0.4655) dagegen ist nicht signifikant. Es wäre falsch, aus diesem Ergebnis schon auf die Äquivalenz der beiden Verfahren zu schließen, aber es liefert schon einen guten Hinweis auf die Vergleichbarkeit der Verfahren.

5.3 Realisation der Theorie in SAS

Es wurden mehrere SAS-Makros erstellt, um potentiellen Anwendern die Gelegenheit zu geben, die vorgestellten Methoden ohne großen Programmieraufwand einzusetzen. Alle Beispiele im vorhergehenden Kapitel wurden mit diesen Programmen

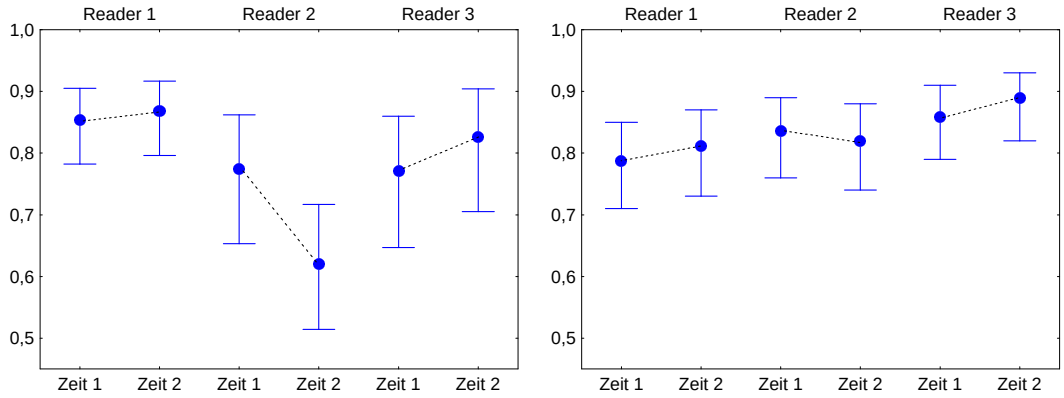

Abbildung 5.3: Konfidenzintervalle für AUC (links) und Trefferquote (rechts)

ausgewertet.

5.3.1 Modell 1: diag.sas

Zunächst einmal wird das Programm **diag.sas** vorgestellt, welches die Anwendung der nichtparametrischen Methoden für Designs mit multiplen Readern und mehreren Methoden ermöglicht. Hierzu wird ein Datensatz (**data**) mit folgender Struktur benötigt: Patienten ID (**subject**), Goldstandard (0,1) (**state**), Reader (**rater**), Methode (**method**), Messwert (**var**). Die Messungen stehen also alle untereinander in einer langen Spalte, während die Beschreibung der Herkunft der Werte in den anderen Spalten steht. Fehlende Werte im Datensatz führen zu einer Fehlermeldung. Der Output des Programmes gliedert sich wie folgt: Zunächst werden alle Punktschätzer ausgegeben, sowohl die einfachen, als auch die Schätzer für den gemittelten Methoden-Effekt und für die Differenz zweier Methoden. Außerdem gibt das Programm einfache und nach der Delta-Methode transformierte Konfidenzintervalle an. Das Niveau ist hier per Default auf 5% gesetzt, kann allerdings auch manuell geändert werden (**alpha**). Der Aufruf des Programmes sieht dann folgendermaßen aus:

```
%diag(data=,
      var=,
      state=,
      rater=,
      method=,
      subject=,
      alpha=0.05);
```

Weiterhin kann man ROC-Kurven mit individueller Beschriftung ausgeben lassen. Allerdings ist die Auflösung dieser Grafiken allenfalls zum internen Gebrauch geeig-

net, für Veröffentlichungen sollte man auf andere Programme zurückgreifen.
Das Programm wurde anhand veröffentlichter und ausgewerteter Beispieldatensätze validiert. Mithilfe geringfügiger Änderungen im Quelltext ist es weiterhin möglich, beliebige andere lineare Kontraste zu testen bzw. Konfidenzintervalle dafür zu erhalten.

5.3.2 Modell 3: cluster.sas und cluster2F.sas

Außerdem wurde die Methodik für clustered data in zwei weiteren Programmen realisiert, je eines für das einfaktorielle (**cluster.sas**) und zweifaktorielle (**cluster2f.sas**) Design. Ein Faktor kann entweder durch verschiedene Untersucher oder verschiedene Methoden dargestellt werden. Die Anwendung des Programmes unterscheidet sich kaum von der im unabhängigen Fall, der Datensatz muss ganz ähnlich aufgebaut sein. Die Variablenbezeichnungen lauten: Patienten ID (**subject**), Goldstandard (0,1) (**gold**), Reader (**rater**), Methode (**method**), Messwert (**value**). Zusätzlich wird noch eine Variable **wdh** benötigt, die die einzelnen ROIs innerhalb eines Clusters einander eindeutig zuordnet. Es ist wichtig, dass die Patientennummer eindeutig ist. Es ist unproblematisch, wenn ein Patient nur einen der beiden Goldstandards hat. Man muss hierfür keine Leer-Beobachtungen einfügen, das Programm ist in der Lage, diesen Fall zu erkennen. Der Aufruf des Programmes sieht dann folgendermaßen aus:

```
%cluster2F(data=,
           gold=,
           subject=,
           rater=,
           method=,
           value=,
           wdh=,
           alpha=0.05);
```

Im Output des Programmes findet man wiederum alle Punkt- und Intervallschätzer für die individuellen Accuracies. Außerdem sind die Teststatistiken und p-Werte für die ANOVA- und die Wald-Typ-Statistik angegeben, zum Vergleich werden auch die gewichteten Schätzer (inklusive Teststatistiken) für die Accuracy mit angegeben.
Für den Fall fehlender Werte, d.h. wenn ein Reader ein Bild nicht ausgewertet hat oder ein Patient mit einer Methode nicht untersucht wurde oder weil eine Aufnahme unbrauchbar war, sind die Makros bisher nicht anwendbar. Man sollte dann im Rahmen einer Sensitivitätsanalyse verschiedene Ersetzungsmethoden der fehlenden Werte (worst case, best case, mittlerer Wert) untersuchen und mit den Ergebnissen des vollständigen Patientengutes vergleichen und auf Widersprüche untersuchen.
Im Folgenden wird zur besseren Illustration ein Auszug aus dem Originaloutput des

Programmes dargestellt. Die verwendeten Variablen stehen hier für die folgenden Werte:

- NN0: Anzahl der Beobachtungen an gesunden ROIs (m_0)
- NN1: Anzahl der Beobachtungen an betroffenen ROIs (m_1)
- D : Anzahl der Faktorstufen insgesamt
- P_MR: Ungewichteter Schätzer der Accuracy
- AREA: gewichteter Schätzer der Accuracy
- LOW: untere Grenze des 95%-Konfidenzintervalls
- UP: obere Grenze des 95%-Konfidenzintervalls
- SUM0: Anzahl der Patienten mit gesunden ROIs (n_0)
- SUM1: Anzahl der Patienten mit betroffenen ROIs (n_1)
- SUML: Anzahl der Patienten mit sowohl gesunden als auch kranken ROIs (n_c)
- N: Anzahl der Patienten insgesamt (n)

Zunächst werden die Schätzer für die Effekte, sowohl ungewichtet, als auch gewichtet ausgegeben. Die angegebenen Konfidenzintervalle beziehen sich allerdings nur auf die ungewichteten Effekte.

```
                    NN0        NN1          D
                     48         48         10
-----------------------------------------------------
 the estimators for area and confidence intervals

        METHOD RATER P_MR AREA  LOW   UP

             1     1 .642 .640 .572 .712
             1     2 .701 .717 .605 .797
             1     3 .658 .641 .564 .753
             1     4 .665 .677 .594 .735
             1     5 .587 .577 .523 .650
             2     1 .846 .833 .769 .923
             2     2 .841 .850 .739 .942
             2     3 .875 .869 .798 .952
             2     4 .853 .833 .760 .945
             2     5 .739 .740 .657 .821

               SUM0       SUM1       SUML       N
                 30         29         19      40
```

Anschließend sieht man die Teststatistiken für den ungewichteten Schätzer, jeweils ATS und WTS für die drei verschiedenen Haupteffekte, inklusive geschätztem Freiheitsgrad in Klammern und dem p-Wert.

```
----------Unweighted mean estimator-----------------

                        METHOD
ANOVA Type Statistic: 32.519 (df= 1.65 ); p=  0.0000
Wald Type Statistic:  50.955 (df= 2.00 ); p=  0.0000

                        READER
ANOVA Type Statistic:  3.389 (df= 2.47 ); p=  0.0246
Wald Type Statistic:  14.706 (df= 4.00 ); p=  0.0054

                     INTERACTION
ANOVA Type Statistic: 11.473 (df= 4.62 ); p=  0.0000
Wald Type Statistic:  30.089 (df= 14.0 ); p=  0.0074
```

Abschließend folgt noch ein Output der Teststatistiken für den gewichteten Schätzer, der Aufbau ist analog zum oben gezeigten Output für die ungewichteten Schätzer.

6 Simulationsergebnisse

Um das Verhalten der neu entwickelten Methode in verschiedenen Studiendesigns zu untersuchen, wurden Simulationen durchgeführt. Die Einhaltung des Niveaus bei einer Teststatistik ist hierbei von zentralem Interesse. Außerdem muss natürlich auch das Verhalten der Teststatistik unter bestimmten Alternativen betrachtet werden. Da prinzipiell eine unendliche Zahl von Konstellationen und Datenbeispielen denkbar ist, wurde versucht, einige typische Designs und Datenlagen auszuwählen, um die Simulationsergebnisse darzustellen.

6.1 Simulation der AUC

Zunächst einmal sollte das Verhalten der AUC an sich untersucht werden. Da in den Simulationen nicht direkt die Fläche unter der Kurve simuliert wird, sondern nur die einzelnen unterliegenden Verteilungen F_0 der Gesunden und F_1 der Kranken, ist es wichtig, zu untersuchen, welchen Einfluss die einzelnen Parameter in den Simulationen auf die Größe der AUC haben. Die Parameter, die variiert werden sollen, sind

- die Clustergröße *rep*, also die gesamte Anzahl der ROIs pro Cluster,
- die Korrelation ρ der Beobachtungen innerhalb eines Clusters, manchmal auch repräsentiert durch c, wobei $c^2/(1+c^2) = \rho$ gilt,
- die Verschiebung a, mit der die Verteilungsfunktion der Kranken transformiert wird.

Wie man in Abbildung 6.1 sieht, wird die Fläche für größere Verschiebungseffekte zwischen Gesunden und Kranken immer größer. Je weiter die Verteilungen gegeneinander verschoben sind, desto leichter ist es, zwischen ihnen zu diskriminieren. Dieser Effekt wird allerdings mit steigender Clusterkorrelation abgeschwächt. Außerdem ist zu sehen, dass die Anzahl der Wiederholungen innerhalb des Clusters keinen nennenswerten Einfluss auf die Größe der Fläche hat. Das war auch zu erwarten, schließlich sollte die Anzahl der Wiederholungen die Verteilung der Daten nicht beeinflussen, ganz im Gegenteil zur Korrelation und der Verschiebung, welche Form und Position der Verteilung bestimmen.

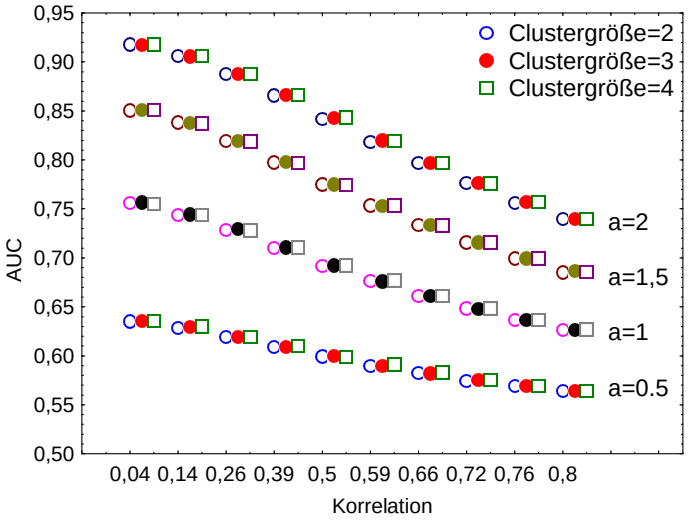

Abbildung 6.1: Erwartungswerte der AUC bei Variation der freien Parameter Korrelation, Verschiebung der ursprünglichen Verteilungen (a) und Clustergröße

6.2 Niveausimulationen

6.2.1 Modell 1

Die Simulationen im Modell 1 sollen vor allem die Unterschiede in der Güte der Approximation in Abhängigkeit vom Stichprobenumfang, der Stichprobenbalanciertheit und der Größe des Effekts zeigen. Hierzu wurden die Stichprobenumfänge $n_i \in \{15, 20, 30, 50, 100\}$ einmal identisch und in einem zweiten Durchgang verschieden ausgewählt, sodass auch unbalancierte Designs entstanden. Außerdem wurden die beiden simulierten Verteilungen so gegeneinander verschoben, dass empirische Accuracies AUC $\in \{0.5, 0.7, 0.85, 0.9, 0.95\}$ entstanden.

