Nichtparametrische Analyse diagnostischer Gütemaße bei Clusterdaten

Dissertation zur Erlangung des mathematisch-
naturwissenschaftlichen Doktorgrades „Doctor rerum naturalium“
der Georg-August-Universität Göttingen

vorgelegt von
Katharina Lange
aus Detmold

Göttingen 2011

D7
Referent: Prof. Dr. Edgar Brunner
Korreferent: Prof. Dr. Martin Schlather

Tag der mündlichen Prüfung: 04. März 2011

Danksagung

Für die Vergabe des Themas und die engagierte Betreuung dieser Arbeit gebührt mein besonderer Dank Herrn Prof. Dr. Edgar Brunner, der die Entstehung dieser Arbeit mit vielen wertvollen Ratschlägen und Hinweisen begleitete und durch die Bereitstellung der Arbeitsmöglichkeiten in der Abteilung für Medizinische Statistik die Arbeit in der vorliegenden Form ermöglichte. Des Weiteren danke ich ihm für die Ermutigung zu Konferenzbeiträgen und für die Möglichkeit, in der statistischen Beratung wertvolle praktische Erfahrungen zu sammeln.
Für die Übernahme des Korreferats und für die Unterstützung im Rahmen des Promotionsstudiengangs *„Applied Statistics and Empirical Methods“* danke ich Herrn Prof. Dr. Martin Schlather.
Des Weiteren bedanke ich mich bei bei Herrn Dr. Frank Konietschke für seine stete und ausdauernde Diskussionsbereitschaft und das sorgfältige Korrekturlesen. Ein herzlicher Dank gebührt auch meiner Mutter, die mir bei der Beseitigung der sprachlichen Fehler dieser Arbeit eine besondere Hilfe war. Bedanken möchte ich mich auch für die Hinweise der Korrekturleser Frau Inga Knorr und Herrn Michael Reese.
Ich danke Herrn Dr. Jörg Kaufmann und Frau PD Dr. Silvia Obenauer für die Bereitstellung der Datensätze, sowie der Abteilung für Medizinische Statistik unter der Leitung von Herrn Prof. Dr. Tim Friede für die stets freundliche und angenehme Arbeitsatmosphäre.
Ein besonderer Dank gilt meinen Eltern und meinem Freund, Michael Reese, für die stetige Unterstützung.

Katharina Lange
Göttingen, im Januar 2011

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Motivation

Da die schnelle und richtige Diagnose einer Erkrankung eine der wichtigsten Voraussetzungen für eine effektive Therapie bildet, rückt die Entwicklung neuer diagnostischer Verfahren zunehmend in den Fokus der Wissenschaft. Neben Therapiestudien entwickeln Diagnosestudien sich daher seit einigen Jahren zum zweiten Standbein klinischer Forschung. Die Evaluation neuer diagnostischer Verfahren soll dabei „*nach denselben wissenschaftlichen und regulatorischen Richtlinien erfolgen, wie auch bei der Entwicklung von Arzneimitteln und anderen medizinischen Produkten*“ (EMEA, 2009). Diesen Anforderungen werden diagnostische Studien jedoch selten gerecht, da auf dem relativ jungen Forschungsgebiet der Diagnosestudien die statistische Methodik zur Analyse nur punktuell existiert. Eine der Hauptursachen hierfür liegt in der Vielfalt der Analysekonzepte und Datenstrukturen. In diagnostischen Studien gibt es zum einen eine Vielzahl an Möglichkeiten zur Definition diagnostischer Güte: Am häufigsten werden Sensitivität und Spezifität verwendet, aber man findet auch alternative Ansätze wie die Fläche unter der ROC-Kurve (AUC) oder prädiktive Werte. Zum anderen existieren – insbesondere im Bereich der bildgebenden Diagnostik – verschiedene Versuchsdesigns. Bei der Evaluation bildgebender Diagnoseverfahren wird von den Zulassungsbehörden verlangt, dass die Auswertung des erhobenen Bildmaterials durch zwei (vorzugsweise sogar drei) unterschiedliche Ärzte (sogenannte Reader) erfolgt, um auf diese Art und Weise die Güte des diagnostischen Verfahrens unabhängig von der Qualität des Arztes beurteilen zu können. In Abhängigkeit von der Frage, ob für die Auswertung verschiedener Diagnoseverfahren unterschiedliches Fachpersonal erforderlich ist und von der ethischen Vertretbarkeit der Anwendung unterschiedlicher Verfahren am gleichen Patienten entstehen allein hier bereits vier Basisdesigns für Studien.
Die Literatur zur Analyse von Diagnosestudien beschränkt sich nun häufig auf spezielle Designs und ein einzelnes diagnostisches Gütemaß. So findet man eine umfangreiche – allerdings auf die AUC begrenzte – Übersicht über die verschiedenen (parametrischen wie nichtparametrischen) Analysemethoden in faktoriellen Versuchsanlagen bei Obuchowski u. a. (2004) und Obuchowski (2007). Der erste rein nichtparametrische Analyseansatz geht hierbei auf Song (1997) zurück, der die von DeLong u. a. (1988) entwickelte Theorie der U-Statistiken auf die Evaluation zweifaktorieller gekreuzter Versuchspläne erweitert. Eine weitere nichtparametrische Methodik wird von Kaufmann u. a. (2005) vorgeschlagen, in der, basierend auf den Ideen von Brunner u. a. (2002), die Theorie der multivariaten Rangstatistiken zur AUC-Analyse verwendet wird. Ein alternatives Konzept besteht in der Anwendung von Resampling-Methoden wie Bootstrapping (siehe zum Beispiel Beiden u. a., 2000) oder Jackknifing (beispielsweise bei Dorfmann u. a., 1992). Zur Analyse von prädiktiven Werten oder Sensitivität und Spezifität werden diese Ansätze jedoch nicht erweitert: in faktoriellen Versuchsplänen wird zur Evaluation von Sensitivität und Spezifität in der Regel die von Liang und Zeger (1986) entwickelte Methodik der Generalized Estimating Equations (GEE) verwendet (siehe zum Beispiel Smith und Hadgu, 1992). Analysemethoden zur Evaluation prädiktiver Werte findet man in der Literatur kaum, da diese auf Grund ihrer Abhängigkeit von der Erkrankungswahrscheinlichkeit selten als Gütemaße diagnostischer Studien herangezogen werden können. Neuere Methoden findet man mit einem Ansatz zur Berechnung von Konfidenzintervallen für den Ein-Stichproben-Fall bei Mercaldo u. a. (2007), Analyseme-

thoden für den Zwei-Stichproben-Vergleich beispielsweise bei Leisenring u. a. (2000) oder Moskowitz und Pepe (2006).
Häufig erhöht sich der Komplexitätsgrad diagnostischer Studien zusätzlich durch das Vorliegen sogenannter Clusterdaten. Diese treten immer dann auf, wenn zur Sicherung der Diagnose oder zur genaueren Lokalisation der Erkrankung mehrere Proben oder Körperregionen des gleichen Patienten untersucht werden. Da ein Patient in diesem Fall sowohl gesunde als auch kranke Beobachtungseinheiten aufweisen kann, sind hier die Kollektive der Gesunden und der Kranken nicht mehr unabhängig, was die Komplexität der zu verwendenden Methodik deutlich ansteigen lässt. Obwohl bei vielen diagnostischen Tests auf natürlichem Weg Clusterdaten entstehen, gibt es keinen einheitlichen Leitfaden zum Umgang mit Daten dieses Typs. So kritisieren Gönen u. a. (2001), dass die Reduktion von Clusterdaten auf eine Beobachtung pro Subjekt – trotz des hiermit verbundenen Verlustes an Information – in der Praxis noch immer Anwendung findet (beispielsweise bei Liu u. a., 2004).
Auch die Literatur zur Analyse von Diagnosestudien mit Clusterdaten ist häufig auf ausgewählte Designs und Gütemaße beschränkt. Zur Evaluation der Fläche unter der ROC-Kurve präsentiert Obuchowski (1997) durch eine Weiterentwicklung der Methodik von DeLong u. a. (1988) erstmals ein Verfahren zur AUC-Analyse von Clusterdaten im Zwei-Stichproben-Fall. Einen alternativen Ansatz mit Hilfe der Theorie stochastischer Prozesse stellen Li und Zhou (2008) vor: Für einen einfaktoriellen Versuchsplan wird eine Analysemethode für die ROC-Kurve selbst sowie die zugehörigen AUCs mittels Monte-Carlo-Simulationen entwickelt. Werner und Brunner (2007) hingegen folgen bei der Evaluation von Clusterdaten der Methodik von Kaufmann u. a. (2005) und erweitern diese Ideen der multivariaten Rankstatistiken auf Clusterdaten. Die Methodik gilt speziell für einfaktorielle sowie zweifaktorielle gekreuzte Versuchspläne. Konietschke und Brunner (2009) präsentieren hierzu schließlich eine Approximation für kleine Stichproben, welche eine wichtige theoretische Grundlage für diese Arbeit bildet. Sämtliche präsentierte Methoden bleiben jedoch auf die Analyse der Fläche unter der ROC-Kurve beschränkt. Mit einem Bootstrap-Ansatz zur Evaluation eines verbunden Zwei-Stichproben-Designs mit Clusterdaten findet man bei Rutter (2000) einen der seltenen Ansätze zur einheitlichen Analyse von AUC, Sensitivität und Spezifität, jedoch keine Erweiterung auf prädiktive Werte. Diese finden für den Fall von Clusterdaten in der Literatur keine gesonderte Beachtung.

Mit dieser Arbeit soll der präsentierten Vielzahl an analytischen Konzepten zur Evaluation von Diagnosestudien eine einheitliche, nichtparametrische Analysemethode für AUC, Sensitivität, Spezifität und prädiktive Werte in Studien mit und ohne Clusterdaten gegenübergestellt werden. Die hier präsentierte Methodik lässt sich dabei insbesondere auf alle vier Basisdesigns anwenden und sogar auf höher faktorielle Designs erweitern. Mit dem vorgestellten Ansatz kann somit ein Großteil der Diagnosestudien übersichtlich auf vergleichbare Art und Weise evaluiert werden.

1.2 Aufbau der Arbeit

Nach einer Präsentation der verschiedenen diagnostischen Gütemaße in den Abschnitten 2.1–2.3 wird durch die Darstellung von Sensitivität und Spezifität als Flächen unter speziellen ROC-Kurven (Abschnitt 2.4) die Grundlage für den einheitlichen Analyseansatz gelegt. Eine Präsentation der verschiedenen Basisdesigns in Studien mit und ohne Clusterdaten folgt in Kapitel 3. Den theoretischen Kern dieser Arbeit bildet die in Kapitel 4 präsentierte Theorie der nichtparametrischen Analyse verbundener Stichproben. Als Anwendungsmöglichkeit dieser Methodik werden in Kapitel 5 Analysemethoden für die vier Basisdesigns vorgestellt, deren Möglichkeiten und Grenzen in der praktischen Anwendung in Kapitel 6 mit Hilfe von Simulationen aufgezeigt werden. Die Beispiele in Kapitel 7 dienen zur Veranschaulichung der präsentierten Methodik. Die Arbeit schließt mit einem Ausblick auf mögliche Erweiterungen und Entwicklungen der in dieser Arbeit dargelegten Konzepte.

2 Maßzahlen diagnostischer Güte

Bei der Evaluation eines neuen diagnostischen Verfahrens stellt sich, wie auch bei Entwicklung neuer Therapiemethoden, im ersten Planungsschritt stets die Frage, wie die Effizienz des Verfahrens zu bemessen ist. Bei Therapiestudien wird diese Frage durch die Wahl eines adäquaten (primären) Endpunktes beantwortet, indem die Güte der Therapiemethode durch das Maß an Veränderung in diesem Endpunkt definiert wird. Diagnosestudien stellen hier einen Sonderfall klinischer Studien dar. Neben dem Ergebnis des diagnostischen Tests wird auch der sogenannte Goldstandard bestimmt, welcher stellvertretend für den wahren Gesundheitszustand des Patienten erhoben wird. Definitionsgemäß ist ein Patient demnach gesund, wenn der Goldstandard negativ ist, und krank, wenn der Goldstandard positiv ist. Damit steht und fällt die Qualität einer Diagnosestudie mit der Verlässlichkeit des gewählten Goldstandards, weswegen dieser stets auf Grundlage des besten derzeit zur Verfügung stehenden Verfahrens ermittelt werden muss. Die verschiedenen diagnostischen Gütemaße erfassen auf unterschiedliche Art und Weise die Übereinstimmung zwischen dem Goldstandard und dem Ergebnis des diagnostischen Verfahrens. Das Ergebnis des diagnostischen Tests kann hierbei dichotom (positiv / negativ), ordinal (beispielsweise in Form eines Schweregrades) oder gar metrisch sein. In Abhängigkeit vom Skalentyp des diagnostischen Testergebnisses und der konkreten klinischen Fragestellung an das diagnostische Verfahren ergeben sich verschiedene Möglichkeiten der Definition diagnostischer Gütemaße, welche in diesem Kapitel vorgestellt werden. Zur mathematischen Beschreibung dieser Effektmaße wird dabei einheitlich folgende Notation verwendet.

Notation 2.1 Für einen diagnostischen Test folge das Testresultat im Kollektiv der Gesunden der Verteilung der Zufallsvariablen $X_0 \sim F_0$, im Kollektiv der Kranken der Verteilung von $X_1 \sim F_1$ und im Gesamtkollektiv der Verteilung von $X \sim F$, wobei F_0, F_1 und F beliebige Verteilungsfunktionen in ihrer normalisierten Form seien (vergleiche Ruymgaart, 1980).
Es sei D eine Zufallsvariable, die den Goldstandard eines jeden Patienten beschreibe, das heißt, eine Zufallsvariable, die den Wert 0 annimmt, falls ein Patient tatsächlich gesund ist und den Wert 1, falls der Patient tatsächlich erkrankt ist.
Analog zur $(0,1)$-Kodierung des Goldstandards kennzeichne bei dichotomen Testresultaten die 0 ein negatives oder auch normales Testergebnis, die 1 beschreibe ein positives, beziehungsweise abnormales Testergebnis.
Bei nicht dichotomen, das heißt bei ordinalen oder metrischen Endpunkten, sei angenommen, dass große Werte im Testergebnis für ein höheres Maß an Abnormalität stehen als kleine. [1]

2.1 Sensitivität und Spezifität

Sensitivität und Spezifität sind die am häufigsten verwendeten Maße diagnostischer Güte, wenn der Endpunkt eines diagnostischen Verfahrens dichotom ist. In diesem Fall sind die Verteilungen F_0 und F_1 Bernoulliverteilungen und jeder erhobene Datenpunkt kann einem der vier Felder der Kontingenztafel in Tabelle 2.1

[1] Dieses kann ohne Beschränkung der Allgemeinheit angenommen werden, da durch eine geeignete Transformation eine derartige Ordnung der Daten stets erreicht werden kann.

zugeordnet werden.

Tabelle 2.1: Vierfeldertafel möglicher diagnostischer Testergebnisse

	Goldstandard: krank (D=1)	Goldstandard: gesund (D=0)
Test positiv (X=1)	richtig positiv	falsch positiv
Test negativ (X=0)	falsch negativ	richtig negativ

Die Sensitivität *se* ist nun definiert als die Wahrscheinlichkeit, dass eine tatsächlich vorliegende Erkrankung durch den diagnostischen Test erkannt wird, welche sich mathematisch durch

$$se = \mathrm{P}(\text{Test richtig positiv}) = \mathrm{P}(\text{Test positiv}|\text{krank}) = \mathrm{P}(\mathrm{X}=1|\mathrm{D}=1) = \mathrm{P}(\mathrm{X}_1 = 1)$$

darstellen lässt. Die Spezifität *sp* beschreibt als Gegenstück zur Sensitivität die Wahrscheinlichkeit, dass das Nicht-Vorhandensein einer Erkrankung durch das diagnostische Verfahren korrekterweise festgestellt wird

$$sp = \mathrm{P}(\text{Test richtig negativ}) = \mathrm{P}(\text{Test negativ}|\text{gesund}) = \mathrm{P}(\mathrm{X}=0|\mathrm{D}=0) = \mathrm{P}(\mathrm{X}_0 = 0).$$

Sensitivität und Spezifität sind somit die Erfolgswahrscheinlichkeiten der bernoulliverteilten Zufallsvariablen X_1 und X_0. Die Sensitivität beschreibt hierbei die Güte des diagnostischen Tests für kranke Patienten, während die Spezifität die Effizienz des Tests für gesunde Patienten erfasst. Eine umfassende Charakterisierung der diagnostischen Güte erfordert somit die Berechnung beider Maßzahlen; auf ein einzelnes, zusammenfassendes Effizienzkriterium muss bei dem Gütekonzept von Sensitivität und Spezifität verzichtet werden.

2.2 Receiver Operating Characteristic Curves und die zugehörigen AUCs

Die **R**eceiver **O**perating **C**haracteristic Curves oder kurz ROC-Kurven bilden eine Erweiterung des Konzeptes von Sensitivität und Spezifität auf diagnostische Tests mit ordinalen oder metrischen Endpunkten. Ist das Ergebnis eines diagnostischen Tests nicht dichotom, so sind Sensitivität und Spezifität nicht definiert, da für das Testergebnis mehr Ausprägungen als die Kategorien *„positiv“* und *„negativ“* existieren. Sollen trotzdem Sensitivität und Spezifität berechnet werden, muss ein Schwellenwert oder sogenannter Cut-Off γ gewählt werden. Durch diesen wird der Endpunkt künstlich dichotomisiert, indem ein Testergebnis als positiv definiert wird, wenn das Testresultat einen Wert größer als γ annimmt und als negativ, falls das Ergebnis kleiner als γ ist. Für jedes γ erhält man auf diese Art und Weise ein Wertepaar $\big(se(\gamma), sp(\gamma)\big)$. [2]
Die ROC-Kurve stellt nun die Veränderung von Sensitivität und Spezifität bei Variation des Schwellenwertes dar: Zur Berechnung dieser Kurve werden für jeden möglichen Schwellenwert γ Sensitivität und Spezifität in Abhängigkeit von γ bestimmt:

$$se(\gamma) = \mathrm{P}(\mathrm{X}_1 > \gamma) = 1 - \mathrm{F}_1(\gamma) \tag{2.1}$$

$$sp(\gamma) = \mathrm{P}(\mathrm{X}_0 < \gamma) = \mathrm{F}_0(\gamma). \tag{2.2}$$

[2] Ohne Beschränkung der Allgemeinheit sei hier angenommen, dass $\mathrm{P}(\mathrm{X}_i = \gamma) = 0,\ i = 0,1$. Diese Voraussetzung ist nur bei nichtstetigen Verteilungen nötig. In diesem Fall kann aber für jedes Paar aus Sensitivität und Spezifität auch stets ein Schwellenwert gefunden werden, der nicht auf der Originalskala enthalten ist und für den somit $\mathrm{P}(\mathrm{X}_i = \gamma) = 0,\ i = 0,1$ gilt.

Die so entstehenden Paare $\big(se(\gamma), sp(\gamma)\big)$ werden in einem Graphen mit der Sensitivität als Ordinate und 1-Spezifität als Abszisse aufgetragen. Die Verbindung dieser Punkte definiert schließlich die ROC-Kurve $\mathrm{ROC}(F_0, F_1)$ des diagnostischen Tests. Abbildung 2.1 verdeutlicht das Entstehen der ROC-Kurve am Beispiel zweier Verteilungsfunktionen, für die die zugehörigen Dichten existieren.

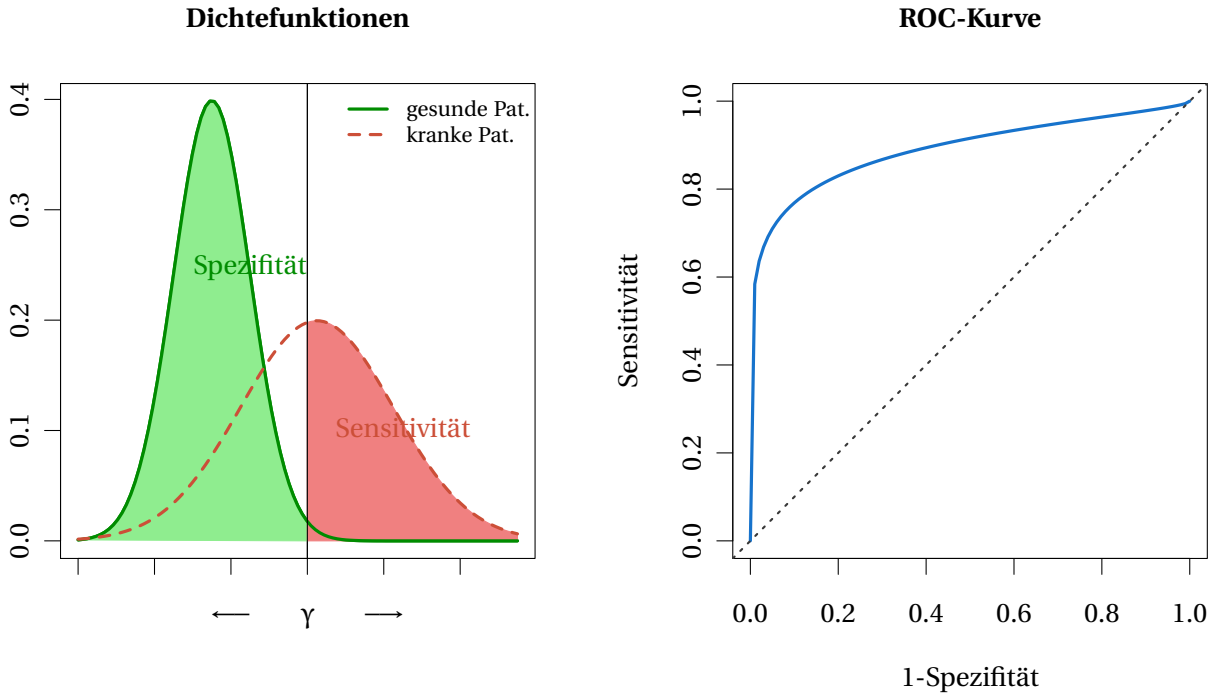

Abbildung 2.1: Entstehungsprozess der ROC-Kurve aus den Dichtefunktionen: Durch Variation von γ ergeben sich die verschiedenen Wertepaare von Sensitivität und 1-Spezifität, die durch die ROC-Kurve gegeneinander aufgetragen werden.

Die Hauptaufgabe eines diagnostischen Verfahrens besteht darin, Gesunde und Kranke möglichst gut zu unterscheiden. In dem Bereich, in dem sich die Dichtefunktionen der Testergebnisse der Gesunden und der Kranken überlappen, wird es – unabhängig vom gewählten Schwellenwert – stets zu Fehldiagnosen kommen: Das Überschneiden der Dichtefunktionen zeigt an, dass ein zufälliger gesunder Patient potenziell ein höheres Maß an Abnormalität aufweisen kann als ein zufälliger kranker Patient. (In diesem Fall wird der gesunde Patient kranker eingestuft als der tatsächlich Kranke.) Je nach Wahl des Schwellenwertes wird hier entweder der Gesunde oder der Kranke falsch diagnostiziert. Da mit kleiner werdendem Schwellenwert die Sensitivität steigt, wohingegen die Spezifität fällt (vergleiche Abbildung 2.1), kann durch Variation von γ also gelegentlich ein Austausch zwischen Sensitivität und Spezifität geschaffen werden. Es kann jedoch nicht verhindert werden, dass im Überschneidungsbereich Fehldiagnosen gestellt werden. Der Abstand der Dichtefunktionen voneinander stellt somit ein Maß diagnostischer Güte dar. Da nun mit größer werdendem Abstand der Dichtefunktionen auch der Abstand der ROC-Kurve zur Hauptwinkelhalbierenden steigt, lässt sich die Verlässlichkeit des diagnostischen Tests an der Form der ROC-Kurve erkennen (vergleiche Abbildung 2.2 auf der folgenden Seite). Hierdurch qualifiziert sich die ROC-Kurve als ein vom Schwellenwert γ unabhängiges Maß diagnostischer Güte.

Insbesondere im Hinblick auf spätere Inferenzstatistik stellt es nun einen großen Vorteil dar, wenn die Güte eines diagnostischen Tests mit Hilfe einer einzigen Größe zusammengefasst werden kann. Zur besseren Handhabbarkeit ist deswegen ein Index zu finden, der die ROC-Kurve auf adäquate Art und Weise auf eine einzelne Kennzahl reduziert. Häufig wird hierbei die Fläche unter der ROC-Kurve (im Englischen **A**rea **u**nder

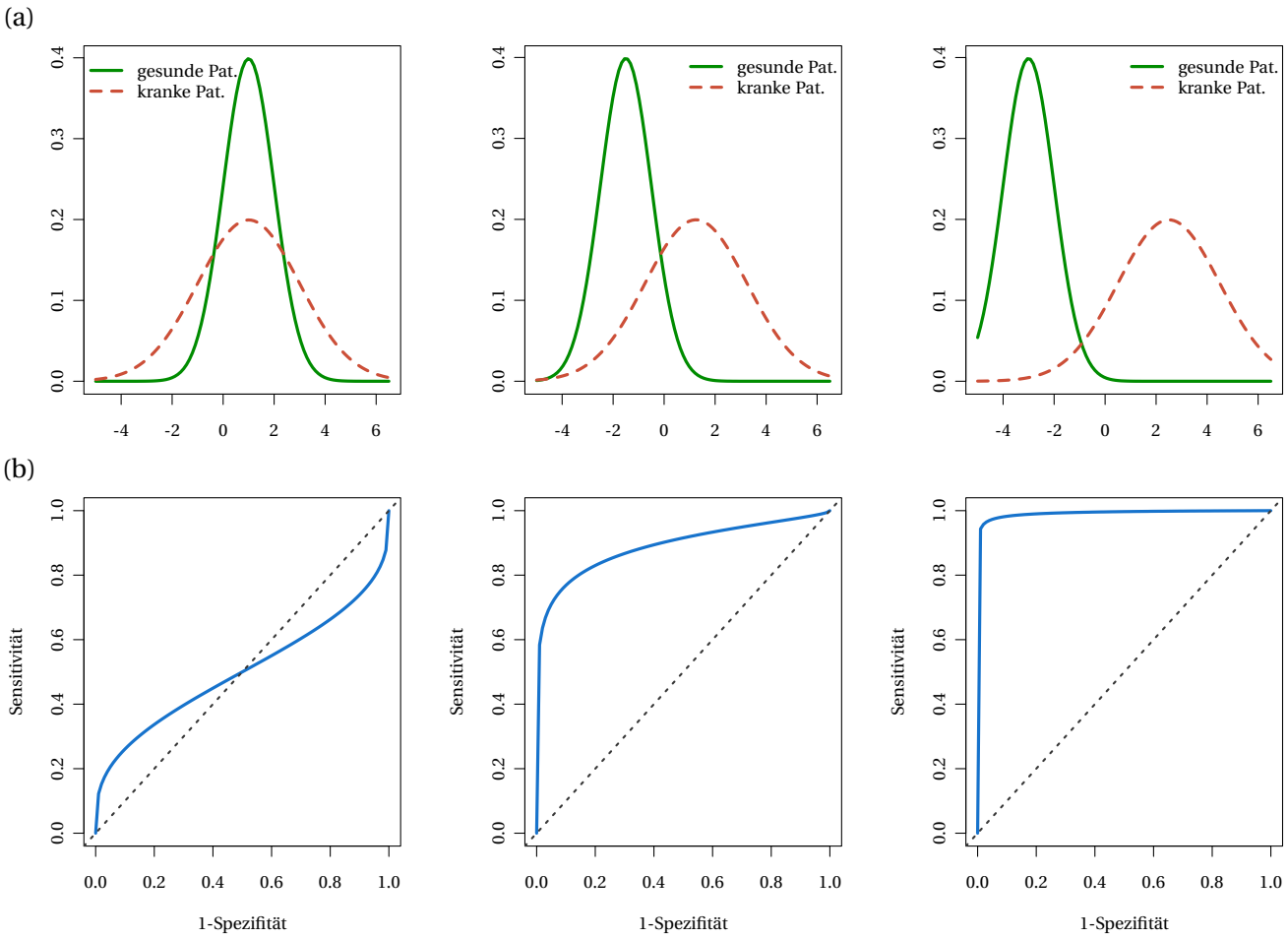

Abbildung 2.2: Zusammenhang zwischen Dichtefunktionen (a) und den zugehörigen ROC-Kurven (b): Je weiter die Dichtefunktionen auseinander liegen, desto größer ist der Abstand der ROC-Kurve zur Hauptwinkelhalbierenden.

the **C**urve, oder kurz AUC) verwendet:

$$\begin{aligned} \mathrm{AUC} = \mathrm{AUC}(\mathrm{F}_0, \mathrm{F}_1) &= \int_{\infty}^{-\infty} se(\gamma)\, d\left(1 - sp(\gamma)\right) = \int_{\infty}^{-\infty} \left(1 - \mathrm{F}_1(\gamma)\right) d\left(1 - \mathrm{F}_0(\gamma)\right) \\ &= 1 - \int_{-\infty}^{\infty} \mathrm{F}_1(\gamma)\, d\mathrm{F}_0(\gamma) = \int_{-\infty}^{\infty} \mathrm{F}_0(\gamma)\, d\mathrm{F}_1(\gamma). \end{aligned} \tag{2.3}$$

Diese Kennzahl der ROC-Kurve hat eine sehr anschauliche Interpretation: Sie beschreibt die Wahrscheinlichkeit, dass ein zufälliger kranker Patient ein höheres Maß an Abnormalität aufweist als ein zufälliger gesunder Patient. Mathematisch wird diese Wahrscheinlichkeit (bei stetigen Zufallsvariablen X_0 und X_1) durch

$$\mathrm{P}(\mathrm{X}_0 < \mathrm{X}_1)$$

beschrieben und für den Fall nicht-stetiger Verteilungsfunktionen F_0 und F_1 zu

$$\mathrm{P}(\mathrm{X}_0 < \mathrm{X}_1) + \frac{1}{2}\mathrm{P}(\mathrm{X}_0 = \mathrm{X}_1)$$

erweitert. In Kombination mit Gleichung (2.3) lässt sich daher zusammenfassen:

$$\mathrm{AUC} = \int \mathrm{F}_0 d\mathrm{F}_1 = \mathrm{P}(\mathrm{X}_0 < \mathrm{X}_1) + \frac{1}{2}\mathrm{P}(\mathrm{X}_0 = \mathrm{X}_1). \tag{2.4}$$

Die Fläche unter der ROC-Kurve entspricht damit dem durch den Zähler der Mann-Whitney Statistik geschätzten nichtparametrischen relativen Effekt (Mann und Whitney, 1947). Dieser bereits von Bamber (1975) und Hanley und McNeil (1982) erkannte Zusammenhang lässt die Methodik der nichtparametrischen Statistik im Bereich der Diagnosestudien einen natürlichen Anwendungsbereich finden.

2.3 Prädiktive Werte

Während Sensitivität, Spezifität und die Fläche unter der ROC-Kurve bedeutende Maßzahlen in diagnostischen Studien darstellen, sind prädiktive Werte ein wichtiges Handwerkszeug in der klinischen Praxis. Sie beantworten die beiden wohl entscheidendsten diagnostischen Fragen der Kliniker: „*Mit welcher Wahrscheinlichkeit ist ein Patient nach einem positiven Testresultat wirklich erkrankt?*“ und „*Wie wahrscheinlich ist es, dass ein Patient nach negativen Testbefund tatsächlich gesund ist?*“ Die erste der beiden Fragen wird durch den positiv prädiktiven Wert p_+ beantwortet

$$p_+ = \mathrm{P}(\text{krank}|\text{Test positiv}) = \mathrm{P}(\mathrm{D} = 1|\mathrm{X} = 1),$$

während die zweite Frage ihre Antwort im negativ prädiktiven Wert p_- findet

$$p_- = \mathrm{P}(\text{gesund}|\text{Test negativ}) = \mathrm{P}(\mathrm{D} = 0|\mathrm{X} = 0).$$

Prädiktive Werte setzen somit, ebenso wie Sensitivität und Spezifität, einen dichotomen Endpunkt des diagnostischen Tests voraus. Der entscheidende Vorteil prädiktiver Werte im Vergleich zu Sensitivität und Spezifität liegt dabei in deren praktischer Relevanz: Die Aussagekraft eines negativen oder positiven Testergebnisses lässt sich durch den Kliniker hier unmittelbar beurteilen. Prädiktive Werte bringen leider aber auch einen nicht zu vernachlässigenden Nachteil mit sich: Sie sind nicht allein von der Güte des diagnostischen Tests abhängig, sondern ebenfalls von der Prävalenz der Erkrankung. Die Prävalenz oder Voraberkrankungswahrscheinlichkeit π ist hierbei die Erkrankungswahrscheinlichkeit eines zufälligen Patienten (bevor der diagnostische Test durchgeführt wurde):

$$\pi = \mathrm{P}(\mathrm{D} = 1).$$

Die Bayes'sche Formel (Bayes, 1763), welche später ein wichtiges theoretisches Fundament der Analyse bilden wird, veranschaulicht diese Abhängigkeit der prädiktiven Werte von der Prävalenz. Sie setzt Sensitivität (se), Spezifität (sp) und Prävalenz (π) mit den prädiktiven Werten in Zusammenhang, indem sie diese als Funktionen von se, sp und π darstellt:

$$p_+ = f_+(se, sp, \pi) = \frac{se \cdot \pi}{se \cdot \pi + (1 - sp) \cdot (1 - \pi)} \tag{2.5}$$

$$p_- = f_-(se, sp, \pi) = \frac{sp \cdot (1 - \pi)}{sp \cdot (1 - \pi) + (1 - se) \cdot \pi}. \tag{2.6}$$

Sensitivität und Spezifität selbst sind hierbei unabhängig von der Voraberkrankungswahrscheinlichkeit, da diese separate Gütemaße für das Kollektiv der Gesunden oder der Kranken sind.

2.4 Eine einheitliche Interpretation von Sensitivität, Spezifität und AUC

Die Fläche unter der ROC-Kurve sowie Sensitivität und Spezifität sind auf den ersten Blick grundsätzlich verschiedene Gütekonzepte zur Beurteilung diagnostischer Verfahren. Lange und Brunner (2011) zeigen jedoch, dass Sensitivität und Spezifität sich als Flächen unter speziellen ROC-Kurven interpretieren lassen. Die beiden ursprünglich verschiedenen Beurteilungskriterien sind somit zwei Seiten derselben Medaille. Insbesondere folgt aus der Arbeit von Lange und Brunner (2011), dass für das Gütekonzept von Sensitivität und Spezifität und für das diagnostische Effizienzmaß der Fläche unter der ROC-Kurve eine einheitliche Analysemethodik existiert.
Zur Darstellung dieser Ergebnisse sei angenommen, dass der zur Berechnung von Sensitivität und Spezifität erforderliche optimale Cut-Off γ aus Pilotstudien bekannt ist. Für den Fall, dass der Endpunkt des diagnostischen Tests bereits per Konstruktion dichotom ist, definiere man $\gamma = 0.5$, um auch für binäre Endpunkte die verallgemeinerte Definition

$$se = P(X_1 > \gamma) = 1 - F_1(\gamma)$$
$$sp = P(X_0 < \gamma) = F_0(\gamma)$$

von Sensitivität und Spezifität verwenden zu können. Hierbei sei analog zu Abschnitt 2.2 angenommen, dass ohne Beschränkung der Allgemeinheit $P(X_i = \gamma) = 0,\ i = 0,1$ gelte. Zur Darstellung von Sensitivität und Spezifität als Flächen unter speziellen ROC-Kurven sei des Weiteren durch

$$\Gamma(x) = \begin{cases} 0, & \text{für } x < \gamma \\ \frac{1}{2} & \text{für } x = \gamma \\ 1, & \text{für } x > \gamma \end{cases}$$

die normalisierte Version einer Ein-Punkt-Verteilung mit Masse bei γ definiert. Mit dieser Notation gilt nun folgende Behauptung:

Satz 2.1 *Unter den obigen Voraussetzungen sind Sensitivität und Spezifität Flächen unter speziellen ROC-Kurven, das heißt*

$$se = \int \Gamma \, dF_1 = AUC(\Gamma, F_1) \tag{2.7}$$
$$sp = \int F_0 \, d\Gamma = AUC(F_0, \Gamma). \tag{2.8}$$

BEWEIS. Nach Gleichung (2.1) ist die Sensitivität definiert als

$$se = P(X_1 > \gamma) = 1 - F_1(\gamma) = 1 - \int F_1 \, d\Gamma = \int \Gamma \, dF_1 = AUC(\Gamma, F_1).$$

Analog gilt für die Spezifität

$$sp = P(X_0 < \gamma) = F_0(\gamma) = \int F_0 \, d\Gamma = AUC(F_0, \Gamma).$$

□

Abbildung 2.3 veranschaulicht diesen Zusammenhang zwischen Sensitivität, Spezifität und AUC durch eine Darstellung der Dichtefunktionen der diagnostischen Testergebnisse und der zugehörigen speziellen ROC-Kurven, deren AUCs Sensitivität und Spezifität entsprechen.

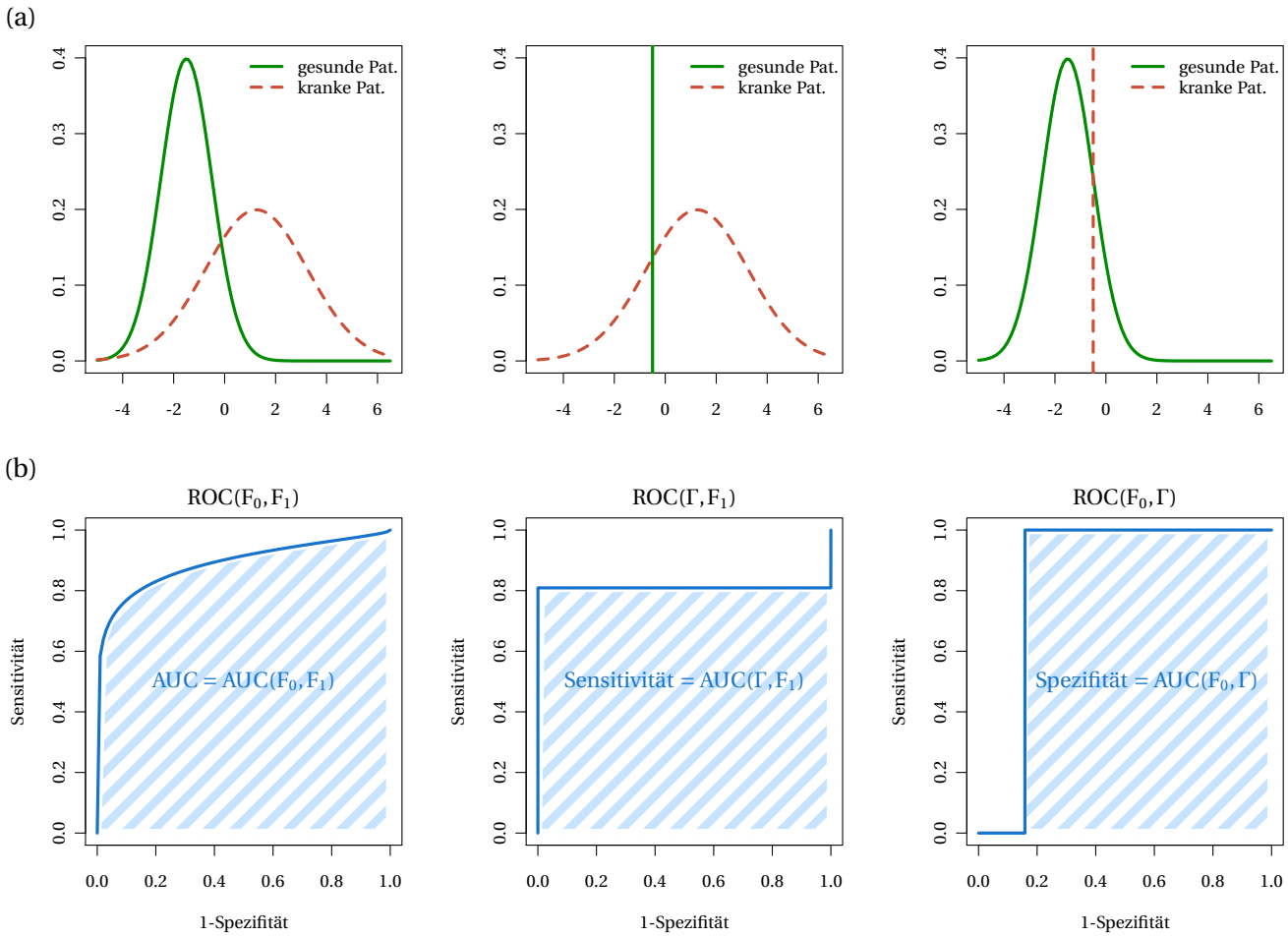

Abbildung 2.3: Zusammenhang zwischen Sensitivität, Spezifität und AUC: (a) zeigt die Dichtefunktionen der Verteilungsfunktionen; die vertikale Linie repräsentiert hierbei die „*Dichte*“ der Ein-Punkt-Verteilung Γ. In (b) sind die zu (a) gehörige ROC-Kurven abgebildet, deren AUCs die gewöhnliche AUC, Sensitivität und Spezifität sind.

Es sei an dieser Stelle angemerkt, dass – in Anlehnung an Lévy (1925) – Γ per Definition eine Verteilungsfunktion ist und $\mathrm{ROC}(F_0, \Gamma)$ und $\mathrm{ROC}(\Gamma, F_1)$ daher definitionsgemäß ROC-Kurven sind. Sensitivität und Spezifität sind unter diesem besonderen Blickwinkel also gewöhnliche AUCs. Diese Tatsache wird im Folgenden den zentralen Ansatzpunkt der einheitlichen Analysemethodik für die verschiedenen diagnostischen Gütemaße bilden.

2.5 Die Wahl des richtigen Gütemaßes

Die Vielfalt diagnostischer Gütemaße erfordert in der Planung und Auswertung von Diagnosestudien die Entscheidung, mit Hilfe welcher Maßzahl diagnostische Güte letztendlich zu bemessen ist. Auf der einen Seite stehen hierbei jene Gütemaße, die unabhängig von der Prävalenz die Zuverlässigkeit des diagnostischen Tests bestimmen: Sensitivität, Spezifität und die Fläche unter der ROC-Kurve. Diese Maßzahlen sind ausschließlich Eigenschaften des diagnostischen Tests selbst und eignen sich daher besonders bei der Zulassung diagnostischer Verfahren. Unabhängig von der Erkrankungswahrscheinlichkeit in der Studienpopulation erfassen sie die Zuverlässigkeit des diagnostischen Tests; die Güte kann hier nicht durch eine gezielte Rekrutierung von mehr gesunden oder kranken Patienten manipuliert werden.

In der klinischen Praxis folgt nun jeder Diagnose eine Therapieentscheidung. Diese Therapieentscheidungen sind in der Regel binärer Natur, das heißt, ein Patient wird entweder behandelt oder nicht behandelt. Durch die Entscheidung für oder gegen eine Therapie wird somit schließlich jedes diagnostische Testergebnis dichotomisiert. Am Ende der Entwicklung eines neuen diagnostischen Verfahrens sollte daher stets ein diagnostischer Test mit binärem Endpunkt stehen, welcher mittels Sensitivität und Spezifität evaluiert wird. Dieses Vorgehen wird von den Zulassungsbehörden indirekt verlangt, indem Sensitivität und Spezifität als primäre Zielkriterien gefordert werden (vergleiche EMEA, 2009; FDA, 2004). Solange die Wahl eines optimalen Schwellenwertes jedoch noch nicht erfolgt ist und der Endpunkt daher noch nicht binär ist, ist im Zuge der Entwicklung des diagnostischen Tests (das heißt klinische Studien in Phase I und II) die AUC-Analyse ein angemessenes Auswertungskriterium – auch um Schwellenwerte letztendlich festzulegen.
Auf der anderen Seite geben prädiktive Werte im Gegensatz zu Sensitivität, Spezifität und der AUC einen unmittelbaren Einblick in die Zuverlässigkeit des diagnostischen Verfahrens aus dem Blickwinkel der klinischen Praxis. Während Sensitivität, Spezifität und die Fläche unter der ROC-Kurve sich als Gütemaße bei der Zulassung neuer diagnostischer Verfahren anbieten, sind prädiktive Werte in diesem Bereich auf Grund ihrer Prävalenzabhängigkeit weniger geeignet: durch eine gezielte Rekrutierung zusätzlicher gesunder oder kranker Patienten lässt sich – unabhängig von der tatsächlichen Güte des diagnostischen Verfahrens – jeder beliebige prädiktive Wert erreichen. Um dieser Manipulation vorzubeugen, verlangt die EMEA (European Medicines Agency) in einem Guideline-Entwurf aus dem Jahr 2008, dass prädiktive Werte nur mit besonderer Vorsicht angegeben werden dürfen und nur wenn angenommen werden kann, dass die Prävalenz der Studienpopulation der Prävalenz der wirklichen Welt entspricht („*predictive values must be reported with caution and only when the study sample is considered to be representative of the prevalence in the real world*“, EMEA, 2008). In der späteren Guideline heißt es schließlich, dass die Studienpopulation der späteren Anwendungspopulation entsprechen soll („*subjects included in confirmatory trials should be representative of the population in which the diagnostic agent is intended to be used*“, EMEA, 2009). Ähnliche Forderungen finden sich in der entsprechenden FDA (U.S. Food and Drug Administration) Guideline („*the trials include the intended population in the appropriate clinical setting*“, FDA, 2004). Folgt man diesen Leitlinien der Zulassungsbehörden, so werden prädiktive Werte für Patienten mit durchschnittlicher Erkrankungswahrscheinlichkeit berechnet. Die Ergebnisse der Evaluation sind jedoch falsch für einen Patienten mit bekanntermaßen höherer oder niedriger Voraberkrankungswahrscheinlichkeit. Der Vorteil prädiktiver Werte im Vergleich zu Sensitivität und Spezifität besteht nun aber einzig darin, dass die Aussagekraft eines individuellen positiven oder negativen Testergebnisses eingeschätzt werden kann. Werden prädiktive Werte lediglich für den Durchschnittspatienten berechnet, verlieren sie den Vorteil der individuellen Interpretierbarkeit und werden somit redundant. Prädiktive Werte sollten daher stets in Abhängigkeit der Voraberkrankungswahrscheinlichkeit angegeben werden.[3] Dieses kann entweder durch eine Darstellung der prädiktiven Werte als Funktionen $p_{\pm}(\pi)$ der Prävalenz erfolgen oder indem man anhand bekannter Risikofaktoren, wie beispielsweise Alter, Geschlecht oder genetischer Prädisposition, die Voraberkrankungswahrscheinlichkeit für verschiedene Risikogruppen schätzt. Auf diese Art und Weise lassen sich für jede Gruppe separat prädiktive Werte berechnen, und für jeden Patienten erhält man so in Abhängigkeit vom persönlichen Riskoprofil einen individuellen prädiktiven Wert.
Die Evaluation neuer diagnostischer Verfahren sollte daher in zwei Schritten erfolgen: Zur Bestimmung der diagnostischen Güte sind Sensitivität, Spezifität und die Fläche unter der ROC-Kurve zu betrachten. Im Hinblick auf die Anwendbarkeit des neuen diagnostischen Tests in der klinischen Praxis sollten jedoch zusätz-

[3] Diese Prävalenzabhängigheit der prädiktiven Werte diskutieren bereits Diamond und Forrester (1979) am Beispiel der koronaren Herzkrankheit, wobei auch in dem dort präsentierten Ansatz die Bayes'sche Formel zur Berechnung der prädiktiven Werte verwendet wird.

lich risikoabhängige prädiktive Werte berechnet werden, um – ganz im Sinne der Definition evidenzbasierter Medizin – *„die beste externe Evidenz für die Entscheidung der medizinischen Versorgung des individuellen Patienten“* (Sackett u. a., 1996) heranziehen zu können.