Die vollständigen Simulationsergebnisse in Tabelle 6.1 zeigen, dass ab einer AUC von 0.9 die Approximation zusammenbricht und selbst die WTS hier weit ins Konservative übergeht. Die Ursachen hierfür liegen im Rand des Definitionsbereichs. Schließlich ist die AUC nur bis 1 definiert, deshalb ist die Verteilung hier gewissermaßen abgeschnitten. Bei steigenden Stichprobenumfängen ist dieses Phänomen nicht mehr zu beobachten, was auch ein Zeichen dafür ist, dass die Konservativität aus der Approximationsgeschwindigkeit der Verteilung resultiert. Um diese Probleme bei der Approximation zu beheben, könnte eventuell eine Transformation der Daten zunächst auf die reelle Achse helfen. Durch Rücktransformation nach Berechnung der Effekte und Varianzen würde man wiederum einen interpretierbaren relativen Effekt erhalten, dessen Verteilung aber besser approximiert ist. Bei der

Tabelle 6.1: Verschiedene Flächen bei balancierten Designs, *kursive* Ziffern stehen für bedingte empirische Niveaus

		Anzahl der		Durchschnittliche Größe des Effekts $\widehat{p}_i$				
n_0	n_1	Faktorstufen	Statistik	0.5	0.7	0.85	0.9	0.95
15	15	2	ATS	0.0500	0.0501	0.0417	0.0341	0.0160
15	15	2	WTS	0.0500	0.0501	*0.0441*	*0.0461*	-
15	15	5	ATS	0.0400	0.0333	0.0222	0.0169	0.0023
15	15	5	WTS	0.0929	0.0726	*0.0476*	*0.0329*	-
15	15	10	ATS	0.0335	0.0353	0.0236	0.0167	0.0049
15	15	10	WTS	0.0811	0.0715	*0.0417*	*0.0362*	-
30	30	2	ATS	0.0513	0.0500	0.0496	0.0431	0.0303
30	30	2	WTS	0.0513	0.0501	*0.0517*	*0.0380*	-
30	30	5	ATS	0.0450	0.0394	0.0325	0.0266	0.0136
30	30	5	WTS	0.0695	0.0594	*0.0440*	*0.0338*	-
30	30	10	ATS	0.0414	0.0388	0.0355	0.0287	0.0157
30	30	10	WTS	0.0663	0.0583	*0.0355*	*0.0338*	-
50	50	2	ATS	0.0529	0.0456	0.0478	0.0428	0.0353
50	50	2	WTS	0.0529	0.0456	*0.0459*	*0.0294*	-
50	50	5	ATS	0.0459	0.0417	0.0396	0.0369	0.0229
50	50	5	WTS	0.0573	0.0508	*0.0485*	*0.0882*	-
50	50	10	ATS	0.0434	0.0435	0.0402	0.0359	0.0279
50	50	10	WTS	0.0581	0.0536	*0.0417*	*0.0147*	-
100	100	2	ATS	0.0496	0.0472	0.0504	0.0462	0.0418
100	100	2	WTS	0.0496	0.0472	*0.0483*	*0.0714*	-
100	100	5	ATS	0.0469	0.0465	0.0450	0.0446	0.0360
100	100	5	WTS	0.0543	0.0528	*0.0438*	-	-
100	100	10	ATS	0.0458	0.0470	0.0431	0.0435	0.0410
100	100	10	WTS	0.0508	0.0530	*0.0438*	*0.0714*	-
200	200	2	ATS	0.0511	0.0498	0.0489	0.0541	0.0482
200	200	2	WTS	0.0511	0.0498	*0.0502*	-	-
200	200	5	ATS	0.0544	0.0493	0.0466	0.0467	0.0386
200	200	5	WTS	0.0569	0.0513	*0.0502*	-	-
200	200	10	ATS	0.0462	0.0504	0.0460	0.0480	0.0437
200	200	10	WTS	0.0497	0.0526	*0.0437*	-	-
300	300	2	ATS	0.0530	0.0505	0.0456	0.0451	0.0491
300	300	2	WTS	0.0530	0.0505	*0.0447*	-	-
300	300	5	ATS	0.0506	0.0485	0.0498	0.0450	0.0477
300	300	5	WTS	0.0528	0.0490	*0.0517*	-	-
300	300	10	ATS	0.0476	0.0488	0.0519	0.0513	0.0473
300	300	10	WTS	0.0522	0.0507	*0.0502*	-	-
500	500	2	ATS	0.0538	0.0512	0.0464	0.0533	0.0469
500	500	2	WTS	0.0538	0.0512	*0.0432*	-	-
500	500	5	ATS	0.0473	0.0469	0.0492	0.0496	0.0484
500	500	5	WTS	0.0478	0.0490	*0.0474*	-	-
500	500	10	ATS	0.0503	0.0491	0.0500	0.0456	0.0480
500	500	10	WTS	0.0510	0.0523	*0.0495*	-	-

Tabelle 6.2: Prozentualer Anteil der Simulationsdurchläufe (von 10000) ohne singulären Kovarianzmatrixschätzer für verschiedene Effektgrößen und Stichprobenumfänge.

	Größe des Effekts		
Stichprobenumfang	0.7	0.85	0.9
15-15	96.23	25.63	3.04
30-30	99.90	37.75	2.37
50-50	99.98	46.00	0.68
100-100	100.00	53.42	0.14
200-200	100.00	61.60	0
300-300	100.00	65.79	0
500-500	100.00	71.77	0
15-30	99.43	38.60	4.47
30-15	99.34	38.45	4.68
20-100	100.00	89.05	29.30
100-20	100.00	88.36	28.73
100-200	100.00	86.01	0.66
100-500	100l00	100l00	82l25

WTS kommt außerdem hinzu, dass die Kovarianzmatrix speziell bei kleinen Stichprobenumfängen fast immer singulär wird. Dann lehnt die WTS überhaupt nicht mehr ab. Es ist in diesem Fall also gar nicht möglich, das empirische Niveau der WTS zu bestimmen. In Tabelle 6.2 sieht man, dass mit zunehmender Fläche die Anzahl der Singularitäten der Kovarianzmatrix schnell zunehmen, die Werte für 0.95 wurden weggelassen, denn hier ist die Matrix immer singulär gewesen.

Bei den unbalancierten Designs in Tabelle 6.3 fällt die Problematik bei hohen Accuracies besonders ins Gewicht. Außerdem ist zu beobachten, dass gerade im Bereich niedriger Stichprobenumfänge und/oder hoher Faktorstufenanzahlen die WTS wie erwartet liberal ist, während die ATS konservativ ist. Dieses Verhältnis dreht sich allerdings in manchen extremen Bereichen der Verteilungen um. Zum Beispiel ist bei den extrem unbalancierten Fällen mit sehr hoher Accuracy (0.95) das empirische Niveau der ATS höher als das der WTS. Auch in diesem Fall liegt das ungewöhnliche Verhalten der WTS daran, dass in den meisten Simulationsschritten die Kovarianzmatrix singulär wird. Bei einer simulierten Accuracy von 0.95 ist die Determinante der Kovarianzmatrix immer kleiner als 10^{-10}, also verschwindend gering. Bedingt man die Simulationsergebnisse der WTS darauf, dass die Kovarianzmatrix nicht singulär ist, so erhält man wiederum nicht das eigentliche empirische Niveau, sondern nur ein bedingtes. Somit ist abschließend zu sagen, dass man die WTS bei Flächen ab 0.8 nicht einsetzen sollte, da die Kovarianzmatrix hier bereits in 50% der Fälle singulär wird.

Tabelle 6.3: Verschiedene Flächen bei unbalancierten Designs, *kursive* Ziffern stehen für bedingte empirische Niveaus

		Anzahl der		Durchschnittliche Größe des Effekts $\widehat{p}_i$				
n_0	n_1	Faktorstufen	Statistik	0.5	0.7	0.85	0.9	0.95
15	30	2	ATS	0.0556	0.0517	0.0484	0.0401	0.0230
15	30	2	WTS	0.0556	0.0516	*0.0451*	*0.0425*	-
15	30	5	ATS	0.0393	0.0415	0.0308	0.0236	0.0065
15	30	5	WTS	0.0850	0.0831	*0.0536*	*0.0313*	-
15	30	10	ATS	0.0397	0.0355	0.0312	0.0235	0.0078
15	30	10	WTS	0.0859	0.0746	*0.0453*	*0.0447*	-
30	15	2	ATS	0.0564	0.0496	0.0418	0.0405	0.0233
30	15	2	WTS	0.0564	0.0495	*0.0450*	*0.0299*	-
30	15	5	ATS	0.0436	0.0377	0.0296	0.0226	0.0071
30	15	5	WTS	0.0850	0.0780	*0.0559*	*0.0299*	-
30	15	10	ATS	0.0411	0.0395	0.0311	0.0279	0.0084
30	15	10	WTS	0.0832	0.0820	*0.0445*	*0.0256*	-
20	100	2	ATS	0.0623	0.0589	0.0551	0.0522	0.0388
20	100	2	WTS	0.0623	0.0589	*0.0533*	*0.0430*	-
20	100	5	ATS	0.0465	0.0442	0.0316	0.0302	0.0157
20	100	5	WTS	0.0978	0.0924	*0.0732*	*0.0662*	-
20	100	10	ATS	0.0422	0.0369	0.0373	0.0323	0.0190
20	100	10	WTS	0.0964	0.0891	*0.0715*	*0.0590*	-
100	20	2	ATS	0.0571	0.0575	0.0555	0.0518	0.0383
100	20	2	WTS	0.0571	0.0575	*0.0554*	*0.0522*	-
100	20	5	ATS	0.0457	0.0436	0.0367	0.0331	0.0164
100	20	5	WTS	0.1023	0.0912	*0.0732*	*0.0630*	-
100	20	10	ATS	0.0438	0.0440	0.0382	0.0337	0.0176
100	20	10	WTS	0.1001	0.0902	*0.0719*	*0.0536*	-
100	200	2	ATS	0.0523	0.0505	0.0504	0.0497	0.0452
100	200	2	WTS	0.0523	0.0505	*0.0506*	*0.0455*	-
100	200	5	ATS	0.0466	0.0503	0.0476	0.0443	0.0402
100	200	5	WTS	0.0542	0.0544	*0.0484*	*0.0758*	-
100	200	10	ATS	0.0505	0.0509	0.0484	0.0472	0.0405
100	200	10	WTS	0.0570	0.0573	*0.0521*	*0.0455*	-
100	500	2	ATS	0.0517	0.0491	0.0486	0.0473	0.0475
100	500	2	WTS	0.0517	0.0491	*0.0486*	*0.0483*	-
100	500	5	ATS	0.0509	0.0472	0.0493	0.0432	0.0361
100	500	5	WTS	0.0625	0.0556	*0.0556*	*0.0508*	-
100	500	10	ATS	0.0549	0.0490	0.0524	0.0459	0.0422
100	500	10	WTS	0.0653	0.0597	*0.0606*	*0.0513*	-

Srivastava (2005) schlägt eine Methode zur Korrektur der Schätzung des Freiheitsgrades vor, die auf Dempster (1958) zurückgeht. Hierbei wird die Verzerrung in der Schätzung des Nenners $Sp(\mathbf{TVTV})$ des Freiheitsgrades korrigiert. Diese Methode wird für die Analyse von hochdimensionalen Daten vorgeschlagen, bei der die ATS auch sehr konservativ wird. Im hier vorliegenden Design beseitigt diese Korrektur die Konservativität leider gar nicht, deshalb werden die Simulationsergebnisse auch nicht dargestellt.

6.2.2 Modell 3

Im Modell 3 wurden vier verschiedene typische Designs für die Simulationen ausgewählt.

1. pro Patient eine gesunde und eine erkrankte Beobachtung,
2. pro Patient eine gesunde und zwei erkrankte Beobachtungen,
3. pro Patient 6 Beobachtungen mit zufälliger Zuordnung zur Gruppe der Gesunden oder Kranken.
4. eine zufällige Anzahl von Beobachtungen zwischen 5 und 15 pro Patient mit zufälliger Zuordnung zu gesund und krank

Design 1 könnte zum Beispiel die Brustkrebsstudie sein, in der die gesunde Brust jeweils als Kontrolle für die erkrankte Brust dient. Beispielhaft für Design 2 ist die Knöcheldickenbestimmung von zwei arthritischen Fingergelenken, bei der von jedem Patienten außerdem ein gesundes Gelenk zum Vergleich gemessen wird. Design 3 steht für einen Versuch, in dem 6 bestimmte Lymphgefäße untersucht werden, bei denen der Gesundheitsstatus vorher nicht bekannt ist. Design 4 steht schließlich für eine Studie, in der alle Leberflecken eines Patienten am linken Arm auf Melanome untersucht werden. Hier ist im Vorfeld weder die Anzahl noch die Verteilung auf die beiden Gesundheitszustände bekannt.
Das erste Design ist balanciert, während die drei anderen unbalanciert sind. Als Stichprobengrößen wurden 50 und 100 Patienten gewählt. Die abhängigen Daten wurden mit Korrelationen von 0, 0.5 und 0.8 simuliert. Die Verschiebung zwischen gesunden und kranken Beobachtungen wurde so gewählt, dass sich folgende Flächen ergaben:
Um das Verhalten der Teststatistiken bei verschiedenen Faktorstufenanzahlen zu untersuchen, wurden hier zwei, sechs und zehn Faktorstufen simuliert. Tabelle 6.5 zeigt die Simulationsergebnisse, man sieht hier bereits im einfaktoriellen Design sehr gut, dass die WTS der ATS in allen Konstellationen unterlegen ist. Die Unterschiede zwischen der gewichteten und der ungewichteten Teststatistik dagegen sind vernachlässigbar gering, die Übersichtstabelle wird deshalb nur im Anhang (Seite 92) dargestellt.

Tabelle 6.4: Größe der AUC bei verschiedenen Korrelationen und Größen der Verschiebung

c	$a = 0$	$a = 1$	$a = 2$	$a = 3$
0	0.5	0.66	0.8	0.9
1	0.5	0.63	0.76	0.86
2	0.5	0.6	0.7	0.8

Zusammenfassend kann man sagen, dass die ATS fast immer ein niedrigeres empirisches Niveau als die WTS hat. Außerdem ist die WTS immer sehr liberal, erst bei sehr großen Stichprobenumfängen und moderaten Faktorstufenzahlen hält sie das Niveau wirklich ein. Dagegen wird die ATS nur in geringem Maße konservativ.

6.2.3 Dichotome Daten

Um das Verhalten der vorgeschlagenen Methodik für dichotome Daten zu untersuchen, wurden Simulationen für die Modelle 1 und 3 bei dichotomen Daten durchgeführt. Es wurden niedrige (AUC=0.65) und hohe (AUC=0.9) Accuracies simuliert. Außerdem wurden verschiedene gleiche und unterschiedliche Stichprobenumfänge zwischen 10 und 100 betrachtet. Bei den Simulationen wird das Verhalten unter Hypothese für die ANOVA-Typ Statistik mit der Wald-Typ-Statistik verglichen. Da es für die Trefferquote keine etablierten Testverfahren gibt, wird diese nicht mitsimuliert.
Zunächst wird das Verhalten im einfachen Multireader-Multimethoden Modell (also in Modell 1) betrachtet. Als Faktorstruktur wird beispielhaft der Vergleich zweier Methoden mit 5 Readern betrachtet, somit hat man für die Wechselwirkung 10 Faktorstufen. Dies reflektiert eine realistische Situation, außerdem hat man so direkt den Vergleich zwischen 2, 5 und 10 Faktorstufen. In Tabelle 6.7 ist deutlich zu sehen, dass die ATS auch bei dichotomen Daten sehr viel besser abschneidet als die WTS. Außerdem ist zu sehen, dass die ATS das Niveau auch bei moderaten Stichprobenumfängen gut einhält und erst bei sehr unbalancierten Designs und für große AUC (also am Rand der Verteilung) zunehmend konservativ wird. Diese Eigenschaft ist aber allen Tests bei dichotomen Daten gemeinsam. Hierfür wurden bereits diverse Korrekturmöglichkeiten vorgeschlagen, eine Zusammenfassung findet sich in Newcombe (2006).
Weiterhin wurden Simulationen für Modell 3, also für clustered data, durchgeführt. Hier wurde nur das einfaktorielle Modell betrachtet, aber Faktorstufenanzahlen von 2, 6 und 10 betrachtet. Die Clusteranzahlen in den verschiedenen Simulationen wurden mit 50 und 100 festgelegt. Verschiedene AUCs wurden durch die Verschiebung der Verteilungen durch Faktor a erreicht: je größer a, desto größer die Accuracy. Verschiedene Korrelationen innerhalb der Cluster werden wider durch c wiederge-

Tabelle 6.5: Clustered data, Niveau-Simulationen für Design 1 und 2 (für 2 Faktorlevel wurde die WTS weggelassen)

			$n = 50$				$n = 100$			
Level	c	Test	$a = 0$	$a = 1$	$a = 2$	$a = 3$	$a = 0$	$a = 1$	$a = 2$	$a = 3$
Design 1										
2	0	ATS	0.0578	0.0535	0.0553	0.0503	0.0531	0.0564	0.0553	0.0521
2	1	ATS	0.0543	0.0503	0.0563	0.0547	0.0541	0.0516	0.0523	0.0515
2	2	ATS	0.0467	0.0482	0.0536	0.0548	0.0504	0.0541	0.0485	0.0527
6	0	ATS	0.0498	0.0494	0.0408	0.0385	0.0489	0.0484	0.049	0.0426
6	0	WTS	0.0959	0.0951	0.0962	0.1159	0.0734	0.0725	0.0749	0.0798
6	1	ATS	0.0451	0.0448	0.0427	0.0424	0.0467	0.0467	0.0474	0.0451
6	1	WTS	0.0911	0.0916	0.094	0.1074	0.067	0.0668	0.073	0.0759
6	2	ATS	0.0406	0.038	0.0385	0.0438	0.0411	0.0469	0.0467	0.0461
6	2	WTS	0.0737	0.0731	0.0788	0.1007	0.0557	0.0649	0.0667	0.0693
10	0	ATS	0.0438	0.0411	0.0389	0.0332	0.0469	0.0488	0.0413	0.04
10	0	WTS	0.1636	0.1588	0.1672	0.1932	0.0947	0.1007	0.0992	0.1141
10	1	ATS	0.0401	0.0379	0.0392	0.0339	0.0484	0.0444	0.0408	0.0449
10	1	WTS	0.1481	0.1481	0.1619	0.1873	0.0979	0.0874	0.0935	0.1145
10	2	ATS	0.0365	0.0334	0.0339	0.035	0.0385	0.0425	0.0446	0.0417
10	2	WTS	0.1117	0.1183	0.1367	0.1604	0.0688	0.0794	0.0909	0.0967
Design 2										
2	0	ATS	0.0561	0.0557	0.0584	0.0532	0.0556	0.0503	0.0498	0.0485
2	1	ATS	0.0546	0.0556	0.0547	0.0552	0.0527	0.0526	0.0547	0.0521
2	2	ATS	0.0509	0.0544	0.0511	0.0593	0.0516	0.0529	0.0536	0.0547
6	0	ATS	0.05	0.043	0.0478	0.0409	0.0483	0.0449	0.0473	0.0427
6	0	WTS	0.0961	0.0942	0.097	0.1	0.0703	0.0705	0.0719	0.0726
6	1	ATS	0.0442	0.0458	0.0429	0.0425	0.0485	0.0453	0.0471	0.049
6	1	WTS	0.0834	0.092	0.0912	0.0983	0.07	0.0637	0.066	0.0727
6	2	ATS	0.0392	0.0389	0.0436	0.044	0.0454	0.0459	0.0464	0.0507
6	2	WTS	0.0702	0.0763	0.0846	0.0934	0.0612	0.0639	0.0668	0.0705
10	0	ATS	0.0454	0.0406	0.0378	0.0355	0.0436	0.043	0.0432	0.0442
10	0	WTS	0.159	0.1649	0.1648	0.1711	0.094	0.094	0.0988	0.1014
10	1	ATS	0.0385	0.0394	0.0386	0.0372	0.0444	0.0467	0.0419	0.0431
10	1	WTS	0.1453	0.1518	0.162	0.1692	0.0924	0.0962	0.0949	0.1012
10	2	ATS	0.0307	0.0331	0.0353	0.0391	0.039	0.0424	0.0429	0.0441
10	2	WTS	0.1078	0.1222	0.1348	0.1566	0.073	0.0788	0.0828	0.089

Tabelle 6.6: Clustered data. Niveau-Simulationen für Design 3 und 4 (für 2 Faktorlevel wurde die WTS weggelassen)

			$n = 50$				$n = 100$			
Level	c	Test	$a = 0$	$a = 1$	$a = 2$	$a = 3$	$a = 0$	$a = 1$	$a = 2$	$a = 3$
Design 3										
2	0	ATS	0.058	0.0563	0.0568	0.068	0.0523	0.0548	0.0569	0.0595
2	1	ATS	0.0512	0.0544	0.0596	0.0592	0.0494	0.0526	0.0514	0.0536
2	2	ATS	0.0535	0.0521	0.0487	0.0544	0.0494	0.052	0.0515	0.0518
6	0	ATS	0.0458	0.0446	0.0487	0.0597	0.0457	0.0478	0.0459	0.0631
6	0	WTS	0.0926	0.092	0.1014	0.1372	0.0638	0.0689	0.0731	0.0991
6	1	ATS	0.0427	0.0459	0.0454	0.0476	0.0431	0.0467	0.0485	0.055
6	1	WTS	0.0876	0.0917	0.0979	0.1071	0.0658	0.0666	0.0725	0.0822
6	2	ATS	0.0395	0.0391	0.0466	0.0497	0.044	0.0421	0.0495	0.0527
6	2	WTS	0.0755	0.0781	0.0899	0.1019	0.0583	0.0595	0.0701	0.077
10	0	ATS	0.0336	0.0389	0.0402	0.0594	0.0444	0.0444	0.0486	0.0668
10	0	WTS	0.15	0.1548	0.173	0.2336	0.0916	0.0933	0.0995	0.147
10	1	ATS	0.0373	0.0381	0.0409	0.048	0.0445	0.0433	0.0445	0.0565
10	1	WTS	0.1464	0.1549	0.1713	0.2042	0.0903	0.0929	0.1018	0.1171
10	2	ATS	0.0297	0.0309	0.0354	0.0388	0.0385	0.0388	0.0396	0.0428
10	2	WTS	0.1069	0.1239	0.1403	0.1695	0.0704	0.0721	0.0867	0.1036
Design 4										
2	0	ATS	0.0501	0.0513	0.0572	0.0683	0.0496	0.0532	0.0593	0.064
2	1	ATS	0.0528	0.0535	0.0561	0.0601	0.0494	0.0524	0.0559	0.0542
2	2	ATS	0.0495	0.0479	0.0521	0.0564	0.0499	0.0487	0.0513	0.0523
6	0	ATS	0.0418	0.0472	0.0504	0.0604	0.0448	0.0459	0.0559	0.0662
6	0	WTS	0.0863	0.0956	0.1044	0.1354	0.0671	0.0671	0.0791	0.1027
6	1	ATS	0.0412	0.0448	0.0512	0.0521	0.0448	0.0472	0.0484	0.0606
6	1	WTS	0.0811	0.0864	0.1049	0.1133	0.0634	0.0688	0.0729	0.0897
6	2	ATS	0.0356	0.0413	0.0472	0.0494	0.0432	0.044	0.0454	0.051
6	2	WTS	0.0714	0.0764	0.0842	0.0974	0.0575	0.0611	0.0665	0.0741
10	0	ATS	0.0356	0.0386	0.0413	0.0597	0.0435	0.0426	0.0461	0.0598
10	0	WTS	0.1461	0.1618	0.1642	0.2413	0.0886	0.0909	0.1027	0.1401
10	1	ATS	0.0381	0.0379	0.0406	0.0477	0.0426	0.0399	0.0457	0.0575
10	1	WTS	0.1402	0.1509	0.1632	0.1917	0.0898	0.0885	0.1006	0.1245
10	2	ATS	0.0319	0.0366	0.0331	0.0375	0.04	0.0412	0.0436	0.0447
10	2	WTS	0.1211	0.1243	0.1403	0.1701	0.0778	0.0764	0.0864	0.0997

Tabelle 6.7: Ergebnisse von 10000 Simulationen, zweifaktorielles Design, binomiale Bewertungsdaten (Modell 1)

			AUC=0.65		AUC=0.9	
n_0	n_1	Faktorstufen	ATS	WTS	ATS	WTS
15	15	5	0.0502	0.1127		
		2	0.0618	0.0618		
		10	0.0502	0.1142		
30	30	5	0.0504	0.0738	0.0476	0.0696
		2	0.0543	0.0543	0.0510	0.0510
		10	0.0470	0.0724	0.0448	0.0741
50	50	5	0.0482	0.0584	0.0470	0.0654
		2	0.0554	0.0555	0.0535	0.0535
		10	0.0490	0.0609	0.0443	0.0674
100	100	5	0.0481	0.0515	0.0506	0.0566
		2	0.0478	0.0478	0.0498	0.0498
		10	0.0474	0.0530	0.0498	0.0585
15	30	5	0.0513	0.1099	0.0461	0.0913
		2	0.0634	0.0635	0.0588	0.0588
		10	0.0521	0.1107	0.0414	0.0973
30	15	5	0.0512	0.0886		
		2	0.0587	0.0589		
		10	0.0488	0.0940		
20	100	5	0.0494	0.1095	0.0441	0.0864
		2	0.0620	0.0620	0.0568	0.0568
		10	0.0479	0.1083	0.0414	0.0897
100	20	5	0.0468	0.0904	0.0388	0.0638
		2	0.0576	0.0577	0.0522	0.0524
		10	0.0465	0.0936	0.0356	0.0637

geben: je größer c, desto größer die Korrelation innerhalb des Clusters, wobei $c = 2$ einer Korrelation von $\rho = 0.8$ entspricht und $c = 1$ einer Korrelation von $\rho = 0.5$ entspricht. Wie vorne gezeigt, hängt die Größe der simulierten Fläche von der Korrelation und der Verschiebung ab. Es gelten wieder ähnliche Zusammenhänge wie bei den ordinalen Daten im Modell 3 (siehe Tabelle 6.4).
Die Ergebnisse der Simulationen sind in Tabelle 6.8 dargestellt. Es wiederholt sich das gleiche Schema wie oben bereits gesehen: die WTS wird bei kleineren Stichproben und höheren Faktorstufenanzahlen zunehmend liberal, während die ATS nur in geringem Maße konservativ wird.

6.3 Powersimulationen

Um das Verhalten der vorgestellten Teststatistiken unter Alternative zu untersuchen, wurden Powersimulationen durchgeführt. Als erste Alternative wurde eine ansteigende Folge von AUCs gewählt, deren Steigung durch einen Parameter δ charakterisiert werden kann: der Verschiebungsvektor

$$(0, 1, 2, 3, ..., d)',$$

der auf die simulierten Verteilungen addiert wird, wird jeweils mit einem ansteigenden δ multipliziert.
Weitere Simulationen wurden mit einer „Ein-Punkt-Alternative“ durchgeführt, hierbei diente

$$(0, 0, 0, 0, ..., 1)'$$

als Verschiebungsvektor. Die Ergebnisse dieser Simulationen zeigten aber keine Unterschiede zum ersten Setting und werden deshalb nicht graphisch dargestellt.