3 Clustering und Versuchsdesigns – Datenstrukturen bei Diagnosestudien

Wie auch bei Therapiestudien können bei Diagnosestudien Datensätze vielfältiger Strukturen entstehen; die dabei in der Praxis am häufigsten vorkommenden Formen werden in diesem Kapitel präsentiert. Auf der einen Seite stehen hierbei die Datenstrukturen, welche durch das Auftreten sogenannter Clusterdaten entstehen (Abschnitt 3.1), die man immer dann vorfindet, wenn zur Sicherung der Diagnose oder zur genaueren Lokalisation der Erkrankung mehrere Proben oder Körperregionen des gleichen Patienten untersucht werden. Auf der anderen Seite existieren – insbesondere im Bereich der bildgebenden Diagnostik – verschiedene Versuchsdesigns. Die Zulassungsbehörden verlangen in diesem Fall, dass die Auswertung des erhobenen Bildmaterials durch mehrere Ärzte erfolgt. In Abhängigkeit von der Frage, ob für die Auswertung verschiedener Diagnoseverfahren unterschiedliches Fachpersonal erforderlich ist, und von der ethischen Vertretbarkeit der Anwendung unterschiedlicher Verfahren am gleichen Patienten, entstehen allein hier bereits vier Konstellationsmöglichkeiten für Studien, welche ausführlich in Abschnitt 3.2 vorgestellt werden.

3.1 Clustering

Häufig reicht in der Diagnostik die Beantwortung der einfachen Frage *„Ist der Patient gesund oder krank?“* nicht aus. Vielmehr werden oftmals mehrere Proben des gleichen Patienten auf die vermutete Erkrankung untersucht. Werden hierbei Proben verschiedener Körperregionen untersucht, kann die Erkrankung durch dieses Vorgehen genauer lokalisiert werden. Werden hingegen mehrere Proben an der gleichen Stelle erhoben, soll durch dieses Vorgehen die Zuverlässigkeit des diagnostischen Tests im Vergleich zur Einfachmessung erhöht werden. In beiden Fällen führt das Erheben mehrerer Proben pro Patient dazu, dass die kleinste Beobachtungseinheit nicht mehr der Patient selbst, sondern die Probe ist. Der Unterschied zwischen der Untersuchung verschiedener Körperregionen und der mehrfachen Untersuchung der gleichen Region liegt lediglich darin, dass die verschiedenen Proben im ersten Fall einen unterschiedlichen Goldstandard aufweisen können, im zweiten Fall muss der Goldstandard jeder Beobachtungseinheit der gleiche sein. Hier kann noch vom Goldstandard oder tatsächlichen Gesundheitszustand des Patienten gesprochen werden, während im anderen Fall vom tatsächlichen Gesundheitszustand der Probe gesprochen werden muss.
Diese unterschiedlichen Arten der Datenstrukturen führen zu folgender Definition:

Definition 3.1 *[Diagnosestudien mit Einfachmessungen und Clusterdaten] Eine Diagnosestudie heißt Diagnosestudie mit Einfachmessungen, wenn eine einzelne Körperregion ohne Messwiederholungen untersucht wird. Sie heißt Diagnosestudie mit Clusterdaten, wenn mehrere Proben des gleichen Patienten untersucht werden. Dabei bezeichne der Begriff Diagnosestudie mit Clusterdaten aus Messwiederholungen eine Diagnosestudie, in der die verschiedenen Proben des gleichen Patienten stets den gleichen Goldstandard aufweisen müssen.*

An dieser Stelle ist der Unterschied zwischen Messwiederholungen und verbundenen Stichproben zu betonen: Wird in der Diagnostik die gleiche Körperregion mit verschiedenen diagnostischen Tests oder durch

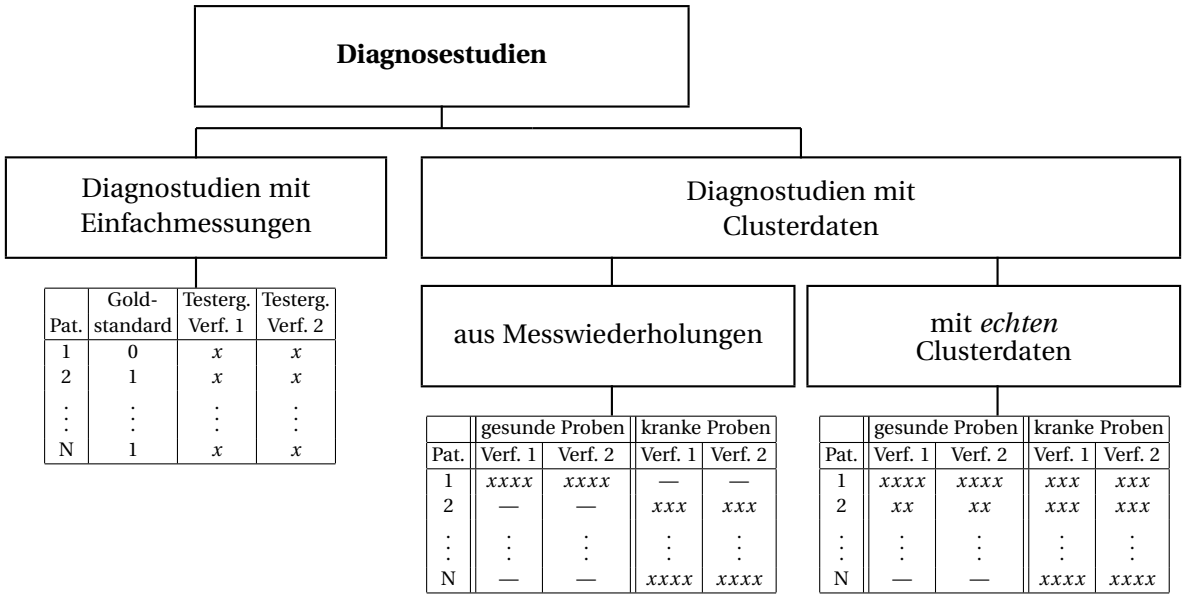

Abbildung 3.1: Übersicht der unterschiedlichen Datenstrukturen bei Diagnosestudien, jedes *x* entspricht hierbei einem erhobenen Messwert

unterschiedliche Ärzte untersucht, so handelt es sich nicht um Messwiederholungen, da die Untersuchungen unter verschiedenen Bedingungen stattfinden. Da die Beobachtungen jedoch am gleichen Patienten erhoben wurden, liegen hingegen durchaus verbundene Messungen vor. Abbildung 3.1 verdeutlicht die unterschiedlichen Datenstrukturen anhand von Diagnosestudien, bei denen an jeder Beobachtungseinheit jeweils zwei diagnostische Verfahren durchgeführt wurden.
Liegt eine Diagnosestudie ohne Clusterdaten vor, so können die Subjekte in gesunde und kranke Patienten eingeteilt werden. Bei Diagnosestudien mit Clusterdaten hingegen werden die Subjekte anhand ihrer Beobachtungseinheiten gemäß nachstehender Definition in eine der folgenden Kategorien eingeteilt.

Definition 3.2 *[Vollständige und unvollständige Fälle] In einer Diagnosestudie heißt ein Subjekt vollständig, wenn es sowohl gesunde als auch kranke Beobachtungseinheiten besitzt; es heißt unvollständig, wenn alle zugehörigen Beobachtungseinheiten den gleichen Gesundheitszustand aufweisen.*

Diese Idee der Kategorisierung der Subjekte mittels ihrer Beobachtungseinheiten geht auf Konietschke und Brunner (2009) zurück, deren Ansätze später auch den Ausgangspunkt der Analysemethodik in dieser Arbeit bilden werden.

Die Definition vollständiger und unvollständiger Subjekte erlaubt nun den Unterschied zwischen Diagnosestudien mit Clusterdaten aus Messwiederholungen und Studien mit echten Clusterdaten aus einem neuen Blickwinkel zu betrachten: Bei Studien mit Clusterdaten aus Messwiederholungen sind alle Patienten bereits nach Voraussetzung unvollständige Subjekte, wohingegen in einer beliebigen Diagnosestudie sowohl vollständige als auch unvollständige Fälle vorkommen können, jedoch nicht müssen.

3.2 Versuchsdesigns

Die Entwicklung neuer Diagnostika erfolgt schwerpunktmäßig im Bereich der bildgebenden Diagnostik. Durch bessere Kontrastmittel, höhere Auflösung der Kameras, computergestütze Bildanalyse und andere

medizintechnische Fortschritte können inzwischen viele invasive Diagnoseverfahren (wie beispielsweise die Angiographie oder eine Biopsie von potenziellem Tumorgewebe) wenigstens für Teile der Patienten durch ein nicht invasives bildgebendes Verfahren ersetzt werden. In diesem Abschnitt werden daher die typischen Versuchsdesigns von Diagnosestudien in der bildgebenden Diagnostik vorgestellt. Die Zulassungsbehörden verlangen in diesem Fall, dass die Auswertung des erhobenen Bildmaterials durch zwei (vorzugsweise sogar drei) unterschiedliche Ärzte (sogenannte Reader) erfolgt, um auf diese Art und Weise die Güte des diagnostischen Verfahrens unabhängig von der Qualität des Arztes beurteilen zu können (vergleiche EMEA, 2009; FDA, 2004). Da Diagnosestudien häufig dem Zweck der Evaluation neuer Verfahren im Vergleich zu bereits bekannten Methoden oder der Gegenüberstellung verschiedener, bereits bekannten Verfahren dienen, ist für alle in die Studie involvierten (bildgebenden) Diagnostika die Auswertung des Bildmaterials durch mehrere Reader vorgeschrieben. Nun kann es sein, dass bei verschiedenen Diagnoseverfahren unterschiedliches Fachpersonal für das Stellen der Diagnose erforderlich ist, wodurch sich folgende Möglichkeiten ergeben:

- Verschiedene Verfahren werden von den gleichen Readern ausgewertet.
- Verschiedene Verfahren werden von verschiedenen Readern ausgewertet.

Des Weiteren kann es auf Grund ethischer Einwände oder der medizinischen Unvereinbarkeit zweier Verfahren vorkommen, dass für verschiedene Diagnoseverfahren unterschiedliche Patientenkollektive erforderlich sind, das heißt, auch hier ergeben sich zwei Konstellationen:

- Verschiedene Verfahren werden an den gleichen Patienten getestet.
- Verschiedene Verfahren werden an verschiedenen Patienten getestet.

Durch diese Unterscheidungen ergeben sich nun vier mögliche faktorielle Designs, die in den folgenden Abschnitten präsentiert werden. Eine schematische Darstellung der Versuchsdesigns bei Einfachmessungen findet man beispielsweise bei Lange (2008, Kapitel 2.3), sodass an dieser Stelle nur die Schemata für Diagnosestudien mit Clusterdaten dargestellt werden.

3.2.1 Design 1

Design 1 betrachtet den einfachsten Fall, in welchem an allen Patienten alle Methoden getestet werden und jede Methode auch von jedem Reader ausgewertet werden kann. Die Ergebnisse eines solchen Designs lassen sich in folgender Tabelle übersichtlich darstellen (in Anlehnung an Konietschke und Brunner, 2009).

Tabelle 3.1: Schematische Darstellung Design 1

	Methode 1						Methode 2					
	Reader 1		Reader 2		Reader 3		Reader 1		Reader 2		Reader 3	
Pat.	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	xx	xxx	xx	xxx	xx	xxx	xx	xxx	xx	xxx	xx	xxx

Dieses Design tritt dann auf, wenn es ethisch bedenkenlos ist, die Patienten mit allen Diagnosemethoden zu untersuchen. Dabei müssen die auswertenden Reader allerdings geschult sein, die Ergebnisse aller Verfahren zu bewerten. Ein Beispiel hierzu wäre der Vergleich verschiedener Aufnahmegeschwindigkeiten und Auflösungen eines Gerätes (zum Beispiel CT oder MRT).

3.2.2 Design 2

Die Annahme verschiedener Reader für verschiedene Diagnoseverfahren aber gleicher Patienten für alle Verfahren führt zum zweiten faktoriellen Design, welches sich gleichermaßen als Tabelle darstellen lässt.

Tabelle 3.2: Schematische Darstellung Design 2

	Methode 1						Methode 2					
	Reader 1		Reader 2		Reader 3		**Reader 4**		**Reader 5**		**Reader 6**	
Pat.	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	xx	xxx	xx	xxx	xx	xxx	xx	xxx	xx	xxx	xx	xxx

Wie auch Design 1 tritt dieses Design auf, wenn es ethisch vertretbar ist, jeden Patienten mit allen Methoden zu untersuchen. Allerdings sind hier die auswertenden Ärzte der verschiedenen Methoden nicht die gleichen, wie es beispielsweise beim Vergleich einer MRT- mit einer Ultraschalluntersuchung der Fall sein kann. Die Ärzte, die das Bildmaterial der zweiten Methode auswerten, sind daher Reader 4 bis 6, und nicht wie ersten Design Reader 1 bis 3.

3.2.3 Design 3

Das dritte Design liegt vor, wenn unterschiedliche Diagnoseverfahren, welche von den gleichen Readern ausgewertet werden, an verschiedenen Patienten getestet werden. Dieses lässt sich durch folgende Tabelle darstellen.

Tabelle 3.3: Schematische Darstellung Design 3

	Methode 1						Methode 2					
	Reader 1		Reader 2		Reader 3		Reader 1		Reader 2		Reader 3	
Pat.	gesund	krank	gesund	krank	gesund	krank	gesund	krank	gesund	krank	gesund	krank
1	xxx	xx	xxx	xx	xxx	xx						
⋮	⋮	⋮	⋮	⋮	⋮	⋮						
N_1	xx	xxx	xx	xxx	xx	xxx						
N_1+1							xxxx	xxx	xxxx	xxx	xxxx	xxx
⋮							⋮	⋮	⋮	⋮	⋮	⋮
N							x	xx	x	xx	x	xx

Ein Design dieser Form liegt beispielsweise dann vor, wenn verschiedene Kontrastmittel miteinander verglichen werden sollen, wobei diese auf Grund ethischer Einwände nicht an den gleichen Patienten getestet werden dürfen.

3.2.4 Design 4

Werden verschiedene Methoden an unterschiedlichen Patienten getestet und die einzelnen Verfahren dabei nicht von den gleichen Readern ausgewertet, liegt das vierte und letzte Design vor.

Tabelle 3.4: Schematische Darstellung Design 4

	Methode 1						Methode 2					
	Reader 1		Reader 2		Reader 3		**Reader 4**		**Reader 5**		**Reader 6**	
Pat.	gesund	krank	gesund	krank	gesund	krank	gesund	krank	gesund	krank	gesund	krank
1	xxx	xx	xxx	xx	xxx	xx						
⋮	⋮	⋮	⋮	⋮	⋮	⋮						
N_1	xx	xxx	xx	xxx	xx	xxx						
N_1+1							xxxx	xxx	xxxx	xxx	xxxx	xxx
⋮							⋮	⋮	⋮	⋮	⋮	⋮
N							x	xx	x	xx	x	xx

Der Vergleich eines Kontrastmittels für ein MRT mit einem Kontrastmittel für eine Ultraschalluntersuchung liefert ein Beispiel für eine derartige Versuchsstruktur.

4 Analyse verbundener Stichproben

Die statistische Methodik zur Analyse von Diagnosestudien findet – unabhängig von dem Versuchsdesign, der Datenstruktur oder dem Gütemaß – ihr theoretisches Fundament stets in der Theorie der multivariaten, nichtparametrischen Analyse verbundener Stichproben. Daher wird in diesem Kapitel in allgemeiner Notation ein Ansatz zur nichtparametrischen Evaluation verbundener Stichproben vorgestellt. Hierbei wird zwischen verbundenen Stichproben aus Einfachmessungen (Kapitel 4.1) und verbundenen Stichproben aus Clusterdaten – inklusive Clusterdaten aus Messwiederholungen – (Kapitel 4.2) unterschieden. Im nachfolgenden Kapitel 5 wird diese Theorie schließlich zur Herleitung der Analysemethode der verschiedenen faktoriellen Designs verwendet.
Zu Gunsten einer besseren Lesbarkeit wird an dieser Stelle auf eine Einführung in die Standardnotation aus den Bereichen Nichtparametrik und lineare Modelle verzichtet; diese befindet sich im Kapitel A im Anhang.

4.1 Analyse verbundener Stichproben bei Einfachmessungen

4.1.1 Modell und Notation

Im Folgenden wird davon ausgegangen, dass für jedes an der Diagnosestudie teilnehmende Versuchssubjekt unter $l = 1,\dots,d$ verschiedenen Bedingungen Daten erhoben werden, das heißt, dass d verbundene Stichproben vorliegen. Verschiedene Bedingungen könnten hierbei beispielsweise als differierende Reader-Methoden-Kombinationen oder auch nur als unterschiedliche Reader oder Diagnosemethoden interpretiert werden. Die auf diese Art und Weise pro Patient erhobenen Messwerte werden gemäß dem in Abbildung 4.1 dargestellten Schema in einem Vektor $\mathbf{X}_{ik} = (\mathrm{X}_{ik}^{(1)},\dots,\mathrm{X}_{ik}^{(d)})'$ zusammengefasst.

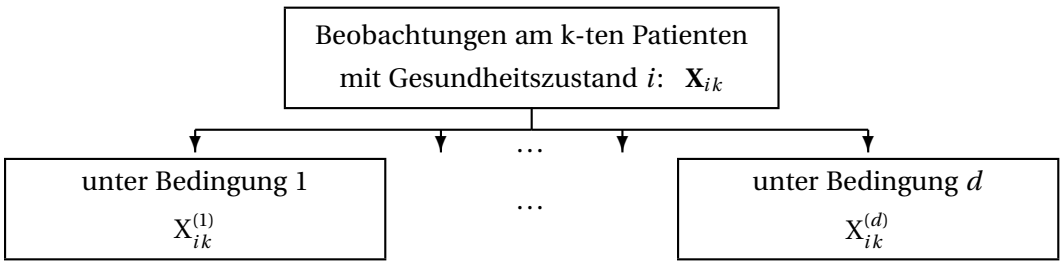

Abbildung 4.1: Notationsdarstellung bei Einfachmessungen

Diese Notation führt nun zu folgendem statistischen Modell.

Modell 1
Gegeben sei eine Diagnosestudie mit $\mathrm{N} = n_0 + n_1$ Subjekten, wobei n_0 die Anzahl der gesunden und n_1 die Anzahl der kranken Patienten kennzeichne. Es seien $\mathbf{X}_{ik} = (\mathrm{X}_{ik}^{(1)},\dots,\mathrm{X}_{ik}^{(d)})'$, $i = 0,1$, $k = 1,\dots,n_i$ die, wie vorangehend erläutert, strukturierten Zufallsvektoren mit Randverteilungen $\mathrm{X}_{ik}^{(l)} \sim \mathrm{F}_i^{(l)}$, $l = 1,\dots,d$, $i = 0,1$, $k = 1,\dots,n_i$, wobei die Funktionen $\mathrm{F}_i^{(l)}$ beliebige Verteilungsfunktionen (in normalisierter Version) mit Ausnahme des trivialen Falles einer Ein-Punkt-Verteilung seien.

Zur Bestimmung von Sensitivität und Spezifität sei $\gamma^{(l)}$ der Schwellenwert der l-ten Komponente. Das heißt, die Beobachtung $X_{ik}^{(l)}$ wird als gesund klassifiziert, falls sie kleiner als $\gamma^{(l)}$ ist und als krank, falls sie größer ist. Ohne Beschränkung der Allgemeinheit gelte hierbei stets $P(X_{ik}^{(l)} = \gamma^{(l)}) = 0$.
Zur Berechnung der prädiktiven Werte sei weiter angenommen, dass das Kollektiv in $g = 1,\ldots,G$ verschiedene Risikogruppen unterteilt werden kann, wobei die Erkrankungswahrscheinlichkeit in der g-ten Gruppe π_g betrage. π_g sei in Vorabstudien bereits durch $\widehat{\pi}_g = \frac{k_g}{q_g}$ geschätzt worden, wobei q_g der Gesamtstichprobenumfang der g-ten Gruppe in dieser Vorabstudie sei und k_g die Anzahl der in dieser Gruppe erkrankten Patienten. [1]
Es seien weiter

- $AUC^{(l)} = \int F_0^{(l)} dF_1^{(l)}$ die Fläche unter der ROC-Kurve der l-ten Komponente,
- $se^{(l)} = P(X_{1k}^{(l)} > \gamma^{(l)})$ und $sp^{(l)} = P(X_{0k}^{(l)} < \gamma^{(l)})$ die Sensitivität und die Spezifität der l-ten Komponente und
- $p_{g,+}^{(l)}$ und $p_{g,-}^{(l)}$ der positive und der negative prädiktive Wert der g-ten Risikogruppe ($g = 1,\ldots,G$) unter der l-ten Bedingung, $l = 1,\ldots,d$.

Diese Gütemaße werden schließlich durch

$$\begin{aligned} \mathbf{AUC} &= (AUC^{(1)},\ldots,AUC^{(d)})' \\ \mathbf{se} &= (se^{(1)},\ldots,se^{(d)})' \\ \mathbf{sp} &= (sp^{(1)},\ldots,sp^{(d)})' \quad \text{und} \\ \mathbf{p}_{g,\pm} &= (p_{g,\pm}^{(1)},\ldots,p_{g,\pm}^{(d)})' \end{aligned}$$

in Vektoren zusammengefasst. Analoges gelte für die Verteilungsfunktionen $\mathbf{F}_i = (F_i^{(1)},\ldots,F_i^{(d)})'$ sowie deren Schätzer $\widehat{\mathbf{F}}_i = (\widehat{F}_i^{(1)},\ldots,\widehat{F}_i^{(d)})'$, $i = 0,1$, wobei $\widehat{F}_i^{(l)}$ die empirische Verteilungsfunktion von $X_{i1}^{(l)},\ldots,X_{in_i}^{(l)}$ in ihrer normalisierten Form sei.

Zur späteren Beweisführung müssen folgende Regularitätsannahmen an das Modell gestellt werden:

Voraussetzung 4.1

(1) Für alle $l,r = 1,...,d$ sei die bivariate Verteilung von $(X_{ik}^{(l)},X_{ik}^{(r)})$ gleich für alle $k = 1,...,n_i$ in der jeweiligen Gruppe $i = 0,1$.
(2) Es gelte: $\tilde{N} = \min(n_0,n_1,q_1,\ldots,q_G) \to \infty$, derart, dass $\frac{N}{n_i} \to d_i$, $i = 0,1$, und $\frac{N}{q_g} \to e_g$, $g = 1,\ldots,G$.
(3) $se^{(l)}, sp^{(l)}, l = 1,\ldots,d$ und $\pi_g, g = 1,\ldots,G$ liegen in $(0,1)$, das heißt, sie sind echt kleiner 1 und echt größer 0.

Diese Voraussetzungen lassen sich wie folgt interpretieren: Voraussetzung (1) bedeutet, dass die Abhängigkeitsstruktur der Beobachtungen unter den verschiedenen Bedingungen für alle Patienten gleich ist. Diese Voraussetzung ist unmittelbar einsichtig, da sie die Annahme widerspiegelt, dass unterschiedliche Patienten gleichen Gesundheitszustandes die unabhängigen Messwiederholungen des Versuches darstellen. Annahme (2) stellt sicher, dass die realisierten, reziproken relativen Stichprobenumfänge $\frac{N}{n_0}$, $\frac{N}{n_1}$ und $\frac{N}{q_g}$, $g = 1,\ldots,G$ repräsentativ für die reziproken relativen Stichprobenumfänge der Gesamtpopulation sind. Diese Annahme folgt in intuitiver Art und Weise aus der Repräsentativität der Stichprobe. Die dritte Voraussetzung muss primär aus technischen Gründen gestellt werden, um eine positive Varianz zu erhalten. Da Experimente ohne Variabilität und der hieraus resultierenden Varianz von 0 jedoch selten Gegenstand der klinischen Forschung sind, sind diese Randfälle im Allgemeinen von untergeordnetem Interesse.

[1] Die zur Berechnung der prädiktiven Werte erforderlichen Daten kann man häufig aus der Literatur erhalten. Die Prävalenzen der unterschiedlichen Risikogruppen können dabei durchaus in verschiedenen Studien geschätzt worden sein.

4.1.2 Schätzung der diagnostischen Gütemaße

In diesem Abschnitt werden Schätzer für die verschiedenen diagnostischen Gütemaße präsentiert. Die in Satz 2.1 vorgestellte neue Interpretation von Sensitivität und Spezifität als Flächen unter speziellen ROC-Kurven wird dabei von zentraler Bedeutung sein: Es wird eine Methodik präsentiert, mittels derer die gewöhnlichen Schätzer von Sensitivität und Spezifität als normale AUC-Schätzer dargestellt werden können.

Schätzung der AUC

Es sei $R_{ik}^{(l)}$ der (Mittel-)Rang von $X_{ik}^{(l)}$ unter allen $N = n_0 + n_1$ Beobachtungen der l-ten Komponente und es sei $\overline{R}_{i\cdot}^{(l)} = n_i^{-1} \sum_{k=1}^{n_i} R_{ik}^{(l)}$ der Mittelwert der $R_{ik}^{(l)}$. Dann gilt:

Satz 4.1 *Für alle $l = 1, \ldots, d$ wird die Fläche unter der ROC-Kurve durch*

$$\widehat{\mathrm{AUC}}^{(l)} = \int \widehat{F}_0^{(l)} d\widehat{F}_1^{(l)} = \frac{1}{n_0} \left(\overline{R}_{1\cdot}^{(l)} - \frac{n_1 + 1}{2} \right) \tag{4.1}$$

konsistent geschätzt.

BEWEIS. Das Ersetzen der Verteilungsfunktionen $F_0^{(l)}$ und $F_1^{(l)}$ durch ihre empirischen Entsprechungen $\widehat{F}_0^{(l)}$ und $\widehat{F}_1^{(l)}$ führt zu diesem nichtparametrischen Schätzer der AUC. Dieser Ansatz wurde erstmals von Bamber (1975) für den eindimensionalen Fall verwendet. Zum Nachweis der Konsistenz und der Rangdarstellung (insbesondere für den allgemeinen Fall nicht-stetiger, mehrdimensionaler Verteilungen) sei auf die Arbeit von Brunner u. a. (2002) verwiesen. □

Schätzung von Sensitivität und Spezifität

Zur Schätzung von Sensitivität und Spezifität wird, analog zur Schätzung der gewöhnlichen AUC, zunächst ein Plug-In-Ansatz verwendet, wodurch man folgende Schätzer erhält:

$$\begin{aligned} \widehat{se}^{(l)} &= \int \Gamma^{(l)} d\widehat{F}_1^{(l)} \text{ für die Sensitivität und} \\ \widehat{sp}^{(l)} &= \int \widehat{F}_0^{(l)} d\Gamma^{(l)} \text{ für die Spezifität } (l = 1, \ldots, d), \end{aligned}$$

wobei $\Gamma^{(l)}$ eine Ein-Punkt-Verteilung mit Punktmasse bei $\gamma^{(l)}$ bezeichne. Zur Verallgemeinerung der gewöhnlichen AUC-Schätzung auf Sensitivität und Spezifität sei nun $\boldsymbol{\gamma} = (\gamma^{(1)}, \ldots, \gamma^{(d)})'$ der Vektor der Cut-Off-Punkte für jede Komponente. Es sei weiter $\{\boldsymbol{\gamma}, \ldots, \boldsymbol{\gamma}\}$ eine *Pseudostichprobe*, das heißt eine Stichprobe, in der jede Pseudobeoachtung den Wert $\boldsymbol{\gamma}$ annimmt. Anschaulich kann diese Stichprobe als eine Realisation einer (multivariaten) ein-Punkt-verteilten Zufallsvariablen mit Punktmasse bei $\boldsymbol{\gamma}$ interpretiert werden. Im Folgenden bezeichne n_{ps} den Stichprobenumfang dieser Pseudostichprobe. Mit dieser Notation gilt schließlich:

$$\widehat{\Gamma}^{(l)}(x) = \frac{1}{n_{ps}} \sum_{k=1}^{n_{ps}} c(x - \gamma^{(l)}) = c(x - \gamma^{(l)}) = \Gamma^{(l)}(x). \tag{4.2}$$

Obige Gleichung lässt sich wie folgt interpretieren: Die (geschätzte) empirische Verteilungsfunktion $\widehat{\Gamma}^{(l)}$ der Ein-Punkt-Verteilung $\Gamma^{(l)}$ konvergiert nicht nur (wie im gewöhnlichen Fall keiner Ein-Punkt-Verteilungen) gegen die tatsächliche Verteilungsfunktion, sondern $\widehat{\Gamma}^{(l)}$ und $\Gamma^{(l)}$ sind bereits finit gleich. Damit lassen sich sich die Schätzer von Sensitivität und Spezifität als gewöhnliche AUC-Schätzer darstellen.

Satz 4.2 *Für alle $l = 1,\dots,d$ werden Sensitivität und Spezifität durch*

$$\widehat{se}^{(l)} = \int \Gamma^{(l)} d\widehat{\mathrm{F}}_1^{(l)} = \int \widehat{\Gamma}^{(l)} d\widehat{\mathrm{F}}_1^{(l)} \tag{4.3}$$

$$\widehat{sp}^{(l)} = \int \widehat{\mathrm{F}}_0^{(l)} d\Gamma^{(l)} = \int \widehat{\mathrm{F}}_0^{(l)} d\widehat{\Gamma}^{(l)} \tag{4.4}$$

konsistent geschätzt.

BEWEIS. Da $\Gamma^{(l)} = \widehat{\Gamma}^{(l)}$ gelten die Gleichungen

$$\int \Gamma^{(l)} d\widehat{\mathrm{F}}_1^{(l)} = \int \widehat{\Gamma}^{(l)} d\widehat{\mathrm{F}}_1^{(l)} \text{ und} \tag{4.5}$$

$$\int \widehat{\mathrm{F}}_0^{(l)} d\Gamma^{(l)} = \int \widehat{\mathrm{F}}_0^{(l)} d\widehat{\Gamma}^{(l)}. \tag{4.6}$$

Die Konsistenz der Schätzer ergibt sich direkt aus den Sätzen 2.1 und 4.1: Nach Satz 2.1 sind Sensitivität und Spezifität Flächen unter speziellen ROC-Kurven, deren konsistente Schätzer in Satz 4.1 präsentiert werden: Während die gewöhnliche AUC mittels der Stichprobe der gesunden und der kranken Patienten geschätzt wird, werden für die Sensitivität nur die Stichprobe der Kranken und für die Spezifität nur die Stichprobe der Gesunden benötigt. Werden Sensitivität und Spezifität nun als Flächen unter speziellen ROC-Kurven betrachtet, so wird die nicht benötigte Stichprobe durch eine Ein-Punkt-verteilte Pseudostichprobe ersetzt, sodass aus Sensitivität und Spezifität *Pseudo-Zwei-Stichproben-Größen* werden. Die Konsistenz und Rangdarstellung dieser Schätzer ergibt sich daher unmittelbar aus Satz 4.1, denn für den Beweis wird die Bedingung, dass weder $\mathrm{F}_0^{(l)}$ noch $\mathrm{F}_1^{(l)}$ Ein-Punkt-Verteilungen sind, nicht benötigt (vergleiche Brunner u. a., 2002). Der spezielle Blickwinkel, der Sensitivität und Spezifität als AUCs (aus Ein-Punkt-Verteilungen) betrachtet, macht keinen weiteren Beweis erforderlich. □

Drei Anmerkungen seien an dieser Stelle noch besonders hervorgehoben:

(1) Die Schätzer $\widehat{se}^{(l)}$ und $\widehat{sp}^{(l)}$ werden zugunsten des einheitlichen Analyseansatzes mittels Rängen oder Integralen dargestellt. Dennoch entsprechen sie den gewöhnlichen Schätzern für Erfolgswahrscheinlichkeiten bernoulliverteilter Zufallsvariablen $\frac{\text{Anzahl richtig Diagnostizierter}}{\text{Gesamtanzahl}}$, sodass gelegentlich die Darstellung der Schätzer – nicht aber der Schätzer selbst – durch die einheitliche Interpretation von Sensitivität, Spezifität und AUC komplexer wird.

(2) Im Gegensatz zu $\int \mathrm{F}_0^{(l)} d\widehat{\mathrm{F}}_1^{(l)}$ und $\int \widehat{\mathrm{F}}_0^{(l)} d\widehat{\mathrm{F}}_1^{(l)}$ sind die Ausdrücke $\int \Gamma^{(l)} d\widehat{\mathrm{F}}_1^{(l)}$ und $\int \widehat{\Gamma}^{(l)} d\widehat{\mathrm{F}}_1^{(l)}$ (beziehungsweise $\int \widehat{\mathrm{F}}_0^{(l)} d\Gamma^{(l)}$ und $\int \widehat{\mathrm{F}}_0^{(l)} d\widehat{\Gamma}^{(l)}$) bereits finit gleich. Dieses bedeutet, dass durch die Pseudoschätzung, das heißt durch die Verwendung von $\widehat{\boldsymbol{\Gamma}} = (\widehat{\Gamma}^{(1)},\dots,\widehat{\Gamma}^{(d)})'$ an Stelle von $\boldsymbol{\Gamma} = (\Gamma^{(1)},\dots,\Gamma^{(d)})'$, kein Nachteil, wie beispielsweise eine Vergrößerung der Varianz, entsteht.

(3) Da die Gleichheit $\boldsymbol{\Gamma} = \widehat{\boldsymbol{\Gamma}}$, unabhängig von n_{ps}, für alle empirischen Verteilungsfunktionen $\widehat{\boldsymbol{\Gamma}}$ angenommen werden kann, kann der Stichprobenumfang n_{ps} beliebig gewählt werden. Die Schätzer $\widehat{se}^{(l)}$ und $\widehat{sp}^{(l)}$ sind unabhängig von der Wahl von n_{ps}.

Schätzung der prädiktiven Werte

Die Schätzer für die prädiktiven Werte ergeben sich nach der Schätzung von Sensitivität und Spezifität ohne weiteren Aufwand. Durch die Bayes'sche Formel (vergleiche Gleichung (2.5) und (2.6), Seite 7) werden die Funktionen f_+ und f_-, als Abbildungen von Sensitivität, Spezifität und Prävalenz auf die prädiktiven Werte p_+ und p_-, definiert. Setzt man in f_+ und f_- nun die entsprechenden Schätzer für Sensitivität, Spezifität und Prävalenz (der g-ten Riskogruppe) ein, so erhält man für jede Risikogruppe g konsistente Schätzer für die prädiktiven Werte $p_{g,+}$ und $p_{g,-}$:

Satz 4.3 *Für alle $l = 1,\dots,d$ und für alle Risikogruppen $g = 1,\dots,\mathrm{G}$ werden der positiv und der negativ prädiktive Wert durch*

$$\widehat{p}_{g,+}^{(l)} = f_+(\widehat{se}^{(l)}, \widehat{sp}^{(l)}, \widehat{\pi}_g) = \frac{\widehat{se}^{(l)} \cdot \widehat{\pi}_g}{\widehat{se}^{(l)} \cdot \widehat{\pi}_g + (1 - \widehat{sp}^{(l)}) \cdot (1 - \widehat{\pi}_g)} \tag{4.7}$$

$$\widehat{p}_{g,-}^{(l)} = f_-(\widehat{se}^{(l)}, \widehat{sp}^{(l)}, \widehat{\pi}_g) = \frac{\widehat{sp}^{(l)} \cdot (1 - \widehat{\pi}_g)}{\widehat{sp}^{(l)} \cdot (1 - \widehat{\pi}_g) + (1 - \widehat{se}^{(l)}) \cdot \widehat{\pi}_g}. \tag{4.8}$$

konsistent geschätzt.

BEWEIS. Die Gleichheit gilt nach der Bayes'schen Formel (Bayes, 1763). Die Konsistenz folgt aus der Konsistenz von $\widehat{se}^{(l)}, \widehat{sp}^{(l)}$ und $\widehat{\pi}_g$ und unter Voraussetzung 4.1/(3) aus dem Slutsky'schen Satz (Slutsky, 1925). □

4.1.3 Asymptotische Verteilung der Schätzer

In diesem Paragraphen wird die asymptotische Verteilung der im letzten Abschnitt vorgestellten Schätzer hergeleitet. Auch hier werden die Ergebnisse der AUC auf Sensitivität und Spezifität erweitert. Im Gegensatz zu Satz 4.1 wird für den Beweis der asymptotischen Normalität der AUC aber die Voraussetzung keiner Ein-Punkt-Verteilungen benötigt, sodass an dieser Stelle eine Erweiterung der Resultate auf Ein-Punkt-Verteilungen erforderlich ist.

Asymptotische Verteilung der AUC

Es sei $\widehat{\mathbf{AUC}} = (\widehat{\mathrm{AUC}}^{(1)},\dots,\widehat{\mathrm{AUC}}^{(d)})'$ der nach Satz 4.1 konsistente Schätzer von $\mathbf{AUC} = (\mathrm{AUC}^{(1)},\dots,\mathrm{AUC}^{(d)})'$. Weiter sei

$$\mathrm{B}^{(l)} = \int \mathrm{F}_0^{(l)} d\widehat{\mathrm{F}}_1^{(l)} - \int \mathrm{F}_1^{(l)} d\widehat{\mathrm{F}}_0^{(l)} + 1 - 2 \cdot \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)},\ l = 1,\dots,d$$

und $\mathbf{B} = (\mathrm{B}^{(1)},\dots,\mathrm{B}^{(d)})'$. Es bezeichne weiter $\mathbf{V}_{\mathrm{AUC}}$ die Kovarianzmatrix von $\sqrt{\mathrm{N}}\mathbf{B}$, und $\lambda_1,\dots,\lambda_d$ seien die Eigenwerte von $\mathbf{V}_{\mathrm{AUC}}$. Zur Herleitung der asymptotischen Verteilung wird eine zusätzliche Regularitätsannahme benötigt:

Voraussetzung 4.2
Es sei $\lambda_{min} = \min\{\lambda_1,\dots,\lambda_d\}$ das Minimum der Eigenwerte von $\mathbf{V}_{\mathrm{AUC}}$. Für dieses Minimum gelte:
$\lambda_{min} \geq \lambda_0 > 0$, wobei $\lambda_0 > 0$ eine beliebige Konstante sei.[2]

Satz 4.4 *Unter den Voraussetzungen 4.1 und 4.2 ist die Statistik $\sqrt{\mathrm{N}}(\widehat{\mathbf{AUC}} - \mathbf{AUC})$ asymptotisch normalverteilt mit Erwartungswert $\mathbf{0}$ und Kovarianzmatrix $\mathbf{V}_{\mathrm{AUC}}$.*

BEWEIS. Der Beweis geht hauptsächlich auf den zentralen Grenzwertsatz nach Lindeberg (Lindeberg, 1922) zurück. Eine detaillierte Ausarbeitung findet man bei Brunner u. a. (2002). □

[2] Da in einem nächsten Schritt gezeigt wird, dass der Ausdruck $\sqrt{\mathrm{N}}(\widehat{\mathbf{AUC}} - \mathbf{AUC})$ asymptotisch $\mathrm{N}(\mathbf{0}, \mathbf{V}_{\mathrm{AUC}})$-verteilt ist, ist die aus dieser Voraussetzung folgende Regularität von $\mathbf{V}_{\mathrm{AUC}}$ nach der klassischen Definition der multivariten Normalverteilung eine notwendige Annahme. Erweitert man die Definition der multivariaten Normalverteilung derart, dass 0-Eigenwerte zugelassen sind (siehe zum Beispiel Munzel, 1996, Abschnitt 5.3), so lässt sich obige Bedingung analog zu Kulle (1999, Kapitel 3.4) etwas abschwächen. Aus Gründen der Übersichtlichkeit wird in dieser Arbeit allerdings die stärkere Voraussetzung 4.2 zugrunde gelegt.

Asymptotische Verteilung von Sensitivität und Spezifität

Zur Herleitung der asymptotischen Verteilung der Sensitivität und Spezifität müssen die Resultate aus Satz 4.4 auf Ein-Punkt-Verteilungen erweitert werden. Es bezeichne $\widehat{\mathbf{se}} = (\widehat{se}^{(1)}, \ldots, \widehat{se}^{(d)})'$ den nach Satz 4.2 konsistenten Schätzer für $\mathbf{se} = (se^{(1)}, \ldots, se^{(d)})'$ und $\widehat{\mathbf{sp}} = (\widehat{sp}^{(1)}, \ldots, \widehat{sp}^{(d)})'$ den entsprechenden Schätzer für $\mathbf{sp} = (sp^{(1)}, \ldots, sp^{(d)})'$. Unter diesen Voraussetzungen gelten folgende asymptotische Resultate.

Satz 4.5 *Unter Voraussetzung 4.1 ist die Statistik* $\sqrt{n_1}(\widehat{\mathbf{se}} - \mathbf{se})$ *asymptotisch normalverteilt mit Erwartungswert* $\mathbf{0}$ *und Kovarianzmatrix* $\mathbf{V}_{se} = \mathrm{Cov}(\mathbf{\Gamma}(\mathbf{X}_{11})) = \mathrm{Cov}(\sqrt{n_1}\widehat{\mathbf{se}})$.

BEWEIS. Es gilt:

$$\widehat{se}^{(l)} - se^{(l)} = \int \Gamma^{(l)} d\widehat{\mathrm{F}}_1^{(l)} - se^{(l)} = \frac{1}{n_1} \sum_{k=1}^{n_1} \Gamma^{(l)}(\mathrm{X}_{1k}^{(l)}) - se^{(l)}, \quad \forall\, l = 1, \ldots d.$$

Da nun die $\mathbf{\Gamma}(\mathbf{X}_{1k})$, $k = 1, \ldots, n_1$ unabhängig und identisch verteilte Zufallsvektoren mit Erwartungswert $\mathbf{se}$ sind, folgt aus dem multivariaten zentralen Grenzwertsatz

$$\sqrt{n_1} \cdot (\widehat{\mathbf{se}} - \mathbf{se}) \xrightarrow{\mathscr{L}} \mathbf{U} \sim \mathrm{N}\Big(\mathbf{0}, \mathrm{Cov}\big(\mathbf{\Gamma}(\mathbf{X}_{11})\big)\Big).$$

Da

$$Cov\big(\mathbf{\Gamma}(\mathbf{X}_{11})\big) = Cov\big(\sqrt{n_1}\widehat{\mathbf{se}}\big)$$

ergibt sich schließlich die Behauptung. □

Analoges gilt (mit gleicher Beweisidee) für die Spezifität.

Satz 4.6 *Unter Voraussetzung 4.1 ist die Statistik* $\sqrt{n_0}(\widehat{\mathbf{sp}} - \mathbf{sp})$ *asymptotisch normalverteilt mit Erwartungswert* $\mathbf{0}$ *und Kovarianzmatrix* $\mathbf{V}_{sp} = \mathrm{Cov}(\mathbf{\Gamma}(\mathbf{X}_{01})) = \mathrm{Cov}(\sqrt{n_0}\widehat{\mathbf{sp}})$.

Die Sätze 4.5 und 4.6 lassen sich als Ein-Punkt-Fall von Satz 4.4 interpretieren. Um später zur Analyse von Sensitivität, Spezifität und AUC aber einen einzigen Analyseansatz – insbesondere ein einziges Programm – verwenden zu können, sollen Sensitivität und Spezifität wie bereits im letzten Abschnitt als Pseudo-Zwei-Stichproben-Größen aufgefasst werden. Das heißt, die Sensitivität kann als die AUC des Kollektivs der Kranken und einer Ein-Punkt-verteilten Population aus Gesunden angesehen werden. Diese wird aus einer Zufallsstichprobe (der Größe n_1) aus der kranken Population und einer Pseudo-Zufalls-Stichprobe (der Größe n_{ps}) aus der Ein-Punkt-verteilten Population der Gesunden geschätzt. Analoges gilt für die Spezifität.
Es soll daher im Folgenden gezeigt werden, dass sich die Resultate der Sätze 4.5 und 4.6 als Resultate eines Pseudo-Zwei-Stichproben-Falles darstellen lassen:

Korollar 4.1 *Es sei* $n_{ps} \in \mathbb{N}$ *der Stichprobenumfang der Pseudostichprobe.*[3] *Es gelte*

1. $n_{ps} + n_1 = \mathrm{N}_{se} \to \infty$ *derart, dass* $\frac{\mathrm{N}_{se}}{n_1} \leq \mathrm{M}_1 < \infty$ *und*
2. $n_0 + n_{ps} = \mathrm{N}_{sp} \to \infty$ *derart, dass* $\frac{\mathrm{N}_{sp}}{n_0} \leq \mathrm{M}_0 < \infty$,

wobei M_1 *und* M_0 *beliebige Konstanten seien.*

[3] Die Größe der Pseudostichprobe, welche zur Schätzung der Sensitivität herangezogen wird, muss selbstverständlich nicht der Größe der Pseudostichprobe entsprechen, die zur Schätzung der Spezifität verwendet wird. Die einheitliche Bezeichnung beider Pseudo-Stichprobenumfänge mit n_{ps} dient einer besseren Lesbarkeit.

Unter Voraussetzung 4.1 gilt dann

1. $\sqrt{\mathrm{N}_{se}} \cdot (\widehat{\mathbf{se}} - \mathbf{se})$ *ist asymptotisch normalverteilt mit Erwartungswert* $\mathbf{0}$ *und Kovarianzmatrix* $\frac{\mathrm{N}_{se}}{n_1} \cdot \mathbf{V}_{se} = \mathbf{V}_{se}^{ps}$.

Des Weiteren gilt

2. $\sqrt{\mathrm{N}_{sp}} \cdot (\widehat{\mathbf{sp}} - \mathbf{sp})$ *ist asymptotisch normalverteilt mit Erwartungswert* $\mathbf{0}$ *und Kovarianzmatrix* $\frac{\mathrm{N}_{sp}}{n_0} \cdot \mathbf{V}_{sp} = \mathbf{V}_{sp}^{ps}$.

BEWEIS. Da $\frac{\mathrm{N}_{se}}{n_1}$ und $\frac{\mathrm{N}_{sp}}{n_0}$ nach Voraussetzung beschränkt sind, folgt die Aussage direkt aus den Sätzen 4.5 und 4.6. □

Mit Hilfe des obigen Korollars finden Sensitivität und Spezifität nun auch hinsichtlich der asymptotischen Verteilung ihren Platz in der Gruppe der AUCs: Wie bereits zuvor wird zur Analyse der Sensitivität die Zufallsstichprobe der Gesunden durch eine Pseudostichprobe $\{\boldsymbol{\gamma}, \ldots, \boldsymbol{\gamma}\}$ ersetzt, zur Analyse der Spezifität die Stichprobe der Kranken. Korollar 4.1 zeigt, dass Sensitivität und Spezifität durch die Verwendung dieser Pseudostichprobe zu gewöhnlichen AUCs werden.

Asymptotische Verteilung der prädiktiven Werte

In diesem Abschnitt wird die asymptotische Verteilung der prädiktiven Werte hergeleitet. Bereits bei der Berechnung konsistenter Schätzer diente die Bayes'sche Formel als Ausgangspunkt der Methodik. Auch bei der Herleitung der asymptotischen Verteilung wird sie das theoretische Fundament bilden. Wie in den Gleichungen (2.5) und (2.6) seien f_+ und f_- die Bayes'schen Funktionen, die den Zusammenhang zwischen Sensitivität, Spezifität und Prävalenz (der g-ten Risikogruppe, $g = 1, \ldots, \mathrm{G}$) beschreiben:

$$p_+ = f_+(se, sp, \pi) = \frac{se \cdot \pi}{se \cdot \pi + (1 - sp) \cdot (1 - \pi)}$$
$$p_- = f_-(se, sp, \pi) = \frac{sp \cdot (1 - \pi)}{sp \cdot (1 - \pi) + (1 - se) \cdot \pi}$$

Im Folgenden seien

$$\mathbf{f}_+\left((\mathbf{se}', \mathbf{sp}', \pi_g)'\right) = \left(f_+(se^{(1)}, sp^{(1)}, \pi_g), \ldots, f_+(se^{(d)}, sp^{(d)}, \pi_g)\right) \text{ und}$$
$$\mathbf{f}_-\left((\mathbf{se}', \mathbf{sp}', \pi_g)'\right) = \left(f_-(se^{(1)}, sp^{(1)}, \pi_g), \ldots, f_-(se^{(d)}, sp^{(d)}, \pi_g)\right)$$

f_+ und f_- in ihrer multivariaten Form und

$$\mathbf{Df}_+ = \mathbf{Df}_+\left((\mathbf{se}', \mathbf{sp}', \pi_g)'\right) = \left(\frac{\partial \mathbf{f}_+}{\partial \mathbf{se}} \middle| \frac{\partial \mathbf{f}_+}{\partial \mathbf{sp}} \middle| \frac{\partial \mathbf{f}_+}{\partial \pi_g}\right) \begin{pmatrix} \mathbf{se} \\ \mathbf{sp} \\ \pi_g \end{pmatrix} \text{ und}$$

$$\mathbf{Df}_- = \mathbf{Df}_-\left((\mathbf{se}', \mathbf{sp}', \pi_g)'\right) = \left(\frac{\partial \mathbf{f}_-}{\partial \mathbf{se}} \middle| \frac{\partial \mathbf{f}_-}{\partial \mathbf{sp}} \middle| \frac{\partial \mathbf{f}_-}{\partial \pi_g}\right) \begin{pmatrix} \mathbf{se} \\ \mathbf{sp} \\ \pi_g \end{pmatrix}$$

seien die zu $\mathbf{f}_+$ und $\mathbf{f}_-$ gehörigen Jacobi-Matrizen der ersten partiellen Ableitungen an der Stelle $(\mathbf{se}', \mathbf{sp}', \pi_g)'$. Es seien weiter $\mathbf{p}_\pm^g = \mathbf{f}_\pm\left((\mathbf{se}', \mathbf{sp}', \pi_g)'\right)$ die prädiktiven Werte der g-ten Risikogruppe und $\widehat{\mathbf{p}}_\pm^g = \mathbf{f}_\pm\left((\widehat{\mathbf{se}}', \widehat{\mathbf{sp}}', \widehat{\pi}_g)'\right)$ seien die nach Satz 4.3 zugehörigen konsistenten Schätzer. Es seien weiter $\mathbf{V}_{se}$ und $\mathbf{V}_{sp}$ die im letzten Abschnitt definierten Kovarianzmatrizen von $\sqrt{n_1}\widehat{\mathbf{se}}$ und $\sqrt{n_0}\widehat{\mathbf{sp}}$. Des Weiteren sei $\sigma_g^2 = \pi_g(1 - \pi_g)$ die Varianz von $\sqrt{m_g}\widehat{\pi}_g$, $g = 1, \ldots, \mathrm{G}$. Für die prädiktiven Werte lässt sich nun folgendes Resultat zeigen.

Satz 4.7 *Unter Voraussetzung 4.1 und mit obiger Notation gilt*

$$\sqrt{\mathrm{N}}(\widehat{\mathbf{p}}_{+}^{g}-\mathbf{p}_{+}^{g})=\sqrt{\mathrm{N}}\left[\mathbf{f}_{+}\begin{pmatrix}\widehat{\mathbf{se}}\\ \widehat{\mathbf{sp}}\\ \widehat{\pi}_g\end{pmatrix}-\mathbf{f}_{+}\begin{pmatrix}\mathbf{se}\\ \mathbf{sp}\\ \pi_g\end{pmatrix}\right]\xrightarrow{\mathscr{L}}\mathbf{B}_{+}^{g}\sim\mathrm{N}\left(\mathbf{0},\mathbf{Df}_{+}\left[d_1\mathbf{V}_{se}\oplus d_0\mathbf{V}_{sp}\oplus e_g\sigma_g^2\right]\mathbf{Df}_{+}'\right)\quad \textit{und}$$

$$\sqrt{\mathrm{N}}(\widehat{\mathbf{p}}_{-}^{g}-\mathbf{p}_{-}^{g})=\sqrt{\mathrm{N}}\left[\mathbf{f}_{-}\begin{pmatrix}\widehat{\mathbf{se}}\\ \widehat{\mathbf{sp}}\\ \widehat{\pi}_g\end{pmatrix}-\mathbf{f}_{-}\begin{pmatrix}\mathbf{se}\\ \mathbf{sp}\\ \pi_g\end{pmatrix}\right]\xrightarrow{\mathscr{L}}\mathbf{B}_{-}^{g}\sim\mathrm{N}\left(\mathbf{0},\mathbf{Df}_{-}\left[d_1\mathbf{V}_{se}\oplus d_0\mathbf{V}_{sp}\oplus e_g\sigma_g^2\right]\mathbf{Df}_{-}'\right),\ g=1,\ldots,\mathrm{G},$$

wobei $\frac{\mathrm{N}}{n_i}\to d_i,\ i=0,1,\ \frac{\mathrm{N}}{q_g}\to e_g,\ g=1,\ldots,\mathrm{G}$ *(vergleiche Voraussetzung 4.1).*

BEWEIS. Nach den Sätzen 4.5 und 4.6 gilt:

$$\sqrt{n_1}\cdot(\widehat{\mathbf{se}}-\mathbf{se})\xrightarrow{\mathscr{L}}\mathbf{U}\sim\mathrm{N}(\mathbf{0},\mathbf{V}_{se})\text{ und}\tag{4.9}$$

$$\sqrt{n_0}\cdot(\widehat{\mathbf{sp}}-\mathbf{sp})\xrightarrow{\mathscr{L}}\mathbf{W}\sim\mathrm{N}(\mathbf{0},\mathbf{V}_{sp}).\tag{4.10}$$

Mit Hilfe des zentralen Grenzwertsatzes lässt sich für die Prävalenz in der g-ten Risikogruppe ($g=1,\ldots\mathrm{G}$) nun ein zu (4.9) und (4.10) vergleichbares Resultat zeigen:

$$\sqrt{q_g}\cdot(\widehat{\pi}_g-\pi_g)\xrightarrow{\mathscr{L}}\mathrm{Q}_g\sim\mathrm{N}(0,\sigma_g^2),\ g=1,\ldots,\mathrm{G}.\tag{4.11}$$

Da die reziproken relativen Stichprobenumfänge nach Voraussetzung 4.1 konvergieren ($\frac{\mathrm{N}}{n_i}\to d_i,\ i=0,1$ und $\frac{\mathrm{N}}{q_g}\to e_g,\ g=1,\ldots,\mathrm{G}$), können die Ausdrücke (4.9), (4.10) und (4.11) wie folgt dargestellt werden:

$$\begin{aligned}&\sqrt{\mathrm{N}}\cdot(\widehat{\mathbf{se}}-\mathbf{se})\xrightarrow{\mathscr{L}}\mathbf{U}''\sim\mathrm{N}(\mathbf{0},d_1\cdot\mathbf{V}_{se}),\\ &\sqrt{\mathrm{N}}\cdot(\widehat{\mathbf{sp}}-\mathbf{sp})\xrightarrow{\mathscr{L}}\mathbf{W}''\sim\mathrm{N}(\mathbf{0},d_0\cdot\mathbf{V}_{sp}),\text{ und}\\ &\sqrt{\mathrm{N}}\cdot(\widehat{\pi}_g-\pi_g)\xrightarrow{\mathscr{L}}\mathrm{Q}_g''\sim\mathrm{N}(0,e_g\cdot\sigma_g^2),\ g=1,\ldots,\mathrm{G}\end{aligned}$$

Sensitivität, Spezifität und Prävalenz werden im Fall der Einfachmessungen auf der Basis verschiedener Patientenkollektive geschätzt und sind daher stochastisch unabhängig. Daher ergibt sich

$$\sqrt{\mathrm{N}}\left[\begin{pmatrix}\widehat{\mathbf{se}}\\ \widehat{\mathbf{sp}}\\ \widehat{\pi}_g\end{pmatrix}-\begin{pmatrix}\mathbf{se}\\ \mathbf{sp}\\ \pi_g\end{pmatrix}\right]\xrightarrow{\mathscr{L}}\mathbf{B}\sim\mathrm{N}\left(\mathbf{0},d_1\mathbf{V}_{se}\oplus d_2\mathbf{V}_{sp}\oplus e_g\sigma_g^2\right),$$

wobei $\oplus$ die Kronecker-Summe bezeichne.
Die Anwendung von Cramers multivariatem δ-Satz mit den Funktionen $\mathbf{f}_{+}$ und $\mathbf{f}_{-}$ als Transformationsfunktionen führt schließlich zu dem Resultat:

$$\sqrt{\mathrm{N}}(\widehat{\mathbf{p}}_{\pm}^{g}-\mathbf{p}_{\pm}^{g})=\sqrt{\mathrm{N}}\left[\mathbf{f}_{\pm}\begin{pmatrix}\widehat{\mathbf{se}}\\ \widehat{\mathbf{sp}}\\ \widehat{\pi}_g\end{pmatrix}-\mathbf{f}_{\pm}\begin{pmatrix}\mathbf{se}\\ \mathbf{sp}\\ \pi_g\end{pmatrix}\right]\xrightarrow{\mathscr{L}}\mathbf{B}_{\pm}^{g}\sim\mathrm{N}\left(\mathbf{0},\mathbf{Df}_{\pm}\left[d_1\mathbf{V}_{se}\oplus d_0\mathbf{V}_{sp}\oplus e_g\sigma_g^2\right]\mathbf{Df}_{\pm}'\right),\ g=1,\ldots,\mathrm{G}.$$

□

Die Verwendung des δ-Satzes zum Beweis der asymptotischen Normalität der Schätzer der prädiktiven Werte geht hierbei auf Mercaldo u. a. (2007) zurück. Im Gegensatz zu den hier gezeigten Ideen führen Mercaldo u. a. (2007) den Beweis allerdings nur univariat und für eine feste Prävalenz π, die als bekannt vorausgesetzt

wird und nicht geschätzt werden kann.

4.1.4 Schätzung der Kovarianzmatrix

Kovarianzmatrix der AUC

Es sei $\mathrm{R}_{ik}^{(l)}$ der Rang von $\mathrm{X}_{ik}^{(l)}$ unter allen N Beobachtungen der l-ten Komponente und es sei $\mathbf{R}_{ik} = (\mathrm{R}_{ik}^{(1)}, \ldots, \mathrm{R}_{ik}^{(d)})'$ der Vektor dieser sogenannten Globalränge. Es sei weiter $\mathrm{Q}_{ik}^{(l)}$, $k = 1, \ldots, n_i$ der Rang von $\mathrm{X}_{ik}^{(l)}$ unter allen n_i Beobachtungen der i-ten Stichprobe der l-ten Komponente und es sei $\mathbf{Q}_{ik} = (\mathrm{Q}_{ik}^{(1)}, \ldots, \mathrm{Q}_{ik}^{(d)})'$ der Vektor dieser sogenannten Internränge $\mathrm{Q}_{ik}^{(l)}$, $l = 1, \ldots, d$, $i = 0, 1$. Weiter bezeichnen

$$\overline{\mathbf{R}}_{i\cdot} = \frac{1}{n_i} \sum_{k=1}^{n_i} \mathbf{R}_{ik} \text{ und } \overline{\mathbf{Q}}_{i\cdot} = \frac{1}{n_i} \sum_{k=1}^{n_i} \mathbf{Q}_{ik} = \frac{n_i + 1}{2} \mathbf{1}_d, \ i = 0, 1$$

die Mittelwerte dieser Vektoren, wobei $\mathbf{1}_d = (1, \ldots, 1)'$ der d-dimensionale Einservektor sei. Es sei schließlich $\mathbf{P}_{ik} = \mathbf{R}_{ik} - \mathbf{Q}_{ik}$ und $\overline{\mathbf{P}}_{i\cdot} = \overline{\mathbf{R}}_{i\cdot} - \overline{\mathbf{Q}}_{i\cdot}$. Mit dieser Notation gilt:

Satz 4.8 *Die Kovarianzmatrix* $\mathbf{V}_{\mathrm{AUC}}$ *von* $\sqrt{\mathrm{N}}(\widehat{\mathbf{AUC}} - \mathbf{AUC})$ *wird durch* $\widehat{\mathbf{V}}_{\mathrm{AUC}} = \widehat{\mathbf{V}}_{\mathrm{AUC},0} + \widehat{\mathbf{V}}_{\mathrm{AUC},1}$ *konsistent geschätzt, wobei*

$$\widehat{\mathbf{V}}_{\mathrm{AUC},i} = \frac{\mathrm{N}}{(\mathrm{N} - n_i)^2 n_i (n_i - 1)} \sum_{k=1}^{n_i} (\mathbf{P}_{ik} - \overline{\mathbf{P}}_{i\cdot})(\mathbf{P}_{ik} - \overline{\mathbf{P}}_{i\cdot})'. \tag{4.12}$$

BEWEIS. Für den Beweis der L_2-Konsistenz von $\widehat{\mathbf{V}}_{\mathrm{N}}$ sei auf die Arbeit von Brunner u. a. (2002) verwiesen. An dieser Stelle sei angemerkt, dass Brunner u. a. (2002) für den Beweis der Konsistenz des obigen Schätzers die Voraussetzung, dass keine der Marginalverteilungen eine Ein-Punkt-Verteilung ist, nicht benötigen. Daher lassen sich die Resultate dieses Satzes ebenfalls als Grundlage für eine konsistente Schätzung von $\mathbf{V}_{se}$ und $\mathbf{V}_{sp}$ verwenden. □

Kovarianzmatrix von Sensitivität und Spezifität

Da Sensitivität und Spezifität als spezielle AUCs angesehen werden können, lassen sich die Ergebnisse von Satz 4.8 auf Sensitivität und Spezifität einfach erweitern. Hierfür wird bei der Auswertung der Sensitivität die Stichprobe der Gesunden und bei der Spezifität die Stichprobe der Kranken durch eine Pseudostichprobe ersetzt, um dann mittels Satz 4.8 die Kovarianzmatrix $\widehat{\mathbf{V}}_{se}^{ps}$ von $\sqrt{\mathrm{N}_{se}} \cdot (\widehat{\mathbf{se}} - \mathbf{se})$ und die Kovarianzmatrix $\widehat{\mathbf{V}}_{sp}^{ps}$ von $\sqrt{\mathrm{N}_{sp}} \cdot (\widehat{\mathbf{sp}} - \mathbf{sp})$ zu schätzen.[4]
Es gilt $\mathrm{Cov}(\sqrt{\mathrm{N}_{se}} \cdot (\widehat{\mathbf{se}} - \mathbf{se})) = \mathbf{V}_{se}^{ps} = \frac{\mathrm{N}_{se}}{n_1} \cdot \mathbf{V}_{se}$ und $\sqrt{\mathrm{N}_{sp}} \cdot (\widehat{\mathbf{sp}} - \mathbf{sp}) = \mathbf{V}_{sp}^{ps} = \frac{\mathrm{N}_{sp}}{n_0} \cdot \mathbf{V}_{sp}$. Die Schätzer dieser Kovarianzmatrizen hängen daher von der Wahl von n_{ps} ab. Es gilt jedoch:

$$\widehat{\mathbf{V}}_{se}^{ps} = \frac{\mathrm{N}_{se}}{n_1} \cdot \frac{1}{n_1 - 1} \sum_{k=1}^{n_1} (\mathbf{\Gamma}(\mathbf{X}_{1k}) - \widehat{\mathbf{se}})(\mathbf{\Gamma}(\mathbf{X}_{1k}) - \widehat{\mathbf{se}})' \text{ sowie}$$
$$\widehat{\mathbf{V}}_{sp}^{ps} = \frac{\mathrm{N}_{sp}}{n_0} \cdot \frac{1}{n_0 - 1} \sum_{k=1}^{n_0} (\mathbf{\Gamma}(\mathbf{X}_{0k}) - \widehat{\mathbf{sp}})(\mathbf{\Gamma}(\mathbf{X}_{0k}) - \widehat{\mathbf{sp}})'.$$

[4] Es sei an dieser Stelle angemerkt, dass $n_{ps} > 1$ gewählt werden muss, da Gleichung (4.12) sonst nicht definiert ist.

Die Ausdrücke

$$\frac{1}{n_1-1}\sum_{k=1}^{n_1}(\boldsymbol{\Gamma}(\mathbf{X}_{1k})-\widehat{\mathbf{se}})(\boldsymbol{\Gamma}(\mathbf{X}_{1k})-\widehat{\mathbf{se}})'=\widehat{\mathbf{V}}_{se} \text{ und} \tag{4.13}$$

$$\frac{1}{n_0-1}\sum_{k=1}^{n_0}(\boldsymbol{\Gamma}(\mathbf{X}_{0k})-\widehat{\mathbf{sp}})(\boldsymbol{\Gamma}(\mathbf{X}_{0k})-\widehat{\mathbf{sp}})'=\widehat{\mathbf{V}}_{sp} \tag{4.14}$$

sind nun aber gerade die Stichprobenkovarianzmatrizen von $\boldsymbol{\Gamma}(\mathbf{X}_{1k})$ und $\boldsymbol{\Gamma}(\mathbf{X}_{0k})$ und damit als natürliche Schätzer der Varianz von $\sqrt{n_1}\widehat{\mathbf{se}}$ und $\sqrt{n_0}\widehat{\mathbf{sp}}$ unabhängig von n_{ps}. Der Stichprobenumfang der Pseudostichprobe taucht daher nur als konstanter Faktor auf und wird sich bei der späteren Berechnung von Teststatistiken und Konfidenzintervallen stets wieder herauskürzen, sodass die endgültigen Resultate schließlich unabhängig von n_{ps} sind. Die Verwendung der Pseudostichprobe gestattet an dieser Stelle zur Analyse von Sensitivität und Spezifität die gewöhnliche AUC-Software zu verwenden, was den Umweg über die komplexere Darstellung der Kovarianzmatrix rechtfertigt.

Kovarianzmatrix der prädiktiven Werte

Ähnlich wie bereits bei der Herleitung der Schätzer und der asymptotischen Verteilung der prädiktiven Werte wird auch bei der Schätzung der Kovarianzmatrix die Bayes'sche Formel in Verbindung mit Cramers multivariatem δ-Satz verwendet.

Satz 4.9 *Die Kovarianzmatrix von* $\sqrt{\mathrm{N}}(\widehat{\mathbf{p}}_{\pm}^{g}-\mathbf{p}_{\pm}^{g})$ *wird durch*

$$\widehat{\mathbf{V}}_{\pm}^{g}=\widehat{\mathbf{Df}}_{\pm}\left[\frac{\mathrm{N}}{n_1}\widehat{\mathbf{V}}_{se}\oplus\frac{\mathrm{N}}{n_0}\widehat{\mathbf{V}}_{sp}\oplus\frac{\mathrm{N}}{q_g}\widehat{\sigma}_g^2\right]\widehat{\mathbf{Df}}_{\pm}'$$

konsistent geschätzt. Hierbei bezeichne $\widehat{\mathbf{Df}}_{\pm}=\mathbf{Df}_{\pm}\left((\widehat{\mathbf{se}}',\widehat{\mathbf{sp}}',\widehat{\pi}_g)'\right)$ *die Jacobimatrix der partiellen Ableitungen von* $\mathbf{f}_{\pm}$ *an der Stelle* $(\widehat{\mathbf{se}}',\widehat{\mathbf{sp}}',\widehat{\pi}_g)'$ *und* $\widehat{\mathbf{V}}_{se}$ *und* $\widehat{\mathbf{V}}_{sp}$ *seien die in (4.13) und (4.14) angegebenen Stichprobenkovarianzmatrizen von* $\sqrt{n_1}\widehat{\mathbf{se}}$ *und* $\sqrt{n_0}\widehat{\mathbf{sp}}$. *Weiter sei* $\widehat{\sigma}_g^2=\frac{q_g}{q_g-1}\cdot\widehat{\pi}_g\cdot(1-\widehat{\pi}_g)$ *der unverzerrte Schätzer der Varianz* σ_g^2.

Beweis. Nach Satz 4.7 gilt:

$$\sqrt{\mathrm{N}}(\widehat{\mathbf{p}}_{\pm}^{g}-\mathbf{p}_{\pm}^{g})=\sqrt{\mathrm{N}}\left[\mathbf{f}_{\pm}\begin{pmatrix}\widehat{\mathbf{se}}\\ \widehat{\mathbf{sp}}\\ \widehat{\pi}_g\end{pmatrix}-\mathbf{f}_{\pm}\begin{pmatrix}\mathbf{se}\\ \mathbf{sp}\\ \pi_g\end{pmatrix}\right]\xrightarrow{\mathscr{L}}\mathbf{B}_{\pm}^{g}\sim\mathrm{N}\left(\mathbf{0},\mathbf{Df}_{\pm}\left[d_1\mathbf{V}_{se}\oplus d_0\mathbf{V}_{sp}\oplus e_g\sigma_g^2\right]\mathbf{Df}_{\pm}'\right),\ g=1,\ldots,\mathrm{G}.$$

$\mathbf{Df}_{\pm}=\mathbf{Df}_{\pm}\left((\mathbf{se}',\mathbf{sp}',\pi_g)'\right)$ wird nun nach dem Slutzky'schen Satz durch $\widehat{\mathbf{Df}}_{\pm}=\mathbf{Df}_{\pm}\left((\widehat{\mathbf{se}}',\widehat{\mathbf{sp}}',\widehat{\pi}_g)'\right)$ konsistent geschätzt. Die Grenzwerte d_i und e_g werden durch $\frac{\mathrm{N}}{n_i}$, $i=0,1$ beziehungsweise $\frac{\mathrm{N}}{q_g}$, $g=1,\ldots,\mathrm{G}$ *„geschätzt"*. Des Weiteren werden $\mathbf{V}_{se}$ und $\mathbf{V}_{sp}$ durch die zugehörigen Stichprobenkovarianzmatrizen geschätzt und σ_g^2 durch die Stichprobenvarianz $\widehat{\sigma}_g^2=\frac{q_g}{q_g-1}\cdot\widehat{\pi}_g\cdot(1-\widehat{\pi}_g)$. Da alle diese Schätzer konsistent sind (beziehungsweise die Folgen $\frac{\mathrm{N}}{n_i}$ $(i=0,1)$ und $\frac{\mathrm{N}}{q_g}$ $(g=1,\ldots,\mathrm{G})$ konvergieren), folgt die Konsistenz des Plug-In-Schätzers

$$\widehat{\mathbf{V}}_{\pm}^{g}=\widehat{\mathbf{Df}}_{\pm}\left[\frac{\mathrm{N}}{n_1}\widehat{\mathbf{V}}_{se}\oplus\frac{\mathrm{N}}{n_0}\widehat{\mathbf{V}}_{sp}\oplus\frac{\mathrm{N}}{q_g}\widehat{\sigma}_g^2\right]\widehat{\mathbf{Df}}_{\pm}'$$

aus dem Slutzky'schen Satz. □

4.1.5 Teststatistiken und Konfidenzintervalle

Die Tatsache, dass alle betrachteten diagnostischen Gütemaße asymptotisch normalverteilt sind, erlaubt eine einheitliche Darstellung von Teststatistiken und Konfidenzintervallen. Zu diesem Zweck sei $\mathbf{e} \in \{\mathbf{AUC}, \mathbf{se}, \mathbf{sp}, \mathbf{p}_-^g, \mathbf{p}_+^g,\ g = 1, \ldots, \mathrm{G}\}$ das zu analysierende diagnostische Gütemaß und $\widehat{\mathbf{e}}$ der zugehörige konsistente Schätzer. Weiter sei $\widehat{\mathbf{V}}_e$ der Schätzer der Kovarianzmatrix von $\sqrt{\mathrm{N}}(\widehat{\mathbf{e}} - \mathbf{e})$. Mit dieser Notation lassen sich die bisher präsentierten Resultate kurz durch

$$\sqrt{\mathrm{N}}(\widehat{\mathbf{e}} - \mathbf{e}) \stackrel{\cdot}{\sim} \mathrm{N}\left(\mathbf{0}, \widehat{\mathbf{V}}_e\right), \qquad \mathbf{e} \in \{\mathbf{AUC}, \mathbf{se}, \mathbf{sp}, \mathbf{p}_-^g, \mathbf{p}_+^g,\ g = 1, \ldots, \mathrm{G}\}$$

zusammenfassen. Da sich die bisherigen Ergebnisse auf diese Art und Weise kurz und prägnant einheitlich darstellen lassen, wird im Folgenden stets von einem beliebigen Gütemaß *e* gesprochen, welches als AUC, Sensitivität, Spezifität oder prädiktiver Wert einer bestimmten Risikogruppe interpretiert werden kann.

Teststatistiken

In diesem Abschnitt werden verschiedene Statistiken zum Testen standardmäßiger Hypothesen präsentiert. In nichtparametrischen Modellen zur multivariaten Datenanalyse wird als Teststatistik für gewöhnlich eine Rangversion der Wald-Typ-Statistik (Puri und Sen, 1971) verwendet. Eine Rangversion der sogenannten ANOVA-Typ-Statistik wurde von Brunner u. a. (1997) und Munzel und Brunner (2000) als Alternative vorgestellt, da diese in Simulationsstudien das Niveau deutlich besser einhält als die standardmäßig verwendete Wald-Typ-Statistik von Puri und Sen (1971). Auf Grund dieser Überlegenheit der ANOVA-Typ-Statistik wird in dieser Arbeit auf eine Darstellung der Wald-Typ-Statistik verzichtet.

Satz 4.10 (Lineares Modell) *Es sei* $\mathbf{e}$ *ein Vektor diagnostischer Gütemaße, welcher durch* $\widehat{\mathbf{e}}$ *konsistent geschätzt wird. Es gelte (in der üblichen Notation)* $\sqrt{\mathrm{N}}(\widehat{\mathbf{e}} - \mathbf{e}) \stackrel{\cdot}{\sim} \mathrm{N}\left(\mathbf{0}, \widehat{\mathbf{V}}_e\right)$*, dann gilt unter* $\mathrm{H}_0 : \mathbf{Ce} = \mathbf{0}$*:*

$$\textbf{\textit{ANOVA-Typ-Statistik}} \qquad \mathrm{Q}^{\mathrm{ATS}}(\mathbf{T}) = \mathrm{N}\frac{\mathrm{Sp}(\mathbf{T}\widehat{\mathbf{V}}_e)}{\mathrm{Sp}(\mathbf{T}\widehat{\mathbf{V}}_e\mathbf{T}\widehat{\mathbf{V}}_e)}\,\widehat{\mathbf{e}}'\mathbf{T}\widehat{\mathbf{e}} \qquad \stackrel{\cdot}{\sim} \chi_f^2 \qquad \textit{mit } f = \frac{\mathrm{Sp}(\mathbf{T}\widehat{\mathbf{V}}_e)^2}{\mathrm{Sp}(\mathbf{T}\widehat{\mathbf{V}}_e\mathbf{T}\widehat{\mathbf{V}}_e)},$$

wobei $\mathbf{T} = \mathbf{C}'(\mathbf{CC}')^-\mathbf{C}$ *die standardisierte Hypothesenmatrix zur Hypothese* $\mathbf{Ce} = \mathbf{0}$ *sei.* $\mathrm{Sp}(\cdot)$ *bezeichne hierbei die Spur einer Matrix.*

BEWEIS. Die Beweis ist nachzulesen bei Brunner u. a. (1999), Brunner u. a. (1997) und Munzel und Brunner (2000). □

Das hier verwendete Modell zum Testen der Hypothesen ist ein einfaches lineares Modell, da die Hypothesen in der Form $\mathbf{Ce} = \mathbf{0}$ formuliert werden. Die Annahme linearer Effekte ist jedoch bestenfalls diskussionswürdig. Da alle diagnostischen Gütemaße Wahrscheinlichkeiten sind, wäre – analog zur Theorie generalisierter linearer Modelle – ein logistisches Modell, welches sich als ein multiplikatives Modell der Chancen oder Odds interpretieren lässt, eher adäquat.

Satz 4.11 (Logistisches Modell) *Es gelten die Voraussetzungen von Satz 4.10. Es sei weiter* $\mathbf{g}(\mathbf{e}) = \mathit{logit}(\mathbf{e}) = \left(\log(\frac{e^{(l)}}{1-e^{(l)}})\right)_{l=1,\ldots,d}$ *die logit-Funktion in ihrer multivariaten Form. Es sei weiter* $\widehat{\mathbf{V}}_e^g = \mathbf{Dg}(\widehat{\mathbf{e}})\widehat{\mathbf{V}}_e\mathbf{Dg}(\widehat{\mathbf{e}})$*, wobei* $\mathbf{Dg}(\widehat{\mathbf{e}})$ *die Jacobimatrix der partiellen Ableitungen* $\frac{\partial \mathbf{g}}{\partial \mathbf{e}}$ *an der Stelle* $\widehat{\mathbf{e}}$ *sei. Dann gilt unter* $\mathrm{H}_0 : \mathbf{C} \cdot \mathbf{g}(\mathbf{e}) = \mathbf{0}$*:*

$$\textbf{\textit{ANOVA-Typ-Statistik}} \qquad \mathrm{Q}_g^{\mathrm{ATS}}(\mathbf{T}) = \mathrm{N}\frac{\mathrm{Sp}(\mathbf{T}\widehat{\mathbf{V}}_e^g)}{\mathrm{Sp}(\mathbf{T}\widehat{\mathbf{V}}_e^g\mathbf{T}\widehat{\mathbf{V}}_e^g)}\,\mathbf{g}(\widehat{\mathbf{e}})'\mathbf{T}\mathbf{g}(\widehat{\mathbf{e}}) \qquad \stackrel{\cdot}{\sim} \chi_f^2 \qquad \textit{mit } f = \frac{\mathrm{Sp}(\mathbf{T}\widehat{\mathbf{V}}_e^g)^2}{\mathrm{Sp}(\mathbf{T}\widehat{\mathbf{V}}_e^g\mathbf{T}\widehat{\mathbf{V}}_e^g)}.$$

BEWEIS. Nach Cramers multivariatem δ-Satz gilt

$$\sqrt{\mathrm{N}}\left(\mathbf{g}(\hat{\mathbf{e}})-\mathbf{g}(\mathbf{e})\right) \dot{\sim} \mathrm{N}\left(\mathbf{0}, \hat{\mathbf{V}}_e^g\right),$$

wodurch die Aussage direkt aus der Anwendung von Satz 4.10 auf die Statistik $\sqrt{\mathrm{N}}\left(\mathbf{g}(\hat{\mathbf{e}})-\mathbf{g}(\mathbf{e})\right)$ folgt. □

Es sei an dieser Stelle angemerkt, dass neben der logit-Funktion auch jede beliebige andere Transformationsfunktion, welche den Anforderungen von Cramers δ-Satz genügt[5], verwendet werden kann. Die logit-Funktion wird hier einzig wegen ihrer aus der logistischen Regression bekannten, praxisnahen Interpretierbarkeit in den Vordergrund gerückt.

Konfidenzintervalle

Mit Hilfe der asymptotischen Verteilung $\sqrt{\mathrm{N}}(\hat{\mathbf{e}}-\mathbf{e}) \dot{\sim} \mathrm{N}\left(\mathbf{0}, \hat{\mathbf{V}}_e\right)$ lassen sich Konfidenzintervalle für Linearkombinationen $\mathbf{c}'\mathbf{e}$ von $e^{(1)}, \ldots, e^{(d)}$ auf die übliche Art und Weise berechnen.

Satz 4.12 (Konfidenzintervalle des linearen Modells) *Es gelte* $\sqrt{\mathrm{N}}(\hat{\mathbf{e}}-\mathbf{e}) \dot{\sim} \mathrm{N}\left(\mathbf{0}, \hat{\mathbf{V}}_e\right)$ *und es sei* $u_{1-\frac{\alpha}{2}}$ *das* $(1-\frac{\alpha}{2})$*-Quantil der Standardnormalverteilung, dann sind*

$$\mathbf{c}'\hat{\mathbf{e}} \pm \frac{\sqrt{\mathbf{c}'\hat{\mathbf{V}}_e\mathbf{c}} \cdot u_{1-\frac{\alpha}{2}}}{\sqrt{\mathrm{N}}}, \tag{4.15}$$

asymptotische $(1-\alpha)$*-Konfidenzintervallgrenzen für* $\mathbf{c}'\mathbf{e}$*.*

BEWEIS. Das Resultat ergibt sich durch die Anwendung der Pivotmethode auf die studentisierten Teststatistiken $t_{\mathbf{c}} = \sqrt{\mathrm{N}} \cdot \frac{(\mathbf{c}'\hat{\mathbf{e}}-\mathbf{c}'\mathbf{e})}{\sqrt{\mathbf{c}'\hat{\mathbf{V}}_e\mathbf{c}}}$, deren asymptotische Standardnormalverteilung direkt aus den Voraussetzungen folgt. □

Analog zu den logistischen Teststatistiken werden nun die Konfidenzintervalle des logistischen Modells präsentiert. Da die den Effekten des logistischen Modells entsprechenden Konfidenzintervalle für $\mathbf{c}'\mathbf{g}(\mathbf{e})$ in der praktischen Anwendung selten zu interpretieren sind, werden an dieser Stelle lediglich die logistischen Konfidenzintervalle der Gütemaße $e^{(1)}, \ldots, e^{(d)}$ selbst angegeben. [6]

Satz 4.13 (Konfidenzintervalle des logistischen Modells) *Es gelte die Notation von Satz 4.11 und es sei* $\hat{\mathbf{V}}_e^g[l,l]$*, das* l*-te Diagnonalelement von* $\hat{\mathbf{V}}_e^g$*, dann sind*

$$g^{-1}\left(g(\hat{e}^{(l)}) \pm \frac{\sqrt{\hat{\mathbf{V}}_e^g[l,l]} \cdot u_{1-\frac{\alpha}{2}}}{\sqrt{\mathrm{N}}}\right), \tag{4.16}$$

die logistischen $(1-\alpha)$*-Konfidenzintervallgrenzen für* $e^{(l)}$*,* $l = 1, \ldots, d$*, wobei* $g^{-1}(x) = \frac{exp(x)}{1+exp(x)}$ *die Umkehrfunktion der logit-Funktion sei.*

BEWEIS. Wie im Beweis von Satz 4.11 folgt das Resultat aus Cramers δ-Satz und ist beispielsweise bei Domhof (2001) nachzulesen. □

[5] Hinreichend hierfür ist, dass die verwendete Funktion streng monoton steigend, stetig differenzierbar und bijektiv ist.

[6] Eine Ausnahme bildet ein aus der logistischen Regression bekannter Zusammenhang: Ist $\mathbf{c}$ die Differenz zweier Einheitsvektoren, so lässt sich $\exp\left(\mathbf{c}'\mathbf{g}(\mathbf{e})\right)$ als Odds-Ratio der zu diesen Einheitsvektoren gehörigen Gruppen interpretieren.

Die Konfidenzintervalle der $e^{(l)},\ l = 1,\ldots,d$ können somit entweder über den logistischen oder über den einfachen linearen Ansatz berechnet werden. Während im linearen Ansatz die Konfidenzintervallgrenzen bei kleinen Stichprobenumfängen jedoch größer als 1 oder kleiner als 0 werden können, kann dieses im logistischen Modell nicht passieren, da der Definitionsbereich der logit-Funktion das offene Intervall $(0,1)$ ist. Die logistischen Konfidenzintervalle sind somit im Gegenteil zu den Konfidenzintervallen des linearen Modells bereits per Konstruktion im Bereich $(0,1)$ enthalten.

4.2 Analyse verbundener Stichproben bei Clusterdaten

In diesem Kapitel werden die Resultate des letzten Abschnittes auf die Analyse von Clusterdaten erweitert. Im Fall von Einfachmessungen folgt die Theorie zur Evaluation von Sensitivität, Spezifität und prädiktiven Werten nach dem in Abbildung 4.2 dargestellten Schema direkt aus der Methodik der AUC-Analyse.

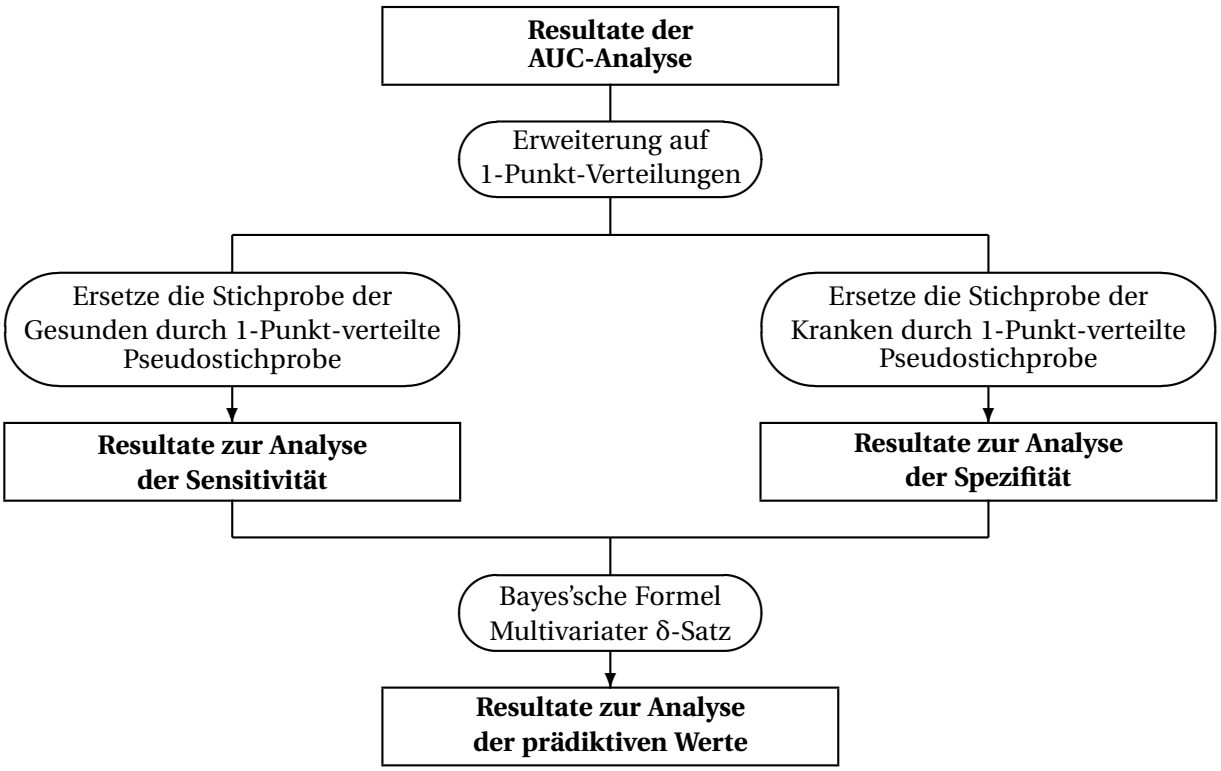

Abbildung 4.2: Schematische Darstellung des Zusammenhangs der diagnostischen Gütemaße bei Einfachmessungen

Liegt eine Studie mit Clusterdaten vor, resultiert das Verfahren zur Auswertung von Sensitivität, Spezifität, positiv und negativ prädiktivem Wert in analoger Weise aus den Ergebnissen der AUC-Analyse. Eine einzige Ausnahme besteht in der Verteilung der Statistik der prädiktiven Werte: Anders als bei Einfachmessungen sind bei Clusterdaten die Schätzer von Sensitivität und Spezifität nicht mehr unabhängig, da am selben Patienten sowohl gesunde als auch kranke Beobachtungen vorliegen können. Die asymptotische Verteilung des Ausdrucks

$$\sqrt{\mathrm{N}}\left[\begin{pmatrix}\widehat{\mathbf{se}}\\ \widehat{\mathbf{sp}}\end{pmatrix}-\begin{pmatrix}\mathbf{se}\\ \mathbf{sp}\end{pmatrix}\right]$$

lässt sich daher nicht wie in Abschnitt 4.1.3 (Seite 25) kanonisch aus der asymptotischen Normalität der bei-

den Teilvektoren herleiten. Um der Abhängigkeit zwischen $\widehat{\mathbf{se}}$ und $\widehat{\mathbf{sp}}$ Rechnung zu tragen, muss im Fall von Clusterdaten für den Erweiterungsschritt von Sensitivität und Spezifität auf die prädiktiven Werte der Vektor $(\mathbf{se}', \mathbf{sp}')'$ als Ganzes betrachtet werden. Er kann hier nicht – wie bei einer Studie mit Einfachmessungen – in die Vektoren **se** und **sp** zerlegt werden, da auf Grund der fehlenden Unabhängigkeit die Kovarianzmatrix von $\sqrt{\mathrm{N}}\left(\widehat{\mathbf{se}}', \widehat{\mathbf{sp}}'\right)'$ nicht der Kronecker-Summe von $\mathrm{Cov}(\sqrt{\mathrm{N}} \cdot \widehat{\mathbf{se}})$ und $\mathrm{Cov}(\sqrt{\mathrm{N}} \cdot \widehat{\mathbf{sp}})$ entspricht. Dennoch lässt sich der Vektor $(\mathbf{se}', \mathbf{sp}')'$ als $2d$-dimensionaler Vektor von Flächen unter speziellen ROC-Kurven interpretieren, wodurch die Methodik der AUC-Analyse auch auf den gesamten Vektor $(\mathbf{se}', \mathbf{sp}')'$ angewendet werden kann. Insbesondere lässt sich die Kovarianz zwischen Sensitivität und Spezifität somit als Kovarianz zwischen speziellen AUCs interpretieren und schätzen. Das heißt, dass sich auch bei Clusterdaten die Methodik zur Evaluation prädiktiver Werte direkt aus den Verfahren zur AUC-Analyse ergibt. Auf Grund dessen soll in diesem Kapitel lediglich die Methodik der AUC-Analyse präsentiert werden, da sich – trotz des beschriebenen Unterschiedes zu den Einfachmessungen – die anderen Gütemaße gemäß Abbildung 4.3 auf die AUC-Analyse zurückführen lassen.