Ein direkter Vergleich der Power von WTS und ATS ist nicht gut möglich, da die WTS das gewählte Niveau in den meisten Fällen enorm überschreitet. Deshalb könnte man in einem weiteren Simulationsschritt die empirische Verteilung und somit die empirischen 95%-Quantile der WTS bestimmen. Mithilfe dieser bestimmt man dann die Power der WTS, die sie bei Niveaueinhaltung unter Hypothese hätte. Diese Adjustierung des Niveaus der WTS führt allerdings dazu, dass die ATS wiederum relativ schlecht abschneidet. Denn in den meisten Bereichen, in denen die WTS sehr liberal wird, ist die ATS zumindest leicht konservativ. Insgesamt haben die adjustierten Powersimulationen aber gezeigt, dass die Power für ATS und WTS vergleichbar ist. Es sind keine großen Unterschiede zu beobachten gewesen. Da die Adjustierung in einer Verschiebung der Powerkurven resultiert, ist dieses Ergebnis optisch auch schon an den Originalpowerkurven zu beobachten: sie weisen keine nennenswerten Unterschiede im Verlauf auf, die Kurven der WTS sind lediglich nach oben verschoben.
In Abbildung 6.2 ist exemplarisch eine Kurve dargestellt. Das zugrunde liegende

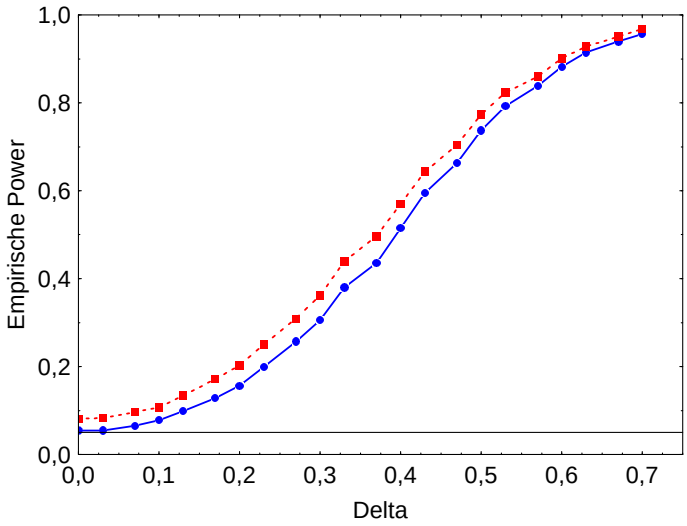

Abbildung 6.2: Exemplarische Powerfunktion für ATS (durchgezogene Linie) und WTS (gestrichelte Linie) im Design 3 für sechs Faktorlevel, $a = 1.5$, $\rho = 0.5$

Design 3 wurde mit sechs Faktorstufen, einer Verschiebung von 1.5 und einer Korrelation von 0.5 simuliert. Es ist deutlich zu sehen, dass die beiden Kurvenverläufe sehr ähnlich sind, der Hauptunterschied liegt im verschobenen Beginn der Kurven. Im Appendix sind außerdem verschiedene Powerkurven für die Designs 1-3 dargestellt (ab Seite 93), wobei die Faktorstufen, Verschiebungen, Korrelationen und Stichprobenumfänge variiert wurden. Die Power wurde hier für die ATS und die WTS bestimmt. Im wesentlichen zeigen alle Kurven, dass die beiden Statistiken das gleiche Powerverhalten haben.

Tabelle 6.8: Ergebnisse von 10000 Simulationen, einfaktorielles Design, binomiale clustered Bewertungsdaten (Modell 3)

AUC	Faktorstufen	ρ	n=50		n=100	
			ATS	WTS	ATS	WTS
0	2	1	0.0520	0.0521	0.0507	0.0508
0	2	2	0.0552	0.0552	0.0490	0.0490
0	2	3	0.0527	0.0527	0.0562	0.0562
2	2	1	0.0549	0.0549	0.0509	0.0510
2	2	2	0.0550	0.0550	0.0524	0.0524
2	2	3	0.0526	0.0526	0.0513	0.0514
3	2	1	0.0563	0.0564	0.0506	0.0506
3	2	2	0.0581	0.0581	0.0523	0.0523
3	2	3	0.0552	0.0552	0.0544	0.0544
0	6	1	0.0458	0.0909	0.0476	0.0658
0	6	2	0.0460	0.0889	0.0476	0.0718
0	6	3	0.0392	0.0827	0.0460	0.0655
2	6	1	0.0457	0.0918	0.0456	0.0647
2	6	2	0.0445	0.0873	0.0491	0.0705
2	6	3	0.0400	0.0821	0.0420	0.0644
3	6	1	0.0471	0.0891	0.0510	0.0684
3	6	2	0.0440	0.0856	0.0471	0.0692
3	6	3	0.0414	0.0834	0.0456	0.0676
0	10	1	0.0409	0.1579	0.0451	0.0939
0	10	2	0.0365	0.1470	0.0435	0.0919
0	10	3	0.0346	0.1476	0.0393	0.0851
2	10	1	0.0386	0.1505	0.0447	0.0938
2	10	2	0.0347	0.1483	0.0438	0.0914
2	10	3	0.0327	0.1373	0.0407	0.0849
3	10	1	0.0381	0.1509	0.0418	0.0915
3	10	2	0.0358	0.1461	0.0452	0.0903
3	10	3	0.0328	0.1442	0.0413	0.0912

7 Zusammenfassung und Ausblick

Diagnostische Studien mit clustered data sind bisher nur von Obuchowski (1997) untersucht worden. Dort wird lediglich der Vergleich von zwei Stichproben, also von zwei Readern oder zwei Methoden, betrachtet. Die vorliegende Arbeit bietet dagegen ein Verfahren, welches auf mehrfaktorielle Designs mit beliebig vielen Faktoren und Faktorstufen angewendet werden kann.
Es wurde gezeigt, dass die nichtparametrischen Methoden zur Auswertung des Behrens-Fisher-Problems in der Diagnostik angewendet werden können, wobei wir insbesondere drei Modelle unterschieden haben. Das einfache Multi-Reader Multi-Methoden Design (Modell 1) ist eine direkte Anwendung des multivariaten nichtparametrischen Behrens-Fisher-Problems, wie es in Brunner *et al.* (2002) vorgestellt wurde. Die explizite Anwendung auf diagnostische Studien wurde in Kaufmann *et al.* (2005) beschrieben.
Die Erweiterung der Theorie auf verbundene Messwiederholungen (Modell 2) wurde in Analogie zu der Arbeit von Brunner *et al.* (1999) durchgeführt. Dort wurde für das allgemeine nichtparametrische Modell (mit der Hypothese $F_1 = F_2$) die Verwendung beliebiger Messwiederholungen vorgestellt. Diese Techniken wurden hier auf Diagnosestudien übertragen.
Schließlich haben wir die Theorie des multivariaten Behrens-Fisher-Problems auf clustered data (Modell 3) erweitert. Die zusätzliche Einführung einer Abhängigkeit zwischen Gesunden und Kranken (entspricht in der Theorie des multivariaten Behrens-Fisher-Problems den zwei Gruppen) wurde in Werner & Brunner (2006) hergeleitet. Die Vorteile der ungewichteten Schätzung des Effektes gegenüber der gewichteten Schätzung wurden aufgezeigt.
Test-Statistiken und Konfidenzintervalle wurden für die Theorie entwickelt, sodass nun für alle drei Modelle entsprechende statistische Methoden zur Verfügung stehen. Das Verhalten der vorgestellten Verfahren bei der Anwendung auf dichotome Testergebnisse wurde untersucht und es konnte gezeigt werden, dass die Fläche unter der ROC-Kurve auch in dieser Situation ein sinnvolles, globales Maß darstellt.
An verschiedenen Beispielen wurde die Anwendung und Interpretation der Methoden dargestellt und verglichen. Außerdem wurden SAS-Makros entwickelt, die die einfache Verwendung der statistischen Methoden ermöglichen.
Schließlich wurde das Verhalten der entwickelten Theorie für kleine Stichproben untersucht. Es wurde gezeigt, dass die ATS auch in Bereichen, in denen die WTS aufgrund singulärer Kovarianzmatrizen nicht mehr anwendbar ist, immer noch akzeptable empirische Niveaus aufweist. Dies ist speziell der Fall am Rand der Vertei-

lung, also bei hohen Accuracies ab 0.8. Nur im Bereich der extrem hohen Accuracies (0.95) ist auch die Approximation der ATS bei Stichproben unter 100 nicht mehr zufriedenstellend.

Ausblick

Es sollte untersucht werden, ob geeignete Transformationen existieren, die die Probleme der Approximation bei sehr hohen Accuracies beheben können. Eventuell bietet die *logit*-Transformation, die bereits für die Konfidenzintervalle vorgestellt wurde, eine Lösung.
Außerdem wäre es interessant zu untersuchen, wie sich die Verfahren auf Teilflächen unter der ROC-Kurve anwenden lassen. Es gibt Studien, in denen von vornherein klar ist, dass der entsprechende diagnostische Test eine Mindestsensitivität von 90% haben muss. Dann macht es natürlich keinen Sinn, Gesamtflächen zu berechnen und zu vergleichen.
Weiterhin ist zu prüfen, wie die Verfahren auf fehlende Werte erweitert werden können. Hiermit sollen jetzt die fehlenden Werte gemeint sein, die dadurch entstehen, dass Aufnahmen manchmal nicht lesbar sind, oder dass einige Reader nicht alle Bilder bewertet haben. Das sind dann Missings, die nur in einzelnen Reader-Methoden-Kombinationen vorkommen. Es ist zu erwarten, dass die Theorie, ähnlich wie im allgemeinen nichtparametrischen Modell, auf diese Situationen auszuweiten ist. Allerdings muss dann vermutlich vorausgesetzt werden, dass die Werte vollständig zufällig (MCAR) fehlen. Wenn eine Aufnahme zum Beispiel deshalb nicht lesbar ist, weil das Verfahren Bilder von einer geringen Qualität liefert, ist diese Annahme sicherlich verletzt. Es sollte jedoch im allgemeinen möglich sein, die MCAR-Annahme zu machen.
Es wäre auch interessant zu überprüfen, wie sich die vorgestellten Verfahren mit den bereits bestehenden Lösungsansätzen für das Problem des fehlenden Goldstandards kombinieren lassen. Das Problem des fehlenden Goldstandards wird bei Zhou & Castelluccio (2003) mithilfe einer logistischen Regression modelliert. Die Vorgehensweise soll helfen, den „verification bias“ zu vermeiden. Um zu zeigen, dass die ML-Verfahren auch bei clustered data anwendbar sind, müssen die Algorithmen auf ihr Verhalten bei clustered data untersucht werden.
Schließlich könnte man die ROC-Oberfläche bei clustered data betrachten, die man erhält, wenn der Goldstandard nicht dichotom, sondern ordinal oder nominal ist. In einer der vorgestellten Methoden (Obuchowski *et al.* , 2001; Obuchowski, 2005) werden gewichtete Mittelwerte aus allen paarweisen AUCs gebildet. Dieses Verfahren wäre somit direkt auf die Schätzer der vorliegenden Arbeit anwendbar. Allerdings vernachlässigt das Verfahren die multivariate Natur der ROC-Oberfläche. Nakas & Yiannoutsos (2004) dagegen verwenden die multivariate Information, indem sie das Volumen unter der ROC-Oberfläche schätzen. Dieses Verfahren lässt sich jedoch nicht so direkt auf clustered data erweitern wie die Idee von Obuchowski.

A Appendix

A.1 Definitionen

Asymptotische Äquivalenz

Zwei Folgen von Zufallsvariablen Y_N und Z_N sind asymptotisch äquivalent, wenn $Y_N - Z_N \xrightarrow{p} 0$ für $N \to \infty$ gilt. Es ist zumeist einfacher, das stärkere Resultat $E(Y_N - Z_N)^2 \to 0$ zu zeigen. Asymptotische Äquivalenz impliziert asymptotische Verteilungsgleichheit.

Zählfunktion

Die hier verwendete normierte Version der Zählfunktion $c : \mathbb{R} \to \mathbb{R}$ ist definiert durch:

$$c(x) = \begin{cases} 0 & : x < 0 \\ \frac{1}{2} & : x = 0 \\ 1 & : x > 0 \end{cases} .$$

Kroneckersumme und -Produkt

1. Für beliebige Matrizen $\mathbf{A}$ und $\mathbf{B}$ heißt $\mathbf{A} \oplus \mathbf{B} = \left(\begin{array}{c|c} \mathbf{A} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{B} \end{array}\right)$ **Kronecker-Summe** von $\mathbf{A}$ und $\mathbf{B}$.

2. Für beliebige Matrizen $\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$ und $\mathbf{B} = \begin{pmatrix} b_{11} & \cdots & b_{1q} \\ \vdots & & \vdots \\ b_{p1} & \cdots & b_{pq} \end{pmatrix}$ heißt $\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix}_{mp \times nq}$ **Kronecker-Produkt** von $\mathbf{A}$ und $\mathbf{B}$.

A.2 Beweise

Beweis von Theorem 3.4

Dieser Beweis orientiert sich an dem von Theorem 3.3 in Brunner *et al.* (2002). Es reicht aus, den Beweis elementweise für die Elemente $p^{(l)}$ von $\mathbf{p}$ durchzuführen. Der Index l wird zur besseren Lesbarkeit weggelassen.
Es ist allerdings zu berücksichtigen, dass der Schätzer $\widehat{p}$, der über die Differenz der ungewichteten Rangmittelwerte gebildet wird, nicht direkt als $\int \widehat{F}_0 d\widehat{F}_1$ mit

$$\widehat{F}_i(x) = \frac{1}{n_i} \sum_{k=1}^{n_i} \frac{1}{m_{ik}} \sum_{s=1}^{m_{ik}} c(x - X_{iks}).$$

darstellbar ist, da die empirische Verteilungsfunktion $\widehat{H}(x)$ nur die gewichteten empirischen Verteilungsfunktionen beinhaltet, die mit $\widetilde{F}(x)$ bezeichnet werden:

$$\widehat{H}(x) = \frac{1}{N} \sum_{i=0}^{1} \sum_{k=1}^{n_i} \sum_{s=1}^{m_{ik}} c(x - X_{iks}) = \frac{1}{N}(m_0 \widetilde{F}_0(x) + m_1 \widetilde{F}_1(x))$$

mit

$$\widetilde{F}_i(x) = \frac{1}{m_i} \sum_{k=1}^{n_i} \sum_{s=1}^{m_{ik}} c(x - X_{iks}).$$

Diese Summe lässt sich offensichtlich nicht als Summe der $\widehat{F}(x)$ darstellen.
Zunächst wird nun $\widehat{p} - p$ zerlegt:

$$\begin{aligned}
(\widehat{p} - p) &= \int \widehat{H} d\widehat{F}_1 - \int \widehat{H} d\widehat{F}_0 - \int H dF_1 + \int H dF_0 \\
&= \int \frac{m_0}{N} \widetilde{F}_0 d\widehat{F}_1 - \int \frac{m_0}{N} F_0 dF_1 + \int \frac{m_1}{N} \widetilde{F}_1 d\widehat{F}_1 - \int \frac{m_1}{N} F_1 dF_1 \\
&\quad - \int \frac{m_0}{N} \widetilde{F}_0 d\widehat{F}_0 + \int \frac{m_0}{N} F_0 dF_0 - \int \frac{m_1}{N} \widetilde{F}_1 d\widehat{F}_0 + \int \frac{m_1}{N} F_1 dF_0 \\
&\quad + \int F_0 d\widehat{F}_1 - \int F_1 d\widehat{F}_0 - \int F_0 d\widehat{F}_1 + \int F_1 d\widehat{F}_0 \\
&= \int F_0 d\widehat{F}_1 - \int F_1 d\widehat{F}_0 + 1 - 2p \\
&\quad \underbrace{+\frac{m_0}{N} \int (\widetilde{F}_0 - F_0) d\widehat{F}_1 - \frac{m_0}{N} \int (\widehat{F}_0 - F_0) dF_1}_{A_1} \\
&\quad \underbrace{+\frac{m_0}{N} \int F_0 dF_0 - \frac{m_0}{N} \int \widetilde{F}_0 d\widehat{F}_0}_{A_2}
\end{aligned}$$

$$\underbrace{-\frac{m_1}{N}\int(\widetilde{F}_1-F_1)d\widehat{F}_0+\frac{m_1}{N}\int(\widehat{F}_1-F_1)dF_0}_{A_3}$$
$$\underbrace{-\frac{m_1}{N}\int F_1dF_1+\frac{m_1}{N}\int\widetilde{F}_1d\widehat{F}_1}_{A_4}$$
$$= B_N+A_1+A_2+A_3+A_4.$$

Hier steht in der ersten Zeile der letzten Gleichung genau B_N, für die Terme A_1-A_4 in der zweiten bis fünften Zeile muss die asymptotische Konvergenz gezeigt werden. Dafür werden die Terme mithilfe der c_r-Ungleichung umgeformt. Da A_1 und A_3 bzw. A_2 und A_4 jeweils symmetrisch in F_0 und F_1 sind, genügt es, die Konvergenz nur für A_1 und A_2 zu zeigen.
Im Folgenden bezeichne $\overline{m}_i = m_i/n_i$ den Mittelwert der Clustergrößen für $i = 0, 1$. A_1 muss zunächst nochmals umgeschrieben werden:

$$\begin{aligned}
&E(\sqrt{N}A_1)^2\\
&= E\left(\frac{m_0}{\sqrt{N}}\int(\widetilde{F}_0-F_0)d\widehat{F}_1-\frac{m_0}{\sqrt{N}}\int(\widehat{F}_0-F_0)dF_1\right)^2\\
&= \frac{m_0^2}{N}E\left(\frac{1}{n_1}\sum_{k=1}^{n_1}\frac{1}{m_{1k}}\sum_{s=1}^{m_{1k}}\left[\widetilde{F}_0(X_{1ks})-F_0(X_{1ks})-\int\left(\widehat{F}_0(x)-F_0(x)\right)dF_1\right]\right)^2\\
&= \frac{m_0^2}{N}E\left(\frac{1}{n_1}\sum_{k=1}^{n_1}\frac{1}{m_{1k}}\sum_{s=1}^{m_{1k}}\left[\left(\frac{1}{m_0}\sum_{k'=1}^{n_0}\sum_{s'=1}^{m_{0k'}}(c(X_{1ks}-X_{0k's'})-F_0(X_{1ks}))\right)\right.\right.\\
&\qquad\left.\left.-\left(\frac{1}{n_0}\sum_{k'=1}^{n_0}\frac{1}{m_{0k'}}\sum_{s'=1}^{m_{0k'}}\int(c(x-X_{0k's'})-F_0(x))\,dF_1\right)\right]\right)^2\\
&= \frac{m_0^2}{N}E\left(\frac{1}{n_1n_0}\sum_{k=1}^{n_1}\frac{1}{m_{1k}}\sum_{s=1}^{m_{1k}}\left[\sum_{k'=1}^{n_0}\sum_{s'=1}^{m_{0k'}}\frac{1}{\overline{m}_0}(c(X_{1ks}-X_{0k's'})-F_0(X_{1ks}))\right.\right.\\
&\qquad\left.\left.-\frac{1}{m_{0k'}}\int(c(x-X_{0k's'})-F_0(x))\,dF_1\right]\right)^2.
\end{aligned}$$

Wenn man das Quadrat ausmultipliziert, entstehen gemischte Terme, deren Erwartungswert immer genau dann 0 ist, wenn einer der Indizes k von den anderen drei verschieden ist. Denn dann kann man mithilfe von Fubinis Theorem die Integrationsreihenfolge vertauschen. Der Erwartungswert der einzelnen Summanden ist dann genau 0, denn man bildet immer die Differenz zwischen Schätzer und Erwartungswert. Das heißt, es bleiben nur die Summanden übrig, bei denen die Indizes entweder paarweise gleich oder alle vier gleich sind. Dafür gibt es aber nur vier verschiedene Möglichkeiten. Diese Terme können dann mit 4 abgeschätzt werden, da sie Differenzen von Verteilungsfunktionen sind, die durch 1 beschränkt sind. Die folgenden

Abschätzungen zeigen dann die Behauptung, dass $\sqrt{N}A_1$ in L_2-Norm gegen 0 konvergiert.

$$\begin{aligned} E(\sqrt{N}A_1)^2 &\leq \frac{m_0^2}{Nn_1^2n_0^2}\sum_{k=1}^{n_1}\sum_{k'=1}^{n_0}\sum_{s=1}^{m_{1k}}\sum_{s'=1}^{m_{1k}}\sum_{s''=1}^{m_{0k'}}\sum_{s'''=1}^{m_{0k'}}\frac{4}{m_{1k}^2\overline{m}_0^2} \\ &\leq \frac{4m_0^2}{Nn_0n_1} = O\left(\frac{1}{N}\right). \end{aligned}$$

Bei A_2 ist zu zeigen, dass das Integral mit $\widetilde{F}(x)$ und $\widehat{F}(x)$ nah genug an $\frac{1}{2}$ ist. Die Aussage ist anschaulich klar, da $\widetilde{F}(x)$ und $\widehat{F}(x)$ zwei asymptotisch erwartungstreue Schätzer für $F(x)$ sind.

$$\begin{aligned} E(\sqrt{N}A_2)^2 &= E\left(\frac{m_0}{\sqrt{N}}\int F_0 dF_0 - \frac{m_0}{\sqrt{N}}\int \widetilde{F}_0 d\widehat{F}_0\right)^2 \\ &= \frac{m_0^2}{N}E\left(\frac{m_0}{n_0}\sum_{k=1}^{n_0}\frac{1}{m_{0k}}\sum_{s=1}^{m_{0k}}\widetilde{F}_0(X_{0ks}) - \frac{1}{2}\right)^2 \\ &= \frac{m_0^2}{N}E\left(\frac{m_0}{n_0}\sum_{k=1}^{n_0}\frac{1}{m_{0k}}\sum_{s=1}^{m_{0k}}\frac{1}{m_0}\sum_{k'=1}^{n_0}\sum_{s'=1}^{m_{0k'}}c(X_{0ks} - X_{0k's'}) - \frac{1}{2}\right)^2 \\ &= \frac{m_0^2}{N}E\left(\frac{m_0}{n_0^2}\sum_{k=1}^{n_0}\sum_{s=1}^{m_{0k}}\sum_{k'=1}^{n_0}\sum_{s'=1}^{m_{0k'}}\frac{1}{m_{0k}\overline{m}_0}(c(X_{0ks} - X_{0k's'}) - \frac{1}{2})\right)^2. \end{aligned}$$

Beim Ausmultiplizieren des Quadrates fällt auch hier wieder auf, dass die gemischten Terme wegfallen, d.h. wenn der Index k von allen drei anderen Indizes verschieden ist. Die Abschätzung des Erwartungswertes verläuft hier analog wie oben, deshalb gilt dann folgende Ungleichung und somit auch die Konvergenz von A_2:

$$E(\sqrt{N}A_2)^2 \leq \frac{m_0^2}{Nn_0^4}\sum_{k=1}^{n_0}\sum_{k'=1}^{n_0}\sum_{s=1}^{m_{0k}}\sum_{s'=1}^{m_{0k}}\sum_{s''=1}^{m_{0k'}}\sum_{s'''=1}^{m_{0k'}}\frac{4}{m_{0k}^2\overline{m}_0^2} = O\left(\frac{1}{N}\right).$$

Diese Ergebnisse lassen sich direkt auf A_3 und A_4 übertragen. Somit ist die Behauptung gezeigt. □

Beweis von Theorem 3.5

Im Folgenden soll die Konsistenz des Varianzschätzers gezeigt werden. Der Beweis kann für jeden Summanden getrennt geführt werden und für die gesamte Matrix elementweise, da die Matrix von endlicher Dimension ist.
Der Beweis erfolgt in zwei Schritten: zunächst wird gezeigt, dass $\widetilde{\mathbf{V}}$ ein konsistenter Schätzer für $\mathbf{V}_N$ ist, dann wird gezeigt, dass

$$E(\widetilde{v}(l,l') - \widehat{v}(l,l'))^2 \to 0$$

gilt. Die Technik des Beweises ist der des Beweises von Theorem 4.1 in Brunner *et al.* (2002) nachempfunden.

1. Schritt:
Verwende folgende Gleichung:

$$\frac{1}{n-1} = \frac{1}{n} + \frac{1}{n(n-1)}.$$

Mit den Bezeichnungen $s_{ik}^{(l,l')} = Cov(\overline{Y}_{ik\cdot}^{(l)}, \overline{Y}_{ik\cdot}^{(l')})$ und $\mu_i^{(l)} = E(\overline{Y}_{ik\cdot}^{(l)})$ gilt:

$$\begin{aligned}
&\widetilde{v}_i(l,l') - v_i(l,l') \\
&= \frac{N}{n_i(n_i-1)} \sum_{k=1}^{n_i} (\overline{Y}_{ik\cdot}^{(l)} - \overline{Y}_{i\cdot\cdot}^{(l)})(\overline{Y}_{ik\cdot}^{(l')} - \overline{Y}_{i\cdot\cdot}^{(l')}) - \frac{N}{n_i^2} \sum_{k=1}^{n_i} s_{ik}^{(l,l')} \\
&= \frac{N}{n_i^2} \sum_{k=1}^{n_i} \left[(\overline{Y}_{ik\cdot}^{(l)} - \mu_i^{(l)} + \mu_i^{(l)} - \overline{Y}_{i\cdot\cdot}^{(l)})(\overline{Y}_{ik\cdot}^{(l')} - \mu_i^{(l')} + \mu_i^{(l')} - \overline{Y}_{i\cdot\cdot}^{(l')}) - s_{ik}^{(l,l')} \right] \\
&\qquad + \frac{N}{n_i^2(n_i-1)} \sum_{k=1}^{n_i} (\overline{Y}_{ik\cdot}^{(l)} - \overline{Y}_{i\cdot\cdot}^{(l)})(\overline{Y}_{ik\cdot}^{(l')} - \overline{Y}_{i\cdot\cdot}^{(l')}) \\
&= \frac{N}{n_i^2} \sum_{k=1}^{n_i} \left[(A_k + B)(C_k + D) - s_{ik}^{(l,l')} \right] + \frac{N}{n_i^2(n_i-1)} \sum_{k=1}^{n_i} E_k
\end{aligned}$$

mit den folgenden Abkürzungen:

$$\begin{aligned}
A_k &= \overline{Y}_{ik\cdot}^{(l)} - \mu_i^{(l)} \\
B &= \mu_i^{(l)} - \overline{Y}_{i\cdot\cdot}^{(l)} \\
C_k &= \overline{Y}_{ik\cdot}^{(l')} - \mu_i^{(l')} \\
D &= \mu_i^{(l')} - \overline{Y}_{i\cdot\cdot}^{(l')} \\
E_k &= (\overline{Y}_{ik\cdot}^{(l)} - \overline{Y}_{i\cdot\cdot}^{(l)})(\overline{Y}_{ik\cdot}^{(l')} - \overline{Y}_{i\cdot\cdot}^{(l')}).
\end{aligned}$$

Es gilt offensichtlich, dass

$$E(A_k C_k) = E((\overline{Y}_{ik\cdot}^{(l)} - \mu_i^{(l)})(\overline{Y}_{ik\cdot}^{(l')} - \mu_i^{(l')}) = s_{ik}^{(l,l')}.$$