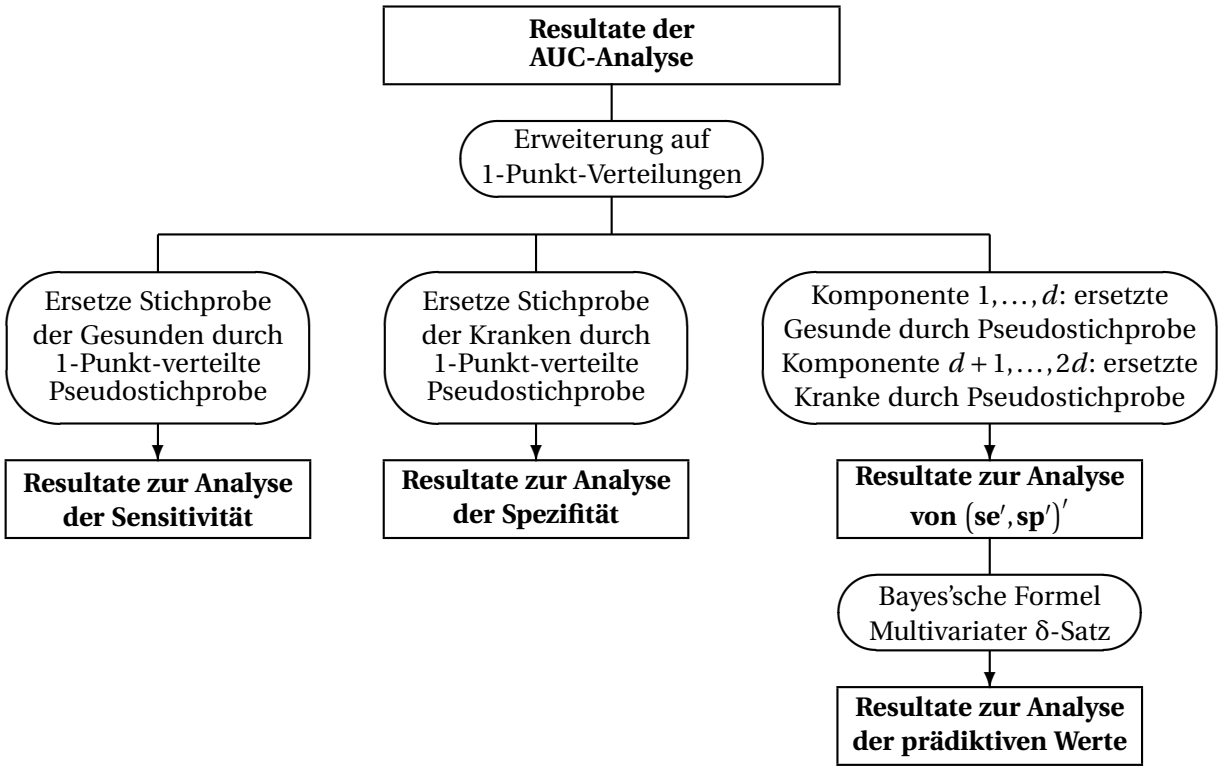

Abbildung 4.3: Schematische Darstellung des Zusammenhangs der diagnostischen Gütemaße bei Clusterdaten

Diese Erweiterung der Methodik von Sensitivität und Spezifität hin zur Analyse prädiktiver Werte wird in Abschnitt 4.2.2 exemplarisch anhand der gewichteten Effektschätzer illustriert, wodurch die Parallelität zum Fall der Einfachmessungen deutlich wird und die Präsentation in den übrigen Abschnitten auf die AUC-Analyse reduziert werden kann. Insbesondere wird hierbei exemplarisch erläutert, wie sich der Vektor $(\mathbf{se}', \mathbf{sp}')'$ durch Ergänzung von Pseudobeobachtungen durch einen einzelnen gewöhnlichen AUC-Schätzer schätzen lässt.

4.2.1 Modell und Notation

Im Folgenden wird davon ausgegangen, dass für jedes an der Diagnosestudie teilnehmende Versuchssubjekt unter $l = 1,\ldots,d$ verschiedenen Bedingungen Daten erhoben werden. Verschiedene Bedingungen können hierbei beispielsweise differierende Reader-Methoden-Kombinationen oder auch nur unterschiedliche Reader oder Diagnosemethoden bedeuten. Bei Diagnosestudien mit Clusterdaten werden die Subjekte gemäß Definition 3.2 (siehe Seite 14) in vollständige und unvollständige Fälle unterschieden. Die auf diese Art und Weise pro Patient erhobene Fülle an Messwerten wird unter Berücksichtigung der unterschiedlichen Datenstruktur bei vollständigen und unvollständigen Fällen gemäß dem folgenden Schema in einem Vektor zusammengefasst. Die Struktur des Vektors wird in Abbildung 4.4 für ein vollständiges Subjekt vorgestellt. Die graphische Darstellung der vektoriellen Struktur der unvollständigen Subjekte erfolgt exemplarisch anhand eines Patienten mit nur gesunden Beobachtungseinheiten (Abbildung 4.5). Die Datenstruktur eines unvollständigen Subjektes mit nur kranken Beobachtungen ist analog.

Vollständige Fälle:

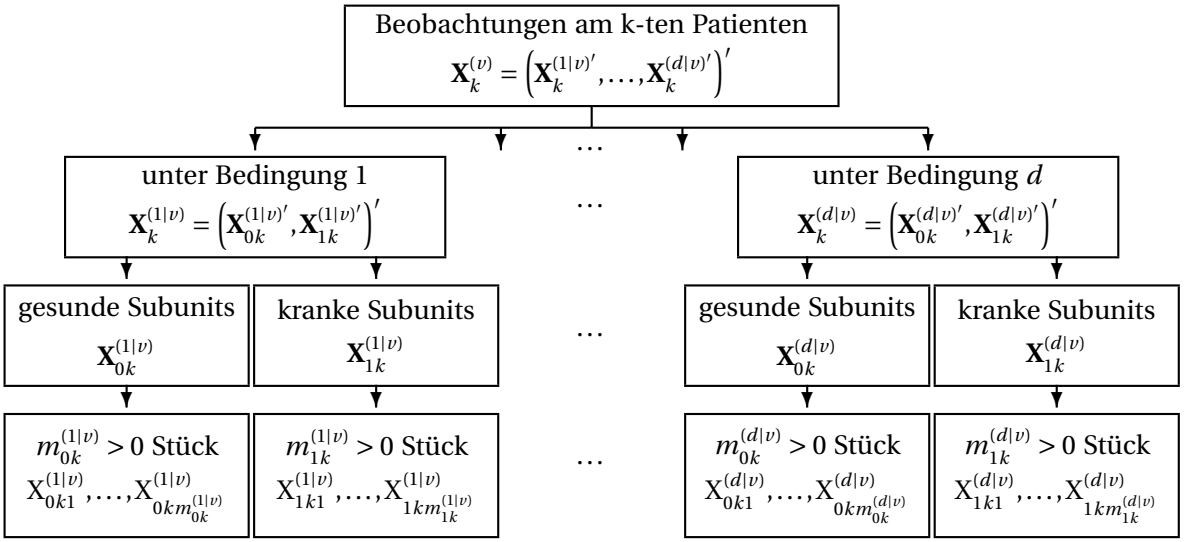

Abbildung 4.4: Notationsdarstellung: Vollständige Fälle

Unvollständige Fälle:

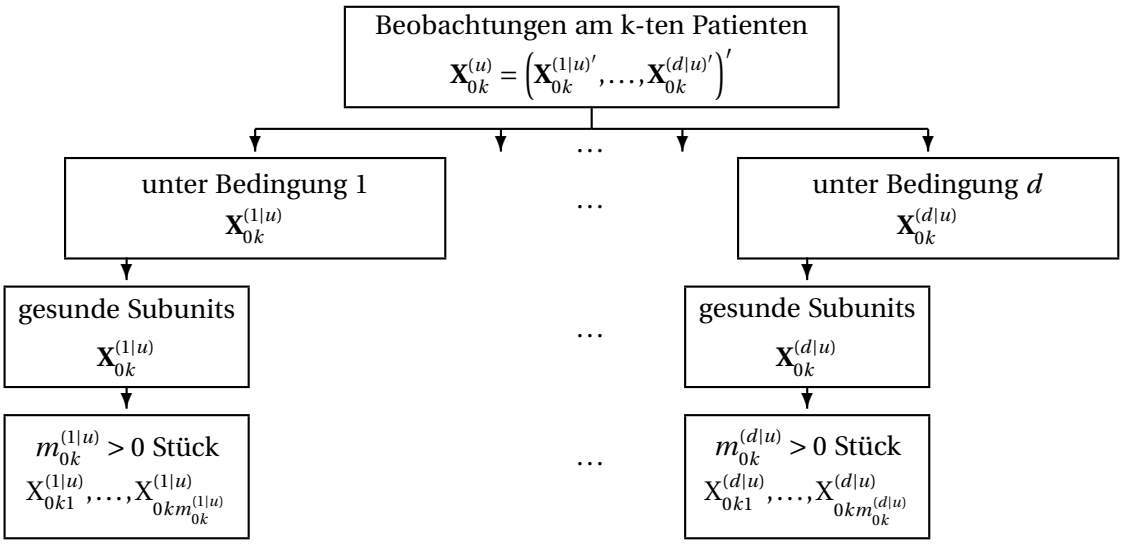

Abbildung 4.5: Notationsdarstellung: Unvollständige Fälle, exemplarisch für den Fall nur gesunder Subunits

Diese Notation führt nun zum statistischen Modell.

Modell 2
Gegeben sei eine Diagnosestudie mit N Subjekten, die sich gemäß Definition 3.2 in $n^{(v)}$ vollständige und $n^{(u)} = n_0^{(u)} + n_1^{(u)}$ unvollständige Subjekte aufteilt ($n_0^{(u)}$ sei dabei die Anzahl der Patienten mit nur gesunden, $n_1^{(u)}$ die Anzahl der Patienten mit nur kranken Beobachtungseinheiten). Für eine spätere einheitliche Notation seien analog zu $n_0^{(u)}$ und $n_1^{(u)}$ ebenfalls $n_0^{(v)} = n^{(v)}$ und $n_1^{(v)} = n^{(v)}$ definiert. Es sei weiter $n_0 = n^{(v)} + n_0^{(u)}$ die Anzahl aller Patienten mit mindestens einer gesunden Beobachtungseinheit, analog sei $n_1 = n^{(v)} + n_1^{(u)}$ die Anzahl an Patienten, die mindestens eine kranke Subunit besitzen. Es seien in Anlehnung an die Abbildungen 4.4 und 4.5

$$\begin{aligned}
\mathbf{X}_k^{(v)} &= \left(\left(\mathbf{X}_{0k}^{(1|v)'}, \mathbf{X}_{1k}^{(1|v)'}\right), \ldots, \left(\mathbf{X}_{0k}^{(d|v)'}, \mathbf{X}_{1k}^{(d|v)'}\right)\right)', && k = 1, \ldots n^{(v)}, \\
\mathbf{X}_{ik}^{(u)} &= \left(\mathbf{X}_{ik}^{(1|u)'}, \mathbf{X}_{ik}^{(2|u)'}, \ldots, \mathbf{X}_{ik}^{(d|u)'}\right)', && i = 0,1,\ k = 1, \ldots, n_i^{(u)},
\end{aligned}$$

mit

$$\begin{aligned}
\mathbf{X}_{ik}^{(l|v)} &= \left(\mathrm{X}_{ik1}^{(l|v)}, \ldots, \mathrm{X}_{ikm_{ik}^{(l|v)}}^{(l|v)}\right)', && l = 1, \ldots, d,\ i = 0,1,\ k = 1, \ldots, n^{(v)}, \\
\mathbf{X}_{ik}^{(l|u)} &= \left(\mathrm{X}_{ik1}^{(l|u)}, \ldots, \mathrm{X}_{ikm_{ik}^{(l|u)}}^{(l|u)}\right)', && l = 1, \ldots, d,\ i = 0,1,\ k = 1, \ldots, n_i^{(u)}
\end{aligned}$$

die, wie vorangehend erläutert, strukturierten Zufallsvektoren mit Randverteilungen

$$\begin{aligned}
&\mathrm{X}_{iks}^{(l|v)} \sim \mathrm{F}_i^{(l|v)}, && l = 1, \ldots, d,\ i = 0,1,\ k = 1, \ldots, n^{(v)},\ s = 1, \ldots, m_{ik}^{(l|v)}, && \text{(vollständige Fälle)} \\
&\mathrm{X}_{iks}^{(l|u)} \sim \mathrm{F}_i^{(l|u)}, && l = 1, \ldots, d,\ i = 0,1,\ k = 1, \ldots, n_i^{(u)},\ s = 1, \ldots, m_{ik}^{(l|u)} && \text{(unvollständige Fälle).}
\end{aligned}$$

Hierbei seien die Funktionen $\mathrm{F}_i^{(l|v)}$ beziehungsweise $\mathrm{F}_i^{(l|u)}$ die Verteilungsfunktionen in normalisierter Form. Es sei weiter

$$\mathrm{M}^{(l)} = m_{\cdot\cdot}^{(l|\cdot)} = \sum_{k=1}^{n^{(v)}} \left(m_{0k}^{(l|v)} + m_{1k}^{(l|v)}\right) + \sum_{k=1}^{n_0^{(u)}} m_{0k}^{(l|u)} + \sum_{k=1}^{n_1^{(u)}} m_{1k}^{(l|u)}, \qquad l = 1, \ldots, d$$

die Gesamtanzahl an Beobachtungseinheiten der $\mathrm{N} = n^{(v)} + n_0^{(u)} + n_1^{(u)}$ Subjekte unter der l-ten Bedingung, sowie

$$\begin{aligned}
m_{0\cdot}^{(l|\cdot)} &= \sum_{k=1}^{n^{(v)}} m_{0k}^{(l|v)} + \sum_{k=1}^{n_0^{(u)}} m_{0k}^{(l|u)}, && l = 1, \ldots, d, \\
m_{1\cdot}^{(l|\cdot)} &= \sum_{k=1}^{n^{(v)}} m_{1k}^{(l|v)} + \sum_{k=1}^{n_1^{(u)}} m_{1k}^{(l|u)}, && l = 1, \ldots, d
\end{aligned}$$

die Summe aller gesunden beziehungsweise kranken Subunits unter Bedingung l. Es seien schließlich

$$\begin{aligned}
m_{i\cdot}^{(l|v)} &= \sum_{k=1}^{n^{(v)}} m_{ik}^{(l|v)}, && l = 1, \ldots, d,\ i = 0,1, \\
m_{i\cdot}^{(l|u)} &= \sum_{k=1}^{n_i^{(u)}} m_{ik}^{(l|u)}, && l = 1, \ldots, d,\ i = 0,1
\end{aligned}$$

die Summen aller Beobachtungseinheiten im Zustand i getrennt nach vollständigen und unvollständigen

Fällen. Es sei weiter

$$m_{\max} = \max\left\{m_{ik}^{(l|t)}, l = 1,\ldots,d, t \in \{u,v\}, i = 0,1, k = 1,\ldots,n_i^{(t)}\right\}$$

die Größe des Clusters mit den meisten Beobachtungen.
An dieser Stelle sei angemerkt, dass für ein Subjekt k in der aller Regel $m_{ik}^{(1|t)} = m_{ik}^{(2|t)} = \ldots = m_{ik}^{(l|t)},\ i = 0,1$ gelten wird, da die Bezeichnungen gesund ($i = 0$) und krank ($i = 1$) immer den Gesundheitszustand der Beobachtungseinheit nach Goldstandard beschreiben. Die Anzahl an gesunden Subunits pro Patient wird daher in der Regel unter allen Bedingungen gleich sein. Gleiches ist auch für die Anzahl der kranken Beobachtungen zu erwarten. Lediglich fehlende Werte können an dieser Stelle Ursache für ungleiche Stichprobenumfänge in den verschiedenen Komponenten sein.
Analog zum Modell der Einfachmessungen sei $\gamma^{(l)}$ der zur Bestimmung von Sensitivität und Spezifität erforderliche Schwellenwert der l-ten Komponente. Ohne Beschränkung der Allgemeinheit gelte hierbei ähnlich zum Fall der Einfachmessungen $\mathrm{P}(\mathrm{X}_{iks}^{(l)} = \gamma^{(l)}) = 0$.
Ebenfalls analog zum ersten Modell sei zur Berechnung der prädiktiven Werte angenommen, dass $g = 1,\ldots,\mathrm{G}$ verschiedene Risikogruppen existieren, wobei die Erkrankungswahrscheinlichkeit in der g-ten Gruppe π_g betrage. π_g sei in Vorabstudien bereits durch $\widehat{\pi}_g = \frac{k_g}{q_g}$ geschätzt worden, wobei q_g der Gesamtstichprobenumfang der g-ten Gruppe in dieser Vorabstudie sei und k_g die Anzahl der in dieser Gruppe erkrankten Patienten. Analog zu Modell 1 werden Gütemaße wieder in Vektoren zusammengefasst:

$$\begin{aligned}
\mathbf{AUC} &= (\mathrm{AUC}^{(1)},\ldots,\mathrm{AUC}^{(d)})' \\
\mathbf{se} &= (se^{(1)},\ldots,se^{(d)})' \\
\mathbf{sp} &= (sp^{(1)},\ldots,sp^{(d)})' \quad \text{und} \\
\mathbf{p}_{g,\pm} &= (p_{g,\pm}^{(1)},\ldots,p_{g,\pm}^{(d)})'.
\end{aligned}$$

Zur späteren Beweisführung müssen nun folgende Regularitätsannahmen an das Modell gestellt werden:

Voraussetzung 4.3

(1) Die Verteilungsfunktionen $\mathrm{F}_i^{(l|t)},\ l = 1,\ldots,d,\ t \in \{u,v\},\ i = 0,1$ seien keine Ein-Punkt-Verteilungen.
(2) Für alle $l = 1,\ldots,d$ seien die Randverteilungen $\mathrm{F}_i^{(l|v)}$ der vollständigen Fälle und $\mathrm{F}_i^{(l|u)}$ der unvollständigen Fälle gleich, das heißt, es gelte: $\mathrm{F}_i^{(l|v)} = \mathrm{F}_i^{(l|u)} =: \mathrm{F}_i^{(l)},\ \forall l = 1,\ldots,d,\ i = 0,1$.
(3) Für die gewichteten Schätzer gelte: Für alle $l = 1,\ldots,d$ gelte $\min\left(m_{0\cdot}^{(l|\cdot)}, m_{1\cdot}^{(l|\cdot)}, q_0,\ldots,q_{\mathrm{G}}\right) \to \infty$, derart, dass $\frac{m_{\max}^2}{m_{i\cdot}^{(l|\cdot)}} \to 0$.
(4) Für alle $l,r = 1,\ldots,d$ seien die bivariate Verteilung von $\left(\mathrm{X}_{iks}^{(l|v)}, \mathrm{X}_{i'ks'}^{(r|v)}\right)$ beziehungsweise von $\left(\mathrm{X}_{iks}^{(l|u)}, \mathrm{X}_{iks'}^{(r|u)}\right)$ unabhängig von $s,\ s'$ und k. Es seien außerdem die bivariaten Verteilungen von $\left(\mathrm{X}_{iks}^{(l|u)}, \mathrm{X}_{iks'}^{(r|u)}\right)$ und $\left(\mathrm{X}_{ik's}^{(l|v)}, \mathrm{X}_{ik's'}^{(r|v)}\right)$ gleich für alle $s,\ s'$ und $k,\ k'$.
(5) Es gelte $\min\left(n_0, n_1, q_0,\ldots,q_{\mathrm{G}}\right) \to \infty$, derart, dass $\frac{\mathrm{N}}{n_i} \to d_i < \infty$, $(i = 0,1)$, und $\frac{\mathrm{N}}{q_g} \to e_i < \infty$, $(g = 1,\ldots,\mathrm{G})$.
(6) $se^{(l)}, sp^{(l)}, l = 1,\ldots,d$ und $\pi_g, g = 1,\ldots,\mathrm{G}$ liegen in $(0,1)$, das heißt, sie sind echt kleiner 1 und echt größer 0.

Des Weiteren kennzeichne zur Vervollständigung der oben eingeführten Vektornotation $\mathbf{F}_i = (\mathrm{F}_i^{(1)},\ldots,\mathrm{F}_i^{(d)})'$, $i = 0,1$ den Vektor der Verteilungsfunktionen.
Die obigen Voraussetzungen sind dabei wie folgt zu interpretieren: Die Annahme keiner Ein-Punkt-Verteilungen (1) stellt sicher, dass die erhobenen Messwerten über ein Mindestmaß an Variabilität verfügen und die Studie somit nicht vollständig deterministisch ist. Voraussetzung (2) besagt, dass die Verteilung

des diagnostischen Testergebnisses unabhängig davon ist, ob ein Subjekt vollständig oder unvollständig ist, welches die Annahme widerspiegelt, dass alle Beobachtungen des gleichen Gesundheitszustandes Messwiederholungen sind. Die dritte Annahme gewährleistet, dass die Anzahl der Beobachtungen, die pro Patient erhoben werden, beschränkt ist. Für die Praxis bedeutet dieses, dass alle Subjekte über eine vergleichbare Anzahl an Beobachtungseinheiten verfügen sollten. Die Voraussetzungen (4) bis (6) sind analog zu den entsprechenden Voraussetzungen bei den Einfachmessungen (siehe Voraussetzung 4.1 auf Seite 20) zu interpretieren.

4.2.2 Schätzung der diagnostischen Gütemaße

Liegen in einer Studie Clusterdaten vor, so gibt es verschiedene Konzepte bei der Entwicklung konsistenter Schätzer: die sogenannten gewichteten Schätzer auf der einen Seite und die ungewichteten auf der anderen. Während bei den ungewichteten Schätzern jedes Cluster (das heißt jeder Patient) das gleiche Gewicht erhält, können bei den gewichteten Schätzern die verschiedenen Cluster unterschiedlich gewichtet werden. In dieser Arbeit wird dabei für die gewichteten Schätzer stets eine spezielle Gewichtung verwendet. Die hieraus resultierenden Schätzer werden kurz als die gewichteten Schätzer bezeichnet – auch wenn der Begriff in seiner ursprünglichen Bedeutung weiter gefasst ist. Gewichtete Schätzer beschreiben in dieser Arbeit jene Schätzer, bei denen jedes Cluster mit der Anzahl an Beobachtungen in diesem Cluster gewichtet wird. Dieses bedeutet, dass bei den gewichteten Schätzern jede Beobachtung das gleiche Gewicht hat, wohingegen bei den ungewichteten Schätzern jeder Patient das gleiche Gewicht erhält. Konsequenterweise sind beide Schätzer gleich, wenn alle Patienten sowohl die gleiche Anzahl an gesunden Beobachtungenseinheiten besitzen, sowie über die gleiche Anzahl kranker Subunits verfügen. Im Folgenden werden die gewichteten Schätzer stets mit einer Tilde $\widetilde{\ }$ und die ungewichteten stets mit einem Dach $\widehat{\ }$ gekennzeichnet, wodurch sich die folgenden beiden Versionen der empirischen Verteilungsfunktionen ergeben:

$$\widetilde{\mathrm{F}}_i^{(l)}(x) = \frac{1}{m_{i\cdot}^{(l|\cdot)}} \left(\sum_{k=1}^{n_i^{(u)}} \sum_{s=1}^{m_{ik}^{(l|u)}} c(x - \mathrm{X}_{iks}^{(l|u)}) + \sum_{k=1}^{n^{(v)}} \sum_{s=1}^{m_{ik}^{(l|v)}} c(x - \mathrm{X}_{iks}^{(l|v)}) \right) \tag{4.17}$$

$$\widehat{\mathrm{F}}_i^{(l)}(x) = \frac{1}{n_i} \left(\sum_{k=1}^{n_i^{(u)}} \frac{1}{m_{ik}^{(l|u)}} \sum_{s=1}^{m_{ik}^{(l|u)}} c(x - \mathrm{X}_{iks}^{(l|u)}) + \sum_{k=1}^{n^{(v)}} \frac{1}{m_{ik}^{(l|v)}} \sum_{s=1}^{m_{ik}^{(l|v)}} c(x - \mathrm{X}_{iks}^{(l|v)}) \right) \tag{4.18}$$

Die obige Darstellung lässt erkennen, dass die vollständigen und die unvollständigen Fälle zur Schätzung der Verteilungsfunktionen zusammengefasst werden. Durch Anwendung der Plug-In-Methode resultieren aus diesen Schätzern der Verteilungsfunktionen in kanonischer Art und Weise gepoolte Schätzer für die diagnostischen Gütemaße. Werner und Brunner (2007) verfolgen diesen Ansatz der gepoolten Schätzung auch für die Varianzschätzung; Simulationsstudien aus dem Bereich der AUC-Analyse zeigen jedoch, dass schon in balancierten Versuchen bei kleinen Stichprobenumfängen (das heißt $n^{(v)} = n_0^{(u)} = n_1^{(u)} \leq 10$) die geschätzte Kovarianzmatrix des AUC-Schätzers häufig negativ definit ist. Bei unbalancierten Versuchen tritt diese Eigenschaft sogar bei größeren Stichprobenumfängen in Abhängigkeit vom Verhältnis $\frac{\mathrm{N}}{n^{(v)}}$ auf (vergleiche Konietschke und Brunner, 2009). Da die Stichprobenumfänge $n^{(v)}$, $n_0^{(u)}$ und $n_1^{(u)}$ nicht der Versuchskontrolle unterliegen und die Positivdefinitheit der Kovarianzmatrix durch die Werner-Brunner-Schätzung daher nicht per Versuchsplanung gesichert werden kann, wird in dieser Arbeit eine an Konietschke und Brunner (2009) angelehnte Schätzung der Kovarianzmatrix verwendet. Die Schätzung erfolgt hier zunächst getrennt für die vollständigen und unvollständigen Fälle und ergibt sich schließlich als gewichtetes Mittel dieser unabhängigen Varianz-Kovarianzschätzer. Dieses Vorgehen sichert die Positivdefinitheit der Kovarianzschätzer der Gütemaße per Konstruktion (vergleiche Konietschke und Brunner, 2009). Problematisch bei dem Ansatz von Konietschke und Brunner (2009) ist jedoch, dass die Effektschätzer nicht per Konstruktion

bereichserhaltend sind.[7] Das heißt, die Schätzer von Sensitivität, Spezifität, AUC und prädiktiven Werten liegen nur asymptotisch im Einheitsintervall, können aber – insbesondere im Fall kleiner Stichproben – größer als 1 oder kleiner als 0 werden. Dieser als ein Artefakt der Schätzung auftretender Effekt wird mit den in dieser Arbeit entwickelten Schätzern korrigiert, sodass sich schließlich die folgenden gewichteten und ungewichteten Schätzer ergeben.

Gewichtete Schätzer

In dieser Arbeit wird mit der Präsentation der gewichteten Schätzer begonnen. Diese sind – wie später zu erkennen sein wird – leichter handhabbar als die ungewichteten: Während sich für die gewichteten Schätzer und deren empirische Kovarianzmatrizen kompakte Rangdarstellungen finden lassen, ist dieses für die ungewichteten Schätzer nicht möglich.

Schätzung der Verteilungsfunktionen

Der folgende Satz 4.14 zeigt die Konsistenz des in Gleichung (4.17) angegebenen gewichteten Schätzers der Verteilungsfunktion. Gleichung (4.17) unterscheidet sich hierbei von der Repräsentation aus Satz 4.14 lediglich in der Darstellung: Letztere Repräsentation zeigt die empirische Verteilungsfunktion bereits als gewichtetes Mittel der empirischen Verteilungsfunktionen der vollständigen und der unvollständigen Fälle. Diese Zerlegung bekommt – wie bereits beschrieben – bei der späteren Schätzung der Kovarianzenmatrix eine besondere Bedeutung.

Satz 4.14 (Gewichtete empirische Verteilungsfunktion) *Im obigen Modell 2 wird die marginale Verteilung* $\mathrm{F}_i^{(l)}$ *der Beobachtungen der* l*-ten Komponente (*$l = 1,\dots,d$*) im Zustand* i *(*$i = 0,1$*) konsistent durch*

$$\widetilde{\mathrm{F}}_i^{(l)}(x) = \frac{m_{i\cdot}^{(l|v)}}{m_{i\cdot}^{(l|\cdot)}} \cdot \widetilde{\mathrm{F}}_i^{(l|v)}(x) + \frac{m_{i\cdot}^{(l|u)}}{m_{i\cdot}^{(l|\cdot)}} \cdot \widetilde{\mathrm{F}}_i^{(l|u)}(x)$$

geschätzt, wobei

$$\widetilde{\mathrm{F}}_i^{(l|v)}(x) = \frac{1}{m_{i\cdot}^{(l|v)}} \sum_{k=1}^{n^{(v)}} \sum_{s=1}^{m_{ik}^{(l|v)}} c(x - \mathrm{X}_{iks}^{(l|v)}) \quad \textit{und} \quad \widetilde{\mathrm{F}}_i^{(l|u)}(x) = \frac{1}{m_{i\cdot}^{(l|u)}} \sum_{k=1}^{n_i^{(u)}} \sum_{s=1}^{m_{ik}^{(l|u)}} c(x - \mathrm{X}_{iks}^{(l|u)}).$$

BEWEIS. $t \in \{u, v\}$ sei im Folgenden der Index, welcher die unvollständigen ($t = u$) und vollständigen Fälle ($t = v$) kennzeichne. In einem ersten Schritt wird zunächst der Ausdruck $\mathrm{E}[\widetilde{\mathrm{F}}_i^{(l|t)}(x) - \mathrm{F}_i^{(l)}(x)]^2$ abgeschätzt. Es gilt für alle $x \in \mathbb{R}$:

$$\begin{aligned}
\mathrm{E}\left[\widetilde{\mathrm{F}}_i^{(l|t)}(x) - \mathrm{F}_i^{(l)}(x)\right]^2 &= \left(\frac{1}{m_{i\cdot}^{(l|t)}}\right)^2 \sum_{k=1}^{n_i^{(t)}} \sum_{s=1}^{m_{ik}^{(l|t)}} \sum_{k'=1}^{n_i^{(t)}} \sum_{s'=1}^{m_{ik'}^{(l|t)}} \mathrm{E}\left[\left(c(x - \mathrm{X}_{iks}^{(l|t)}) - \mathrm{F}_i^{(l)}(x)\right)\left(c(x - \mathrm{X}_{ik's'}^{(l|t)}) - \mathrm{F}_i^{(l)}(x)\right)\right] \\
&= \left(\frac{1}{m_{i\cdot}^{(l|t)}}\right)^2 \sum_{k=1}^{n_i^{(t)}} \sum_{s',s=1}^{m_{ik}^{(l|t)}} \mathrm{E}\left[\left(c(x - \mathrm{X}_{iks}^{(l|t)}) - \mathrm{F}_i^{(l)}(x)\right)\left(c(x - \mathrm{X}_{iks'}^{(l|t)}) - \mathrm{F}_i^{(l)}(x)\right)\right] \\
&\leq \left(\frac{1}{m_{i\cdot}^{(l|t)}}\right)^2 \sum_{k=1}^{n_i^{(t)}} \sum_{s',s=1}^{m_{ik}^{(l|t)}} 1 = \left(\frac{1}{m_{i\cdot}^{(l|t)}}\right)^2 \sum_{k=1}^{n_i^{(t)}} \left(m_{ik}^{(l|t)}\right)^2 \leq \left(\frac{1}{m_{i\cdot}^{(l|t)}}\right)^2 \sum_{k=1}^{n_i^{(t)}} m_{ik}^{(l|t)} m_{\max} \leq \frac{m_{\max}}{m_{i\cdot}^{(l|t)}},
\end{aligned}$$

[7] Bereichserhaltend bedeutet in diesem Fall, dass die Schätzer im Intervall $[0, 1]$ liegen. Eine genauere Definition findet man bei Efron und Tibshirani (1993, Abschnitt 13.6).

wobei im zweiten Schritt verwendet wird, dass die beiden Terme für verschiedene k unabhängig sind und der Erwartungswert der einzelnen Terme 0 ist. Unter Verwendung der c_r-Ungleichung lässt sich nun zeigen:

$$\begin{aligned}
\mathrm{E}\left[\widetilde{\mathrm{F}}_i^{(l)}(x) - \mathrm{F}_i^{(l)}(x)\right]^2 &\leq 2\left(\frac{m_{i\cdot}^{(l|v)}}{m_{i\cdot}^{(l|\cdot)}}\right)^2 \mathrm{E}\left[\widetilde{\mathrm{F}}_i^{(l|v)}(x) - \mathrm{F}_i^{(l)}(x)\right]^2 + 2\left(\frac{m_{i\cdot}^{(l|u)}}{m_{i\cdot}^{(l|\cdot)}}\right)^2 \mathrm{E}\left[\widetilde{\mathrm{F}}_i^{(l|u)}(x) - \mathrm{F}_i^{(l)}(x)\right]^2 \\
&\leq 2\frac{\left(m_{i\cdot}^{(l|v)}\right)^2 m_{\max}}{\left(m_{i\cdot}^{(l|\cdot)}\right)^2 m_{i\cdot}^{(l|v)}} + 2\frac{\left(m_{i\cdot}^{(l|u)}\right)^2 m_{\max}}{\left(m_{i\cdot}^{(l|\cdot)}\right)^2 m_{i\cdot}^{(l|u)}} \\
&\leq 2\underbrace{\frac{m_{i\cdot}^{(l|v)}}{m_{i\cdot}^{(l|\cdot)}}}_{\leq 1}\frac{m_{\max}}{m_{i\cdot}^{(l|\cdot)}} + 2\underbrace{\frac{m_{i\cdot}^{(l|u)}}{m_{i\cdot}^{(l|\cdot)}}}_{\leq 1}\frac{m_{\max}}{m_{i\cdot}^{(l|\cdot)}} \leq 4\frac{m_{\max}}{m_{i\cdot}^{(l|\cdot)}} \to 0. \qquad (4.19)
\end{aligned}$$

Für spätere Zwecke sei an dieser Stelle angemerkt, dass sich aus diesem Resultat und dem Satz von Fubini direkt folgern lässt, dass ebenfalls

$$\mathrm{E}\left[\widetilde{\mathrm{F}}_0^{(l|t)}(\mathrm{X}_{1ks}^{(l|t')}) - \mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')})\right]^2 \leq \frac{m_{\max}}{m_{0\cdot}^{(l|t)}}, \quad \forall l = 1,\ldots,d,\ t,t' \in \{u,v\}, \quad \text{sowie} \qquad (4.20)$$

$$\mathrm{E}\left[\widetilde{\mathrm{F}}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')}) - \mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')})\right]^2 \leq \frac{4m_{\max}}{m_{0\cdot}^{(l|\cdot)}} \quad \forall l = 1,\ldots,d,\ t' \in \{u,v\} \qquad (4.21)$$

gilt. □

Basierend auf den gewichteten Schätzern der Verteilungsfunktionen, werden im Folgenden mit Hilfe der Plug-In-Methode die gewichteten Schätzer für die verschiedenen diagnostischen Gütemaße hergeleitet.

Schätzung der AUC

In diesem Abschnitt wird zunächst die Konsistenz des Plug-In-Schätzers für die AUC nachgewiesen; im Anschluss wird eine Rechenzeit-effiziente Rangdarstellung dieses Schätzer präsentiert.

Satz 4.15 *Für alle $l = 1,\ldots,d$ wird die Fläche unter der ROC-Kurve konsistent durch*

$$\widehat{\mathrm{AUC}}^{(l)} = \int \widetilde{\mathrm{F}}_0^{(l)} d\widetilde{\mathrm{F}}_1^{(l)}$$

geschätzt.

BEWEIS. Es gilt

$$\begin{aligned}
\int \widetilde{\mathrm{F}}_0^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} &= \frac{m_{0\cdot}^{(l|v)}}{m_{0\cdot}^{(l|\cdot)}}\frac{m_{1\cdot}^{(l|v)}}{m_{1\cdot}^{(l|\cdot)}}\int \widetilde{\mathrm{F}}_0^{(l|v)} d\widetilde{\mathrm{F}}_1^{(l|v)} + \frac{m_{0\cdot}^{(l|v)}}{m_{0\cdot}^{(l|\cdot)}}\frac{m_{1\cdot}^{(l|u)}}{m_{1\cdot}^{(l|\cdot)}}\int \widetilde{\mathrm{F}}_0^{(l|v)} d\widetilde{\mathrm{F}}_1^{(l|u)} \\
&\quad + \frac{m_{0\cdot}^{(l|u)}}{m_{0\cdot}^{(l|\cdot)}}\frac{m_{1\cdot}^{(l|v)}}{m_{1\cdot}^{(l|\cdot)}}\int \widetilde{\mathrm{F}}_0^{(l|u)} d\widetilde{\mathrm{F}}_0^{(l|v)} + \frac{m_{0\cdot}^{(l|u)}}{m_{0\cdot}^{(l|\cdot)}}\frac{m_{1\cdot}^{(l|u)}}{m_{1\cdot}^{(l|\cdot)}}\int \widetilde{\mathrm{F}}_0^{(l|u)} d\widetilde{\mathrm{F}}_0^{(l|u)} \\
&= \sum_{t\in\{u,v\}}\sum_{t'\in\{u,v\}} \frac{m_{0\cdot}^{(l|t)}}{m_{0\cdot}^{(l|\cdot)}} \cdot \frac{m_{1\cdot}^{(l|t')}}{m_{1\cdot}^{(l|\cdot)}} \int \widetilde{\mathrm{F}}_0^{(l|t)} d\widetilde{\mathrm{F}}_1^{(l|t')}.
\end{aligned}$$

Da

$$\sum_{t\in\{u,v\}}\sum_{t'\in\{u,v\}} \frac{m_{0\cdot}^{(l|t)}}{m_{0\cdot}^{(l|\cdot)}} \cdot \frac{m_{1\cdot}^{(l|t')}}{m_{1\cdot}^{(l|\cdot)}} = 1,$$

ist $\int \widetilde{\mathrm{F}}_0^{(l)} d\widetilde{\mathrm{F}}_1^{(l)}$ das gewichtete Mittel der $\int \widetilde{\mathrm{F}}_0^{(l|t)} d\widetilde{\mathrm{F}}_1^{(l|t')}$. Hieraus ergibt sich unter Verwendung der c_r-Ungleichung:

$$\begin{aligned}
\mathrm{E}\left(\int \widetilde{\mathrm{F}}_0^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right)^2 &= \mathrm{E}\left(\sum_{t\in\{u,v\}} \sum_{t'\in\{u,v\}} \frac{m_{0\cdot}^{(l|t)}}{m_{0\cdot}^{(l|\cdot)}} \cdot \frac{m_{1\cdot}^{(l|t')}}{m_{1\cdot}^{(l|\cdot)}} \int \widetilde{\mathrm{F}}_0^{(l|t)} d\widetilde{\mathrm{F}}_1^{(l|t')} - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right)^2 \\
&\leq 4 \cdot \sum_{t\in\{u,v\}} \sum_{t'\in\{u,v\}} \left(\frac{m_{0\cdot}^{(l|t)}}{m_{0\cdot}^{(l|\cdot)}} \cdot \frac{m_{1\cdot}^{(l|t')}}{m_{1\cdot}^{(l|\cdot)}}\right)^2 \mathrm{E}\left(\int \widetilde{\mathrm{F}}_0^{(l|t)} d\widetilde{\mathrm{F}}_1^{(l|t')} - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right)^2. \qquad (4.22)
\end{aligned}$$

Für die einzelnen Summanden gilt nun mit Hilfe der Jensen- und der c_r-Ungleichung:

$$\begin{aligned}
&\mathrm{E}\left(\int \widetilde{\mathrm{F}}_0^{(l|t)} d\widetilde{\mathrm{F}}_1^{(l|t')} - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right)^2 = \mathrm{E}\left(\int \widetilde{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)} d\widetilde{\mathrm{F}}_1^{(l|t')} + \int \mathrm{F}_0^{(l)} d(\widetilde{\mathrm{F}}_1^{(l|t')} - \mathrm{F}_1^{(l)})\right)^2 \\
&\leq 2\mathrm{E}\left(\int \left(\widetilde{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)}\right) d\widetilde{\mathrm{F}}_1^{(l|t')}\right)^2 + 2\mathrm{E}\left(\int \mathrm{F}_0^{(l)} d(\widetilde{\mathrm{F}}_1^{(l|t')} - \mathrm{F}_1^{(l)})\right)^2 \\
&\leq 2\mathrm{E}\int \left(\widetilde{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)}\right)^2 d\widetilde{\mathrm{F}}_1^{(l|t')} + 2\mathrm{E}\left(\int \mathrm{F}_0^{(l)} d(\widetilde{\mathrm{F}}_1^{(l|t')} - \mathrm{F}_1^{(l)})\right)^2 \\
&\leq \frac{2}{m_{1\cdot}^{(l|t')}} \sum_{k=1}^{n_1^{(t')}} \sum_{s=1}^{m_{1k}^{(l|t')}} \mathrm{E}\left(\widetilde{\mathrm{F}}_0^{(l|t)}(\mathrm{X}_{1ks}^{(l|t')}) - \mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')})\right)^2 \\
&\quad + \frac{2}{\left(m_{1\cdot}^{(l|t')}\right)^2} \sum_{k=1}^{n_1^{(t')}} \sum_{s=1}^{m_{1k}^{(l|t')}} \sum_{k'=1}^{n_1^{(t')}} \sum_{s'=1}^{m_{1k'}^{(l|t')}} \mathrm{E}\left(\left[\mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')}) - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right]\left[\mathrm{F}_0^{(l)}(\mathrm{X}_{1k's'}^{(l|t')}) - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right]\right) \\
&\leq \frac{2}{m_{1\cdot}^{(l|t')}} \sum_{k=1}^{n_1^{(t')}} \sum_{s=1}^{m_{1k}^{(l|t')}} \frac{m_{\max}}{m_{0\cdot}^{(l|t)}} + \frac{2}{\left(m_{1\cdot}^{(l|t')}\right)^2} \sum_{k=1}^{n_1^{(t')}} \sum_{s',s=1}^{m_{1k}^{(l|t')}} \mathrm{E}\left[\left(\mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')}) - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right)\left(\mathrm{F}_0^{(l)}(\mathrm{X}_{1ks'}^{(l|t')}) - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right)\right] \\
&\leq \frac{2m_{\max}}{m_{0\cdot}^{(l|t)}} + \frac{2}{\left(m_{1\cdot}^{(l|t')}\right)^2} \sum_{k=1}^{n_1^{(t')}} \sum_{s',s=1}^{m_{1k}^{(l|t')}} 1 \leq \frac{2m_{\max}}{m_{0\cdot}^{(l|t)}} + \frac{2m_{\max}}{m_{1\cdot}^{(l|t')}},
\end{aligned}$$

wobei im vorletzten Schritt für den ersten Summanden die Abschätzung (4.20) verwendet wurde. Für den zweiten Summanden wurde die Tatsache ausgenutzt, dass die beiden Terme $\mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')}) - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}$ und $\mathrm{F}_0^{(l)}(\mathrm{X}_{1k's'}^{(l|t')}) - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}$ für verschiedene k unabhängig sind und dass der Erwartungswert der einzelnen Terme = 0 ist.
Das Einsetzen dieses Resultats in Gleichung 4.22 ergibt nun:

$$\begin{aligned}
\mathrm{E}\left(\int \widetilde{\mathrm{F}}_0^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right)^2 &\leq 4 \cdot \sum_{t\in\{u,v\}} \sum_{t'\in\{u,v\}} \left(\frac{m_{0\cdot}^{(l|t)}}{m_{0\cdot}^{(l|\cdot)}} \cdot \frac{m_{1\cdot}^{(l|t')}}{m_{1\cdot}^{(l|\cdot)}}\right)^2 \left(\frac{2m_{\max}}{m_{0\cdot}^{(l|t)}} + \frac{2m_{\max}}{m_{1\cdot}^{(l|t')}}\right) \\
&\leq 4 \cdot \sum_{t\in\{u,v\}} \sum_{t'\in\{u,v\}} \left(\frac{2m_{\max}}{m_{0\cdot}^{(l|\cdot)}} + \frac{2m_{\max}}{m_{1\cdot}^{(l|\cdot)}}\right) = \frac{32m_{\max}}{m_{0\cdot}^{(l|\cdot)}} + \frac{32m_{\max}}{m_{1\cdot}^{(l|\cdot)}} \to 0.
\end{aligned}$$

Hierbei verfährt man bei der Abschätzung im vorletzten Schritt analog zu Abschätzung (4.19) (Seite 38) im Beweis zu Satz 4.14. □

Es soll nun eine Rangdarstellung für den oben präsentierten Schätzer hergeleitet werden. Dafür wird zunächst folgendes Lemma benötigt:

Lemma 4.1 *Es sei* $\widetilde{\mathrm{H}}^{(l)} = \frac{1}{\mathrm{M}^{(l)}}\left(m_{0\cdot}^{(l|\cdot)}\widetilde{\mathrm{F}}_0^{(l)} + m_{1\cdot}^{(l|\cdot)}\widetilde{\mathrm{F}}_1^{(l)}\right)$ *die mittlere gewichtete empirische Verteilungsfunktion von* $\mathrm{F}_0^{(l)}$ *und* $\mathrm{F}_1^{(l)}$*, dann gilt:*

(1.) $$\widetilde{\mathrm{AUC}}^{(l)} = \int \widetilde{\mathrm{F}}_0^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} = \int \widetilde{\mathrm{H}}^{(l)} d(\widetilde{\mathrm{F}}_1^{(l)} - \widetilde{\mathrm{F}}_0^{(l)}) + \frac{1}{2}, \quad \forall\, l = 1,\ldots,d$$

(2.) $$\widetilde{\mathrm{H}}^{(l)}(\mathrm{X}_{1ks}^{(l|t)}) = \frac{1}{\mathrm{M}^{(l)}}\left(\mathrm{R}_{1ks}^{(l|t)} - \frac{1}{2}\right), \quad \forall\, l = 1,\ldots,d,\ t \in \{u,v\},$$

wobei $\mathrm{R}_{1ks}^{(l|t)}$ *den Rang von* $\mathrm{X}_{1ks}^{(l|t)}$ *unter allen Beobachtungen der* l*-ten Komponente beschreibe.*

BEWEIS.