Außerdem sind die Erwartungswerte von allen vier Größen $A_k - D$ gleich 0 und für $k \neq k'$ sind je zwei Terme unabhängig voneinander. Die zweiten Momente der Terme lassen sich jeweils mit 1 abschätzen. Für den Erwartungswert $E(B^2D^2)$ gilt folgende Abschätzung:

$$E(B^2D^2) \leq \sqrt{E(B^4)E(D^4)} \leq \frac{1}{n_i^2}$$

wegen der Cauchy-Schwarz-Ungleichung. Außerdem gilt für den Erwartungswert $E(E_k E_{k'}) \leq 1$.
Betrachtet man nun den Erwartungswert der quadrierten Differenz, so erhält man folgende Abschätzung für die L_2-Konvergenz des Ausdrucks:

$$
\begin{aligned}
& E(\widetilde{v}_i(l,l') - v_i(l,l'))^2 \\
&= E\left(\frac{N}{n_i^2}\sum_{k=1}^{n_i}\left[(A_k+B)(C_k+D) - s_{ik}^{(l,l')}\right] + \frac{N}{n_i^2(n_i-1)}\sum_{k=1}^{n_i} E_k\right)^2 \\
&\leq \frac{4N^2}{n_i^4}\left(\sum_{k=1}^{n_i}\sum_{k'=1}^{n_i} E(A_kC_k - E(A_kC_k))(A_{k'}C_{k'} - E(A_{k'}C_{k'}))\right. \\
&\quad \left. + E(B^2)\sum_{k=1}^{n_i}\sum_{k'=1}^{n_i} E(C_kC_{k'}) + E(D^2)\sum_{k=1}^{n_i}\sum_{k'=1}^{n_i} E(A_kA_{k'}) + \sum_{k=1}^{n_i}\sum_{k'=1}^{n_i} E(B^2D^2)\right) \\
&\quad + \frac{N^2}{n_i^4(n_i-1)^2}\sum_{k=1}^{n_i}\sum_{k'=1}^{n_i} E(E_kE_{k'}) \\
&\leq \frac{4N^2}{n_i^4}\left(\sum_{k=1}^{n_i} E(A_kC_k - E(A_kC_k))^2 + E(B^2)\sum_{k=1}^{n_i} E(C_k^2) + E(D^2)\sum_{k=1}^{n_i} E(A_k^2)\right. \\
&\quad \left. + n_i^2 E(B^2D^2)\right) + \frac{N^2}{n_i^4(n_i-1)^2}\sum_{k=1}^{n_i}\sum_{k'=1}^{n_i} E(E_kE_{k'}) \\
&\leq \frac{4N^2}{n_i^4}(n_i + n_i + n_i + n_i^2\frac{1}{n_i^2}) + \frac{N^2}{n_i^2(n_i-1)^2} \\
&\leq \frac{16N^2}{n_i^3} + \frac{N^2}{n_i^2(n_i-1)^2} \leq \frac{16N_0^2}{n_i} + \frac{N_0^2}{(n_i-1)^2} = O\left(\frac{1}{n_i}\right).
\end{aligned}
$$

Damit ist gezeigt, dass $\mathbf{V}$ und $\tilde{\mathbf{V}}$ in L_2-Norm gegeneinander konvergieren.

2. Schritt:
Zunächst folgen einige algebraische Überlegungen, in denen die Differenz der Schätzer so umgeformt wird, dass man anschließend ganz leicht die Konvergenz sieht. Wir verwenden folgende einheitliche Schreibweise für die jeweiligen Schätzer:

$$
\begin{aligned}
\widetilde{v}_i(l,l') &= \frac{N}{n_i(n_i-1)}\sum_{k=1}^{n_i}(\overline{Y}_{ik\cdot}^{(l)} - \overline{Y}_{i\cdot\cdot}^{(l)})(\overline{Y}_{ik\cdot}^{(l')} - \overline{Y}_{i\cdot\cdot}^{(l')}) \\
\widehat{v}_i(l,l') &= \frac{N}{(N-m_i)^2 n_i(n_i-1)}\sum_{k=1}^{n_i}(\overline{Z}_{ik\cdot}^{(l)} - \overline{Z}_{i\cdot\cdot}^{(l)})(\overline{Z}_{ik\cdot}^{(l')} - \overline{Z}_{i\cdot\cdot}^{(l')})
\end{aligned}
$$

Nach Voraussetzung gilt $N(n_i-1)^{-1} < N_0$, deshalb reicht es, die Behauptung mit dem Vorfaktor N_0/n_i^{-1} zu zeigen. Mithilfe der c_r-Ungleichung kann man somit zeigen:

$$\begin{aligned}
&E(\widetilde{v}_i(l,l') - \widehat{v}_i(l,l'))^2 \\
&= E\Bigg(\frac{N}{n_i(n_i-1)}\sum_{k=1}^{n_i}\Bigg((\overline{Y}_{ik\cdot}^{(l)} - \overline{Y}_{i\cdot\cdot}^{(l)})(\overline{Y}_{ik\cdot}^{(l')} - \overline{Y}_{i\cdot\cdot}^{(l')}) \\
&\quad - \frac{1}{(N-m_i)^2}(\overline{Z}_{ik\cdot}^{(l)} - \overline{Z}_{i\cdot\cdot}^{(l)})(\overline{Z}_{ik\cdot}^{(l')} - \overline{Z}_{i\cdot\cdot}^{(l')})\Bigg)\Bigg)^2 \\
&\leq \frac{N_0}{n_i}\sum_{k=1}^{n_i} E\Bigg((\overline{Y}_{ik\cdot}^{(l)} - \overline{Y}_{i\cdot\cdot}^{(l)})(\overline{Y}_{ik\cdot}^{(l')} - \overline{Y}_{i\cdot\cdot}^{(l')}) \\
&\quad - \frac{1}{(N-m_i)^2}(\overline{Z}_{ik\cdot}^{(l)} - \overline{Z}_{i\cdot\cdot}^{(l)})(\overline{Z}_{ik\cdot}^{(l')} - \overline{Z}_{i\cdot\cdot}^{(l')})\Bigg)^2 \\
&\leq \frac{2N_0}{n_i}\sum_{k=1}^{n_i} E\left((\overline{Y}_{ik\cdot}^{(l)} - \overline{Y}_{i\cdot\cdot}^{(l)})^2\left[(\overline{Y}_{ik\cdot}^{(l')} - \overline{Y}_{i\cdot\cdot}^{(l')}) - \frac{(\overline{Z}_{ik\cdot}^{(l')} - \overline{Z}_{i\cdot\cdot}^{(l')})}{N-m_i}\right]^2\right) \\
&\quad + \frac{2N_0}{n_i}\sum_{k=1}^{n_i} E\left(\frac{(\overline{Z}_{ik\cdot}^{(l')} - \overline{Z}_{i\cdot\cdot}^{(l')})^2}{(N-m_i)^2}\left[(\overline{Y}_{ik\cdot}^{(l)} - \overline{Y}_{i\cdot\cdot}^{(l)}) - \frac{(\overline{Z}_{ik\cdot}^{(l)} - \overline{Z}_{i\cdot\cdot}^{(l)})}{N-m_i}\right]^2\right).
\end{aligned}$$

Um nun weiter abzuschätzen wird verwendet, dass die Zufallsvariablen $\overline{Y}_{ik\cdot}^{(l)}$ und $\overline{Z}_{ik\cdot}^{(l')}/(N-m_i)$ gleichmäßig durch 1 beschränkt sind und deshalb gilt:

$$|\overline{Y}_{ik\cdot}^{(l)} - \overline{Y}_{i\cdot\cdot}^{(l)}| \leq 1 \text{ und } \left|\frac{\overline{Z}_{ik\cdot}^{(l')} - \overline{Z}_{i\cdot\cdot}^{(l')}}{N-m_i}\right| \leq 1.$$

Im Folgenden wird jetzt nur der erste Summand aus der letzten Ungleichung betrachtet, denn die Behauptung folgt für den zweiten Summanden analog.

$$\begin{aligned}
&E\left((\overline{Y}_{i1\cdot}^{(l)} - \overline{Y}_{i\cdot\cdot}^{(l)})^2\left[(\overline{Y}_{i1\cdot}^{(l')} - \overline{Y}_{i\cdot\cdot}^{(l')}) - \frac{(\overline{Z}_{i1\cdot}^{(l')} - \overline{Z}_{i\cdot\cdot}^{(l')})}{N-m_i}\right]^2\right) \\
&\leq E\left((\overline{Y}_{i1\cdot}^{(l')} - \overline{Y}_{i\cdot\cdot}^{(l')}) - \frac{(\overline{Z}_{i1\cdot}^{(l')} - \overline{Z}_{i\cdot\cdot}^{(l')})}{N-m_i}\right)^2 \\
&= E\left(\left(\overline{Y}_{i1\cdot}^{(l')} - \frac{\overline{Z}_{i1\cdot}^{(l')}}{N-m_i}\right) - \left(\overline{Y}_{i\cdot\cdot}^{(l')} - \frac{\overline{Z}_{i\cdot\cdot}^{(l')}}{N-m_i}\right)\right)^2 \\
&\leq 2E\left(\overline{Y}_{i1\cdot}^{(l')} - \frac{\overline{Z}_{i1\cdot}^{(l')}}{N-m_i}\right)^2 + 2E\left(\overline{Y}_{i\cdot\cdot}^{(l')} - \frac{\overline{Z}_{i\cdot\cdot}^{(l')}}{N-m_i}\right)^2
\end{aligned}$$

$$\leq \quad 4E\left(\overline{Y}_{i1\cdot}^{(l')} - \frac{\overline{Z}_{i1\cdot}^{(l')}}{N - m_i}\right)^2 .$$

Die nun folgende Abschätzung verwendet die Unabhängigkeit von $\overline{Y}_{ik\cdot}^{(l)}$ und $\overline{Y}_{ik'\cdot}^{(l)}$ für verschiedene k, k' und die Tatsache, dass $E\left[c(X_{1ks}^{(l)} - X_{0k's'}^{(l)}) - F_0^{(l)}(X_{1ks}^{(l)})\right] = 0$. Um die Notation sonst überschaubar zu halten, legt man ein spezielles i fest. Es sei also $i = 1$.

$$\begin{aligned}
E\left(\overline{Y}_{11\cdot}^{(l)} - \frac{\overline{Z}_{11\cdot}^{(l)}}{m_0}\right)^2 &= E\left(\frac{1}{m_{11}}\sum_{s=1}^{m_{11}}\left[F_0^{(l)}(X_{11s}^{(l)}) - \widetilde{F}_0^{(l)}(X_{11s}^{(l)})\right]\right)^2 \\
&\leq \frac{1}{m_{11}}\sum_{s=1}^{m_{11}} E\left(\frac{1}{m_0}\sum_{k=1}^{n_0}\sum_{s'=1}^{m_{0k}}\left[F_0^{(l)}(X_{11s}^{(l)}) - c(X_{11s}^{(l)} - X_{0ks'}^{(l)})\right]\right)^2 \\
&= E\left(\frac{1}{m_0}\sum_{k=1}^{n_0}\sum_{s'=1}^{m_{0k}}\left[F_0^{(l)}(X_{111}^{(l)}) - c(X_{111}^{(l)} - X_{0ks'}^{(l)})\right]\right)^2 \\
&\leq \frac{1}{m_0^2}\sum_{k=1}^{n_0}\sum_{s=1}^{m_{0k}}\sum_{s'=1}^{m_{0k}} E\Big[\left(F_0^{(l)}(X_{111}^{(l)}) - c(X_{111}^{(l)} - X_{0ks}^{(l)})\right) \\
&\qquad \left(F_0^{(l)}(X_{111}^{(l)}) - c(X_{111}^{(l)} - X_{0ks'}^{(l)})\right)\Big] \\
&\leq \frac{1}{m_0^2}\sum_{k=1}^{n_0} m_{0k}^2 \leq \frac{M_0}{m_0^2}\sum_{k=1}^{n_0} m_{0k} = \frac{M_0}{m_0} = O\left(\frac{1}{m_0}\right) = O\left(\frac{1}{N}\right).
\end{aligned}$$

Die gleiche Abschätzung liefert für $i = 0$:

$$E\left(\overline{Y}_{01\cdot}^{(l)} - \frac{\overline{Z}_{01\cdot}^{(l)}}{m_1}\right)^2 = O\left(\frac{1}{m_1}\right).$$

Somit ist die Konvergenz für alle Summanden gezeigt und die Behauptung bewiesen. □

Beweis von Satz 3.9

Dass $\widehat{\mathbf{S}}_0$ und $\widehat{\mathbf{S}}_1$ konsistente Schätzer bilden, folgt direkt aus Theorem 3.5, wenn man berücksichtigt, dass Annahme (A1) gilt und die λ_{ik} somit problemlos in die Beweise eingefügt werden können.
Für die Kovarianzmatrizen $\widehat{\mathbf{C}}_0$ und $\widehat{\mathbf{C}}_1$ folgt der Nachweis der Konsistenz analog. □

Beweis von Proposition 3.10

Zur Vereinfachung der Schreibweise setzen wir $\lambda = m_2/m_1$ und schreiben für $m_1 = m$. Aufgrund der Voraussetzung $m_1 \neq m_2$ folgt also, dass $\lambda \neq 1$ gelten muss.

$$\begin{aligned}
1.Var(U) &= \frac{1}{4}\left[\frac{1}{m}(\sigma + (m-1)c) + \frac{1}{\lambda m}(\sigma + (\lambda m - 1)c)\right] \\
&= \frac{1+\lambda}{4\lambda m}\sigma + \frac{2\lambda m - \lambda - 1}{4\lambda m}c \\
&= \sigma\left(\frac{1+\lambda+\rho(2\lambda m - \lambda - 1)}{4\lambda m}\right). \\
2.Var(G) &= \frac{1}{(\lambda+1)^2 m^2}\left[m\sigma + m(m-1)c + \lambda m\sigma + \lambda m(\lambda m - 1)c\right] \\
&= \frac{1}{(\lambda+1)^2 m}\left[\sigma + \lambda\sigma + c(m - 1 + \lambda^2 m - \lambda)\right] \\
&= \sigma\left(\frac{1+\lambda+\rho(m(1+\lambda^2) - \lambda - 1)}{(\lambda+1)^2 m}\right).
\end{aligned}$$

Nun wird untersucht, für welche Werte von ρ die Varianz des gewichteten Schätzers kleiner als die des ungewichteten Schätzers ist:

$$\begin{aligned}
0 &< Var(U) - Var(G) \\
&= \left(\frac{1+\lambda+\rho(2\lambda m - \lambda - 1)}{4\lambda m}\right) - \left(\frac{1+\lambda+\rho(m(1+\lambda^2) - \lambda - 1)}{(\lambda+1)^2 m}\right) \\
&= \frac{(\lambda+1)^2 - 4\lambda}{(\lambda+1)4\lambda m} + \rho\left[\frac{1}{2} - \frac{\lambda+1}{4\lambda m} - \frac{1+\lambda^2}{(\lambda+1)^2} + \frac{1}{m(\lambda+1)}\right] \\
&= \frac{(\lambda-1)^2}{4\lambda m(\lambda+1)} + \rho\left[-\frac{(\lambda-1)^2}{4\lambda m(\lambda+1)} - \frac{(\lambda-1)^2}{2(\lambda+1)^2}\right]
\end{aligned}$$

Die folgende Umformung ist nur der unter Annahme $\lambda \neq 1$ möglich. Da der Fall gleicher Stichprobenumfänge aber trivialerweise gleiche Varianzen ergibt, ist diese Annahme hier nicht einschränkend.

$$\begin{aligned}
\frac{1}{\rho} &> \frac{\frac{(\lambda-1)^2}{4\lambda m(\lambda+1)} + \frac{(\lambda-1)^2}{2(\lambda+1)^2}}{\frac{(\lambda-1)^2}{4\lambda m(\lambda+1)}} \\
&= \frac{\frac{1}{4\lambda m(\lambda+1)} + \frac{1}{2(\lambda+1)^2}}{\frac{1}{4\lambda m(\lambda+1)}} \\
&= 1 + \frac{2\lambda m(\lambda+1)}{(\lambda+1)^2} \\
&= \frac{\lambda + 1 + 2\lambda m}{\lambda + 1}.
\end{aligned}$$

Daraus folgt dann für ρ:

$$\rho \quad < \quad \frac{\lambda+1}{2\lambda m+\lambda+1} = \frac{(\lambda+1)m}{2\lambda m^2+(\lambda+1)m} = \frac{m_1+m_2}{2m_1m_2+m_1+m_2}. \quad \square$$

A.3 Weitere Simulationsergebnisse

Tabelle A.1: Vergleich der ATS mit ungewichtetem und gewichtetem Schätzer in unbalancierten Designs ($n = 50$)

			Design 3		Design 4	
a	Level	c	Ungewichtet	Gewichtet	Ungewichtet	Gewichtet
0	2	0	0.058	0.061	0.050	0.051
1	2	0	0.056	0.057	0.051	0.051
2	2	0	0.057	0.057	0.057	0.054
3	2	0	0.068	0.056	0.068	0.053
0	2	1	0.051	0.057	0.053	0.054
1	2	1	0.054	0.053	0.054	0.058
2	2	1	0.060	0.057	0.056	0.054
3	2	1	0.059	0.055	0.060	0.053
0	2	2	0.054	0.055	0.050	0.049
1	2	2	0.052	0.056	0.048	0.051
2	2	2	0.049	0.051	0.052	0.052
3	2	2	0.054	0.053	0.056	0.055
0	6	0	0.046	0.048	0.042	0.044
1	6	0	0.045	0.046	0.047	0.047
2	6	0	0.049	0.049	0.050	0.045
3	6	0	0.060	0.044	0.060	0.045
0	6	1	0.043	0.046	0.041	0.046
1	6	1	0.046	0.045	0.045	0.046
2	6	1	0.045	0.045	0.051	0.049
3	6	1	0.048	0.042	0.052	0.045
0	6	2	0.040	0.043	0.036	0.038
1	6	2	0.039	0.042	0.041	0.041
2	6	2	0.047	0.046	0.047	0.045
3	6	2	0.050	0.043	0.049	0.043

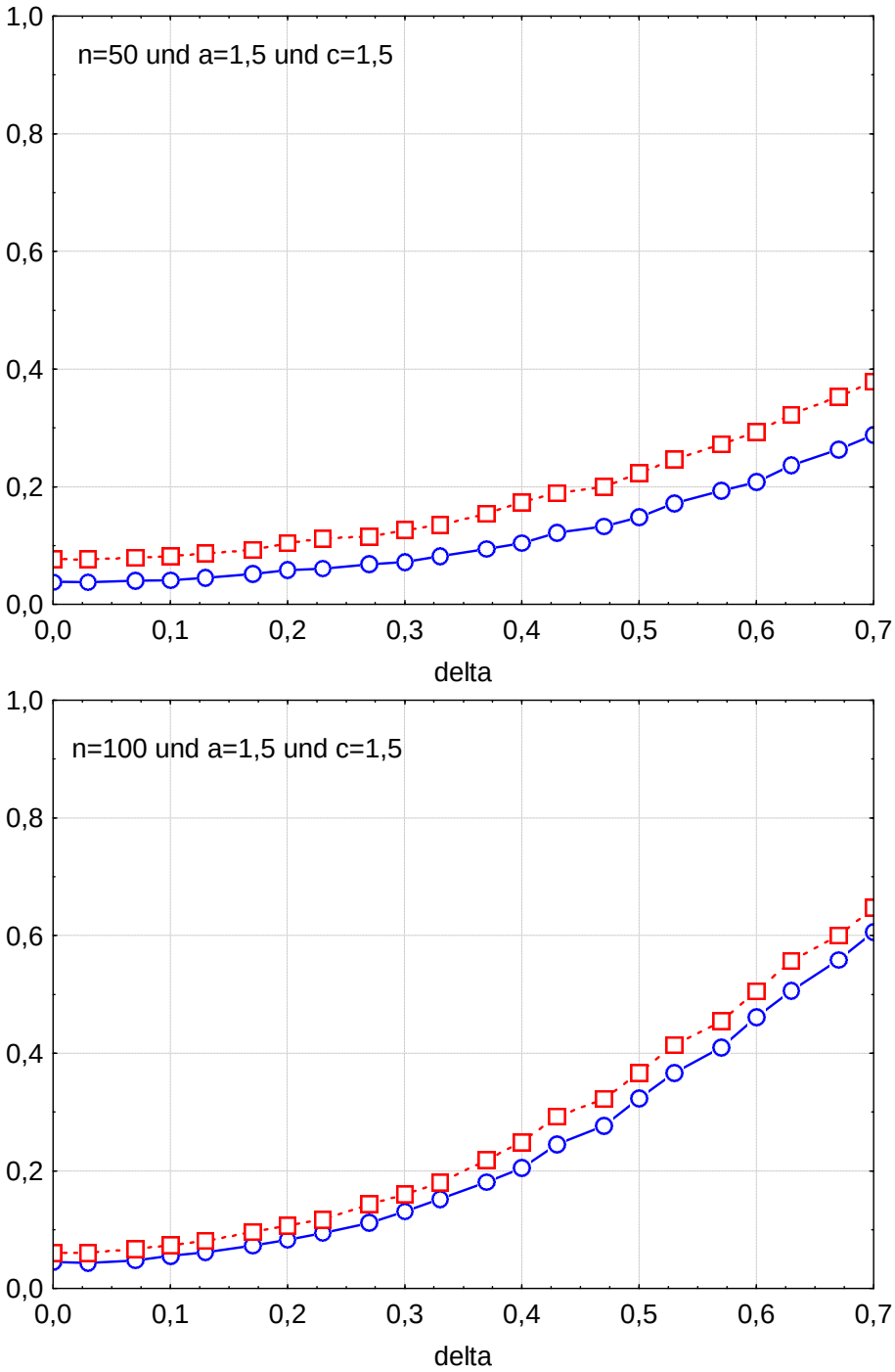

Abbildung A.1: Powersimulationen im Design 1, Anzahl Faktorstufen=6, ATS durchgezogene Linie, WTS gestrichelte Linie

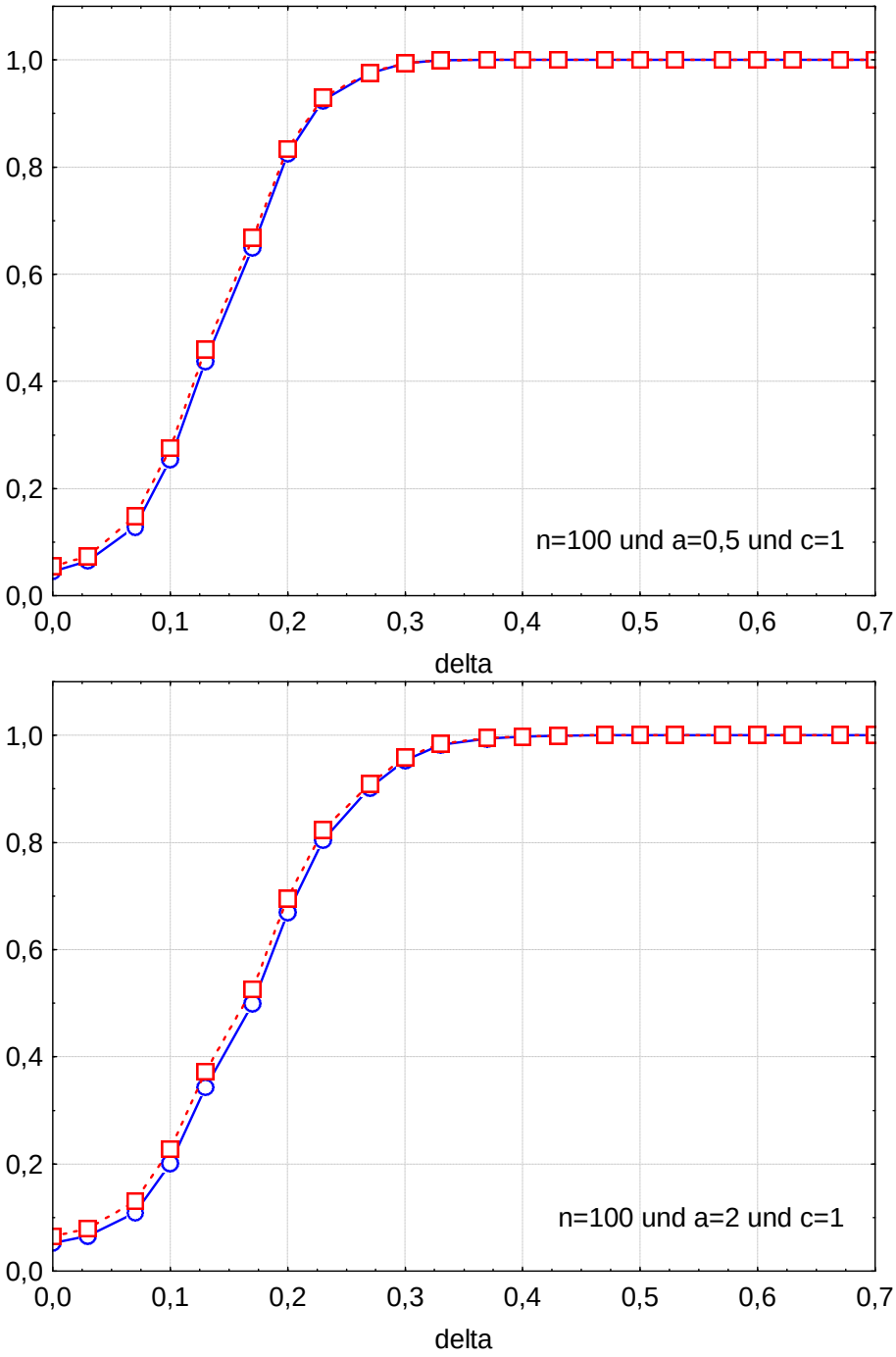

Abbildung A.2: Powersimulationen im Design 2, Anzahl Faktorstufen=4, ATS durchgezogene Linie, WTS gestrichelte Linie

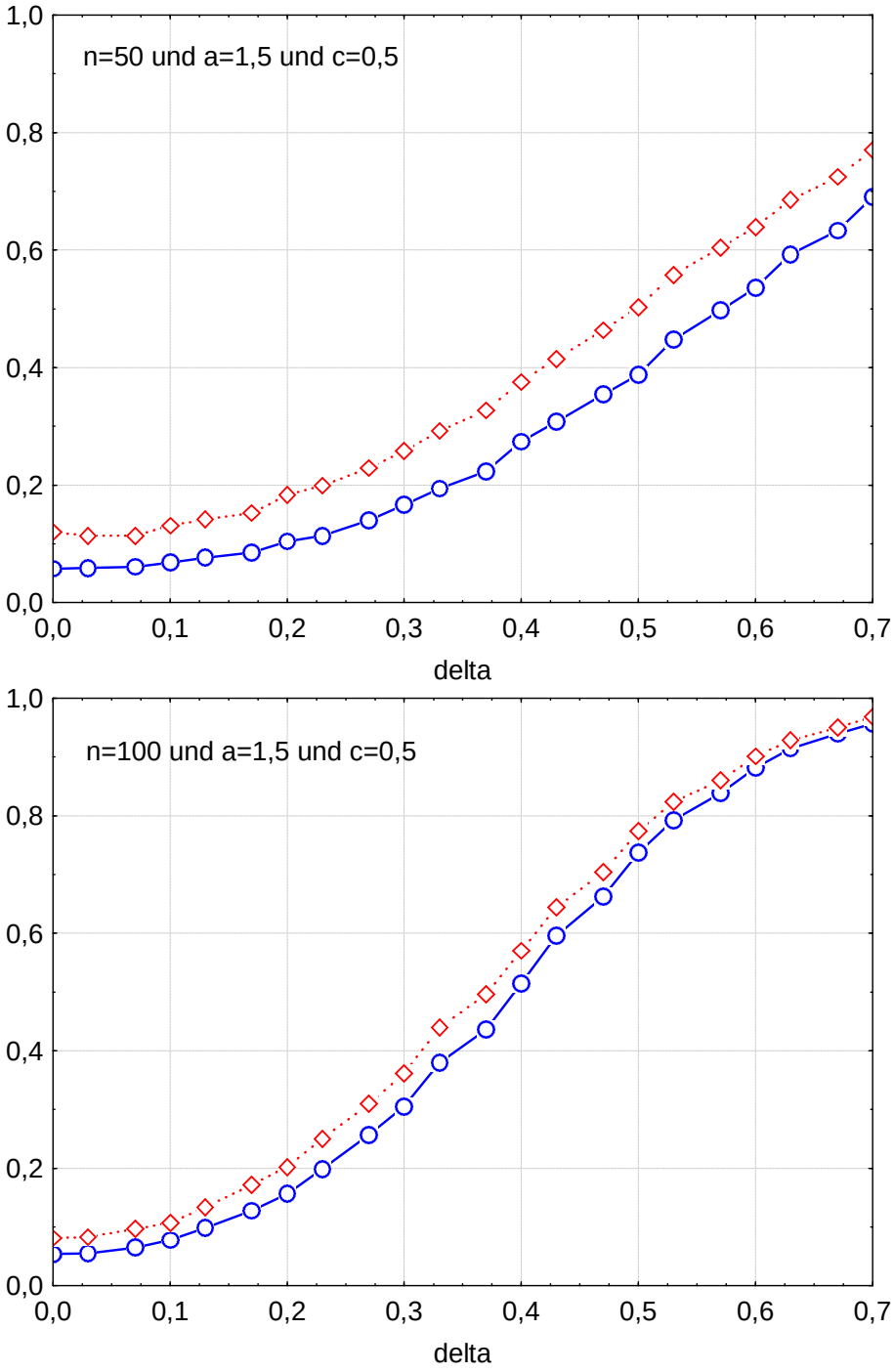

Abbildung A.3: Powersimulationen im Design 3, Anzahl Faktorstufen=6, ATS durchgezogene Linie, WTS gestrichelte Linie

Literaturverzeichnis

Ahn, C. 1997. An Evaluation of Methods for the Estimation of Sensitivity and Specificity of Site-Specific Diagnostic Tests. *Biometrical Journal*, **7**, 793–807.