(1.) Der Beweis erfolgt in Anlehnung an Werner (2006, Kapitel 3, Lemma 3.1), in deren Arbeit ein vergleichbares Resultat nachgewiesen wird.

$$\begin{aligned}
\int \widetilde{\mathrm{F}}_0^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} &= \frac{m_{0\cdot}^{(l|\cdot)}}{\mathrm{M}^{(l)}} \int \widetilde{\mathrm{F}}_0^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} + \frac{m_{1\cdot}^{(l|\cdot)}}{\mathrm{M}^{(l)}} \int \widetilde{\mathrm{F}}_0^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} \\
&= \frac{m_{0\cdot}^{(l|\cdot)}}{\mathrm{M}^{(l)}} \int \widetilde{\mathrm{F}}_0^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} + \frac{m_{1\cdot}^{(l|\cdot)}}{\mathrm{M}^{(l)}} - \frac{m_{1\cdot}^{(l|\cdot)}}{\mathrm{M}^{(l)}} \int \widetilde{\mathrm{F}}_1^{(l)} d\widetilde{\mathrm{F}}_0^{(l)} \\
&= \int \widetilde{\mathrm{H}}^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} - \frac{m_{1\cdot}^{(l|\cdot)}}{\mathrm{M}^{(l)}} \int \widetilde{\mathrm{F}}_1^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} + \frac{m_{1\cdot}^{(l|\cdot)}}{\mathrm{M}^{(l)}} - \int \widetilde{\mathrm{H}}^{(l)} d\widetilde{\mathrm{F}}_0^{(l)} + \frac{m_{0\cdot}^{(l|\cdot)}}{\mathrm{M}^{(l)}} \int \widetilde{\mathrm{F}}_0^{(l)} d\widetilde{\mathrm{F}}_0^{(l)} \\
&= \int \widetilde{\mathrm{H}}^{(l)} d(\widetilde{\mathrm{F}}_1^{(l)} - \widetilde{\mathrm{F}}_0^{(l)}) + \frac{m_{1\cdot}^{(l|\cdot)}}{\mathrm{M}^{(l)}} - \frac{1}{2}\frac{m_{1\cdot}^{(l|\cdot)}}{\mathrm{M}^{(l)}} + \frac{1}{2}\frac{m_{0\cdot}^{(l|\cdot)}}{\mathrm{M}^{(l)}} = \int \widetilde{\mathrm{H}}^{(l)} d(\widetilde{\mathrm{F}}_1^{(l)} - \widetilde{\mathrm{F}}_0^{(l)}) + \frac{1}{2}.
\end{aligned}$$

(2.) Für die mittlere empirische Verteilungsfunktion gilt:

$$\begin{aligned}
\widetilde{\mathrm{H}}^{(l)}(x) &= \frac{1}{\mathrm{M}^{(l)}}\left(m_{0\cdot}^{(l|\cdot)}\widetilde{\mathrm{F}}_0^{(l)} + m_{1\cdot}^{(l|\cdot)}\widetilde{\mathrm{F}}_1^{(l)}\right) = \frac{1}{\mathrm{M}^{(l)}}\left(m_{0\cdot}^{(l|u)}\widetilde{\mathrm{F}}_0^{(l|u)} + m_{0\cdot}^{(l|v)}\widetilde{\mathrm{F}}_0^{(l|v)} + m_{1\cdot}^{(l|u)}\widetilde{\mathrm{F}}_1^{(l|u)} + m_{1\cdot}^{(l|v)}\widetilde{\mathrm{F}}_1^{(l)|v}\right) \\
&= \frac{1}{\mathrm{M}^{(l)}}\left(\sum_{k=1}^{n_0^{(u)}}\sum_{s=1}^{m_{0k}^{(l|u)}} c(x - \mathrm{X}_{0ks}^{(l|u)}) + \sum_{k=1}^{n^{(v)}}\sum_{s=1}^{m_{0k}^{(l|v)}} c(x - \mathrm{X}_{0ks}^{(l|v)}) + \sum_{k=1}^{n_1^{(u)}}\sum_{s=1}^{m_{1k}^{(l|u)}} c(x - \mathrm{X}_{1ks}^{(l|u)}) + \sum_{k=1}^{n^{(v)}}\sum_{s=1}^{m_{1k}^{(l|v)}} c(x - \mathrm{X}_{1ks}^{(l|v)})\right) \\
&= \frac{1}{\mathrm{M}^{(l)}}\left(\sum_{i=0}^{1}\sum_{t\in\{u,v\}}\sum_{k=1}^{n_i^{(t)}}\sum_{s=1}^{m_{iks}^{(l|t)}} c(x - \mathrm{X}_{iks}^{(l|t)})\right).
\end{aligned}$$

Des Weiteren gilt für alle $l = 1,\ldots,d$ und $t \in \{u,v\}$:

$$\begin{aligned}
\mathrm{R}_{iks}^{(l|t)} &= \frac{1}{2} + \sum_{i=0}^{1}\sum_{t'\in\{u,v\}}\sum_{k'=1}^{n_i^{(t')}}\sum_{s'=1}^{m_{ik's'}^{(l|t')}} c(\mathrm{X}_{iks}^{(l|t)} - \mathrm{X}_{ik's'}^{(l|t')}) \\
&= \frac{1}{2} + \mathrm{M}^{(l)} \cdot \widetilde{\mathrm{H}}^{(l)}(\mathrm{X}_{iks}^{(l|t)}),
\end{aligned}$$

woraus das Resulat $\widetilde{\mathrm{H}}^{(l)}(\mathrm{X}_{1ks}^{(l|t)}) = \frac{1}{\mathrm{M}^{(l)}}\left(\mathrm{R}_{1ks}^{(l|t)} - \frac{1}{2}\right)$ direkt folgt.

□

Aus diesem Lemma ergibt sich nun direkt die Rangdarstellung des gewichteten AUC-Schätzers.

Korollar 4.2 (Rangdarstellung des gewichteten AUC Schätzers) *Für den gewichteten Schätzer der Fläche unter der ROC-Kurve* $\widetilde{\mathrm{AUC}}$ *gilt:*

$$\widetilde{\mathrm{AUC}}^{(l)} = \frac{1}{\mathrm{M}^{(l)}}\left(\widetilde{\mathrm{R}}_{1\cdot\cdot}^{(l|\cdot)} - \widetilde{\mathrm{R}}_{0\cdot\cdot}^{(l|\cdot)}\right) + \frac{1}{2},$$

wobei $\widetilde{\mathrm{R}}_{i\cdot\cdot}^{(l|\cdot)} = \frac{1}{m_{i\cdot}^{(l|\cdot)}} \sum_{t\in\{u,v\}} \sum_{k=1}^{n_i^{(t)}} \sum_{s=1}^{m_{ik}^{(l|t)}} \mathrm{R}_{iks}^{(l|t)}$ *den gewichteten Rangmittelwert beschreibe.*

BEWEIS.

$$
\begin{aligned}
\widetilde{\mathrm{AUC}}^{(l)} &= \int \widetilde{\mathrm{H}}^{(l)} d(\widetilde{\mathrm{F}}_1^{(l)} - \widetilde{\mathrm{F}}_0^{(l)}) + \frac{1}{2} = \int \widetilde{\mathrm{H}}^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} - \int \widetilde{\mathrm{H}}^{(l)} d\widetilde{\mathrm{F}}_0^{(l)} + \frac{1}{2} \\
&= \sum_{t\in\{u,v\}} \frac{m_{1\cdot}^{(l|t)}}{m_{1\cdot}^{(l|\cdot)}} \int \widetilde{\mathrm{H}}^{(l)} d\widetilde{\mathrm{F}}_1^{(l|t)} - \sum_{t\in\{u,v\}} \frac{m_{0\cdot}^{(l|t)}}{m_{0\cdot}^{(l|\cdot)}} \int \widetilde{\mathrm{H}}^{(l)} d\widetilde{\mathrm{F}}_0^{(l|t)} + \frac{1}{2} \\
&= \sum_{t\in\{u,v\}} \frac{m_{1\cdot}^{(l|t)}}{m_{1\cdot}^{(l|\cdot)}} \frac{1}{m_{1\cdot}^{(l|t)}} \sum_{k=1}^{n_1^{(t)}} \sum_{s=1}^{m_{1k}^{(l|t)}} \widetilde{\mathrm{H}}^{(l)}(\mathrm{X}_{1ks}^{(l|t)}) - \sum_{t\in\{u,v\}} \frac{m_{0\cdot}^{(l|t)}}{m_{0\cdot}^{(l|\cdot)}} \frac{1}{m_{0\cdot}^{(l|t)}} \sum_{k=1}^{n_0^{(t)}} \sum_{s=1}^{m_{0k}^{(l|t)}} \widetilde{\mathrm{H}}^{(l)}(\mathrm{X}_{0ks}^{(l|t)}) + \frac{1}{2} \\
&= \frac{1}{\mathrm{M}^{(l)}} \left(\sum_{t\in\{u,v\}} \frac{m_{1\cdot}^{(l|t)}}{m_{1\cdot}^{(l|\cdot)}} \frac{1}{m_{1\cdot}^{(l|t)}} \sum_{k=1}^{n_1^{(t)}} \sum_{s=1}^{m_{1k}^{(l|t)}} \left(\mathrm{R}_{1ks}^{(l|t)} - \frac{1}{2} \right) - \sum_{t\in\{u,v\}} \frac{m_{0\cdot}^{(l|t)}}{m_{0\cdot}^{(l|\cdot)}} \frac{1}{m_{0\cdot}^{(l|t)}} \sum_{k=1}^{n_0^{(t)}} \sum_{s=1}^{m_{0k}^{(l|t)}} \left(\mathrm{R}_{0ks}^{(l|t)} - \frac{1}{2} \right) \right) + \frac{1}{2} \\
&= \frac{1}{\mathrm{M}^{(l)}} \left(\frac{1}{m_{1\cdot}^{(l|\cdot)}} \sum_{t\in\{u,v\}} \sum_{k=1}^{n_1^{(t)}} \sum_{s=1}^{m_{1k}^{(l|t)}} \left(\mathrm{R}_{1ks}^{(l|t)} - \frac{1}{2} \right) - \frac{1}{m_{0\cdot}^{(l|\cdot)}} \sum_{t\in\{u,v\}} \sum_{k=1}^{n_0^{(t)}} \sum_{s=1}^{m_{0k}^{(l|t)}} \left(\mathrm{R}_{0ks}^{(l|t)} - \frac{1}{2} \right) \right) + \frac{1}{2} \\
&= \frac{1}{\mathrm{M}^{(l)}} \left(\widetilde{\mathrm{R}}_{1\cdot\cdot}^{(l|\cdot)} - \widetilde{\mathrm{R}}_{0\cdot\cdot}^{(l|\cdot)} \right) + \frac{1}{2}.
\end{aligned}
$$

□

Wie bereits in der Einführung dieses Kapitels angekündigt, erfolgt an dieser Stelle einmalig die Erweiterung der Theorie für die AUC auf die Theorie der übrigen Gütemaße. An den übrigen Stellen sei für diese Weiterentwicklung an die entsprechende Passage aus dem Kapitel für Einfachmessungen verwiesen.

Schätzung von Sensitivität und Spezifität

Analog zur Analyse von Sensitivität und Spezifität bei Einfachmessungen werden auch bei Clusterdaten *se* und *sp* als Flächen unter speziellen ROC-Kurven interpretiert. Um auch hier Sensitivität und Spezifität als Pseudo-Zwei-Stichproben-Größen darstellen zu können, muss gezeigt werden, dass der gewichtete Schätzer einer Ein-Punkt-Verteilung $\widetilde{\mathbf{\Gamma}}$ der Ein-Punkt-Verteilung $\mathbf{\Gamma}$ entspricht (vergleiche Gleichung 4.2 in Abschnitt 4.1). Analog zu Abschnitt 4.1.2 sei $\boldsymbol{\gamma} = (\gamma^{(1)}, \ldots, \gamma^{(d)})'$ der Vektor der Cut-Off-Punkte für jede Komponente. Es sei nun

$$
\gamma_i^{(l)}(\mathrm{X}_{i'ks}^{(l|t)}) = \begin{cases} \gamma^{(l)} \text{ falls } i = i' \\ \mathrm{X}_{i'ks}^{(l|t)} \text{ sonst} \end{cases}
$$

eine Funktion, die jeder Beobachtung im Zustand i auf den Cut-Off der zugehörigen Komponente abbildet und den Wert jeder Beobachtungen im zu i komplementären Zustand beibehält. Die Stichproben

$$
\begin{aligned}
\gamma_0(\mathbf{X}_k^{(v)}) &= \left(\gamma_0^{(1)}\left(\mathbf{X}_{0k}^{(1|v)}, \mathbf{X}_{1k}^{(1|v)}\right), \ldots, \gamma_0^{(d)}\left(\mathbf{X}_{0k}^{(d|v)}, \mathbf{X}_{1k}^{(d|v)}\right) \right), && k = 1, \ldots n^{(v)} \\
\gamma_0(\mathbf{X}_{ik}^{(u)}) &= \left(\gamma_0^{(1)}\left(\mathbf{X}_{ik}^{(1|u)}\right), \gamma_0^{(2)}\left(\mathbf{X}_{ik}^{(2|u)}\right), \ldots, \gamma_0^{(d)}\left(\mathbf{X}_{ik}^{(d|u)}\right) \right), && i = 0,1,\ k = 1, \ldots, n_i^{(u)}
\end{aligned}
$$

können somit als Realisation einem nach $\mathbf{F}_1$ verteilten Kollektiv von kranken Beobachtungseinheiten und einem nach $\mathbf{\Gamma}$ verteilten Kollektiv an gesunden Beobachtungen betrachtet werden. Wird diese Verteilung der gesunden Beobachtungen nun aus obiger Pseudostichprobe durch die gewichtete empirische Vertei-

lungsfunktion geschätzt, so gilt für diesen Schätzer:

$$\begin{aligned}\widetilde{\Gamma}^{(l)}(x) = \widetilde{\mathrm{F}}_0^{(l)}(x) &= \frac{1}{m_{0\cdot}^{(l|\cdot)}} \sum_{k=1}^{n^{(v)}} \sum_{s=1}^{m_{0k}^{(l|v)}} c\left(x - \gamma_0^{(l)}\left(\mathrm{X}_{0ks}^{(l|v)}\right)\right) + \frac{1}{m_{0\cdot}^{(l|\cdot)}} \sum_{k=1}^{n_0^{(u)}} \sum_{s=1}^{m_{0k}^{(l|u)}} c\left(x - \gamma_0^{(l)}\left(\mathrm{X}_{0ks}^{(l|u)}\right)\right) \\ &= \frac{1}{m_{0\cdot}^{(l|\cdot)}} \sum_{k=1}^{n^{(v)}} \sum_{s=1}^{m_{0k}^{(l|v)}} c\left(x - \gamma^{(l)}\right) + \frac{1}{m_{0\cdot}^{(l|\cdot)}} \sum_{k=1}^{n_0^{(u)}} \sum_{s=1}^{m_{0k}^{(l|u)}} c\left(x - \gamma^{(l)}\right) = c\left(x - \gamma^{(l)}\right) = \Gamma^{(l)}(x).\end{aligned}$$

An dieser Gleichheit lässt sich erkennen, dass zur Schätzung von $\Gamma^{(l)}$ auch eine beliebige andere Stichprobengröße an gesunden Patienten hätte verwendet werden können. Lediglich zur Vereinfachung der Notation wurde an dieser Stelle die Stichprobe der gesunden Beobachtungen als Basis für die Pseudostichprobe verwendet. Gleiches gilt, wenn mittels der Funktionen $\gamma_1^{(l)},\ l = 1,\ldots,d$ die kranken Beobachtungen mit dem Cut-Off der entsprechenden Komponente überschrieben werden. Analog zum Fall der Einfachmessungen lässt sich nun auch für Clusterdaten ein AUC-Schätzer für Sensitivität und Spezifität angeben:

Satz 4.16 *Für alle $l = 1,\ldots,d$ werden Sensitivität und Spezifität durch*

$$\widetilde{se}^{(l)} = \int \Gamma^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} = \int \widetilde{\Gamma}^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} \tag{4.23}$$

$$\widetilde{sp}^{(l)} = \int \widetilde{\mathrm{F}}_0^{(l)} d\Gamma^{(l)} = \int \widetilde{\mathrm{F}}_0^{(l)} d\widetilde{\Gamma}^{(l)} \tag{4.24}$$

konsistent geschätzt, wobei $\widetilde{\Gamma}^{(l)}$ wie oben beschrieben geschätzt wird, indem alle gesunden Beobachtungen (bei der Schätzung der Sensitivität) beziehungsweise alle kranken Beobachtungen (bei der Schätzung der Spezifität) durch den Cut-Off der zugehörigen Komponente ersetzt werden.

BEWEIS. Für den Beweis der Konsistenz der AUC-Schätzer (Satz 4.15) wurde die Bedingung, dass keine der Marginalverteilungen von $\mathbf{F}_0$ oder $\mathbf{F}_1$ Ein-Punkt-Verteilungen sind, nicht benötigt. Die Behauptung folgt daher direkt aus Gleichheit von $\widetilde{\mathbf{\Gamma}}$ und $\mathbf{\Gamma}$ und der Tatsache, dass Sensitivität und Spezifität spezielle AUCs sind (Satz 2.1) und Satz 4.15. (Vergleichbar zu dem entsprechenden Beweis im Falle der Einfachmessungen.) Eine Rangdarstellung der Schätzer ergibt sich schließlich aus dem zu Satz 4.15 gehörigen Korollar 4.2. □

Es lässt sich nachweisen, dass die in Satz 4.16 beschriebenen Schätzer gerade den kanonischen gewichteten Schätzern

$$\widetilde{se}^{(l)} = \frac{\#\text{richtig krank diagnostizierte Beobachtungen}}{\#\text{kranke Beobachtungen}}$$

$$\widetilde{sp}^{(l)} = \frac{\#\text{richtig gesund diagnostizierte Beobachtungen}}{\#\text{gesunde Beobachtungen}}$$

für Sensitivität und Spezifität entsprechen. Die komplexere Präsentation in Integralform verdeutlicht lediglich, dass die Schätzer von Sensitivität und Spezifität sich als gewöhnliche AUC-Schätzer darstellen lassen.

Die parallele Schätzung von Sensitivität und Spezifität in einem Vektor

Zur späteren Analyse der prädiktiven Werte soll nun auch der Vektor $\left(\mathbf{se}', \mathbf{sp}'\right)'$ mit Hilfe von Pseudobeobachtungen geschätzt werden. Dabei handelt es sich bei $\left(\mathbf{se}', \mathbf{sp}'\right)'$ um einen Vektor der Länge $2d$, in welchem jeder Eintrag die Fläche unter einer speziellen ROC-Kurve darstellt:

$$\begin{pmatrix}\mathbf{se}\\ \mathbf{sp}\end{pmatrix} = \begin{pmatrix}\int \mathbf{\Gamma} d\mathbf{F}_1\\ \int \mathbf{F}_0 d\mathbf{\Gamma}\end{pmatrix} = \int \mathbf{G}_0 d\mathbf{G}_1 \qquad \text{mit} \qquad \mathbf{G}_0 = \begin{pmatrix}\mathbf{\Gamma}\\ \mathbf{F}_0\end{pmatrix} \text{ und } \mathbf{G}_1 = \begin{pmatrix}\mathbf{F}_1\\ \mathbf{\Gamma}\end{pmatrix}.$$

Die Schätzung dieses Vektors mit Hilfe von Pseudobeobachtungen erfolgt in Anlehnung an die im letzten Abschnitt präsentierte Methodik:

(1.) Ersetze bei einem jeden Subjekt alle gesunden Beobachtungen durch den Cut-off $\gamma^{(l)}$ der entsprechenden Komponente.

(2.) Ersetze bei einem jeden Subjekt alle kranken Beobachtungen durch den Cut-off $\gamma^{(l)}$ der entsprechenden Komponente.

(3.) Die neuen Beobachtungsvektoren aus (1.) und (2.) werden für jedes Subjekt zu einem Beobachtungsvektor doppelter Dimension zusammengefasst.

Dieser Vektor wird nun als Beobachtungsvektor unter $2d$ Bedingungen interpretiert. Die Beobachtungen unter den Bedingungen $1,\dots,d$ bestehen aus den kranken Beobachtungen der $1,\dots,d$ verschiedenen Faktorstufenkombinationen sowie gesunden, ein-Punkt-verteilten Pseudobeobachtungen. Mit Hilfe dieser Stichprobe wird daher die Sensitivität der $1,\dots,d$ verschiedenen Reader-Methoden-Kombinationen geschätzt. Die Beobachtungen unter den Bedingungen $d+1,\dots,2d$ bestehen nun aus den gesunden Beobachtungen der $1,\dots,d$ verschiedenen Faktorstufenkombinationen, sowie kranken, ein-Punkt-verteilten Pseudobeobachtungen, sodass in den Komponenten $d+1,\dots,2d$ die Spezifität unter den $1,\dots,d$ verschiedenen Bedingungen geschätzt wird.
Dieses Vorgehen lässt sich wie folgt formalisieren: Die neuen Beobachtungsvektoren $\gamma(\mathbf{X}_{ik}^{(t)})$ entstehen durch

$$\gamma(\mathbf{X}_{ik}^{(t)}) = \begin{pmatrix} \gamma_0(\mathbf{X}_{ik}^{(t)}) \\ \gamma_1(\mathbf{X}_{ik}^{(t)}) \end{pmatrix}, \quad t \in \{u,v\},\ i=0,1,\ k=1,\dots,n_i^{(t)}.$$

Da die unvollständigen Fälle nur über gesunde oder kranke Beobachtungen verfügen, lässt sich in diesem Fall obiger Ausdruck zu

$$\gamma(\mathbf{X}_{0k}^{(u)}) = \begin{pmatrix} \left(\bigoplus_{l=1}^{d} \mathbf{1}_{m_{0k}^{(l|u)}}\right)\cdot\boldsymbol{\gamma} \\ \mathbf{X}_{0k}^{(u)} \end{pmatrix}, \qquad \text{sowie} \qquad \gamma(\mathbf{X}_{1k}^{(u)}) = \begin{pmatrix} \mathbf{X}_{1k}^{(u)} \\ \left(\bigoplus_{l=1}^{d} \mathbf{1}_{m_{1k}^{(l|u)}}\right)\cdot\boldsymbol{\gamma} \end{pmatrix}, \quad i=0,1,\ k=1,\dots,n_i^{(u)}$$

vereinfachen. Werden diese mit Pseudobeobachtungen ergänzten Untersuchungsergebnisse $\gamma(\mathbf{X}_{ik}^{(t)})$, $t \in \{u,v\}$, $i=0,1$ nun als Beobachtungen unter $2d$ verschiedenen Bedingungen aufgefasst und die zugehörigen AUCs gemäß Satz 4.15 geschätzt, so erhält man als Resultat den Vektor $\left(\widetilde{\mathbf{se}}', \widetilde{\mathbf{sp}}'\right)'$. Es ist unverzichtbar, Sensitivität und Spezifität als einen $2d$-dimensionalen Vektor aus AUCs darzustellen, welcher in einem einzelnen Schritt geschätzt wird. Nur dieses Vorgehen erlaubt es, später die Kovarianz zwischen $\widetilde{\mathbf{se}}$ und $\widetilde{\mathbf{sp}}$ zu schätzen. Dieser Kovarianzschätzer ist erforderlich, wenn mit Hilfe von Cramers δ-Satz schließlich die Verteilung und Kovarianzmatrix der prädiktiven Werte berechnet werden soll. Die Methodik der AUC-Analyse lässt sich somit nicht nur auf die einzelnen Vektoren **se** und **sp** übertragen, sondern – durch die obige Manipulation der Stichprobe – auch direkt auf den Vektor $\left(\mathbf{se}', \mathbf{sp}'\right)'$.

Schätzung der prädiktiven Werte

Wie bereits im Fall der Einfachmessungen werden auch bei Clusterdaten die prädiktiven Werte mit Hilfe der Bayes'schen Formel geschätzt.

Satz 4.17 *Für alle $l=1,\dots,d$ und für alle Risikogruppen $g=1,\dots,\mathrm{G}$ werden der positiv und der negativ prädiktive Wert durch*

$$\widetilde{p}_{g,+}^{(l)} = f_+(\widetilde{se}^{(l)}, \widetilde{sp}^{(l)}, \widehat{\pi}_g) = \frac{\widetilde{se}^{(l)}\cdot\widehat{\pi}_g}{\widetilde{se}^{(l)}\cdot\widehat{\pi}_g + (1-\widetilde{sp}^{(l)})\cdot(1-\widehat{\pi}_g)} \tag{4.25}$$

$$\widetilde{p}_{g,-}^{(l)} = f_-(\widetilde{se}^{(l)}, \widetilde{sp}^{(l)}, \widehat{\pi}_g) = \frac{\widetilde{sp}^{(l)} \cdot (1-\widehat{\pi}_g)}{\widetilde{sp}^{(l)} \cdot (1-\widehat{\pi}_g) + (1-\widetilde{se}^{(l)}) \cdot \widehat{\pi}_g}. \tag{4.26}$$

konsistent geschätzt.

BEWEIS. Analog zum Fall ohne Clusterdaten (Satz 4.3, Seite 23) folgt die Konsistenz aus der Konsistenz von $\widetilde{se}^{(l)}, \widetilde{sp}^{(l)}$ und $\widehat{\pi}_g$, dem Slutsky'schen Satz (Slutsky, 1925) und Voraussetzung 4.3. □

Ungewichtete Schätzer

Im Gegensatz zu den gewichteten Schätzern erhält bei den ungewichteten Schätzern nicht jede Beobachtung, sondern jeder Patient das gleiche Gewicht. Eine ungewichtete Schätzung scheint daher in den meisten Fällen adäquater zu sein als eine gewichtete, da hier keiner der Patienten durch eine besonders große Clustergröße das Ergebnis dominieren kann. Die ungewichteten Schätzer bringen jedoch einen entscheidenden Nachteil mit sich: Es lässt sich keine Rangdarstellung dieser Schätzer finden. Werner (2006) und Konietschke und Brunner (2009) umgehen dieses Problem, indem sie sowohl gewichtete als auch ungewichtete Verteilungsfunktionen für die Berechnung des Plug-In-Schätzers verwenden. Da diese Schätzer aber nicht per Konstruktion im Intervall $[0,1]$ liegen müssen, eignen sie sich lediglich für ein lineares Modell. Sie sind hingegen für ein logistisches Modell ungeeignet, da wegen der nur asymptotischen Bereichserhaltung die Möglichkeit der Transformation nur für $N \to \infty$ gewährleistet werden kann. Für eine ausführliche Präsentation dieser Schätzer sei auf die Originalpublikation von Konietschke und Brunner (2009) verwiesen.

Schätzung der Verteilungsfunktionen

Wie bereits im vorherigen Abschnitt werden in einem ersten Schritt ungewichtete Schätzer für die Verteilungsfunktionen präsentiert, welche als Grundlage für die späteren Plug-In-Schätzer dienen.

Satz 4.18 (Ungewichtete empirische Verteilungsfunktion) *Im obigen Modell 2 wird die marginale Verteilung $F_i^{(l)}$ der Beobachtungen l-ten Komponente ($l = 1,\dots,d$) im Zustand i ($i = 0,1$) konsistent durch*

$$\widehat{F}_i^{(l)}(x) = \frac{n^{(v)}}{n^{(v)} + n_i^{(u)}} \cdot \widehat{F}_i^{(l|v)}(x) + \frac{n_i^{(u)}}{n^{(v)} + n_i^{(u)}} \cdot \widehat{F}_i^{(l|u)}(x) = \frac{n^{(v)}}{n_i} \cdot \widehat{F}_i^{(l|v)}(x) + \frac{n_i^{(u)}}{n_i} \cdot \widehat{F}_i^{(l|u)}(x)$$

geschätzt, wobei

$$\widehat{F}_i^{(l|v)}(x) = \frac{1}{n^{(v)}} \sum_{k=1}^{n^{(v)}} \frac{1}{m_{ik}^{(l|v)}} \sum_{s=1}^{m_{ik}^{(l|v)}} c(x - X_{iks}^{(l|v)}) \quad \textit{und} \quad \widehat{F}_i^{(l|u)}(x) = \frac{1}{n_i^{(u)}} \sum_{k=1}^{n_i^{(u)}} \frac{1}{m_{ik}^{(l|u)}} \sum_{s=1}^{m_{ik}^{(l|u)}} c(x - X_{iks}^{(l|u)}).$$

BEWEIS. Es gilt für alle $x \in \mathbb{R}$ gilt:

$$\begin{aligned}
\mathrm{E}\left[\widehat{F}_i^{(l|t)}(x) - F_i^{(l)}(x)\right]^2 &= \mathrm{E}\left[\frac{1}{n_i^{(t)}} \sum_{k=1}^{n_i^{(t)}} \frac{1}{m_{ik}^{(l|t)}} \sum_{s=1}^{m_{ik}^{(l|t)}} \left(c(x - X_{iks}^{(l|t)}) - F_i^{(l|t)}(x)\right)\right]^2 \\
&= \frac{1}{\left(n_i^{(t)}\right)^2} \sum_{k=1}^{n_i^{(t)}} \sum_{k'=1}^{n_i^{(t)}} \frac{1}{m_{ik}^{(l|t)}} \frac{1}{m_{ik'}^{(l|t)}} \sum_{s=1}^{m_{ik}^{(l|t)}} \sum_{s'=1}^{m_{ik'}^{(l|t)}} \mathrm{E}\left[\left(c(x - X_{iks}^{(l|t)}) - F_i^{(l|t)}(x)\right)\left(c(x - X_{ik's'}^{(l|t)}) - F_i^{(l|t)}(x)\right)\right] \\
&\leq \frac{1}{\left(n_i^{(t)}\right)^2} \sum_{k=1}^{n_i^{(t)}} \left(\frac{1}{m_{ik}^{(l|t)}}\right)^2 \sum_{s=1}^{m_{ik}^{(l|t)}} \sum_{s'=1}^{m_{ik}^{(l|t)}} \underbrace{\mathrm{E}\left[\left(c(x - X_{iks}^{(l|t)}) - F_i^{(l|t)}(x)\right)\left(c(x - X_{iks'}^{(l|t)}) - F_i^{(l|t)}(x)\right)\right]}_{\leq 1} \leq \frac{1}{n_i^{(t)}},
\end{aligned}$$

wobei im letzten Schritt verwendet wird, dass die beiden Terme für verschiedene k unabhängig sind und der Erwartungswert der einzelnen Terme = 0 ist.
Aus diesem Resultat folgt schließlich:

$$\begin{aligned}
\mathrm{E}\left[\widehat{\mathrm{F}}_i^{(l)}(x) - \mathrm{F}_i^{(l)}(x)\right]^2 &= \mathrm{E}\left[\frac{n^{(v)}}{n_i} \cdot \widehat{\mathrm{F}}_i^{(l|v)}(x) + \frac{n_i^{(u)}}{n_i} \cdot \widehat{\mathrm{F}}_i^{(l|u)}(x) - \mathrm{F}_i^{(l)}(x)\right]^2 \\
&\leq 2\left(\frac{n^{(v)}}{n_i}\right)^2 \mathrm{E}\left(\widehat{\mathrm{F}}_i^{(l|v)}(x) - \mathrm{F}_i^{(l)}(x)\right)^2 + 2\left(\frac{n_i^{(u)}}{n_i}\right)^2 \mathrm{E}\left(\widehat{\mathrm{F}}_i^{(l|u)}(x) - \mathrm{F}_i^{(l)}(x)\right)^2 \\
&\leq 2\left(\frac{n^{(v)}}{n_i}\right)^2 \frac{1}{n^{(v)}} + 2\left(\frac{n_i^{(u)}}{n_i}\right)^2 \frac{1}{n_i^{(u)}} \leq \frac{4}{n_i} \to 0
\end{aligned}$$

Analog zum gewichteten Fall sei an dieser Stelle für spätere Zwecke bereits angemerkt, dass aus dem Satz von Fubini folgt, dass ebenfalls

$$\begin{aligned}
&\mathrm{E}\left[\widehat{\mathrm{F}}_{1-i}^{(l|t)}(\mathrm{X}_{iks}^{(l|t')}) - \mathrm{F}_{1-i}^{(l)}(\mathrm{X}_{iks}^{(l|t')})\right]^2 \leq \frac{1}{n_{1-i}^{(t)}} \quad \forall\, t, t' \in \{u, v\},\ l = 1, \ldots, d \text{ sowie} \\
&\mathrm{E}\left[\widehat{\mathrm{F}}_{1-i}^{(l)}(\mathrm{X}_{iks}^{(l|t')}) - \mathrm{F}_{1-i}^{(l)}(\mathrm{X}_{iks}^{(l|t')})\right]^2 \leq \frac{4}{n_{1-i}} \quad \forall\, t' \in \{u, v\},\ l = 1, \ldots, d
\end{aligned} \tag{4.27}$$

gilt. □

Ebenfalls analog zum gewichteten Fall wird im folgenden Abschnitt die L_2-Konsistenz des aus diesen ungewichteten empirischen Verteilungsfunktionen resultieren AUC-Schätzers gezeigt. Auch an dieser Stelle wird zur Schätzung der AUC ein gewöhnlicher Plug-In-Schätzer verwendet.

Schätzung der AUC

Satz 4.19 *Für alle $l = 1, \ldots, d$ wird die Fläche unter der ROC-Kurve konsistent durch*

$$\widehat{\mathrm{AUC}}^{(l)} = \int \widehat{\mathrm{F}}_0^{(l)} d\widehat{\mathrm{F}}_1^{(l)}$$

geschätzt.

Beweis. Man betrachte analog zu Satz 4.15 (Konsistenz des gewichteten AUC-Schätzers) zunächst $\int \widehat{\mathrm{F}}_0^{(l|t)} d\widehat{\mathrm{F}}_1^{(l|t')}$, $t, t' \in \{u, v\}$. Mit Hilfe der Jensen- und der c_r-Ungleichung gilt:

$$\begin{aligned}
\mathrm{E}\left(\int \widehat{\mathrm{F}}_0^{(l|t)} d\widehat{\mathrm{F}}_1^{(l|t')} - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right)^2 &= \mathrm{E}\left(\int \widehat{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)} d\widehat{\mathrm{F}}_1^{(l|t')} + \int \mathrm{F}_0^{(l)} d(\widehat{\mathrm{F}}_1^{(l|t')} - \mathrm{F}_1^{(l)})\right)^2 \\
&\leq 2\mathrm{E}\left(\int \left(\widehat{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)}\right) d\widehat{\mathrm{F}}_1^{(l|t')}\right)^2 + 2\mathrm{E}\left(\int \mathrm{F}_0^{(l)} d(\widehat{\mathrm{F}}_1^{(l|t')} - \mathrm{F}_1^{(l)})\right)^2 \\
&\leq 2\mathrm{E}\int \left(\widehat{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)}\right)^2 d\widehat{\mathrm{F}}_1^{(l|t')} + 2\mathrm{E}\left(\int \mathrm{F}_0^{(l)} d(\widehat{\mathrm{F}}_1^{(l|t')} - \mathrm{F}_1^{(l)})\right)^2 \\
&\leq \frac{2}{n_1^{(t')}} \sum_{k=1}^{n_1^{(t')}} \frac{1}{m_{1k}^{(l|t')}} \sum_{s=1}^{m_{1k}^{(l|t')}} \mathrm{E}\left(\widehat{\mathrm{F}}_0^{(l|t)}(\mathrm{X}_{1ks}^{(l|t')}) - \mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')})\right)^2 \\
&\quad + \frac{2}{\left(n_1^{(t')}\right)^2} \sum_{k=1}^{n_1^{(t')}} \sum_{k'=1}^{n_1^{(t')}} \frac{1}{m_{1k}^{(l|t')}} \frac{1}{m_{1k'}^{(l|t')}} \sum_{s=1}^{m_{1k}^{(l|t')}} \sum_{s'=1}^{m_{1k'}^{(l|t')}} \mathrm{E}\left(\left[\mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')}) - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right]\left[\mathrm{F}_0^{(l)}(\mathrm{X}_{1k's'}^{(l|t')}) - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right]\right)
\end{aligned}$$

$$\leq \frac{2}{n_1^{(t')}} \sum_{k=1}^{n_1^{(t')}} \frac{1}{m_{1k}^{(l|t')}} \sum_{s=1}^{m_{1k}^{(l|t')}} \frac{1}{n_0^{(t)}} + \frac{2}{\left(n_1^{(t')}\right)^2} \sum_{k=1}^{n_1^{(t')}} \frac{1}{\left(m_{1k'}^{(l|t')}\right)^2} \sum_{s,s'=1}^{m_{1k}^{(l|t')}} \mathrm{E}\left(\mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')}) - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right)\left(\mathrm{F}_0^{(l)}(\mathrm{X}_{1ks'}^{(l|t')}) - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right)$$

$$\leq \frac{2}{n_0^{(t)}} + \frac{2}{\left(n_1^{(t')}\right)^2} \sum_{k=1}^{n_1^{(t')}} \frac{1}{\left(m_{1k'}^{(l|t')}\right)^2} \sum_{s,s'=1}^{m_{1k}^{(l|t')}} 1 = \frac{2}{n_0^{(t)}} + \frac{2}{n_1^{(t')}}.$$

Hierbei wurde verwendet, dass die beiden Terme für verschiedene k unabhängig sind und der Erwartungswert der einzelnen Terme = 0 ist. Die vorletzte Zeile folgt nach der soeben bewiesenen Abschätzung (4.27). Es gilt weiter:

$$\begin{aligned}
\mathrm{E}\left(\widehat{\mathrm{AUC}}^{(l)} - \mathrm{AUC}^{(l)}\right)^2 &= \mathrm{E}\left(\sum_{t\in\{u,v\}} \sum_{t'\in\{u,v\}} \left(\frac{n_0^{(t)}}{n_0}\frac{n_1^{(t')}}{n_1}\right)\left[\int \widehat{\mathrm{F}}_0^{(l|t)} d\widehat{\mathrm{F}}_1^{(l|t')} - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right]\right)^2 \\
&\leq 4 \cdot \sum_{t\in\{u,v\}} \sum_{t'\in\{u,v\}} \left(\frac{n_0^{(t)}}{n_0}\frac{n_1^{(t')}}{n_1}\right)^2 \mathrm{E}\left(\int \widehat{\mathrm{F}}_0^{(l|t)} d\widehat{\mathrm{F}}_1^{(l|t')} - \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)}\right)^2 \\
&\leq 4 \cdot \sum_{t\in\{u,v\}} \sum_{t'\in\{u,v\}} \left(\frac{n_0^{(t)}}{n_0}\frac{n_1^{(t')}}{n_1}\right)^2 \left(\frac{2}{n_0^{(t)}} + \frac{2}{n_1^{(t')}}\right) \\
&= \frac{8}{n_0^2 n_1^2} \sum_{t\in\{u,v\}} \sum_{t'\in\{u,v\}} \underbrace{n_0^{(t)} n_1^{(t')}}_{\leq n_0 n_1} (n_0^{(t)} + n_1^{(t')}) \leq \frac{16(n_1+n_0)}{n_1 n_0} \to 0.
\end{aligned}$$

□

4.2.3 Asymptotische Verteilung des AUC-Schätzers

In diesem Abschnitt wird die asymptotische Verteilung der Statistiken $\sqrt{\mathrm{N}}\left(\widetilde{\mathbf{AUC}} - \mathbf{AUC}\right)$ und $\sqrt{\mathrm{N}}\left(\widehat{\mathbf{AUC}} - \mathbf{AUC}\right)$ hergeleitet. Analog zum Fall der Einfachmessungen bildet die asymptotische Normalität dieser Ausdrücke den Ausgangspunkt der späteren Teststatistiken.
Die Beweise gliedern sich hierbei in zwei Etappen: In einem ersten Schritt wird eine zu $\sqrt{\mathrm{N}}\left(\widetilde{\mathbf{AUC}} - \mathbf{AUC}\right)$ beziehungsweise $\sqrt{\mathrm{N}}\left(\widehat{\mathbf{AUC}} - \mathbf{AUC}\right)$ asymptotisch äquivalente Zufallsvariable bestimmt, deren asymptotische Normalität in einem zweiten Schritt gezeigt wird.
Die Beweisideen lassen sich hierbei auch auf Ein-Punkt-Verteilungen übertragen, sodass die Ergebnisse der AUC-Analyse analog zum Fall der Einfachmessungen auch für Sensitivität und Spezifität gelten. Die Ideen dieser Erweiterung werden in Kürze skizziert. Aus dieser Weiterentwicklung der Resultate auf Ein-Punkt-Verteilung folgt unmittelbar die asymptotische Normalität von

$$\sqrt{\mathrm{N}}\left(\begin{pmatrix}\widetilde{\mathbf{se}}\\ \widetilde{\mathbf{sp}}\end{pmatrix} - \begin{pmatrix}\mathbf{se}\\ \mathbf{sp}\end{pmatrix}\right) \qquad \text{beziehungsweise} \qquad \sqrt{\mathrm{N}}\left(\begin{pmatrix}\widehat{\mathbf{se}}\\ \widehat{\mathbf{sp}}\end{pmatrix} - \begin{pmatrix}\mathbf{se}\\ \mathbf{sp}\end{pmatrix}\right),$$

wodurch sich schließlich – wie auch bei den Einfachmessungen – mit Hilfe des multivariaten δ-Satzes und unter Voraussetzung 4.3/(5) die asymptotische Normalität auch für die Schätzer der prädiktiven Werte zeigen lässt.

Gewichtete Schätzer

Satz 4.20 *Es sei*

$$\widetilde{\mathrm{B}}^{(l)} = \int \mathrm{F}_0^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} - \int \mathrm{F}_1^{(l)} d\widetilde{\mathrm{F}}_0^{(l)} + 1 - 2 \cdot \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)},\ l = 1, \ldots, d$$

und es sei weiter $\widetilde{\mathbf{B}} = (\widetilde{\mathrm{B}}^{(1)}, \ldots, \widetilde{\mathrm{B}}^{(d)})'$*, dann gilt:*

$$\left\| \sqrt{\mathrm{N}} \left(\widetilde{\mathbf{AUC}} - \mathbf{AUC} \right) - \sqrt{\mathrm{N}} \widetilde{\mathbf{B}} \right\|_2 \to \mathbf{0}.$$

BEWEIS. Nach Definition der L_2-Konvergenz kann der Beweis komponentenweise geführt werden, sodass es genügt zu zeigen, dass $\mathrm{E}\left[\sqrt{\mathrm{N}} \left(\widetilde{\mathrm{AUC}}^{(l)} - \mathrm{AUC}^{(l)} \right) - \sqrt{\mathrm{N}} \widetilde{\mathrm{B}}^{(l)} \right]^2 \to 0,\ \forall l = 1, \ldots, d$. Es gilt (beispielsweise nachzulesen bei Brunner u. a., 2002, Beweis zu Theorem 3.3.):

$$\begin{aligned} \sqrt{\mathrm{N}} \left(\widetilde{\mathrm{AUC}}^{(l)} - \mathrm{AUC}^{(l)} \right) &= \sqrt{\mathrm{N}} \widetilde{\mathrm{B}}^{(l)} + \sqrt{\mathrm{N}} \int \left(\widetilde{\mathrm{F}}_0^{(l)} - \mathrm{F}_0^{(l)} \right) d \left(\widetilde{\mathrm{F}}_1^{(l)} - \mathrm{F}_1^{(l)} \right) \\ &= \sqrt{\mathrm{N}} \widetilde{\mathrm{B}}^{(l)} + \sqrt{\mathrm{N}} \widetilde{\mathrm{A}}^{(l)}. \end{aligned}$$

Das heißt, es muss gezeigt werden, dass $\mathrm{E}(\sqrt{\mathrm{N}} \widetilde{\mathrm{A}}^{(l)})^2 \to 0$. Nun gilt mit Hilfe der c_r-Ungleichung:

$$\begin{aligned} \mathrm{E}(\sqrt{\mathrm{N}} \widetilde{\mathrm{A}}^{(l)})^2 &= \mathrm{E} \left[\sqrt{\mathrm{N}} \int \left(\widetilde{\mathrm{F}}_0^{(l)} - \mathrm{F}_0^{(l)} \right) d \left(\widetilde{\mathrm{F}}_1^{(l)} - \mathrm{F}_1^{(l)} \right) \right]^2 \\ &= \mathrm{E} \left[\sqrt{\mathrm{N}} \sum_{t \in \{u,v\}} \sum_{t' \in \{u,v\}} \frac{m_{0\cdot}^{(l|t)}}{m_0^{(l|\cdot)}} \frac{m_{1\cdot}^{(l|t')}}{m_1^{(l|\cdot)}} \int \left(\widetilde{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)} \right) d \left(\widetilde{\mathrm{F}}_1^{(l|t')} - \mathrm{F}_1^{(l)} \right) \right]^2 \\ &\leq 4 \sum_{t \in \{u,v\}} \sum_{t' \in \{u,v\}} \left(\frac{m_{0\cdot}^{(l|t)}}{m_{0\cdot}^{(l|\cdot)}} \frac{m_{1\cdot}^{(l|t')}}{m_{1\cdot}^{(l|\cdot)}} \right)^2 \mathrm{E} \left[\sqrt{\mathrm{N}} \int \left(\widetilde{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)} \right) d \left(\widetilde{\mathrm{F}}_1^{(l|t')} - \mathrm{F}_1^{(l)} \right) \right]^2. \end{aligned} \tag{4.28}$$

Betrachte daher $\mathrm{E} \left[\sqrt{\mathrm{N}} \int \left(\widetilde{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)} \right) d \left(\widetilde{\mathrm{F}}_1^{(l|t')} - \mathrm{F}_1^{(l)} \right) \right]^2 = \mathrm{E} \left[\sqrt{\mathrm{N}} \widetilde{\mathrm{A}}_{tt'}^{(l)} \right]^2$. Es gilt $\forall t, t' \in \{u, v\}$:

$$\begin{aligned} \mathrm{E} \left[\sqrt{\mathrm{N}} \widetilde{\mathrm{A}}_{tt'}^{(l)} \right]^2 &= \mathrm{N} \cdot \mathrm{E} \left[\int \left(\widetilde{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)} \right) d\widetilde{\mathrm{F}}_1^{(l|t')} - \int \left(\widetilde{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)} \right) d\mathrm{F}_1^{(l)} \right]^2 \\ &= \mathrm{N} \cdot \mathrm{E} \left[\frac{1}{m_{1\cdot}^{(l|t')}} \sum_{k=1}^{n_1^{(t')}} \sum_{s=1}^{m_{1k}^{(l|t')}} \left(\widetilde{\mathrm{F}}_0^{(l|t)}(\mathrm{X}_{1ks}^{(l|t')}) - \mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')}) \right) - \int \left(\widetilde{\mathrm{F}}_0^{(l|t)}(x) - \mathrm{F}_0^{(l)}(x) \right) d\mathrm{F}_1^{(l)}(x) \right]^2 \\ &= \mathrm{N} \cdot \mathrm{E} \left[\frac{1}{m_{1\cdot}^{(l|t')}} \sum_{k=1}^{n_1^{(t')}} \sum_{s=1}^{m_{1k}^{(l|t')}} \left(\frac{1}{m_{0\cdot}^{(l|t)}} \sum_{k'=1}^{n_0^{(t)}} \sum_{s'=1}^{m_{0k'}^{(l|t)}} c(\mathrm{X}_{1ks}^{(l|t')} - \mathrm{X}_{0k's'}^{(l|t)}) - \mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')}) \right) \right. \\ &\quad \left. - \left(\frac{1}{m_{0\cdot}^{(l|t)}} \sum_{k'=1}^{n_0^{(t)}} \sum_{s'=1}^{m_{0k'}^{(l|t)}} \int \left(c(x - \mathrm{X}_{0k's'}^{(l|t)}) - \mathrm{F}_0^{(l)}(x) \right) d\mathrm{F}_1^{(l)}(x) \right) \right]^2 \end{aligned}$$

Zur besseren Lesbarkeit bezeichnen im Folgenden

$$\begin{aligned} \varphi_1 \left(\mathrm{X}_{1ks}^{(l|t')}, \mathrm{X}_{0k's'}^{(l|t)} \right) &= c(\mathrm{X}_{1ks}^{(l|t')} - \mathrm{X}_{0k's'}^{(l|t)}) - \mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')}) \\ \varphi_2 \left(\mathrm{X}_{0k's'}^{(l|t)} \right) &= \int c(x - \mathrm{X}_{0k's'}^{(l|t)}) - \mathrm{F}_0^{(l)}(x) d\mathrm{F}_1^{(l)}(x). \end{aligned}$$

Mit dieser Notation gilt:

$$
\begin{aligned}
\mathrm{E}\left[\sqrt{\mathrm{N}}\widetilde{\mathrm{A}}_{tt'}^{(l)}\right]^2 &= \mathrm{N}\cdot\mathrm{E}\left[\frac{1}{m_{1\cdot}^{(l|t')}}\sum_{k=1}^{n_1^{(t')}}\sum_{s=1}^{m_{1k}^{(l|t')}}\frac{1}{m_{0\cdot}^{(l|t)}}\sum_{k'=1}^{n_0^{(t)}}\sum_{s'=1}^{m_{0k'}^{(l|t)}}\varphi_1\left(\mathrm{X}_{1ks}^{(l|t')},\mathrm{X}_{0k's'}^{(l|t)}\right)-\varphi_2\left(\mathrm{X}_{0k's'}^{(l|t)}\right)\right]^2 \\
&= \frac{\mathrm{N}}{\left(m_{0\cdot}^{(l|t)}m_{1\cdot}^{(l|t')}\right)^2}\sum_{k_1=1}^{n_1^{(t')}}\sum_{k_2=1}^{n_1^{(t')}}\sum_{s_1=1}^{m_{1k_1}^{(l|t')}}\sum_{s_2=1}^{m_{1k_2}^{(l|t')}}\sum_{k_1'=1}^{n_0^{(t)}}\sum_{k_2'=1}^{n_0^{(t)}}\sum_{s_1'=1}^{m_{0k_1'}^{(l|t)}}\sum_{s_2'=1}^{m_{0k_2'}^{(l|t)}}\underbrace{\mathrm{E}\left[\prod_{w=1}^{2}\left(\varphi_1\left(\mathrm{X}_{1k_ws_w}^{(l|t')},\mathrm{X}_{0k_w's_w'}^{(l|t)}\right)-\varphi_2\left(\mathrm{X}_{0k_w's_w'}^{(l|t)}\right)\right)\right]}_{=0,\text{ wenn } k_1\neq k_2 \text{ oder } k_1'\neq k_2'} \\
&= \frac{\mathrm{N}}{\left(m_{0\cdot}^{(l|t)}m_{1\cdot}^{(l|t')}\right)^2}\sum_{k_1=1}^{n_1^{(t')}}\sum_{s_1,s_2=1}^{m_{1k_1}^{(l|t')}}\sum_{k_1'=1}^{n_0^{(t)}}\sum_{s_1',s_2'=1}^{m_{0k_1'}^{(l|t)}}\mathrm{E}\Big[\underbrace{\prod_{w=1}^{2}\left(\varphi_1\left(\mathrm{X}_{1k_ws_w}^{(l|t')},\mathrm{X}_{0k_w's_w'}^{(l|t)}\right)-\varphi_2\left(\mathrm{X}_{0k_w's_w'}^{(l|t)}\right)\right)}_{\leq 4}\Big] \\
&\leq \frac{\mathrm{N}}{\left(m_{0\cdot}^{(l|t)}m_{1\cdot}^{(l|t')}\right)^2}\sum_{k_1=1}^{n_1^{(t')}}\sum_{s_1,s_2=1}^{m_{1k_1}^{(l|t')}}\sum_{k_1'=1}^{n_0^{(t)}}\sum_{s_1',s_2'=1}^{m_{0k_1'}^{(l|t)}}4 \leq \frac{4\cdot\mathrm{N}\cdot m_{\max}^2}{m_{0\cdot}^{(l|t)}m_{1\cdot}^{(l|t')}}.
\end{aligned}
$$

Das Einsetzen dieses Resultats in Gleichung (4.28) liefert schließlich:

$$
\begin{aligned}
\mathrm{E}(\sqrt{\mathrm{N}}\widetilde{\mathrm{A}}^{(l)})^2 &\leq 4\sum_{t\in\{u,v\}}\sum_{t'\in\{u,v\}}\left(\frac{m_{0\cdot}^{(l|t)}}{m_{0\cdot}^{(l|\cdot)}}\frac{m_{1\cdot}^{(l|t')}}{m_{1\cdot}^{(l|\cdot)}}\right)^2\mathrm{E}\left[\sqrt{\mathrm{N}}\int\left(\widetilde{\mathrm{F}}_0^{(l|t)}-\mathrm{F}_0^{(l)}\right)d\left(\widetilde{\mathrm{F}}_1^{(l|t')}-\mathrm{F}_1^{(l)}\right)\right]^2 \\
&\leq 4\sum_{t\in\{u,v\}}\sum_{t'\in\{u,v\}}\left(\frac{m_{0\cdot}^{(l|t)}}{m_{0\cdot}^{(l|\cdot)}}\frac{m_{1\cdot}^{(l|t')}}{m_{1\cdot}^{(l|\cdot)}}\right)^2\frac{4\cdot\mathrm{N}\cdot m_{\max}^2}{m_{0\cdot}^{(l|t)}m_{1\cdot}^{(l|t')}} = \frac{16\cdot\mathrm{N}\cdot m_{\max}^2}{\left(m_{0\cdot}^{(l|\cdot)}m_{1\cdot}^{(l|\cdot)}\right)^2}\sum_{t\in\{u,v\}}\sum_{t'\in\{u,v\}}\underbrace{m_{0\cdot}^{(l|t)}\cdot m_{1\cdot}^{(l|t')}}_{\leq m_{0\cdot}^{(l|\cdot)}\cdot m_{1\cdot}^{(l|\cdot)}} \\
&\leq \frac{64\cdot\mathrm{N}\cdot m_{\max}^2}{m_{0\cdot}^{(l|\cdot)}m_{1\cdot}^{(l|\cdot)}} \leq \frac{64\cdot m_{\max}^2(n_0+n_1)}{m_{0\cdot}^{(l|\cdot)}m_{1\cdot}^{(l|\cdot)}} = \frac{64\cdot m_{\max}^2\cdot n_0}{m_{0\cdot}^{(l|\cdot)}m_{1\cdot}^{(l|\cdot)}}+\frac{64\cdot m_{\max}^2\cdot n_1}{m_{0\cdot}^{(l|\cdot)}m_{1\cdot}^{(l|\cdot)}} \\
&\leq \frac{64\cdot m_{\max}^2}{m_{1\cdot}^{(l|\cdot)}}+\frac{64\cdot m_{\max}^2}{m_{0\cdot}^{(l|\cdot)}} \to 0,
\end{aligned}
$$

wobei im letzten Schritt verwendet wurde, dass $m_{i\cdot}^{(l|\cdot)} \geq n_i$. □

Im Folgenden wird nun die asymptotische Verteilung von $\sqrt{\mathrm{N}}\widetilde{\mathbf{B}}$ berechnet. Zu Gunsten einer besseren Lesbarkeit sei zu diesem Zweck die (nicht beobachtbare) Zufallsvariable

$$
\mathrm{Y}_{iks}^{(l|t)} = \mathrm{F}_{1-i}^{(l)}\left(\mathrm{X}_{iks}^{(l|t)}\right), \qquad l=1,\ldots,d,\ t\in\{u,v\},\ i=0,1,\ k=1,\ldots,n_i^{(t)},\ s=1,\ldots,m_{ik}^{(l|t)}, \tag{4.29}
$$

definiert. Diese Zufallsvariable wird in Anlehnung an Brunner und Munzel (2002, Abschnitt 4.4.2) als *asymptotische Rangtransformation* bezeichnet. Es bezeichne weiter $\mathbf{V}_{\widehat{\mathrm{AUC}}}$ die Kovarianzmatrix von $\sqrt{\mathrm{N}}\widetilde{\mathbf{B}}$ und $\lambda_1,\ldots,\lambda_d$ seien die Eigenwerte von $\mathbf{V}_{\widehat{\mathrm{AUC}}}$. Zur Herleitung der asymptotischen Verteilung wird nun eine zusätzliche Regularitätsannahme benötigt:

Voraussetzung 4.4
Es sei $\lambda_{min} = \min\{\lambda_1,\ldots,\lambda_d\}$ das Minimum der Eigenwerte von $\mathbf{V}_{\widehat{\mathrm{AUC}}}$. Für dieses Minimum gelte:
$\lambda_{min} \geq \lambda_0 > 0$, wobei $\lambda_0 > 0$ eine beliebige Konstante sei. [8]

[8] Analog zum Fall der Einfachmessungen (siehe Seite 23) lässt sich diese Voraussetzung über die Einführung einer multivariaten degenerierten Normalverteilung etwas abschwächen (analog zu Kulle, 1999, Kapitel 3.4). Aus Gründen der besseren Lesbarkeit wird jedoch diese stärkere Voraussetzung gefordert.

Satz 4.21 *Unter den Voraussetzungen 4.3 und 4.4 ist die Statistik* $\sqrt{\mathrm{N}}(\widetilde{\mathbf{AUC}} - \mathbf{AUC})$ *asymptotisch normalverteilt mit Erwartungswert* $\mathbf{0}$ *und Kovarianzmatrix* $\mathbf{V}_{\widetilde{\mathrm{AUC}}}$.

BEWEIS. Um die asymptotische Normalität von $\sqrt{\mathrm{N}}(\widetilde{\mathbf{AUC}} - \mathbf{AUC})$ nachzuweisen, genügt es nach dem soeben bewiesenen Satz 4.20 zu zeigen, dass $\sqrt{\mathrm{N}}\widetilde{\mathbf{B}}$ asymptotisch normalverteilt ist. Es gilt

$$
\begin{aligned}
\widetilde{\mathrm{B}}^{(l)} &= \int \mathrm{F}_0^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} - \int \mathrm{F}_1^{(l)} d\widetilde{\mathrm{F}}_0^{(l)} + 1 - 2 \cdot \mathrm{AUC}^{(l)} \\
&= \frac{m_{1\cdot}^{(l|v)}}{m_{1\cdot}^{(l|\cdot)}} \left(\frac{1}{m_{1\cdot}^{(l|v)}} \sum_{k=1}^{n^{(v)}} \sum_{s=1}^{m_{1k}^{(l|v)}} \mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|v)}) \right) + \frac{m_{1\cdot}^{(l|u)}}{m_{1\cdot}^{(l|\cdot)}} \left(\frac{1}{m_{1\cdot}^{(l|u)}} \sum_{k=1}^{n_1^{(u)}} \sum_{s=1}^{m_{1k}^{(l|u)}} \mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|u)}) \right) \\
&\quad - \frac{m_{0\cdot}^{(l|v)}}{m_{0\cdot}^{(l|\cdot)}} \left(\frac{1}{m_{0\cdot}^{(l|v)}} \sum_{k=1}^{n^{(v)}} \sum_{s=1}^{m_{0k}^{(l|v)}} \mathrm{F}_1^{(l)}(\mathrm{X}_{0ks}^{(l|v)}) \right) - \frac{m_{0\cdot}^{(l|u)}}{m_{0\cdot}^{(l|\cdot)}} \left(\frac{1}{m_{0\cdot}^{(l|u)}} \sum_{k=1}^{n_0^{(u)}} \sum_{s=1}^{m_{0k}^{(l|u)}} \mathrm{F}_1^{(l)}(\mathrm{X}_{0ks}^{(l|u)}) \right) + 1 - 2 \cdot \mathrm{AUC}^{(l)} \\
&= \sum_{k=1}^{n^{(v)}} \left(\frac{1}{m_{1\cdot}^{(l|\cdot)}} \sum_{s=1}^{m_{1k}^{(l|v)}} \mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|v)}) - \frac{1}{m_{0\cdot}^{(l|\cdot)}} \sum_{s=1}^{m_{0k}^{(l|v)}} \mathrm{F}_1^{(l)}(\mathrm{X}_{0ks}^{(l|v)}) \right) \\
&\quad + \sum_{k=1}^{n_1^{(u)}} \left(\frac{1}{m_{1\cdot}^{(l|\cdot)}} \sum_{s=1}^{m_{1k}^{(l|u)}} \mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|u)}) \right) - \sum_{k=1}^{n_0^{(u)}} \left(\frac{1}{m_{0\cdot}^{(l|\cdot)}} \sum_{s=1}^{m_{0k}^{(l|u)}} \mathrm{F}_1^{(l)}(\mathrm{X}_{0ks}^{(l|u)}) \right) + 1 - 2 \cdot \mathrm{AUC}^{(l)} \\
&= \frac{1}{n^{(v)}} \sum_{k=1}^{n^{(v)}} \left(\frac{n^{(v)}}{m_{1\cdot}^{(l|\cdot)}} \sum_{s=1}^{m_{1k}^{(l|v)}} \mathrm{Y}_{1ks}^{(l|v)} - \frac{n^{(v)}}{m_{0\cdot}^{(l|\cdot)}} \sum_{s=1}^{m_{0k}^{(l|v)}} \mathrm{Y}_{0ks}^{(l|v)} \right) \\
&\quad + \frac{1}{n_1^{(u)}} \sum_{k=1}^{n_1^{(u)}} \left(\frac{n_1^{(u)}}{m_{1\cdot}^{(l|\cdot)}} \sum_{s=1}^{m_{1k}^{(l|u)}} \mathrm{Y}_{1ks}^{(l|u)} \right) - \frac{1}{n_0^{(u)}} \sum_{k=1}^{n_0^{(u)}} \left(\frac{n_0^{(u)}}{m_{0\cdot}^{(l|\cdot)}} \sum_{s=1}^{m_{0k}^{(l|u)}} \mathrm{Y}_{0ks}^{(l|u)} \right) + 1 - 2 \cdot \mathrm{AUC}^{(l)} \\
&= \frac{1}{n^{(v)}} \sum_{k=1}^{n^{(v)}} \left(\frac{n^{(v)}}{m_{1\cdot}^{(l|\cdot)}} \mathrm{Y}_{1k\cdot}^{(l|v)} - \frac{n^{(v)}}{m_{0\cdot}^{(l|\cdot)}} \mathrm{Y}_{0k\cdot}^{(l|v)} \right) + \frac{1}{n_1^{(u)}} \sum_{k=1}^{n_1^{(u)}} \left(\frac{n_1^{(u)}}{m_{1\cdot}^{(l|\cdot)}} \mathrm{Y}_{1k\cdot}^{(l|u)} \right) - \frac{1}{n_0^{(u)}} \sum_{k=1}^{n_0^{(u)}} \left(\frac{n_0^{(u)}}{m_{0\cdot}^{(l|\cdot)}} \mathrm{Y}_{0k\cdot}^{(l|u)} \right) + 1 - 2 \cdot \mathrm{AUC}^{(l)}.
\end{aligned}
\tag{4.30}
$$

$\widetilde{\mathrm{B}}^{(l)}$ ist damit – abgesehen von einer Konstanten – die Summe der Mittelwerte der unabhängigen Zufallsvariablen $\frac{n^{(v)}}{m_{1\cdot}^{(l|\cdot)}} \cdot \mathrm{Y}_{1k\cdot}^{(l|v)} - \frac{n^{(v)}}{m_{0\cdot}^{(l|\cdot)}} \cdot \mathrm{Y}_{0k\cdot}^{(l|v)}$, sowie $\frac{n_1^{(u)}}{m_{1\cdot}^{(l|\cdot)}} \cdot \mathrm{Y}_{1k\cdot}^{(l|u)}$ und $\frac{n_0^{(u)}}{m_{0\cdot}^{(l|\cdot)}} \cdot \mathrm{Y}_{0k\cdot}^{(l|u)}$. Da stets $\mathrm{Y}_{iks}^{(l|t)} \leq 1$ und außerdem $\frac{n_i^{(t)}}{m_{i\cdot}^{(l|\cdot)}} \leq 1$ gilt, sind die einzelnen Summanden nach Voraussetzung durch $m_{\max}$ gleichmäßig beschränkt ($\forall\, l = 1, \ldots, d$, $t \in \{u, v\}$, $i = 0, 1$, $k = 1, \ldots, n_i^{(t)}$, $s = 1, \ldots, m_{ik}^{(l|t)}$). Die asymptotische Normalität von $\sqrt{\mathrm{N}}\widetilde{\mathrm{B}}^{(l)}$ folgt daher unter den Voraussetzungen 4.3 und 4.4 aus dem zentralen Grenzwertsatz nach Lindeberg (Lindeberg, 1922). Die multivariate Normalität des Vektors $\sqrt{\mathrm{N}}\widetilde{\mathbf{B}}$ lässt sich schließlich kanonisch mit der Cramer-Wold-Technik nachweisen. □

Dieser Beweis lässt sich in kanonischer Art und Weise auch auf Ein-Punkt-Verteilungen übertragen. In diesem Fall wird der erste Beweisschritt, in welchem die asymptotische Äquivalenz nachgewiesen wurde, nicht benötigt, da sich im Fall der Ein-Punkt-Verteilungen bereits $\widetilde{\mathbf{se}}$ und $\widetilde{\mathbf{sp}}$ als Mittelwerte unabhängiger und gleichmäßig beschränkter Zufallsvariablen darstellen lassen:

$$
\widetilde{se}^{(l)} = \int \widetilde{\Gamma}^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} = \int \Gamma^{(l)} d\widetilde{\mathrm{F}}_1^{(l)} = \frac{1}{m_{1\cdot}^{(l|\cdot)}} \sum_{t \in \{u,v\}} \sum_{k=1}^{n_1^{(t)}} \sum_{s=1}^{m_{1k}^{(l|t)}} \underbrace{\Gamma^{(l)}(\mathrm{X}_{1ks}^{(l|t)})}_{=:\mathrm{C}_{1ks}^{(l|t)}} = \frac{1}{n_1} \sum_{t \in \{u,v\}} \sum_{k=1}^{n_1^{(t)}} \frac{n_1}{m_{1\cdot}^{(l|\cdot)}} \mathrm{C}_{1k\cdot}^{(l|t)}.
$$

Analoges gilt für die Spezifität. Die asymptotische Normalität von $\sqrt{n_0}(\widetilde{\mathbf{sp}} - \mathbf{sp})$ und $\sqrt{n_1}(\widetilde{\mathbf{se}} - \mathbf{se})$ folgt daher

wie im Beweis des letzten Satzes direkt aus dem Lindeberg'schen zentralen Grenzwertsatz.[9] Die asymptotische Normalität von $\sqrt{N}(\widetilde{\mathbf{sp}} - \mathbf{sp})$ und $\sqrt{N}(\widetilde{\mathbf{se}} - \mathbf{se})$ folgt schließlich analog zum Fall der Einfachmessungen (siehe Korollar 4.1, Seite 24). Aus dieser Erweiterung der Resultate auf Ein-Punkt-Verteilung ergibt sich gleichsam die asymptotische Normalität von $\sqrt{N}\left[\left(\widetilde{\mathbf{se}}', \widetilde{\mathbf{sp}}'\right)' - \left(\mathbf{se}', \mathbf{sp}'\right)'\right]$, woraus sich – analog zu den Einfachmessungen – unter Zuhilfenahme des multivariaten δ-Satzes die asymptotische Normalität für die Schätzer der prädiktiven Werte zeigen lässt.

Ungewichtete Schätzer

Die Beweisideen und Resultate im Falle der ungewichteten Schätzer gleichen denen der gewichteten Schätzer.

Satz 4.22 *Es sei*

$$\widehat{\mathrm{B}}^{(l)} = \int \mathrm{F}_0^{(l)} d\widehat{\mathrm{F}}_1^{(l)} - \int \mathrm{F}_1^{(l)} d\widehat{\mathrm{F}}_0^{(l)} + 1 - 2 \cdot \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)},\ l = 1, \ldots, d$$

und es sei $\widehat{\mathbf{B}} = (\widehat{\mathrm{B}}^{(1)}, \ldots, \widehat{\mathrm{B}}^{(d)})'$*, dann gilt:*

$$\left\| \sqrt{N}\left(\widehat{\mathbf{AUC}} - \mathbf{AUC}\right) - \sqrt{N}\widehat{\mathbf{B}} \right\|_2 \to \mathbf{0}.$$

BEWEIS. Der Beweis lässt sich analog zum gewichteten Fall führen und ist im Anhang B.1.1 (Seite 87) nachzulesen. □

Es bezeichne nun $\mathbf{V}_{\widehat{\mathrm{AUC}}}$ die Kovarianzmatrix von $\sqrt{N}\widehat{\mathbf{B}}$ und $\tau_1, \ldots, \tau_d$ seien die Eigenwerte von $\mathbf{V}_{\widehat{\mathrm{AUC}}}$. Zur Herleitung der asymptotischen Verteilung wird nun eine zusätzliche Regularitätsannahme benötigt:

Voraussetzung 4.5
Es sei $\tau_{min} = \min\{\tau_1, \ldots, \tau_d\}$ das Minimum der Eigenwerte von $\mathbf{V}_{\widehat{\mathrm{AUC}}}$. Für dieses Minimum gelte:
$\tau_{min} \geq \tau_0 > 0$, wobei $\tau_0 > 0$ eine beliebige Konstante sei.[10]

Satz 4.23 *Unter den Voraussetzungen 4.3 und 4.5 ist die Statistik* $\sqrt{N}(\widehat{\mathbf{AUC}} - \mathbf{AUC})$ *asymptotisch normalverteilt mit Erwartungswert* $\mathbf{0}$ *und Kovarianzmatrix* $\mathbf{V}_{\widehat{\mathrm{AUC}}}$.

BEWEIS. Der Beweis befindet sich im Anhang B.1.2 (Seite 89). □

Analog zum gewichteten Fall lässt sich auch dieses Resultat kanonisch auf Ein-Punkt-Verteilungen erweitern, sodass die obigen Ergebnisse auch für Sensitivität und Spezifität gelten und sich daher – wie schon im gewichteten Fall – auf die prädiktiven Werte übertragen lassen.

4.2.4 Schätzung der Kovarianzmatrix des AUC-Schätzers

Gewichtete Schätzung

Analog zum Falle der Einfachmessungen wird bei der Schätzung der Kovarianzmatrix von $\sqrt{N}\left(\widehat{\mathbf{AUC}} - \mathbf{AUC}\right)$ die Voraussetzung, dass weder $\mathbf{F}_0$ noch $\mathbf{F}_1$ in einer ihrer Komponenten Ein-Punkt-Verteilungen sind, nicht

[9] An dieser Stelle muss keine zusätzliche Voraussetzung gestellt werden, da eine Voraussetzung 4.4 entsprechende Annahme bereits automatisch durch Voraussetzung 4.3/(6) (Seite 35) erfüllt ist.

[10] Wie auch schon im gewichteten Fall (siehe Voraussetzung 4.4) gibt es auch im ungewichteten Fall analoge Möglichkeiten, diese Voraussetzung etwas abzuschwächen.

benötigt, sodass die hier präsentierten Resultate uneingeschränkt und ohne zusätzlichen Beweis auch für Sensitivität, Spezifität sowie den Vektor $\left(\mathbf{se}', \mathbf{sp}'\right)'$ gelten. Der Kovarianzschätzer der Statistik $\sqrt{\mathrm{N}}\left(\widetilde{\mathbf{p}}_{\pm}^{g} - \mathbf{p}_{\pm}^{g}\right)$ ergibt sich schließlich – wie bei den Einfachmessungen – unter Anwendung des multivariaten δ-Satzes aus der Schätzung von $\mathrm{Var}(\widehat{\pi}_g)$, $g = 1, \ldots, \mathrm{G}$, und aus dem in diesem Abschnitt präsentierten Schätzer von $\mathrm{Cov}(\sqrt{\mathrm{N}}\left(\widetilde{\mathbf{se}}', \widetilde{\mathbf{sp}}'\right)')$.

Satz 4.24 *Es seien*

$$\begin{aligned}
\mathbf{M}_{i\cdot}^{(\cdot)} &= diag(m_{i\cdot}^{(1|\cdot)}, \ldots, m_{i\cdot}^{(d|\cdot)}) \in \mathbb{R}^{d\times d}, && i = 0, 1 \\
\mathbf{M}_{i\cdot}^{(t)} &= diag(m_{i\cdot}^{(1|t)}, \ldots, m_{i\cdot}^{(d|t)}) \in \mathbb{R}^{d\times d}, && t \in \{u, v\},\ i = 0, 1 \text{ und} \\
\mathbf{M}_{ik}^{(t)} &= diag(m_{ik}^{(1|t)}, \ldots, m_{ik}^{(d|t)}) \in \mathbb{R}^{d\times d}, && t \in \{u, v\},\ i = 0, 1,\ k = 1, \ldots n_i^{(t)}
\end{aligned}$$

Diagonalmatrizen, welche die Gesamtanzahlen an gesunden ($i = 0$) beziehungsweise kranken ($i = 1$) Beobachtungseinheiten der Komponenten $1, \ldots, d$ darstellen. Es seien in Anlehnung an die Definition der asymptotischen Rangtransformationen in Gleichung (4.29) *die beobachtbaren gewichteten Rangtransformationen*

$$\widetilde{\mathrm{Y}}_{iks}^{(l|t)} = \widetilde{\mathrm{F}}_{1-i}^{(l)}\left(\mathrm{X}_{iks}^{(l|t)}\right), \qquad l = 1, \ldots, d,\ t \in \{u, v\},\ i = 0, 1,\ k = 1, \ldots, n_i^{(t)},\ s = 1, \ldots, m_{ik}^{(l|t)}$$

definiert. Die Kovarianzmatrix $\mathbf{V}_{\widetilde{\mathrm{AUC}}}$ *von* $\sqrt{\mathrm{N}}(\widetilde{\mathbf{AUC}} - \mathbf{AUC})$ *wird dann durch* $\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}}} = \widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}}}^{(v)} + \widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}},0}^{(u)} + \widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}},1}^{(u)}$ *konsistent geschätzt, wobei*

$$\begin{aligned}
\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}}}^{(v)} = \frac{\mathrm{N}n^{(v)}}{n^{(v)} - 1} \sum_{k=1}^{n^{(v)}} &\left(\left(\mathbf{M}_{1\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{1k\cdot}^{(v)} - \left(\mathbf{M}_{0\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{0k\cdot}^{(v)} - \left[\mathbf{M}_{1k}^{(v)}\left(\mathbf{M}_{1\cdot}^{(v)}\mathbf{M}_{1\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{1\cdot\cdot}^{(v)} - \mathbf{M}_{0k}^{(v)}\left(\mathbf{M}_{0\cdot}^{(v)}\mathbf{M}_{0\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{0\cdot\cdot}^{(v)}\right]\right) \\
&\cdot\left(\left(\mathbf{M}_{1\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{1k\cdot}^{(v)} - \left(\mathbf{M}_{0\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{0k\cdot}^{(v)} - \left[\mathbf{M}_{1k}^{(v)}\left(\mathbf{M}_{1\cdot}^{(v)}\mathbf{M}_{1\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{1\cdot\cdot}^{(v)} - \mathbf{M}_{0k}^{(v)}\left(\mathbf{M}_{0\cdot}^{(v)}\mathbf{M}_{0\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{0\cdot\cdot}^{(v)}\right]\right)', \\
\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}},i}^{(u)} = \frac{\mathrm{N}n_i^{(u)}}{n_i^{(u)} - 1} \sum_{k=1}^{n_i^{(u)}} &\left(\left(\mathbf{M}_{i\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{ik\cdot}^{(u)} - \mathbf{M}_{ik}^{(u)}\left(\mathbf{M}_{i\cdot}^{(u)}\mathbf{M}_{i\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{i\cdot\cdot}^{(u)}\right)\left(\left(\mathbf{M}_{i\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{ik\cdot}^{(u)} - \mathbf{M}_{ik}^{(u)}\left(\mathbf{M}_{i\cdot}^{(u)}\mathbf{M}_{i\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{i\cdot\cdot}^{(u)}\right)'.
\end{aligned}$$

BEWEIS. Der Beweis ist sehr technisch. Er befindet sich daher im Anhang B.2 auf Seite 90. □

Es soll nun eine Rangdarstellung für den Schätzer der Kovarianzmatrix hergeleitet werden. Hierfür wird zunächst folgendes Lemma benötigt:

Lemma 4.2 *Es bezeichne* $\mathrm{R}_{iks}^{(l|t)}$ *den Globalrang von* $\mathrm{X}_{iks}^{(l|t)}$*, das heißt den Rang von* $\mathrm{X}_{iks}^{(l|t)}$ *unter allen Beobachtungen der l-ten Komponente und es sei* $\mathrm{Q}_{iks}^{(l|t)}$ *der Internrang von* $\mathrm{X}_{iks}^{(l|t)}$ *(der Rang von* $\mathrm{X}_{iks}^{(l|t)}$ *unter allen Beobachtungen der l-ten Komponente im Zustand i), dann gilt:*

$$\widetilde{\mathrm{Y}}_{iks}^{(l|t)} = \widetilde{\mathrm{F}}_{1-i}^{(l)}(\mathrm{X}_{iks}^{(l|t)}) = \frac{1}{m_{(1-i)\cdot}^{(l|\cdot)}}\left(\mathrm{R}_{iks}^{(l|t)} - \mathrm{Q}_{iks}^{(l|t)}\right) \tag{4.31}$$

BEWEIS. Es gilt:

$$\begin{aligned}
&& \widetilde{\mathrm{H}}^{(l)}(\mathrm{X}_{iks}^{(l|t)}) &= \frac{1}{\mathrm{M}^{(l)}}\left(m_{(1-i)\cdot}^{(l|\cdot)}\widetilde{\mathrm{F}}_{1-i}^{(l)}(\mathrm{X}_{iks}^{(l|t)}) + m_{i\cdot}^{(l|\cdot)}\widetilde{\mathrm{F}}_{i}^{(l)}(\mathrm{X}_{iks}^{(l|t)})\right) \\
\Rightarrow && \widetilde{\mathrm{F}}_{1-i}^{(l)}(\mathrm{X}_{iks}^{(l|t)}) &= \frac{\mathrm{M}^{(l)}\widetilde{\mathrm{H}}^{(l)}(\mathrm{X}_{iks}^{(l|t)}) - m_{i\cdot}^{(l|\cdot)}\widetilde{\mathrm{F}}_{i}^{(l)}(\mathrm{X}_{iks}^{(l|t)})}{m_{(1-i)\cdot}^{(l|\cdot)}}
\end{aligned} \tag{4.32}$$

Nun gilt nach Lemma 4.1 (Seite 40):

$$\widetilde{\mathrm{H}}^{(l)}(\mathrm{X}_{iks}^{(l|t)}) = \frac{1}{\mathrm{M}^{(l)}}\left(\mathrm{R}_{iks}^{(l|t)} - \frac{1}{2}\right).$$

Analog zu Lemma 4.1 zeigt man, dass ebenfalls

$$\widetilde{\mathrm{F}}_i^{(l)}(\mathrm{X}_{iks}^{(l|t)}) = \frac{1}{m_{i\cdot}^{(l|\cdot)}}\left(\mathrm{Q}_{iks}^{(l|t)} - \frac{1}{2}\right)$$

gilt. Das Einsetzen dieser Resultate in Gleichung (4.32) vervollständigt schließlich den Beweis. □

Korollar 4.3 (Rangdarstellung des Schätzers der Kovarianzmatrix) *Es sei* $\mathrm{R}_{iks}^{(l|t)}$ *wieder der Globalrang von* $\mathrm{X}_{iks}^{(l|t)}$ *sowie* $\mathrm{Q}_{iks}^{(l|t)}$ *der Internrang von* $\mathrm{X}_{iks}^{(l|t)}$. *Es seien* $\mathbf{R}_{ik\cdot}^{(t)} = (\mathrm{R}_{ik\cdot}^{(1|t)}, \ldots, \mathrm{R}_{ik\cdot}^{(d|t)})'$ *und* $\mathbf{Q}_{ik\cdot}^{(t)} = (\mathrm{Q}_{ik\cdot}^{(1|t)}, \ldots, \mathrm{Q}_{ik\cdot}^{(d|t)})'$ *die zugehörigen Vektoren der Rangsummen des* k*-ten Patienten. Es seien weiter* $\mathbf{D}_{ik}^{(t)} = \mathbf{R}_{ik\cdot}^{(t)} - \mathbf{Q}_{ik\cdot}^{(t)}$ *und* $\mathbf{D}_{i\cdot}^{(t)} = \sum_{k=1}^{n_i^{(t)}} \mathbf{D}_{ik}^{(t)}$, *dann gilt:*

$$\begin{aligned}
\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}}}^{(v)} &= \frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\left(\mathbf{M}_0^{-1}\mathbf{M}_1^{-1}\right)^2 \sum_{k=1}^{n^{(v)}} \left(\mathbf{D}_{1k}^{(v)} - \mathbf{D}_{0k}^{(v)} - \left[\mathbf{M}_{1k}^{(v)}\left(\mathbf{M}_{1\cdot}^{(v)}\right)^{-1}\mathbf{D}_{1\cdot}^{(v)} - \mathbf{M}_{0k}^{(v)}\left(\mathbf{M}_{0\cdot}^{(v)}\right)^{-1}\mathbf{D}_{0\cdot}^{(v)}\right]\right) \\
&\qquad \cdot \left(\mathbf{D}_{1k}^{(v)} - \mathbf{D}_{0k}^{(v)} - \left[\mathbf{M}_{1k}^{(v)}\left(\mathbf{M}_{1\cdot}^{(v)}\right)^{-1}\mathbf{D}_{1\cdot}^{(v)} - \mathbf{M}_{0k}^{(v)}\left(\mathbf{M}_{0\cdot}^{(v)}\right)^{-1}\mathbf{D}_{0\cdot}^{(v)}\right]\right)' \\
\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}},i}^{(u)} &= \frac{\mathrm{N}n_i^{(u)}}{n_i^{(u)}-1}\left(\mathbf{M}_0^{-1}\mathbf{M}_1^{-1}\right)^2 \sum_{k=1}^{n_i^{(u)}} \left(\mathbf{D}_{ik}^{(u)} - \mathbf{M}_{ik}^{(u)}\left(\mathbf{M}_{i\cdot}^{(u)}\right)^{-1}\mathbf{D}_{i\cdot}^{(u)}\right)\left(\mathbf{D}_{ik}^{(u)} - \mathbf{M}_{ik}^{(u)}\left(\mathbf{M}_{i\cdot}^{(u)}\right)^{-1}\mathbf{D}_{i\cdot}^{(u)}\right)'.
\end{aligned}$$

Beweis. Das Resultat folgt durch Einsetzten des obigen Resultates (4.31) in die in Satz 4.24 angegebenen Varianzschätzer. □

Es sei an dieser Stelle angemerkt, dass sich die Kovarianzschätzer im Falle von $m_{ik}^{(1|t)} = \ldots = m_{ik}^{(d|t)} =: m_{ik}^{(t)}$, $\forall\, t \in \{u, v\},\ i = 0,1,\ k = 1, \ldots, n_i^{(t)}$, das heißt, in dem Fall, dass für jeden Patient unter den verschiedenen Bedingungen gleich viele Beobachtungen im Zustand $i = 0,1$ vorliegen, vereinfacht durch

$$\begin{aligned}
\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}}}^{(v)} &= \frac{\mathrm{N}n^{(v)}}{(n^{(v)}-1)(m_{0\cdot}^{(\cdot)}m_{1\cdot}^{(\cdot)})^2} \sum_{k=1}^{n^{(v)}} \left(\mathbf{D}_{1k}^{(v)} - \mathbf{D}_{0k}^{(v)} - \frac{m_{1k}^{(v)}}{m_{1\cdot}^{(v)}}\mathbf{D}_{1\cdot}^{(v)} + \frac{m_{0k}^{(v)}}{m_{0\cdot}^{(v)}}\mathbf{D}_{0\cdot}^{(v)}\right)\left(\mathbf{D}_{1k}^{(v)} - \mathbf{D}_{0k}^{(v)} - \frac{m_{1k}^{(v)}}{m_{1\cdot}^{(v)}}\mathbf{D}_{1\cdot}^{(v)} + \frac{m_{0k}^{(v)}}{m_{0\cdot}^{(v)}}\mathbf{D}_{0\cdot}^{(v)}\right)', \\
\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}},i}^{(u)} &= \frac{\mathrm{N}n_i^{(u)}}{(n_i^{(u)}-1)(m_{0\cdot}^{(\cdot)}m_{1\cdot}^{(\cdot)})^2} \sum_{k=1}^{n_i^{(u)}} \left(\mathbf{D}_{ik}^{(u)} - \frac{m_{ik}^{(u)}}{m_{i\cdot}^{(u)}}\mathbf{D}_{i\cdot}^{(u)}\right)\left(\mathbf{D}_{ik}^{(u)} - \frac{m_{ik}^{(u)}}{m_{i\cdot}^{(u)}}\mathbf{D}_{i\cdot}^{(u)}\right)'.
\end{aligned}$$

darstellen lässt.

Ungewichtete Schätzer

In zum gewichteten Fall ähnlicher Weise lässt sich der ungewichtete Schätzer der Kovarianzmatrix herleiten.

Satz 4.25 *Es seien*

$$\widehat{\mathrm{Y}}_{iks}^{(l|t)} = \widehat{\mathrm{F}}_{1-i}\left(\mathrm{X}_{iks}^{(l|t)}\right), \qquad l = 1, \ldots, d,\ t \in \{u, v\},\ i = 0,1,\ k = 1, \ldots, n_i^{(t)},\ s = 1, \ldots, m_{ik}^{(l|t)}$$

die ungewichteten Rangtransformationen[11] *und*

$$\begin{aligned}\overline{\widehat{\mathrm{Y}}}_{ik\cdot}^{(l|t)} &= \tfrac{1}{m_{ik}^{(l|t)}}\textstyle\sum_{s=1}^{m_{ik}^{(l|t)}}\widehat{\mathrm{Y}}_{iks}^{(l|t)}, \qquad l=1,\ldots,d,\ t\in\{u,v\},\ i=0,1,\ k=1,\ldots,n_i^{(t)}\\ \overline{\widehat{\mathrm{Y}}}_{i\cdot\cdot}^{(l|t)} &= \tfrac{1}{n_i^{(t)}}\textstyle\sum_{k=1}^{n_i^{(t)}}\tfrac{1}{m_{ik}^{(l|t)}}\sum_{s=1}^{m_{ik}^{(l|t)}}\widehat{\mathrm{Y}}_{iks}^{(l|t)}, \quad l=1,\ldots,d,\ t\in\{u,v\},\ i=0,1\end{aligned}$$

die ungewichteten Mittelwerte der ungewichteten Rangtransformationen. Die Kovarianzmatrix $\mathbf{V}_{\widehat{\mathrm{AUC}}}$ *von* $\sqrt{\mathrm{N}}(\widehat{\mathbf{AUC}}-\mathbf{AUC})$ *wird durch* $\widehat{\mathbf{V}}_{\widehat{\mathrm{AUC}}}=\widehat{\mathbf{V}}_{\widehat{\mathrm{AUC}}}^{(v)}+\widehat{\mathbf{V}}_{\widehat{\mathrm{AUC}},0}^{(u)}+\widehat{\mathbf{V}}_{\widehat{\mathrm{AUC}},1}^{(u)}$ *konsistent geschätzt, wobei*

$$\widehat{\mathbf{V}}_{\widehat{\mathrm{AUC}},i}^{(u)}=\frac{\mathrm{N}}{n_i^{(u)}(n_i^{(u)}-1)}\left(\frac{n_i^{(u)}}{n_i}\right)^2\sum_{k=1}^{n_i^{(u)}}\left(\overline{\widehat{\mathbf{Y}}}_{ik\cdot}^{(u)}-\overline{\widehat{\mathbf{Y}}}_{i\cdot\cdot}^{(u)}\right)\left(\overline{\widehat{\mathbf{Y}}}_{ik\cdot}^{u}-\overline{\widehat{\mathbf{Y}}}_{i\cdot\cdot}^{(u)}\right)'\ \textit{und}$$

$$\begin{aligned}\widehat{\mathbf{V}}_{\widehat{\mathrm{AUC}}}^{(v)}=\frac{\mathrm{N}}{n^{(v)}(n^{(v)}-1)}\sum_{k=1}^{n^{(v)}}&\left(\left[\frac{n^{(v)}}{n_1}\overline{\widehat{\mathbf{Y}}}_{1k\cdot}^{(v)}-\frac{n^{(v)}}{n_0}\overline{\widehat{\mathbf{Y}}}_{0k\cdot}^{(v)}\right]-\left[\frac{n^{(v)}}{n_1}\overline{\widehat{\mathbf{Y}}}_{1\cdot\cdot}^{(v)}-\frac{n^{(v)}}{n_0}\overline{\widehat{\mathbf{Y}}}_{0\cdot\cdot}^{(v)}\right]\right)\\ &\cdot\left(\left[\frac{n^{(v)}}{n_1}\overline{\widehat{\mathbf{Y}}}_{1k\cdot}^{(v)}-\frac{n^{(v)}}{n_0}\overline{\widehat{\mathbf{Y}}}_{0k\cdot}^{(v)}\right]-\left[\frac{n^{(v)}}{n_1}\overline{\widehat{\mathbf{Y}}}_{1\cdot\cdot}^{(v)}-\frac{n^{(v)}}{n_0}\overline{\widehat{\mathbf{Y}}}_{0\cdot\cdot}^{(v)}\right]\right)'.\end{aligned}$$

BEWEIS. Der Beweis ist, wie auch im gewichteten Fall, sehr technisch. Die Methodik der Beweisführung ist analog; daher wird in dieser Arbeit auf die Herleitung der Konsistenz des ungewichteten Varianzschätzers verzichtet. Die Beweisideen können jedoch am Beispiel des gewichteten Schätzers (siehe Anhang B.2 auf Seite 90) leicht nachvollzogen werden. □

4.2.5 Teststatistiken und Konfidenzintervalle

Wie auch im Fall der Einfachmessungen lässt sich auch bei den Clusterdaten für alle betrachteten diagnostischen Gütemaße eine einheitliche Darstellung von Konfidenzintervallen und Teststatistiken finden. Diese beruht auf der Tatsache, dass für alle Gütemaße $\mathbf{e}\in\{\mathbf{AUC},\mathbf{se},\mathbf{sp},\mathbf{p}_-^g,\mathbf{p}_+^g,\ g=1,\ldots,\mathrm{G}\}$ gilt:

$$\sqrt{\mathrm{N}}\,(\widehat{\mathbf{e}}-\mathbf{e})\ \dot{\sim}\ \mathrm{N}(\mathbf{0},\widehat{\mathbf{V}}_{\widehat{e}}) \qquad \text{beziehungsweise} \qquad \sqrt{\mathrm{N}}\,(\widetilde{\mathbf{e}}-\mathbf{e})\ \dot{\sim}\ \mathrm{N}(\mathbf{0},\widetilde{\mathbf{V}}_{\widetilde{e}}).$$

Das Vorgehen zur Berechnung von Konfidenzintervallen und Teststatistiken ist hierbei analog zum Fall der Einfachmessungen, sodass an dieser Stelle nur die Resultate in Kürze und ohne Beweise dargestellt werden. Die Darstellung erfolgt aus Gründen einer besseren Lesbarkeit dabei lediglich für die gewichteten Schätzer, entsprechende Resultate gelten aber in gleicher Form auch für die ungewichteten Schätzer.