Bamber, D. 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, **12**, 387–415.

Beam, C. A. 1998. Analysis of clustered data in receiver operating characteristic studies. *Statistical Methods in Medical Research*, **7**, 324–336.

Benhin, E., Rao, J.N.K., & Scott, A.J. 2005. Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika*, **92**, 435–450.

Bernhardt, T.M., Rapp-Bernhardt, U., Lenzen, H., Roehl, F.W., Diederich, S., Papke, K., Ludwig, K., & Heindel, W. 2004. Low-Voltage Digital Selenium Radiography: Detection of Simulated interstitial lung disease, nodules, and catheters – a phantom study. *Radiology*, **232**, 693–700.

Box, G. E. P. 1954. Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification. *The Annals of Mathematical Statistics*, **25**, 290–302.

Brunner, E., & Denker, M. 1994. Rank Statistics under Dependent Observations and Applications to Factorial Designs. *Journal of Statistical Planning and Inference*, **42**, 353–378.

Brunner, E., & Munzel, U. 2000. The Nonparametric Behrens-Fisher Problem: Asymptotic Theory and a Small-Sample Approximation. *Biometrical Journal*, **42**, 1–9.

Brunner, E., Dette, H., & Munk, A. 1997. Box-Type Approximations in Nonparametric Factorial Designs. *Journal of the American Statistical Association*, **92**, 1494–1502.

Brunner, E., Munzel, U., & Puri, M. L. 1999. Rank-Score Tests in Factorial Designs with Repeated Measures. *Journal of Multivariate Analysis*, **70**, 286–317.

Brunner, E., Munzel, U., & Puri, M. L. 2002. The Multivariate Nonparametric Behrens-Fisher Problem. *Journal of Statistical Planning and Inference*, **108**, 37–53.

Cramér, H. 1928. On the composition of elementary errors. *Skandinavisk Aktuarietidskrift*, **11**, 13–74 and 141–180.

Datta, S., & Satten, G.A. 2005. Rank-Sum tests for clustered Data. *Journal of the American Statistical Association*, **100**, 908–915.

Davies, G.M., Koch, G.G., & Beck, J. 1997. Statistical Strategies for event rate comparisons in dental studies. *Journal of Biopharmaceutical Statistics*, **7(4)**, 625–634.

DeLong, E.R., DeLong, D.M., & Clarke-Pearson, D.L. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**, 837–45.

Dempster, A.P. 1958. A High Dimensional Two Sample Significance Test. *Annals of Mathematical Statistics*, **29**, 995–1010.

Dodd, L.E., & Pepe, M.S. 2003. Partial AUC Estimation and Regression. *Biometrics*, **59**, 614–623.

Donner, A., & Klar, N. 2000. *Design and Analysis of Cluster Randomization Trials in Health Research*. Oxford University Press.

Dorfman, D.D., Berbaum, K.S., & Metz, C.E. 1992. Receiver Operating Characteristic rating analysis: Generalization to the population of readers and patients with the Jackknife method. *Investigative Radiology*, **27**, 723–731.

European Medicines Agency (EMEA), Committee for Proprietory Medicinal Products (CPMP). 2001. *Points to consider on the evaluation of diagnostic agents*.

Feinstein, A.R. 1975. Clinical biostatistics XXXI. On the sensitivity, specificity, and discrimination of diagnostic tests. *Clinical pharmacology and therapeutics*, **17(1)**, 104–116.

Hanley, J.A. 1989. Receiver operating characteristic (ROC) methodology: the state of the art. *Critical Reviews in Diagnostic Imaging*, **29**, 307–335.

Hanley, J.A., & Lippman-Hand, A. 1983. If nothing goes wrong, is everything all right? Interpreting zero numerators. *Journal of the American Medical Association*, **249**, 1743–1745.

Hanley, J.A., & McNeil, B.J. 1982. The meaning and the use of the Area under a Receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Hillis, S.L., Obuchowski, N.A., Schartz, K.M., & Berbaum, K.S. 2005. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. *Statistics in Medicine*, **24**, 1579–607.

Hujoel, P.P., Moulton, L.H., & Loesche, W.J. 1990. Estimation of sensitivity and specificity of site-specific diagnostic tests. *Journal of Periodontic Research*, **25**, 193–196.

Kaufmann, J., Werner, C., & Brunner, E. 2005. Nonparametric Methods for Analyzing the Accuracy of Diagnostic Tests with Multiple Readers. *Statistical Methods in Medical Research*, **14**, 129–146.

Köbberling, J., Richter, K., Trampisch, H.J., & Windeler, J. 1991. *Methodologie der medizinischen Diagnostik*. Springer-Verlag.

Lee, E.W. 1996. Two sample comparison for large groups of correlated binary responses. *Statistics in Medicine*, **15**, 1187–1197.

Lee, E.W., & Dubin, N. 1994. Estimation and sample size considerations for clustered binary responses. *Statistics in Medicine*, **13**, 1241–1252.

Lloyd, C.J. 1998. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*, **93**, 1356–1364.

Lloyd, C.J., & Yong, Z. 1999. Kernel estimators of the ROC curve are better than empirical. *Statistics and Probability Letters*, **44**, 221–228.

Lu, Y., Jin, H., & Genant, H.K. 2003. On the non-inferiority of a diagnostic test based on paired observations. *Statistics in Medicine*, **22**, 3029–3044.

Lui, K., & Zhou, X. 2004. Testing non-inferiority (and equivalence) between two diagnostic procedures in paired-sample ordinal data. *Statistics in Medicine*, **23**, 545–559.

McClish, D.K. 1989. Analyzing a portion of the ROC curve. *Medical Decision Making*, **9**, 190–198.

Metz, C.E. 1978. Basic Principles of ROC Analysis. *Seminars in Nuclear Medicine*, **8**, 283–298.

Nakas, C.T., & Yiannoutsos, C.T. 2004. Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine*, **23**, 3437–3449.

Newcombe, R.G. 2006. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part I: General issues and tail-area-based methods. *Statistics in Medicine*, **25**, 543–557.

Obuchowski, N. A. 1997. Nonparametric Analysis of Clustered ROC Curve Data. *Biometrics*, **53**, 567–578.

Obuchowski, N.A. 1998. On the comparison of correlated proportions for clustered data. *Statistics in Medicine*, **17**, 1495–1507.

Obuchowski, N.A. 2000. Sample Size Tables for Receiver Operating Characteristic Studies. *American Journal of Roentgenology*, **175**, 603–608.

Obuchowski, N.A. 2001. Can electronic medical images replace hard-copy film? Defining and testing the equivalence of diagnostic tests. *Statistics in Medicine*, **20**, 2845–2863.

Obuchowski, N.A. 2005. Estimating and comparing diagnostic tests' accuracy when the gold standard is not binary. *Academic Radiology*, **12**, 1198–1204.

Obuchowski, N.A., & Lieber, M.L. 1998. Confidence intervals for the receiver operating characteristic area in studies with small samples. *Academic Radiology*, **5**, 561–571.

Obuchowski, N.A., & Lieber, M.L. 2002. Confidence Bounds when the estimated ROC area is 1.0. *Academic Radiology*, **9**, 526–530.

Obuchowski, N.A., Goske, M.J., & Applegate, K.E. 2001. Assessing the physicians' accuracy in diagnostic paediatric patients with acute abdominal pain: measuring accuracy for multiple diseases. *Statistics in Medicine*, **20**, 3261–3278.

Parker, R.A., & Davis, R.B. 1999. Evaluating whether a Binary decision rule operates better than chance. *Biometrical Journal*, **41**, 25–31.

Peterson, W.W., Birdsall, T.G., & Fox, W.C. 1954. The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, **4**, 171–212.

Ransohoff, D.F., & Feinstein, A.F. 1978. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *The New England Journal of Medicine*, **299**, 926–930.

Rao, J.N.K., & Scott, A.J. 1992. A simple Method for the Analysis of clustered binary data. *Biometrics*, **48**, 577–585.

Roe, C.A., & Metz, C.E. 1997. Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: validation with computer simulation. *Academic Radiology*, **4**, 298–303.

Rosner, B., & Grove, D. 1999. Use of Mann-Whitney U-Test for clustered data. *Statistics in Medicine*, **18**, 1387–1400.

Rosner, B., Glynn, R. J., & Lee, M.T. 2003. Incorparation of Clustering effects for the Wilcoxon rank sum test: A large sample approach. *Biometrics*, **59**, 1089–1098.

Rousseeuw, P. J., & Molenberghs, G. 1993. Transformation of Non Positive Semidefinite Correlation Matrices. *Communications in Statistics – Theory and Methods*, **A22**, 965–984.

Schisterman, E.F., Faraggi, D., & Reiser, B. 2004. Adjusting the generalized ROC curve for covariates. *Statistics in Medicine*, **23**, 3319–3331.

Sen, P.K. 1960. On some convergence properties of U-statistics. *Calcutta Statistical Association Bulletin*, **10**, 1–18.

Shapiro, D.E. 1999. The interpretation of diagnostic tests. *Statistical Methods in Medical Research*, **8**, 113–134.

Srivastava, M.S. 2005. Some Tests Concerning The Covariance Matrix In High Dimensional Data. *Journal of Japanese Statistical Society*, **35**, 251–272.

Stanish, W.M., Gillings, D.B., & Koch, G.G. 1978. An Application of Multivariate Ratio Methods for the Analysis of a longitudinal trial with missing data. *Biometrics*, **34**, 305–317.

Su, C., Gardner, I.A., & Johnson, W.O. 2004. Diagnostic Test accuracy and prevalence inferences based on joint and sequential testing with finite population sampling. *Statistics in Medicine*, **23**, 2237–2255.

Thompson, M.L. 2003. Assesing the diagnostic accuracy of a sequence of tests. *Biostatistics*, **4(3)**, 341–351.

Tsimikas, J.V., Bosch, R.J., Coull, B.A., & El Barmi, H. 2002. Profile-likelihood inference for highly accurate diagnostic tests. *Biometrics*, **58**, 946–956.

van der Vaart, A. W. 1998. *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Walter, S.D. 2003. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. *Statistics in Medicine*, **21**, 1237–1256.

Walter, S.D. 2005. The partial area under the summary ROC curve. *Statistics in Medicine*, **24**, 2025–2040.

Welch, B. L. 1938. The Significance of the Difference Between Two Means when the Population Variances are Unequal. *Biometrika*, **29**, 350–362.

Werner, C., & Brunner, E. 2006. Rank methods for the analysis of clustered data in diagnostic trials. *Computational Statistics and Data Analysis*, **in press**.

Williamson, J.M., Datta, S., & Satten, G.A. 2003. Marginal analyses of clustered data when cluster size is informative. *Biometrics*, **59**, 36–42.

Youden, W.J. 1950. Index for rating diagnostic tests. *Cancer*, **3**, 32–35.

Zhou, X., & Castelluccio, P. 2003. Nonparametric analysis for the ROC areas of two diagnostic tests in the presence of nonignorable verification bias. *Journal of Statistical Planning and Inference*, **115**, 193–213.

Zhou, X., Obuchowski, N.A., & McClish, D.K. 2002. *Statistical Methods in Diagnostic Medicine*. Wiley, New York.

Zweig, M.H., & Campbell, G. 1993. Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. *Clinical Chemistry*, **39**, 561–577.

Lebenslauf

Persönliche Daten

Name	Carola Werner
Anschrift	Rosdorfer Weg 17a, 37073 Göttingen
Geburtsdatum	12. September 1977
Geburtsort	Wolfenbüttel
Familienstand	ledig

Ausbildung

Juni 1997	Abitur (Note: 1,3)
1997-2002	Studium: Diplom-Mathematik an der Georg-August-Universität mit Nebenfach Volkswirtschaftslehre (Spezialisierung in Statistik)
Juni 1999	Vordiplom (Note: „Gut“)
November 2002	Diplom (Note: „Sehr gut“) Titel der Diplomarbeit: „Dimensionsstabile Approximation für Verteilungen von zufälligen quadratischen Formen im Repeated-Measures-Design“
April 2003	Aufnahme in den Promotionsstudiengang „Applied Statistics and Empirical Methods“ der Universität Göttingen

Tätigkeiten

1999 - 2002	Studentische Hilfskraft am Institut für Medizinische Statistik im Bereich Forschung und Lehre bzw. im DFG-Projekt „Longitudinale Daten“
Seit Dez. 2002	Wissenschaftliche Assistentin am Institut für Medizinische Statistik