Teststatistiken

Satz 4.26 (Lineares Modell) *Es sei* $\mathbf{e}$ *ein Vektor diagnostischer Gütemaße, welcher durch* $\widetilde{\mathbf{e}}$ *konsistent geschätzt wird. Es gelte (in der üblichen Notation)* $\sqrt{\mathrm{N}}\,(\widetilde{\mathbf{e}}-\mathbf{e})\ \dot{\sim}\ \mathrm{N}\left(\mathbf{0},\widetilde{\mathbf{V}}_{\widetilde{e}}\right)$*, dann gilt unter* $\mathrm{H}_0:\mathbf{Ce}=\mathbf{0}$*:*

$$\textbf{\textit{ANOVA-Typ-Statistik}} \qquad \mathrm{Q}^{\mathrm{ATS}}(\mathbf{T})=\mathrm{N}\frac{\mathrm{Sp}(\mathbf{T}\widetilde{\mathbf{V}}_{\widetilde{e}})}{\mathrm{Sp}(\mathbf{T}\widetilde{\mathbf{V}}_{\widetilde{e}}\mathbf{T}\widetilde{\mathbf{V}}_{\widetilde{e}})}\,\widetilde{\mathbf{e}}'\mathbf{T}\widetilde{\mathbf{e}} \qquad \dot{\sim}\ \chi_f^2 \qquad \textit{mit}\ f=\frac{\mathrm{Sp}(\mathbf{T}\widetilde{\mathbf{V}}_{\widetilde{e}})^2}{\mathrm{Sp}(\mathbf{T}\widetilde{\mathbf{V}}_{\widetilde{e}}\mathbf{T}\widetilde{\mathbf{V}}_{\widetilde{e}})},$$

wobei $\mathbf{T}=\mathbf{C}'(\mathbf{CC}')^{-}\mathbf{C}$ *die standardisierte Hypothesenmatrix zur Hypothese* $\mathbf{Ce}=\mathbf{0}$ *sei.* $\mathrm{Sp}(\cdot)$ *bezeichne hierbei die Spur einer Matrix.*

[11] Obwohl für die ungewichteten Rangtransformationen keine Rangdarstellung existiert, wird der Begriff der Rangtransformation an dieser Stelle aus Gründen der Einheitlichkeit verwendet.

Satz 4.27 (Logistisches Modell) *Es gelten die Voraussetzungen von Satz 4.10. Es sei weiter* $\mathbf{g}(\mathbf{e}) = logit(\mathbf{e}) = \left(\log(\frac{e^{(l)}}{1-e^{(l)}})\right)_{l=1,\ldots,d}$ *die logit-Funktion in ihrer multivariaten Form. Es sei weiter* $\widetilde{\mathbf{V}}^g_{\tilde{e}} = \mathbf{Dg}(\tilde{\mathbf{e}})\widetilde{\mathbf{V}}_{\tilde{e}}\mathbf{Dg}(\tilde{\mathbf{e}})$*, wobei* $\mathbf{Dg}(\tilde{\mathbf{e}})$ *die Jacobimatrix der partiellen Ableitungen* $\frac{\partial \mathbf{g}}{\partial \mathbf{e}}$ *an der Stelle* $\tilde{\mathbf{e}}$ *sei. Dann gilt unter* $\mathrm{H}_0 : \mathbf{C}\cdot\mathbf{g}(\mathbf{e}) = \mathbf{0}$*:*

$$\textbf{\textit{ANOVA-Typ-Statistik}} \qquad \mathrm{Q}^{\mathrm{ATS}}_g(\mathbf{T}) = \mathrm{N}\frac{\mathrm{Sp}(\mathbf{T}\widetilde{\mathbf{V}}^g_{\tilde{e}})}{\mathrm{Sp}(\mathbf{T}\widetilde{\mathbf{V}}^g_{\tilde{e}}\mathbf{T}\widetilde{\mathbf{V}}^g_{\tilde{e}})}\,\mathbf{g}(\tilde{\mathbf{e}})'\mathbf{T}\mathbf{g}(\tilde{\mathbf{e}}) \qquad \dot{\sim} \chi^2_f \qquad \textit{mit } f = \frac{\mathrm{Sp}(\mathbf{T}\widetilde{\mathbf{V}}^g_{\tilde{e}})^2}{\mathrm{Sp}(\mathbf{T}\widetilde{\mathbf{V}}^g_{\tilde{e}}\mathbf{T}\widetilde{\mathbf{V}}^g_{\tilde{e}})}.$$

Konfidenzintervalle

Satz 4.28 (Konfidenzintervalle des linearen Modells) *Es gelte* $\sqrt{\mathrm{N}}(\tilde{\mathbf{e}} - \mathbf{e}) \dot{\sim} \mathrm{N}\left(\mathbf{0}, \widetilde{\mathbf{V}}_{\tilde{e}}\right)$ *und es sei* $u_{1-\frac{\alpha}{2}}$ *das* $(1-\frac{\alpha}{2})$*-Quantil der Standardnormalverteilung, dann sind*

$$\mathbf{c}'\tilde{\mathbf{e}} \pm \frac{\sqrt{\mathbf{c}'\widetilde{\mathbf{V}}_{\tilde{e}}\mathbf{c}}\cdot u_{1-\frac{\alpha}{2}}}{\sqrt{\mathrm{N}}} \tag{4.33}$$

asymptotische $(1-\alpha)$*-Konfidenzintervallgrenzen für* $\mathbf{c}'\mathbf{e}$*.*

Satz 4.29 (Konfidenzintervalle des logistischen Modells) *Es gelte die Notation von Satz 4.11 und es sei* $\widetilde{\mathbf{V}}^g_{\tilde{e}}[l,l]$*, das* l*-te Diagnonalelement von* $\widetilde{\mathbf{V}}^g_{\tilde{e}}$*, dann sind*

$$g^{-1}\left(g(\tilde{e}^{(l)}) \pm \frac{\sqrt{\widetilde{\mathbf{V}}^g_{\tilde{e}}[l,l]}\cdot u_{1-\frac{\alpha}{2}}}{\sqrt{\mathrm{N}}}\right) \tag{4.34}$$

die logistischen $(1-\alpha)$*-Konfidenzintervallgrenzen für* $e^{(l)}$*,* $l = 1,\ldots,d$*, wobei* $g^{-1}(x) = \frac{exp(x)}{1+exp(x)}$ *die Umkehrfunktion der logit-Funktion sei.*

5 Verbundene Stichproben in der Praxis: die verschiedenen Studiendesigns

Basierend auf der in der Arbeit von Lange (2008) vorgestellten Methodik erfolgt in diesem Kapitel die Anwendung der vorherigen Ergebnisse auf die in Kapitel 3.2 vorgestellten Studiendesigns. Erarbeitet werden dabei die vier Hauptkonstellationen von Readern und diagnostischen Methoden in Diagnosestudien (siehe Tabelle 5.1):

Tabelle 5.1: Hauptkonstellationen diagnostischer Studien

	gleiche Reader bei verschiedenen Methoden	**unterschiedliche Reader** bei verschiedenen Methoden
gleiche Patienten für verschiedenen Methoden	Design 1	Design 2
unterschiedliche Patienten für verschiedenen Methoden	Design 3	Design 4

Lange (2008) reduziert die Analysemethode für die verschiedenen Studiendesigns dabei von den oben präsentierten vier möglichen auf lediglich zwei konzeptionell verschiedene Ansätze: die Analyse abhängiger Beobachtungen bei verschiedener Methoden (Design 1 und 2) sowie die Analyse unabhängiger Beobachtungen bei verschiedenen Methoden (Design 3 und 4). Doch bevor auf die verschiedenen Designs im Detail eingegangen wird, sei an dieser Stelle an die zentralen Ergebnisse des vorherigen Kapitels erinnert, welche sich einheitlich und kompakt in folgendem Resultat zusammenfassen lassen:

Satz 5.1 *Es sei* $\mathbf{e} = \left(e^{(l)}\right)'_{l=1,\ldots,d} = \left(e^{(1)},\ldots,e^{(d)}\right)' \in \{\mathbf{AUC}, \mathbf{se}, \mathbf{sp}, \mathbf{p}_+, \mathbf{p}_-\}$ *der Vektor eines geeigneten diagnostischen Gütemaßes unter verschiedenen Bedingungen* $l = 1,\ldots,d$*, welcher unter den Voraussetzungen von Modell 1 (bei Einfachmessungen, siehe Seite 19) oder von Modell 2 (bei Clusterdaten, siehe Seite 34) durch* $\widehat{\mathbf{e}}$ *beziehungsweise* $\widetilde{\mathbf{e}}$ *geschätzt wird. In der Notation des entsprechenden Modells gilt dann stets*

$$\sqrt{\mathrm{N}}\,(\widehat{\mathbf{e}} - \mathbf{e}) \mathrel{\dot{\sim}} \mathrm{N}(\mathbf{0}, \widehat{\mathbf{V}}_{\widehat{e}}) \qquad \textit{beziehungsweise} \qquad \sqrt{\mathrm{N}}\,(\widetilde{\mathbf{e}} - \mathbf{e}) \mathrel{\dot{\sim}} \mathrm{N}(\mathbf{0}, \widetilde{\mathbf{V}}_{\widetilde{e}}).$$

Lange (2008) zeigt nun, dass sich durch eine Strukturierung des Vektorindex l aus diesem Resultat die Analysemethoden für die faktoriellen Designs ergeben.

Design 1

Es gelten die Voraussetzungen und Notationen von Modell 1 (für den Fall von Einfachmessung, siehe Seite 19) beziehungsweise von Modell 2 (bei Clusterdaten, siehe Seite 34). Um den verschiedenen Reader-Methoden-Kombinationen in der Notation Rechnung zu tragen, wird der in der Modellnotation verwendete Index l durch einen Indexvektor (s,r) ersetzt, sodass jedes $l = 1,\ldots,d$ eine ausgewählte Reader-Methoden-Kombination (s,r), $s = 1,\ldots,\mathrm{S}$, $r = 1,\ldots,\mathrm{R}$ repräsentiert. Dieses Vorgehen soll anhand des Parametervektors

des diagnostischen Gütemaße **e** illustriert werden, der sich durch die Strukturierung von l wie folgt darstellt:

$$\mathbf{e} = \left(e^{(1|1)}, \ldots, e^{(1|\mathrm{R})}, e^{(2|1)}, \ldots, e^{(2|\mathrm{R})}, \ldots, e^{(\mathrm{S}|1)}, \ldots, e^{(\mathrm{S}|\mathrm{R})}\right)'. \tag{5.1}$$

Die verschiedenen zu testenden Hypothesen lassen sich demnach mittels der folgenden Matrizen formulieren.

$$\begin{array}{lll}
\text{(1.) } \textbf{Methodeneffekt} & \mathrm{H}_0^{\mathrm{Meth}} : \mathbf{C}_{\mathrm{Meth}} \cdot \mathbf{e} = \mathbf{0} & \mathbf{C}_{\mathrm{Meth}} = \mathbf{P}_{\mathrm{S}} \otimes \frac{1}{\mathrm{R}} \mathbf{1}_{\mathrm{R}}' \\
\text{(2.) } \textbf{Readereffekt} & \mathrm{H}_0^{\mathrm{Read}} : \mathbf{C}_{\mathrm{Read}} \cdot \mathbf{e} = \mathbf{0} & \mathbf{C}_{\mathrm{Read}} = \frac{1}{\mathrm{S}} \mathbf{1}_{\mathrm{S}}' \otimes \mathbf{P}_{\mathrm{R}} \\
\text{(3.) } \textbf{Interaktion} & \mathrm{H}_0^{\mathrm{Meth}\times\mathrm{Read}} : \mathbf{C}_{\mathrm{Meth}\times\mathrm{Read}} \cdot \mathbf{e} = \mathbf{0} & \mathbf{C}_{\mathrm{Meth}\times\mathrm{Read}} = \mathbf{P}_{\mathrm{S}} \otimes \mathbf{P}_{\mathrm{R}}.
\end{array} \tag{5.2}$$

Für dieses strukturierte **e** in (5.1) gelten nun die Resultate aus Satz 5.1, sodass sich mit Hilfe der Teststatistiken aus den Abschnitten 4.1.5 (für Einfachmessungen, ab Seite 29) beziehungsweise 4.2.5 (bei Clusterdaten auf Seite 53) die obigen Hypothesen testen lassen. Die Methodik zur Konstruktion geeigneter Konfidenzintervalle wird ebenfalls in den jeweiligen Abschnitten beschrieben.

Design 2

Wie beim ersten Design lässt sich die Analysemethodik des zweiten Designs durch eine Umstrukturierung des Vektorindex l herleiten. Analog zu Design 1 gelten zunächst die Voraussetzungen und Notationen von Modell 1 beziehungsweise von Modell 2. Da im Fall von Design 2 die verschiedenen diagnostischen Verfahren $s = 1, \ldots, \mathrm{S}$ von verschiedenen Readern $r(s) = 1, \ldots, \mathrm{R}_s$ evaluiert werden, erhält der Parametervektor der diagnostischen Gütemaße durch die erforderliche Strukturierung von l in $(s, r(s))$ die Form

$$\mathbf{e} = \left(e^{(1|1)}, \ldots, e^{(1|\mathrm{R}_1)}, e^{(2|1)}, \ldots, e^{(2|\mathrm{R}_2)}, \ldots . e^{(\mathrm{S}|1)}, \ldots, e^{(\mathrm{S}|\mathrm{R}_\mathrm{S})}\right)'. \tag{5.3}$$

In analoger Weise werden alle in diesem Modell verwendeten Indizes l durch den zugehörigen Indexvektor $(s, r(s))$ substituiert. Die in diesem hierarchischen Design zu testenden Hypothesen lassen sich in Matrixnotation wie folgt formulieren.

$$\text{(1.) } \textbf{Methodeneffekt} \qquad \mathrm{H}_0^{\mathrm{Meth}} : \mathbf{C}_{\mathrm{Meth}} \cdot \mathbf{e} = \mathbf{0} \qquad \mathbf{C}_{\mathrm{Meth}} = \mathbf{P}_{\mathrm{S}} \bigoplus_{s=1}^{\mathrm{S}} \frac{1}{\mathrm{R}_s} \mathbf{1}_{\mathrm{R}_s}' \tag{5.4}$$

$$\text{(2.) } \textbf{Reader(Methode)-Effekt} \qquad \mathrm{H}_0^{\mathrm{Read(Meth)}} : \mathbf{C}_{\mathrm{Read(Meth)}} \cdot \mathbf{e} = \mathbf{0} \qquad \mathbf{C}_{\mathrm{Read(Meth)}} = \bigoplus_{s=1}^{\mathrm{S}} \mathbf{P}_{\mathrm{R}_s} \tag{5.5}$$

Teststatistiken sowie Konfidenzintervalle erhält man analog zum ersten Design durch Anwendung der Resultate aus den Abschnitten 4.1.5 und 4.2.5.

Design 3

Im dritten Design wird – analog zum ersten – wieder jede Methode $s = 1, \ldots, \mathrm{S}$ von jedem Reader $r = 1, \ldots, \mathrm{R}$ evaluiert, sodass alle Vektoren die schon in Gleichung (5.1) präsentierte Struktur aufweisen und die Hypothesen sich analog zum ersten Design mit Hilfe der Matrizen der Gleichungen (5.2) formulieren lassen. Analog zu Lange (2008) seien die folgenden zusätzlichen Notationen eingeführt und folgende Voraussetzungen erfüllt:

Voraussetzung 5.1 *Es sei* N_s *die Anzahl an Subjekten, welche mittels Diagnoseverfahren* s $(s = 1, \ldots, \mathrm{S})$ *diagnostiziert werden und* $\mathrm{N} = \sum_{s=1}^{\mathrm{S}} \mathrm{N}_s$ *die Gesamtanzahl an Subjekten.*

(1) $\mathrm{N} \to \infty$, *derart, dass* $\frac{\mathrm{N}}{\mathrm{N}_s} \leq \mathrm{C} < \infty$, *für alle* $s = 1, \ldots, \mathrm{S}$.

(2) Für alle $s = 1,\ldots,\mathrm{S}$ *seien die Voraussetzungen von Modell 1 (bei Einfachmessungen) beziehungsweise von Modell 2 (bei Mehrfachmessungen) erfüllt.*

Nach Voraussetzung 5.1/(2) und Satz 5.1 gilt

$$\sqrt{\mathrm{N}_s}\left(\left(\widehat{e}^{(s|1)},\ldots\widehat{e}^{(s|\mathrm{R})}\right)' - \left(e^{(s|1)},\ldots e^{(s|\mathrm{R})}\right)'\right) \dot{\sim} \mathrm{N}(\mathbf{0},\widehat{\mathbf{V}}_{\widehat{e}}^{(s)}), \qquad \forall s = 1,\ldots,\mathrm{S}.$$

Lange (2008) zeigt nun, dass hieraus mit Hilfe von Voraussetzung 5.1/(1) folgt

$$\sqrt{\mathrm{N}}\,(\widehat{\mathbf{e}} - \mathbf{e}) \dot{\sim} \mathrm{N}\left(\mathbf{0},\bigoplus_{s=1}^{\mathrm{S}} \frac{\mathrm{N}}{\mathrm{N}_s}\widehat{\mathbf{V}}_{\widehat{e}}^{(s)}\right)$$

(siehe Lange, 2008, Abschnitt 3.4.2), wodurch sich die Hypothesen dieses 2-faktoriellen hierarchischen Designs nun analog zum ersten Design testen lassen.

Design 4
Da Design 4 sich unter Voraussetzung 5.1/(1) in gleicher Art und Weise aus Design 2 ergibt wie Design 3 aus Design 1, wird an dieser Stelle auf eine Präsentation der Methodik verzichtet.

Ergänzend sei an dieser Stelle angemerkt, dass sich durch die hier präsentierte Strukturierung des Vektors auch höher faktorielle Studien modellieren lassen, indem der Indexvektor $(s|r)$ um weitere Komponenten ergänzt wird. So kann man beispielsweise über einen Indexvektor $(z|s|r)$, mit der erste Komponente z neben Reader und Methode zusätzlich verschiedene Zentren kennzeichnen.

6 Simulationen

Die Möglichkeiten und Grenzen der praktischen Anwendung der entwickelten Verfahren werden in diesem Kapitel mit Hilfe von Simulationen aufgezeigt. Unter verschiedenen ausgewählten Parameterkonstellationen werden sowohl die Einhaltung des Niveaus als auch die Power der Verfahren mit Hilfe der SAS-Prozedur `PROC IML` (SAS 9.2, SAS Institute Inc., Cary, NC, USA) simuliert. Ausführliche Simulationen der verschiedenen Reader-Methoden-Kombinationen bei Einfachmessungen findet man bei Lange (2008). Auf Grund dessen werden in diesem Kapitel lediglich die Simulationsergebnisse der Methodik bei Clusterdaten präsentiert. Da die Analyse verbundener Stichproben das theoretische Fundament aller vier Designs bildet, wird in dieser Arbeit exemplarisch das erste Design simuliert. Die Ergebnisse lassen sich leicht auf die anderen Designs extrapolieren (vergleiche die Simulationsstudie von Lange, 2008). Bevor das Simulationsmodell sowie die simulierten Parameterkonstellationen präsentiert werden, stehen zunächst einige theoretische Überlegungen hinsichtlich des Vergleiches der gewichteten und ungewichteten Schätzer im Vordergrund.

Satz 6.1 *Es sei* $\lambda_{ik}^{(l|v)} \geq 0$ *das Gewicht, mit welchem der* k*-te vollständige Patient in den Schätzungen der Parameter und Verteilungsfunktionen berücksichtigt wird. Es sei weiter* $\lambda_{ik}^{(l|u)} \geq 0$ *das Gewicht, mit welchem der* k*-te unvollständige Patient im Zustand* i *in die Schätzungen eingeht, wobei* $\sum_{t\in\{u,v\}} \sum_{k=1}^{n_i^{(t)}} \lambda_{ik}^{(l|t)} = 1,\ i = 0,1$*. Es sei schließlich*

$$\widetilde{\mathrm{F}}^{(l)}_{\boldsymbol{\lambda}_i^{(l)}\, i}(x) = \sum_{t\in\{u,v\}} \sum_{k=1}^{n_i^{(t)}} \frac{\lambda_{ik}^{(l|t)}}{m_{ik}^{(l|t)}} \sum_{s=1}^{m_{ik}^{(l|t)}} c(x - \mathrm{X}_{iks}^{(l|t)})$$

die zu den Gewichten $\boldsymbol{\lambda}_i^{(l)} = \left(\lambda_{i1}^{(l|v)}, \cdots \lambda_{in^{(v)}}^{(l|v)}, \lambda_{i1}^{(l|u)}, \cdots, \lambda_{in_i^{(u)}}^{(l|u)}\right)'$ *gehörige gewichtete empirische Verteilungsfunktion.* [1]

(1.) Es sei außerdem

$$\tau_{ik}^{(l|t)} = \sum_{s=1}^{m_{ik}^{(l|t)}} \sum_{s'=1}^{m_{ik}^{(l|t)}} \mathrm{Cov}(c(x - \mathrm{X}_{iks}^{(l|t)}), c(x - \mathrm{X}_{iks'}^{(l|t)}))$$

die Varianz der Summe $\sum_{s=1}^{m_{ik}^{(l|t)}} c(x - \mathrm{X}_{iks}^{(l|t)})$*. Mit dieser Notation ist die Varianz von* $\widetilde{\mathrm{F}}^{(l)}_{\boldsymbol{\lambda}_i^{(l)}\, i}(x)$ *für festes* $x \in \mathbb{R}$ *dann minimal, wenn für die Gewichte* $\lambda_{ik}^{(l|i)}$ *gilt:*

$$\lambda_{ik}^{(l|t)} = \frac{\frac{\left(m_{ik}^{(l|t)}\right)^2}{\tau_{ik}^{(l|v)}}}{\sum\limits_{t\in\{u,v\}} \sum\limits_{k=1}^{n_i^{(t)}} \frac{\left(m_{ik}^{(l|t)}\right)^2}{\tau_{ik}^{(l|v)}}}, \quad l = 1,\ldots,d,\ t \in \{u,v\},\ i = 0,1,\ k = 1,\ldots,n_i^{(t)}.$$

[1] Diese Notation bedeutet, dass man mit $\lambda_{ik}^{(l|t)} = \frac{m_{ik}^{(l|t)}}{m_{i\cdot}^{(l|\cdot)}}$ die gewichteten und mit $\lambda_{ik}^{(l|t)} = \frac{1}{n_i}$ gerade die ungewichteten Schätzer erhält.

(2.) Es sei

$$\rho_{ik}^{(l,r|t)} = \sum_{s=1}^{m_{ik}^{(l|t)}} \sum_{s'=1}^{m_{ik}^{(r|t)}} \mathrm{Cov}(c(x - \mathrm{X}_{iks}^{(l|t)}), c(x - \mathrm{X}_{iks'}^{(r|t)}))$$

die Kovarianz von $\sum_{s=1}^{m_{ik}^{(l|t)}} c(x - \mathrm{X}_{iks}^{(l|t)})$ *und* $\sum_{s'=1}^{m_{ik}^{(r|t)}} c(x - \mathrm{X}_{iks'}^{(r|t)})$. *Es sei* $\rho_{ik}^{(l,r|t)} \geq 0$. *Die Kovarianz* $\mathrm{Cov}\left(\widetilde{\mathrm{F}}^{(l)}_{\boldsymbol{\lambda}_i^{(l)} i}(x), \widetilde{\mathrm{F}}^{(l)}_{\boldsymbol{\lambda}_i^{(r)} i}(x)\right)$ *ist für festes* $x \in \mathbb{R}$ *minimal, wenn für die Gewichte gilt:*

$$\lambda_{ik}^{(l|t)} = \frac{\frac{m_{ik}^{(l|t)} m_{ik}^{(r|t)}}{\rho_{ik}^{(l,r|v)}}}{\sum\limits_{t \in \{u,v\}} \sum\limits_{k=1}^{n_i^{(t)}} \frac{m_{ik}^{(l|t)} m_{ik}^{(r|t)}}{\rho_{ik}^{(l,r|v)}}}, \quad l, r = 1, \ldots, d,\ t \in \{u, v\},\ i = 0, 1,\ k = 1, \ldots, n_i^{(t)}.$$

BEWEIS. Die Ergebnisse lassen sich über die Berechnung der Ableitungen der zu diesem Optimierungsproblem unter Nebenbedingungen gehörigen Lagrangefunktion leicht verifizieren, da es sich bei dem Optimierungsproblem um ein konvexes Programm handelt. □

Wird davon ausgegangen, dass die verschiedenen Reader-Methoden-Kombinationen positiv korreliert sind, so beeinflussen unter anderem zwei Größen die Power des verwendeten statistischen Verfahrens.

(1.) Die Varianz des Effektschätzers: Steigt diese an, wird die Power des Verfahrens kleiner.

(2.) Die Kovarianz zwischen den Reader-Methoden-Kombinationen: Steigt diese an, so bekommt das statistische Verfahren mehr Power, da die Varianz der Differenzen kleiner wird.

Die Wahl der optimalen Gewichtung stellt somit ein komplexes Problem dar, da sie von der Struktur der Kovarianzmatrix abhängt: Sind die Beobachtungen innerhalb eines Clusters komplett unabhängig, so gilt $\tau_{ik}^{(l|t)} = m_{ik}^{(l|t)} \cdot \mathrm{Var}(c(x - \mathrm{X}_{ik1}^{(l|t)})$ und somit ist nach Satz 6.1 der gewichtete Schätzer der Verteilungsfunktion der Schätzer minimaler Varianz. Bei vollständiger Korrelation der Beobachtungen innerhalb eines Clusters (Duplizierung der Messwerte) gilt hingegen $\tau_{ik}^{(l|t)} = \left(m_{ik}^{(l|t)}\right)^2 \cdot \mathrm{Var}(c(x - \mathrm{X}_{ik1}^{(l|t)})$, wodurch der ungewichtete Schätzer minimale Varianz aufweist. Bei der Kovarianz beobachtet man ähnliche Effekte: Während bei vollständiger Abhängigkeit der verschiedenen Reader-Methoden-Kombinationen der ungewichtete Schätzer die Kovarianz minimiert, minimiert bei vollständiger Unabhängigkeit der gewichtete Schätzer die Kovarianz. Nimmt man an, dass sich diese Eigenschaften der empirischen Verteilungsfunktionen auf die Schätzer der diagnostischen Gütemaße übertragen, erhält man die in Tabelle 6.1 präsentierten Eigenschaften auch für die präsentierten Statistiken. Das Zeichen $\gtrsim$ ist hierbei das in der Entscheidungstheorie üblicherweise verwen-

Tabelle 6.1: Analytischer Ansatz zur Evaluation der Güte der Schätzer in Abhängigkeit der Korrelation

	vollständige Abhängigkeit der Reader-Methoden-Kombinationen	vollständige Unabhängigkeit der Reader-Methoden-Kombinationen
vollständig abhängige Cluster	Var: ungewichtet $\gtrsim$ gewichtet Cov: ungewichtet $\lesssim$ gewichtet **Gesamt: ?**	Var: ungewichtet $\gtrsim$ gewichtet Cov: ungewichtet $\gtrsim$ gewichtet **Gesamt: ungewichtet $\gtrsim$ gewichtet**
unabhängige Cluster	Var: ungewichtet $\lesssim$ gewichtet Cov: ungewichtet $\lesssim$ gewichtet **Gesamt: ungewichtet $\lesssim$ gewichtet**	Var: ungewichtet $\lesssim$ gewichtet Cov: ungewichtet $\gtrsim$ gewichtet **Gesamt: ?**

dete Symbol für „favorisiert“ oder „überlegen“ und gibt hier an, welcher der beiden Schätzer die höhere Power aufweist.

Da sich diese Überlegungen nun aber lediglich an den Eigenschaften der empirischen Verteilungsfunktionen orientieren, jedoch keinen Beweis für die die diagnostischen Gütemaße darstellen, wird anhand einer Powersimulation der ANOVA-Typ-Statistik für die AUC evaluiert, ob sich die obigen Charakteristika von den empirischen Verteilungsfunktionen auf die Teststatistiken vererben. Exemplarisch wird eine Ein-Punkt-Alternative bei einer Clustergröße von 5 Beobachtungen pro Patient an 50 Patienten bei einer Erkrankungsprävalenz von $\pi = 0.5$ simuliert. Bei vollständiger Abhängigkeit der verschiedenen Reader-Methoden-Kombinationen ist die Durchführung eines statistischen Tests zum Nachweis von Unterschieden unnötig, da die Varianz des Unterschiedes auf 0 geschätzt wird. Bei der Simulation der ANOVA-Typ-Statistik der obigen vier Extremfälle wird daher anstelle einer vollständigen Abhängigkeit der Reader-Methoden-Konstellation nur eine Korrelation von 0.95 angenommen. Man erhält hierbei die folgenden Ergebnisse:

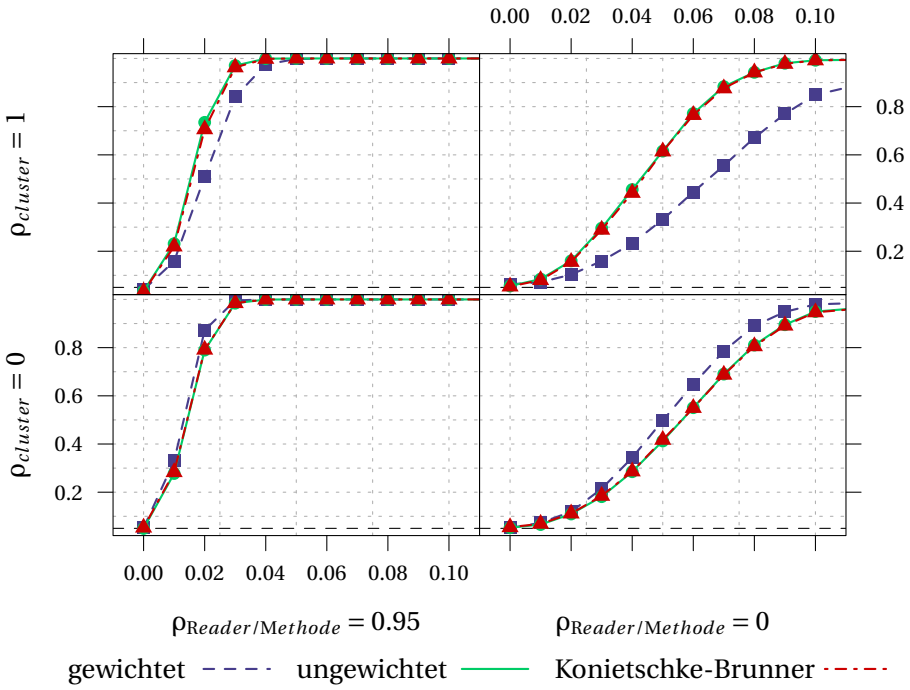

Abbildung 6.1: Powersimulation der ANOVA-Typ-Statistik (AUC) der in der obigen Tabelle präsentierten vier Extremfälle bei einer Clustergröße von 5 Beobachtungen, einem Stichprobenumfang von 50 und einer Erkrankungsprävalenz von $\pi = 0.5$

Abbildung 6.1 zeigt, dass sich die unter Zuhilfenahme von Satz 6.1 plausibilisierten Optimalitätseigenschaften der empirischen Verteilungsfunktionen auf die Optimalität der AUC-Teststatistiken zu übertragen scheinen. Diese Tatsache sollte bei der Wahl des optimalen Schätzers nicht außer Acht gelassen werden. Dennoch sollte die Entscheidung für einen bestimmten Schätzer nicht nur anhand der Powereigenschaften getroffen werden, da die Interpretation der Ergebnisse von der Wahl des Schätzers abhängt: Ist der Patient das Ziel der Diagnose, so ist auch die Güte des diagnostischen Verfahrens am Patienten zu beurteilen, wodurch die ungewichteten Schätzer angemessener erscheinen. Ist hingegen die Untereinheit das Ziel der Diagnose, sollte jede gemessene Subunit das gleiche Gewicht bekommen, wodurch an dieser Stelle die gewichteten Schätzer die Güte adäquater beschreiben.

Die im Rahmen dieser Arbeit durchgeführte Simulationsstudie untersucht sowohl die Methodik der gewichteten als auch die der ungewichteten Schätzer im Vergleich zu den von Konietschke und Brunner (2009) präsentierten Ansätzen. Hierbei bleibt die Analyse auf ausgewählte Korrelationsstrukturen beschränkt. Die Beobachtungen des k-ten Patienten ($k = 1, \ldots, \mathrm{N}$) werden dabei wie folgt generiert:[2]

$$\mathbf{X}_k = \begin{pmatrix} \mathrm{X}_{0k1}^{(1)} \\ \vdots \\ \mathrm{X}_{0k1}^{(d)} \\ \vdots \\ \vdots \\ \mathrm{X}_{0km_{0k}}^{(1)} \\ \vdots \\ \mathrm{X}_{0km_{0k}}^{(d)} \\ \mathrm{X}_{1k1}^{(1)} \\ \vdots \\ \mathrm{X}_{1k1}^{(d)} \\ \vdots \\ \vdots \\ \mathrm{X}_{1k1}^{(1)} \\ \vdots \\ \mathrm{X}_{1km_{1k}}^{(d)} \end{pmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \\ \\ \end{matrix}\right\} \text{Untersuchungsergebnis der ersten gesunden Subunit unter den verschiedenen Bedingungen} \\ \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} \text{Untersuchungsergebnis der } m_{0k}\text{-ten Subunit unter den verschiedenen Bedingungen} \\ \left.\begin{matrix} \\ \\ \\ \\ \\ \end{matrix}\right\} \text{Untersuchungsergebnis der ersten kranken Subunit unter den verschiedenen Bedingungen} \\ \left.\begin{matrix} \\ \\ \\ \\ \\ \end{matrix}\right\} \text{Untersuchungsergebnis der } m_{1k}\text{-ten Subunit unter den verschiedenen Bedingungen} \end{matrix} \sim \mathrm{N}\left(\boldsymbol{\mu}, \mathbf{I}_{m_{\cdot k}\cdot d} + \bigoplus_{s=1}^{m_{\cdot k}} \sigma^2_{cluster}\mathbf{J}_{d\times d} + \sigma^2_{pat}\mathbf{J}_{(m_{\cdot k}\cdot d)\times(m_{\cdot k}\cdot d)}\right).$$

Hierbei bezeichne $m_{\cdot k}$ die Gesamtanzahl an Beobachtungseinheiten des k-ten Patienten. Der Erwartungswertvektor $\boldsymbol{\mu}$ wird entsprechend der gewünschten Parameterkonstellation variiert: der Erwartungswert aller gesunden Beobachtungen wird stets auf 0 gesetzt und der Erwartungswert der kranken Subunits der l-ten Komponente ist $\Phi^{-1}\left(\mathrm{AUC}^{(l)}\right) \cdot \sqrt{2\left(1 + \sigma^2_{pat} + \sigma^2_{cluster}\right)}$, wobei Φ^{-1} die Umkehrfunktion der Standardnormalverteilung bezeichne. Der Goldstandard einer jeden Beobachtung wird hierbei zufällig durch unabhängig $\mathrm{Bern}(\pi)$ verteilte Zufallsvariablen festgesetzt. Die Daten entstehen demnach mithilfe eines linearen Modells mit zufälligen Effekten

$$\begin{aligned} \mathbf{X}_k =& \boldsymbol{\mu} + \mathbf{1}_{d\cdot m_{\cdot k}} \cdot a_{k,pat} + (\mathbf{I}_{m_{\cdot k}} \otimes \mathbf{1}_d) \cdot \mathbf{a}_{k,cluster} + \boldsymbol{\epsilon}, \text{ wobei} \\ & a_{k,pat} \sim \mathrm{N}(0, \sigma^2_{pat}),\ \mathbf{a}_{k,cluster} \sim \mathrm{N}(\mathbf{0}, \sigma^2_{cluster}\mathbf{I}_{m_{\cdot k}}),\ \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_{d\cdot m_{\cdot k}}), \end{aligned}$$

wobei die Beobachtungen verschiedener Patienten stets als unabhängig betrachtet werden. Demnach gilt:

(1.) Die Varianz einer zufälligen Beobachtung beträgt $1 + \sigma^2_{pat} + \sigma^2_{cluster}$.

(2.) Die Korrelation verschiedener Subunits (des gleichen Patienten) ist $\rho_{pat} = \frac{\sigma^2_{pat}}{1+\sigma^2_{pat}+\sigma^2_{cluster}}$.

(3.) Die Korrelation zweier Beobachtungen (unter verschiedenen Bedingungen) an der selben Subunit beträgt $\rho_{beob} = \frac{\sigma^2_{pat}+\sigma^2_{cluster}}{1+\sigma^2_{pat}+\sigma^2_{cluster}}$.

Im Rahmen dieser Simulationstudie wird lediglich die ANOVA-Typ-Statistik simuliert, da Simulationsstudien zeigen, dass diese insbesondere bei kleinen Stichprobenumfängen das Niveau deutlich besser einhält

[2] Es sei an dieser Stelle angemerkt, dass zu Gunsten einer vereinfachten Darstellung der Korrelationsstruktur der Vektor $\mathbf{X}_k$ im Vergleich zur Notation in Modell 2 umstrukturiert wurde.

als die Wald-Typ-Statistik (siehe zum Beispiel Brunner u. a., 2002; Munzel und Brunner, 2000). Simulationsergebnisse aus dem Bereich der Diagnosestudien von Lange (2008) und Werner (2006) untermauern diese Überlegenheit der ANOVA-Typ-Statistik, sodass bei der hier durchgeführten Simulationsstudie auf eine Evaluation der Wald-Typ-Statistik verzichtet wird. Sowohl bei den Niveau- als auch bei den Powersimulation werden folgende Parameter variiert:

(1.) die Clusterkorrelationen $(\rho_{pat}, \rho_{beob}) \in \{(0.25,\ 0.5),\ (0.5,\ 0.75)\}$
(2.) die Clustergrößen $m_{\cdot k} = 2,\ 5$
(3.) die Erkrankungswahrscheinlichkeiten $\pi = 0.5,\ 0.7,\ 0.9$[3]
(4.) die Gesamtanzahl an Patienten N = 50, 80, 100 (bei Powersimulation N = 50, 100)

Für die Niveausimulationen wird des Weiteren die Größe des untersuchten diagnostischen Gütemaßes AUC, $se,\ sp,= 0.7,\ 0.8,\ 0.9$ variiert. Bei den Powersimulationen wird stets auf einen Ausgangswert von 0.7 eine Ein-Punkt-Alternative von δ addiert. Man erhält hierbei die in den folgenden Abschnitten präsentierten Resultate. Dabei wurden in jeder Simulation 10000 Durchläufe durchgeführt.

6.1 AUC

6.1.1 Niveausimulationen

In den Abbildungen 6.2 und 6.3 finden sich die Ergebnisse der Niveausimulation der AUCs. Zu Gunsten ei-

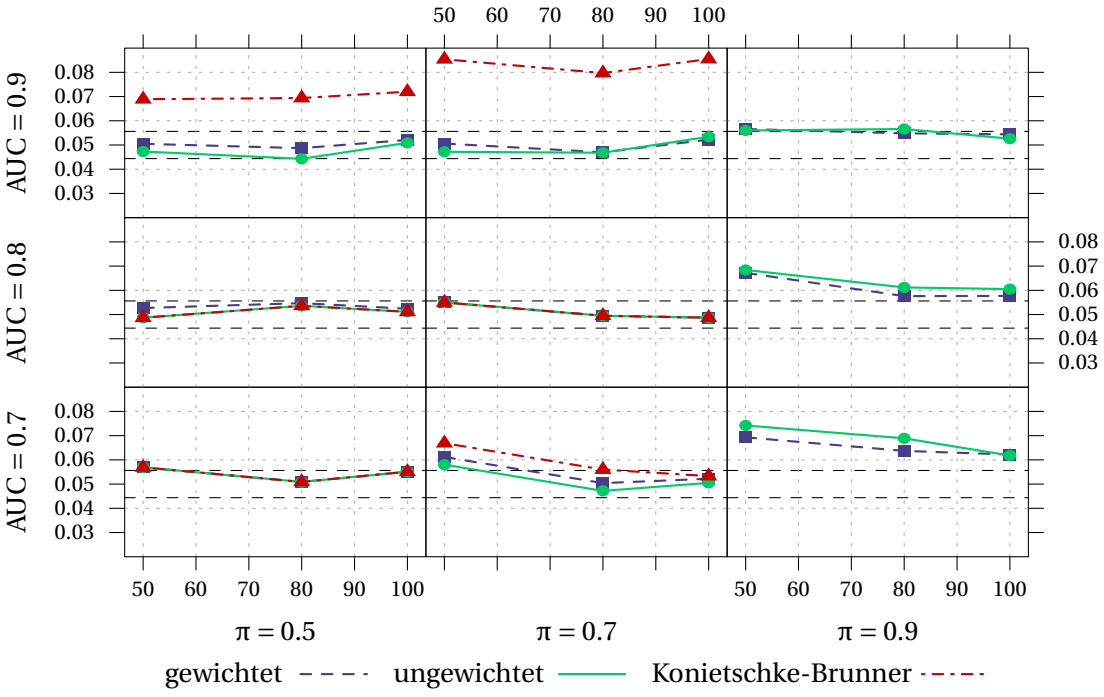

Abbildung 6.2: Niveausimulation des Methodeneffekts der AUC I: Simulationen bei 5% nominellem Niveau für 2 Methoden und 3 Reader bei einer Clusterkorrelation von 0.25 / 0.5 und einer Clustergröße von 2 Beobachtungen; die Ergebnisse der Konietschke-Brunner-Methodik sind auf Grund ihrer starken Liberalität für π = 0.9 nicht dargestellt.

[3] Prävalenzen < 0.5 werden an dieser Stelle nicht simuliert, denn die Ergebnisse sind unter den obigen Simulationsbedingungen symmetrisch um π = 0.5: Da im Kollektiv der Gesunden und und im Kollektiv der Kranken die Varianzen gleich sind, sind die Stichproben der Gesunden und der Kranken austauschbar, wodurch die Resultate an der Stelle π den Ergebnissen bei 1 − π entsprechen.

ner besseren Auflösung zeigen die Graphiken hierbei nur den Bereich von 0.02 bis 0.09. Die von Konietschke und Brunner (2009) vorgestellte Methodik weist bei hohen AUCs oder bei starker Unbalanciertheit zwischen der Anzahl der gesunden und kranken Patienten eine starke Liberalität auf, sodass diese Statistik bei einer Prävalenz von $\pi = 0.9$ nicht mehr in den Abbildungsbereich fällt. Es sei daher an dieser Stelle auf die zugehörigen Powerkurven (siehe Abbildung 6.4) verwiesen, deren Startpunkte das empirische Niveau darstellen. Das Ausmaß der Liberalität der Konieschke-Brunner-Statistik hängt dabei von Clustergröße und Clusterkorrelation ab. Mit höherer Korrelation und größeren Clustern wird die Statistik weniger liberal (vergleiche hierzu auch Abbildung C.1 im Anhang auf Seite 95). Bei den in dieser Arbeit präsentierten gewichteten und ungewichteten Schätzern zeigt sich im Allgemeinen, dass das Niveau gut eingehalten wird. Lediglich bei großer Unbalanciertheit ($\pi = 0.9$) und kleinen Fallzahlen $N = 50, 80$ lässt sich eine leichte Liberalität erkennen. Vergegenwärtigt man sich an dieser Stelle allerdings die Tatsache, dass bei dieser Paramenterkonstellation in Erwartung nur 10% aller Beobachtungen gesund sind, erkennt man, dass die hohe Prävalenz in Verbindung mit dem kleinen Stichprobenumfang zu einem sehr kleinen n_0 führt. Eine leichte Liberalität der Teststatistik ist daher nicht überraschend.

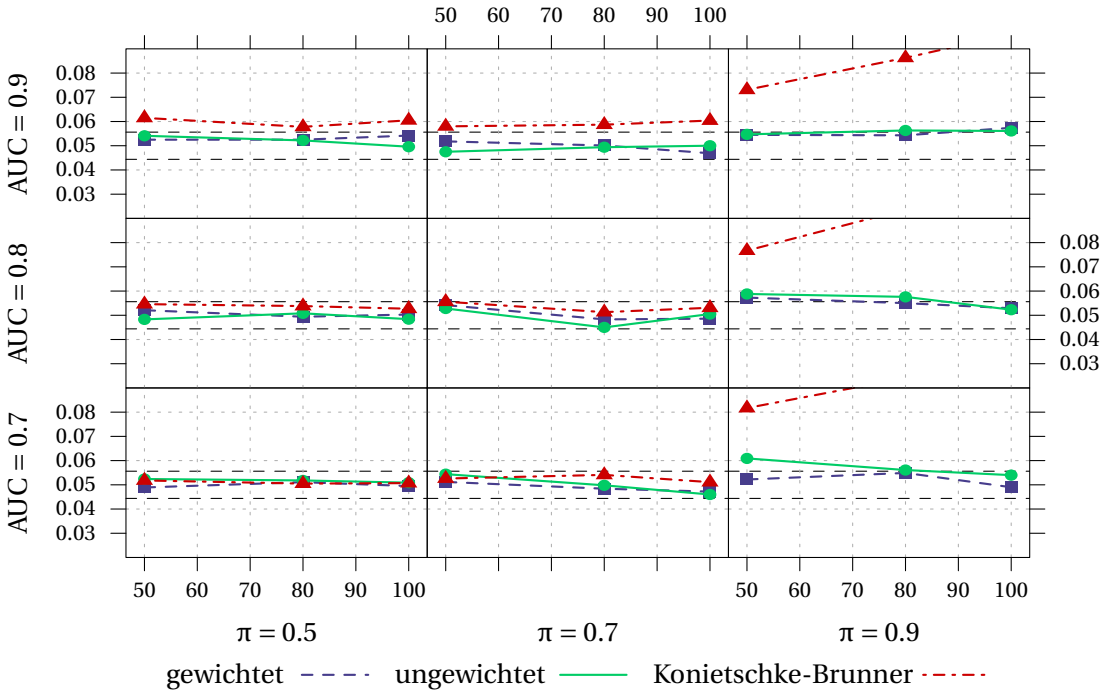

Abbildung 6.3: Niveausimulation des Methodeneffekts der AUC `II`: Simulationen bei 5% nominellem Niveau für 2 Methoden und 3 Reader bei einer Clusterkorrelation von 0.5 / 0.75 und einer Clustergröße von 5 Beobachtungen

Niveausimulationen des Readereffekts zeigen, dass sich dieser als zweiter Haupteffekt analog zum Methodeneffekt verhält und hier ähnliche nominelle Niveaus erzielt werden. Gleiches gilt auch für den Wechselwirkungseffekt.

6.1.2 Powersimulationen

Die Powersimulationen (siehe Abbildung 6.4) zeigen, dass alle drei Methoden im Falle balancierter Designs ein ähnliches Verhalten aufweisen. Lediglich bei starker Unbalanciertheit zeigt die Methodik nach Konietschke und Brunner (2009) trotz anfänglicher starker Liberalität ab einem Effekt von $\delta \approx 0.05$ weniger

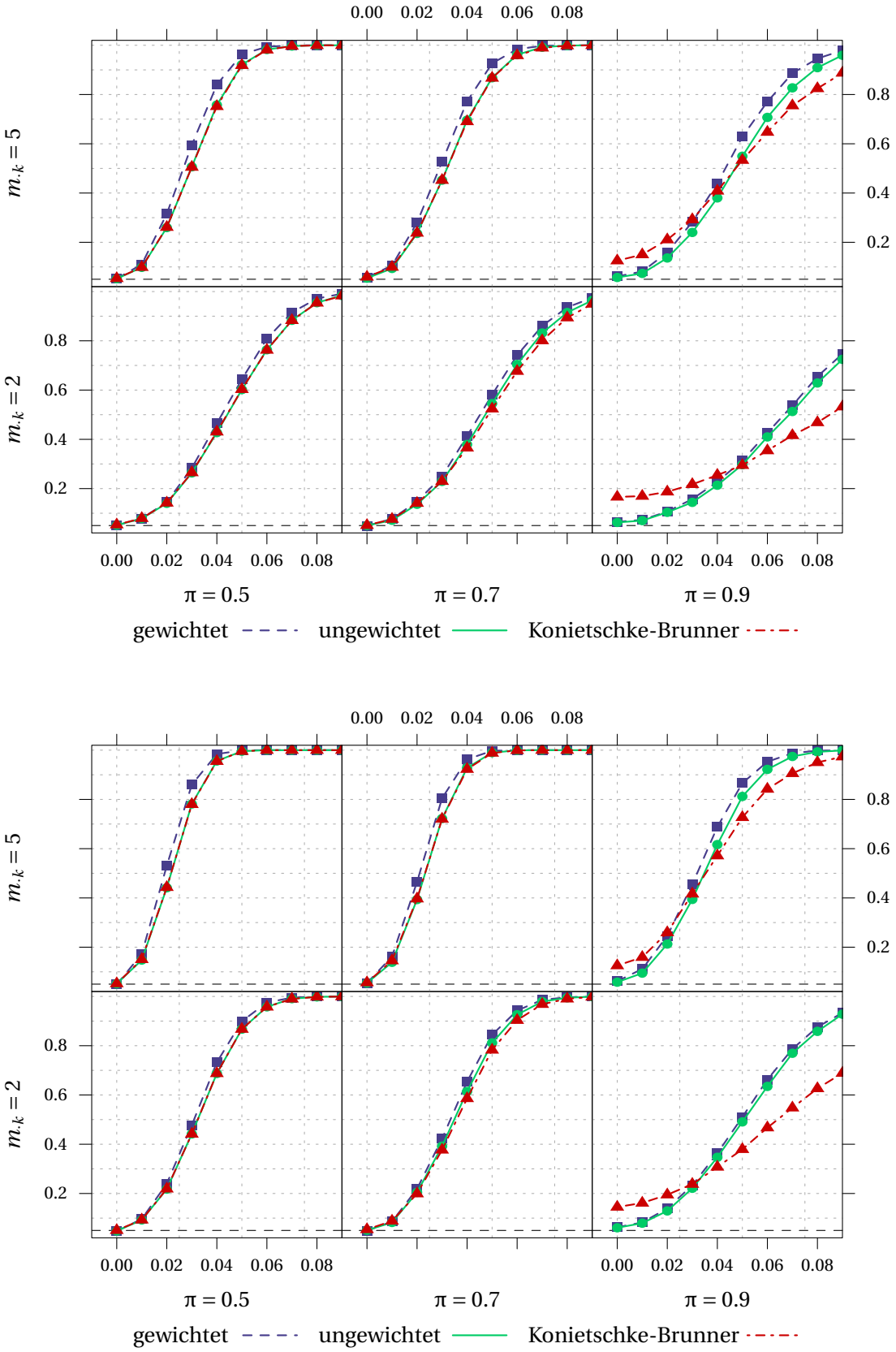

Abbildung 6.4: Powersimulation des Methodeneffekts der AUC: Simulationen bei 5% nominellem Niveau und einem Stichprobenumfang von N = 100 für 2 Methoden und 3 Reader bei einer Clusterkorrelation von 0.25 / 0.5 (obere Graphik) beziehungsweise 0.5 / 0.75 (untere Graphik)

Power. Es lässt sich des Weiteren erkennen, dass mit höherer Korrelation zwischen den Reader-Methoden-Konstellationen die Power aller Verfahren zunimmt. Gleiches gilt, wenn die Anzahl an Beobachtungen pro Patient erhöht wird. Es zeigt sich außerdem, dass unter den simulierten Parameterkonstellationen die ungewichteten Schätzer über eine marginal geringere Power als die gewichteten verfügen.
Exemplarisch werden in dieser Arbeit lediglich die Powerkurven für einen Gesamtstichprobenumfang von N = 100 vorgestellt. Die Powerkurven für N = 50 sind durch gleichen Verlauf auf anderem Niveau gekennzeichnet und werden aus Gründen der Übersichtlichkeit in dieser Arbeit nicht präsentiert. Der Reader-Effekt als zweiter Haupteffekt hat vergleichbare Powereigenschaften wie der hier gezeigte Methoden-Effekt. Ein Wechselwirkungseffekt wird mit geringer Power aufgedeckt, die Kurvenverläufe gleichen auf niedrigerem Niveau jedoch den Powerkurven der Haupteffekte. Zu Gunsten einer besseren Lesbarkeit finden diese Resultate in der vorliegenden Arbeit keine graphische Darstellung.

6.2 Sensitivität und Spezifität

6.2.1 Niveausimulationen

Analog zur AUC-Simulation werden Simulationen von Sensitivität und Spezifität durchgeführt. Die Sensitivität beschreibt die Erfolgswahrscheinlichkeit der verschiedenen Reader-Methoden-Kombinationen auf dem Kollektiv der Kranken, während die Spezifität das gleiche auf dem Kollektiv der Gesunden tut. Somit be-

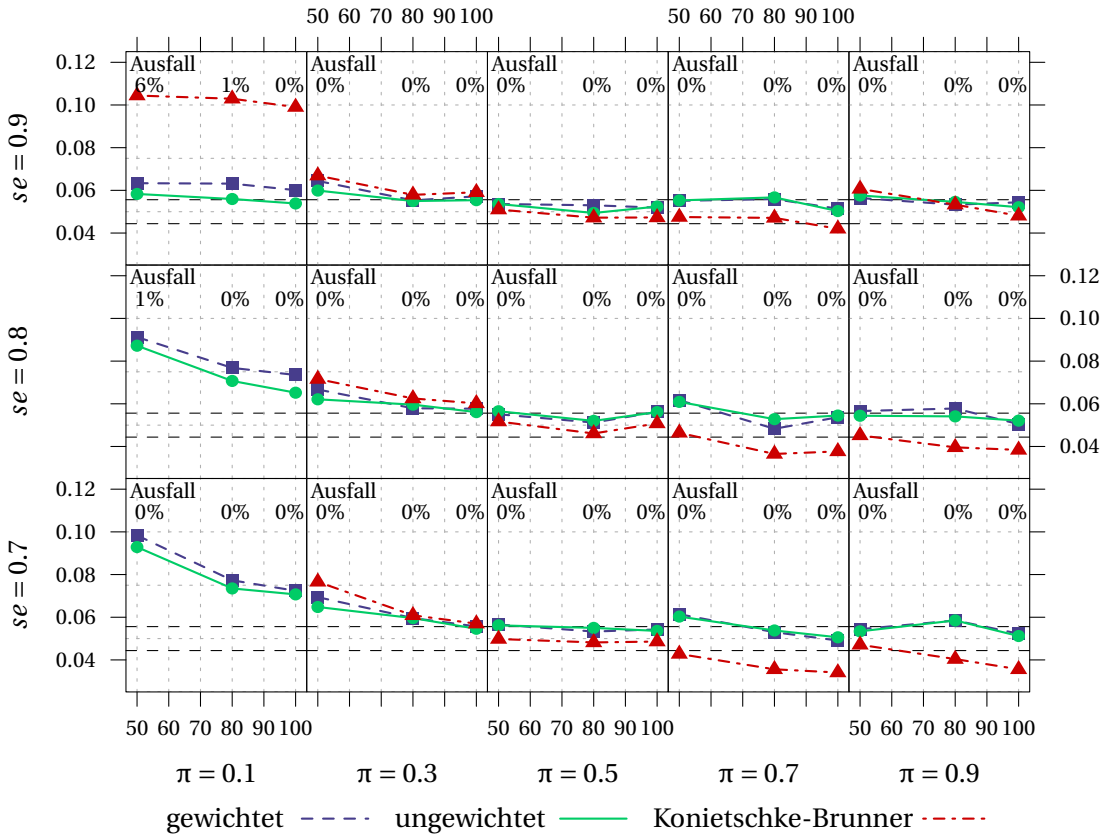

Abbildung 6.5: Niveausimulation des Methodeneffekts der Sensitivität `I`: Simulationen der Sensitivität bei 5% nominellem Niveau für 2 Methoden und 3 Reader bei einer Clusterkorrelation von 0.25 / 0.5 und einer Clustergröße von 2 Beobachtungen

schreiben Sensitivität und Spezifität den gleichen Effekt – lediglich auf verschiedenen Populationen: Die Sensitivität im Kollektiv der Kranken entspricht gerade der Spezifität im Kollektiv der Gesunden. Auf Grund dessen wird in dieser Arbeit nur die Sensitivität simuliert, da sich die Ergebnisse gleichermaßen auch auf die Spezifität übertragen lassen. Diese Ergebnisse sind nun aber im Gegensatz zu den AUC-Simulation nicht mehr symmetrisch um $\pi = 0.5$, sodass zusätzlich die Prävalenzen $\pi = 0.1$ und $\pi = 0.3$ simuliert werden.
Die Resultate dieser Simulationen finden sich in den Abbildungen 6.5 und 6.6. Im Fall niedriger Prävalenzen und kleiner Gesamtstichprobenumfänge (daraus ergibt sich eine kleine Anzahl an kranken Patienten) kommt es vor, dass der Nenner der Teststatistik 0 wird, die zu testende Hypothese lässt sich dann nicht überprüfen. Da die Ausfallquoten für alle drei simulierten Verfahren in etwa gleich sind, wird der Übersicht halber nur die mittlere Ausfallquote in den Graphiken 6.5 und 6.6 dokumentiert. Die Simulationsergebnisse für die Sensitivtät gleichen denen der AUC-Analyse. Für kleine Fallzahlen (geringe Prävalenz, kleine absolute Fallzahl) sind alle Schätzer liberal, wobei das Ausmaß der Liberalität von Clustergröße und Korrelation abhängt (vergleiche hierzu auch Abbildung C.2 im Anhang auf Seite 96). Ab einer Fallzahl von etwa $n_1 = 25$ halten die in dieser Arbeit präsentieren gewichteten und ungewichteten Schätzer das Niveau ein, wohingegen die Konietschke-Brunner-Statistik für große Fallzahlen leicht konservativ wird.

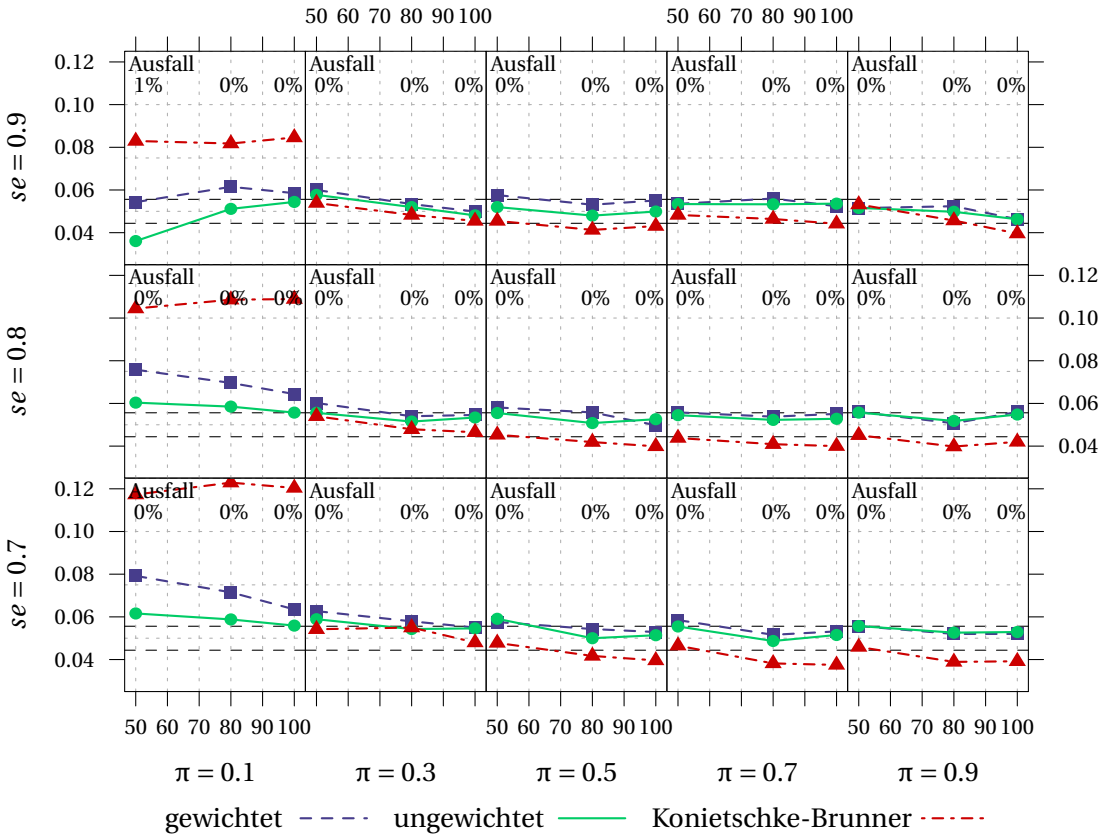

Abbildung 6.6: Niveausimulation des Methodeneffekts der Sensitivität `II`: Simulationen der Sensitivität bei 5% nominellem Niveau für 2 Methoden und 3 Reader bei einer Clusterkorrelation von 0.5 / 0.75 und einer Clustergröße von 5 Beobachtungen

Für den Readereffekt sowie für den Wechselwirkungseffekt zeigen sich – wie auch schon bei den AUC-Simulationen – ähnliche Resultate, die auf Grund einer besseren Übersichtlichkeit in dieser Arbeit allerdings nicht präsentiert werden. Im Falle des Wechselwirkungseffekts erhöht sich die Ausfallquote bei klei-

nen Stichprobenumfängen jedoch merklich.

6.2.2 Powersimulationen

Auf Grund der in der Niveausimulation dokumentierten Ausfallquoten wird im Rahmen der Powersimulation die Prävalenz $\pi = 0.1$ nicht mitsimuliert. Die resultierenden empirischen Powerkurven finden sich in Abbildung 6.7. Es lassen sich ähnliche Effekte wie auch schon bei der AUC-Simulation feststellen: Mit zunehmender Korrelation zwischen den Reader-Methoden-Kombinationen nimmt die Power aller Verfahren zu. Gleiches geschieht bei einer Erhöhung der Clustergröße. Im Falle der Sensitivität wird die zur Schätzung verwendete Stichprobengröße (n_1) über die Erkrankungswahrscheinlichkeit π kontrolliert. Je höher die Prävalenz der Erkrankung ist, umso mehr kranke Patienten nehmen in der Erwartung an der Studie teil, sodass mit erhöhter Erkrankungsprävalenz π die Power für die Analyse der Sensitivität zunimmt.[4] Es sei an dieser Stelle angemerkt, dass bewusst auch im Rahmen der Sensitivitäts- und Spezifitätssimulationen die Erkrankungsprävalenz π und nicht die Stichprobenumfänge n_i variiert wurden, da bei Fallzahlplanungen häufig die erwartete Prävalenz und nicht die zu realisierenden n_i als Grundlage der Studienplanung dienen. Ein Vergleich der verschiedenen simulierten Verfahren zeigt, dass das Verhalten aller Methoden im simulierten Bereich sehr ähnlich ist: Lediglich die auf den Konietschke-Brunner-Schätzern basierende Statistik weist bei hohen Fallzahlen minimal weniger Power als die in dieser Arbeit vorgestellten gewichteten und ungewichteten Schätzer auf. Wie auch bei schon bei der AUC zeigt sich, dass in der simulierten Parameterkonstellation die gewichteten Schätzer über etwas mehr Power verfügen als die ungewichteten, aber auch hier ist der Effekt marginal, sodass die Entscheidung für den gewichteten oder ungewichteten Schätzer eher auf Grundlage der später gewünschten Interpretation getroffen werden sollte.
Wie auch schon bei der Powersimulation der AUC zeigen sich ähnliche Resultate für den Reader- und den Wechselwirkungseffekt. Wie zu erwarten, zeigt die Interaktion dabei weniger Power als die Haupteffekte.

6.3 Prädiktive Werte

Das primäre Interesse der prädiktiven Werte liegt in ihrer intuitiven Interpretierbarkeit in der klinischen Praxis. Daher sind für prädiktive Werte Hypothesentests nur von untergeordnetem Interesse, das Hauptinteresse liegt hierbei vielmehr auf der Konstruktion von Konfidenzintervallen. Für die prädiktiven Werte werden daher anstelle der Teststatistiken die Länge und Überdeckungswahrscheinlichkeiten der Konfidenzintervalle simuliert. Hierbei werden sowohl die Konfidenzintervalle des linearen (Standard-) Modells evaluiert als auch die Konfidenzintervalle des logistischen Modells. Hierbei kann der Ansatz von Konietschke und Brunner (2009) nicht in ein logistisches Modell übersetzt werden, da die Schätzer nicht per Konstruktion im Intervall $[0,1]$ liegen und daher nicht transformiert werden können. Exemplarisch werden in dieser Simulationsstudie folgende Parameterkonstellationen angenommen:

(1) Die Sensitivität liegt konstant bei 0.7.
(2) Die Erkrankungsprävalenz in der Studie liegt bei $\pi = 0.5$.
(3) Die Erkrankungsprävalenz der Studie entspricht nicht der Erkrankungsprävalenz der späteren Anwendungspopulation. Diese liegt bei 0.7 und wird aus einer zusätzlichen Studie aus 500 Patienten geschätzt.
(4) Die Spezifität wird variiert: $sp \in \{0.5,\ 0.7,\ 0.8,\ 0.9\}$, wodurch sich die positiv prädiktiven Werte $p_+ \in \{0.77,\ 0.85,\ 0.89,\ 0.94\}$ ergeben.

[4] Mit einer Erhöhung der Erkrankungsprävalenz sinkt natürlich gleichermaßen die Anzahl an gesunden Patienten in der Studie, weswegen die durch die Veränderung der Prävalenz erreichte Powerzunahme im Bereich der Sensitivität unweigerlich zu einer Abnahme der Power im Bereich der Spezifitätsanalyse führt.

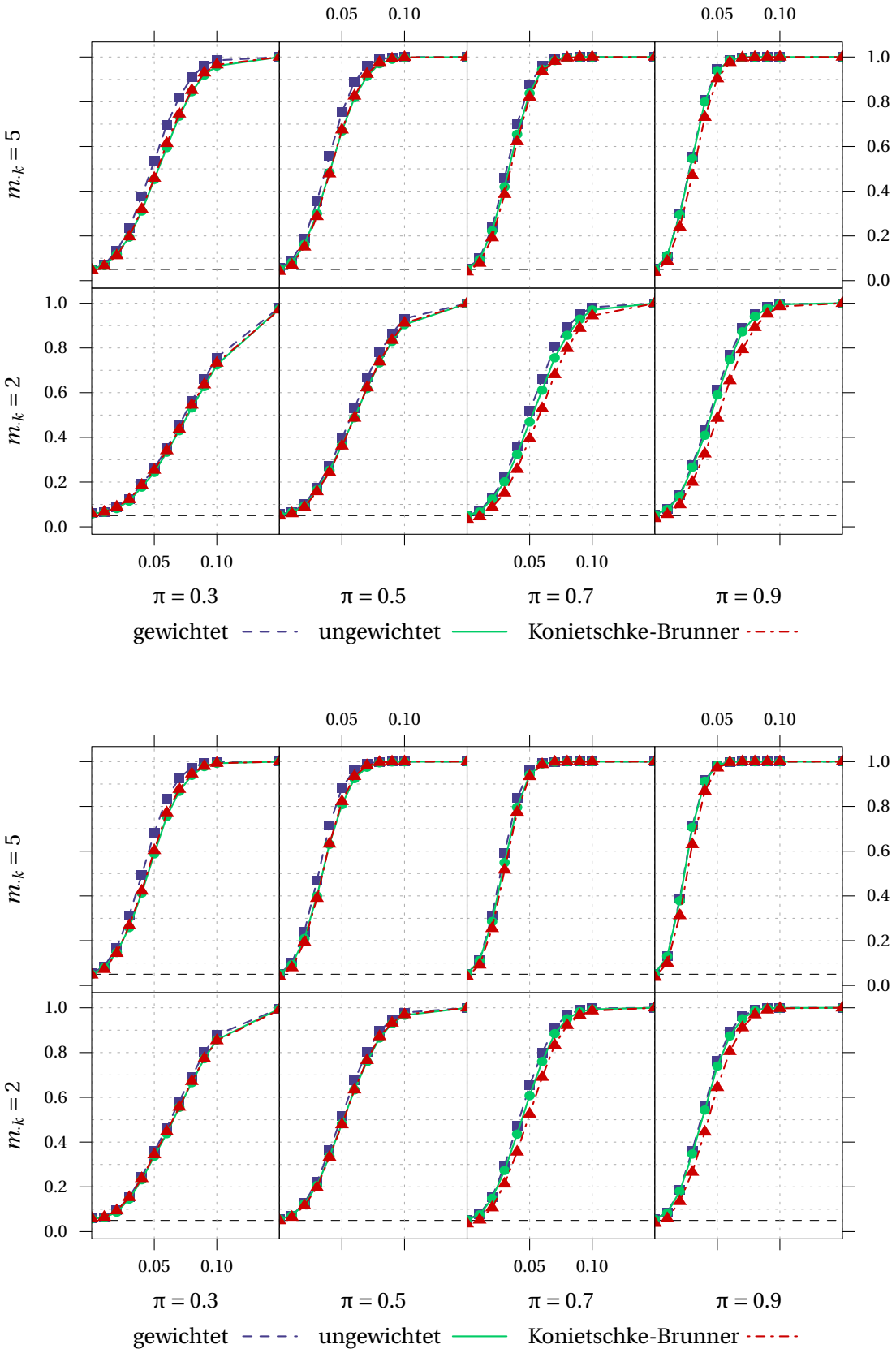

Abbildung 6.7: Powersimulation des Methodeneffekts der Sensitivität: Simulationen bei 5% nominellem Niveau und einem Stichprobenumfang von N = 100 für 2 Methoden und 3 Reader bei einer Clusterkorrelation von 0.25 / 0.5 (obere Graphik) und 0.5 / 0.75 (untere Graphik).

(5) Clusterkorrelation und Clustergröße werden wie bei den übrigen Simulationen variiert.

Es werden hierbei lediglich die positiv prädiktiven Werte simuliert, da zwischen positiv und negativ prädiktivem Wert ein ähnlicher Zusammenhang wie zwischen Sensitivität und Spezifität besteht. Die Simulationsstudie liefert die in den Abbildungen 6.8 und 6.9 dargestellten Resultate. Während die logit-transformierten Konfidenzintervalle stets eine Überdeckungswahrscheinlichkeit von etwa 95% aufweisen, zeigt sich, dass die Standardmethodik des linearen Modells im Bereich hoher Effekte (das heißt hoher prädiktiver Werte) liberal wird.

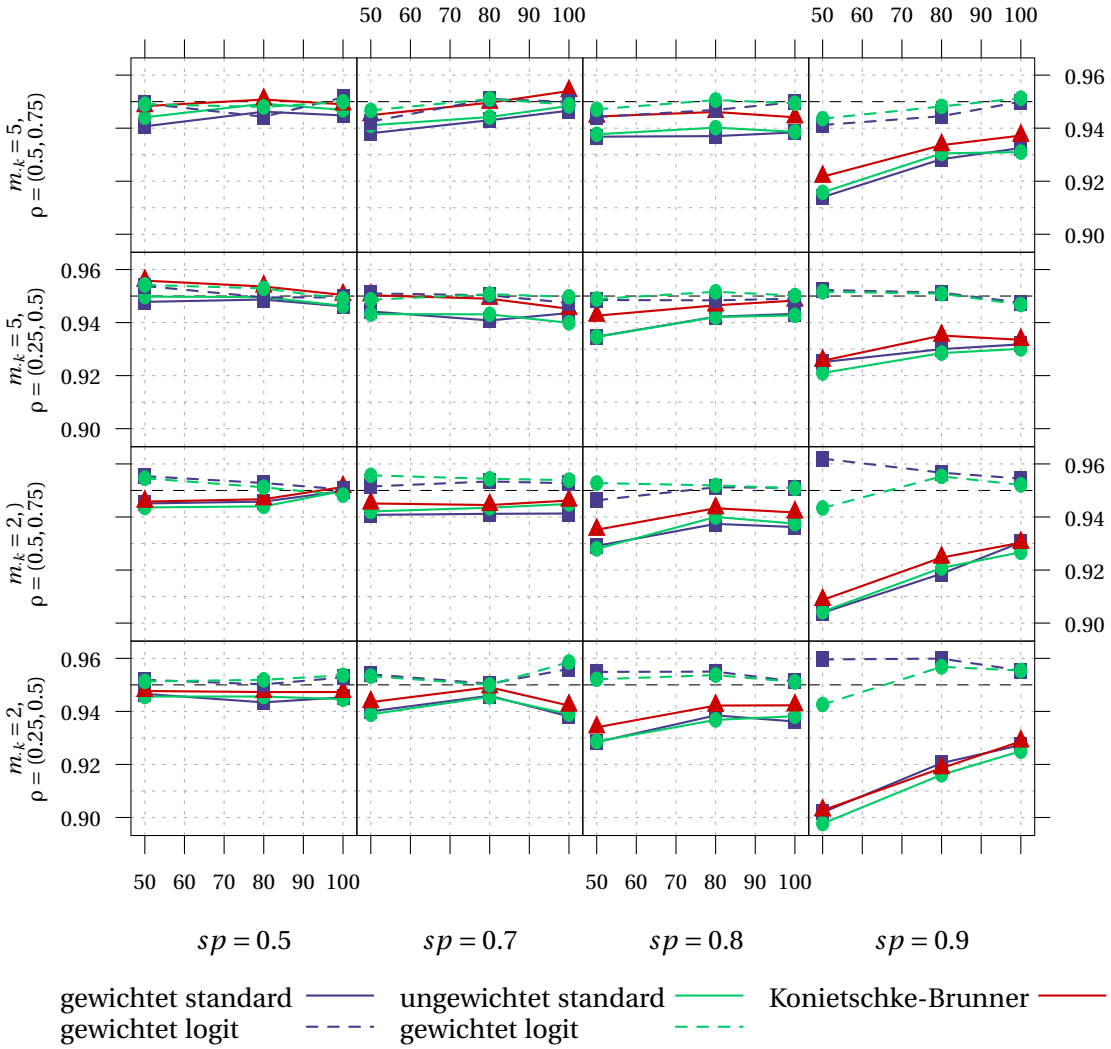

Abbildung 6.8: Coverageprobability der 95%-Konfidenzintervalle der positiv prädiktiven Werte: Die Sensitivität liegt in den Simulationen stets bei $se = 0.7$, die Prävalenz in der Diagnosestudie bei 0.5. Die Bevölkerungsprävalenz zur Berechnung der prädiktiven Werte wurde stets auf 0.7 gesetzt und mit Hilfe von 500 Patienten geschätzt.

Im Hinblick auf die Länge lässt sich erkennen, dass die Konfidenzintervalle mit ansteigender Clustergröße und größer werdendem Effekt kleiner werden. Die verschiedenen Schätzermethoden zeigen hierbei nur marginale Unterschiede in der Länge auf. Lediglich bei hohen prädiktiven Werten sind die logit-Konfidenzintervalle merklich länger als die Standard-Konfindenzintervalle, halten aber im Gegensatz zu diesen in diesem Bereich noch die Überdeckungswahrscheinlichkeit ein, sodass insbesondere bei hohen Werten für p_+

und p_- zu den logistischen Konfidenzinervallen zu raten ist.

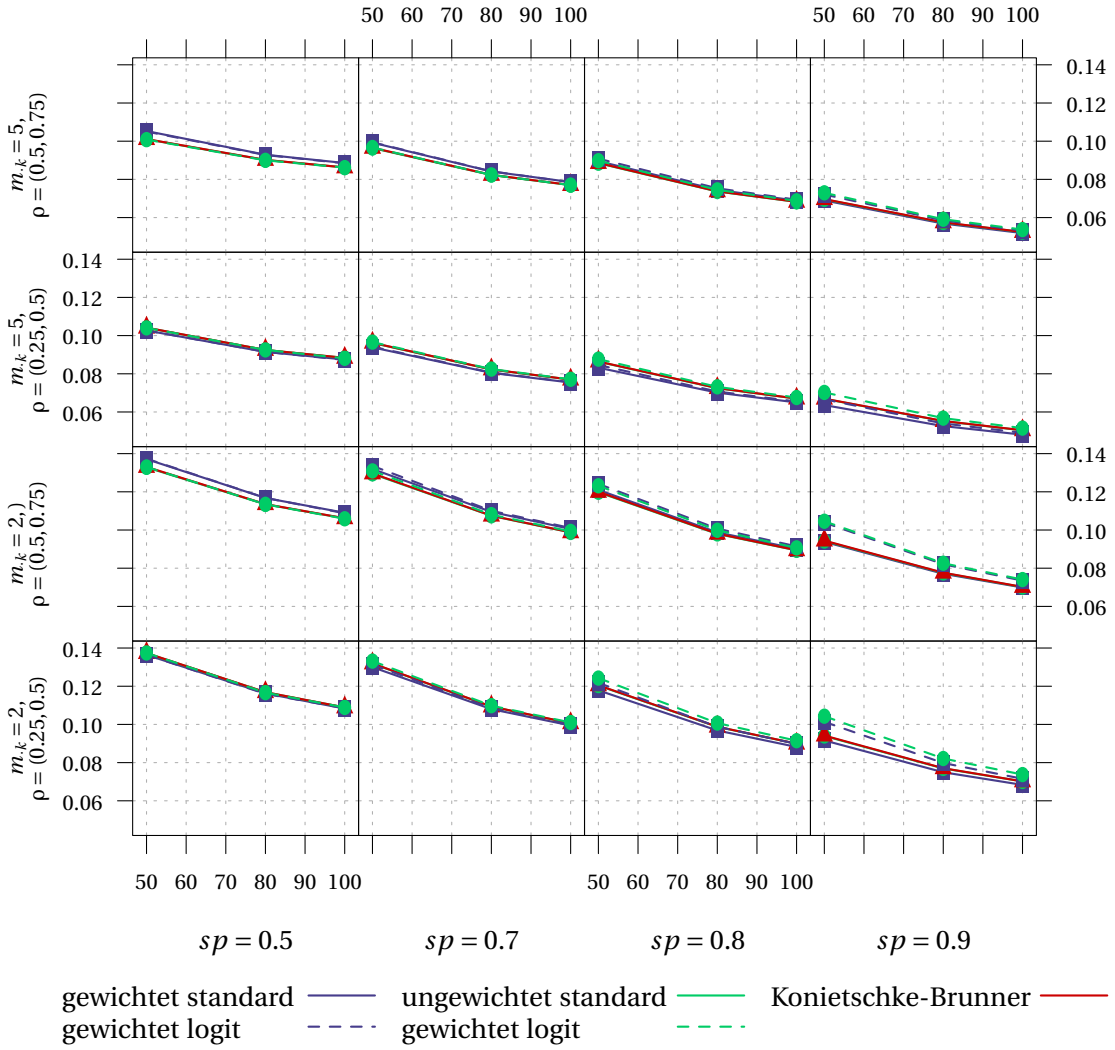

Abbildung 6.9: Länge der 95%-Konfidenzintervalle der positiv prädiktiven Werte: Die Sensitivität liegt in den Simulationen stets bei $se = 0.7$, die Prävalenz in der Diagnosestudie bei 0.5. Die Bevölkerungsprävalenz zur Berechnung der prädiktiven Werte wurde stets auf 0.7 gesetzt und mit Hilfe von 500 Patienten geschätzt.

7 Anwendungsbeispiele aus der klinischen Praxis

7.1 Thrombosediagnostik mittels farbkodierter Dopplersonographie [1]

Zur Diagnose von Thrombose im Bein oder Becken soll die Güte des Kontrastmittels Levovist mit der Güte einer Leeraufnahme in der farbkodierten Dopplersonographie verglichen werden. Dabei werden die sonographischen Bilder beider Methoden von zwei Readern evaluiert. Da beide Reader sowohl die Levovist- als auch die Leeraufnahme auswerten und alle Patienten mittels beider diagnostischer Methoden untersucht werden, entspricht die faktorielle Struktur dieser Studie dem in Kapitel 5 vorgestellten ersten Design. Durch das Vorliegen mehrerer potentiell gefährdeter Gefäße an einem Patienten entstehen in dieser Studie Clusterdaten, sodass an N = 48 Patienten insgesamt M = 77 Gefäße untersucht werden. Hierbei weisen $n^{(v)} = 5$ Patienten sowohl gesunde als auch kranke Gefäße auf, wohingegen $n_0^{(u)} = 23$ Patienten als vollständig gesund diagnostiziert werden und $n_1^{(u)} = 20$ Patienten nur kranke Gefäße aufweisen. Der Endpunkt beider diagnostischen Tests ist in diesem Fall ein Score von 1 (definitiv keine Thrombose) bis 5 (definitive Thrombose). Abbildung 7.1 zeigt die zugehörigen ROC-Kurven.

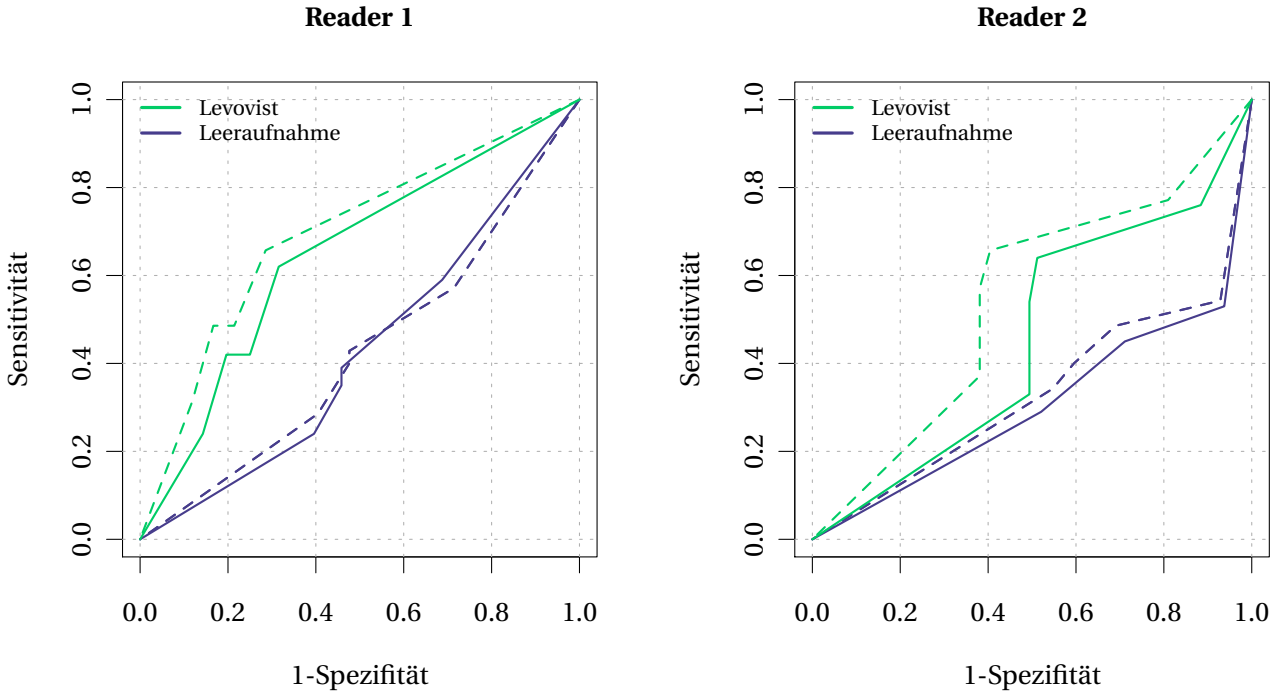

Abbildung 7.1: ROC-Kurven des Levovist Datensatzes: Die durchgezogenen Linien zeigen die ungewichtete Schätzung, die gestrichelten die gewichtete.

[1] Die Daten dieser Studie wurden freundlicherweise von Dr. Jörg Kaufmann von der Firma Bayer Schering zur Verfügung gestellt.

Aus den in Abbildung 7.1 dargestellten ROC-Kurven ergeben sich schließlich die in Abbildung 7.2 dargestellten AUCs.

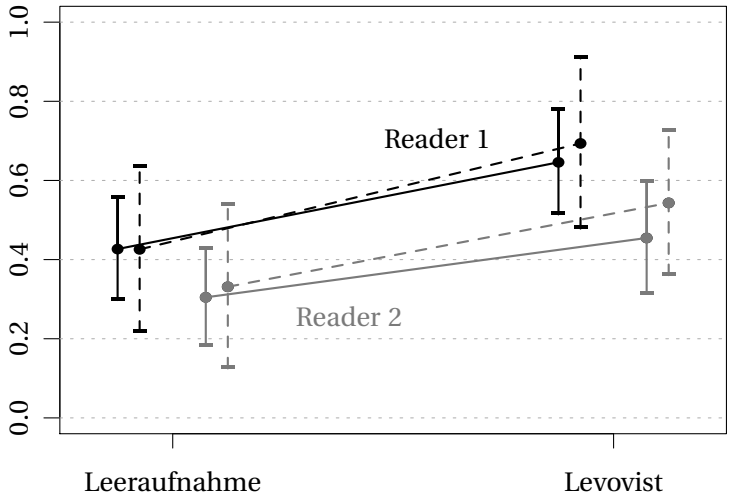

Abbildung 7.2: AUC-Schätzer mit 95%-Konfidenzintervallen der Levovist-Daten: Die durchgezogenen Linien zeigen die ungewichtete Schätzung, die gestrichelten die gewichtete.

Die statistischen Tests zeigen bei der AUC-Analyse schließlich, sowohl im ungewichteten als auch im gewichteten Fall, neben einem signifikanten Methodeneffekt ebenfalls einen signifikanten Einfluss des Readers (siehe Tabelle 7.1).

Tabelle 7.1: Ergebnisse der AUC-Analyse des Levovist-Datensatzes

	p-Werte ungewichtet	p-Werte gewichtet
Methode	0.0006	0.0017
Reader	0.0004	0.0014
Interaktion	0.3665	0.4061

Es lässt sich damit unabhängig von der Schätzmethode eine signifikante Verbesserung der diagnostischen Fähigkeit durch das Levovist nachweisen. Hierbei gibt es zwar Unterschiede in der Evalutionsfähigkeit beider Reader, allerdings ist die Zunahme der diagnostischen Güte durch das Levovist für beide Reader gleich, sodass von einer homogenen Verbesserung durch das Levovist gesprochen werden kann.
Im Folgenden sollen nun durch Wahl eines adäquaten Cut-Offs die Sensitivität und die Spezifität des Levovists im Vergleich zur Leeraufnahme bestimmt werden. Hierbei wird ein Patient als krank diagnostiziert, wenn er einen Score von 3, 4 oder 5 erhält und als gesund, falls er als 1 oder 2 klassifiziert wird (diese Klassifikation lässt sich beispielsweise durch einem Schwellenwert $\gamma = 2.5$ realisieren). Die sich hieraus ergebenden Sensitivitäten und Spezifitäten mit zugehörigen 95%-Konfidenzintervallen zeigt Abbildung 7.3 auf der folgenden Seite. Im Falle der Sensitivität kommt es bei Reader 1 nur zu einer leichten Verbesserung der diagnostischen Fähigkeit durch das Levovist, wohingegen sich bei Reader 2 deutlichere Unterschiede zeigen. Die Spezifität beider Reader hingegen scheint durch das Levovist homogen verbessert zu werden. Dabei zeigen sich sowohl im Hinblick auf Sensitivität als auch im Hinblick auf Spezifität deutliche Unterschiede hinsichtlich der diagnostischen Fähigkeit der beiden Reader. Diese deskriptiven Resultate werden durch die in Tabelle 7.2 (siehe nachfolgende Seite) dargestellten Ergebnisse untermauert.

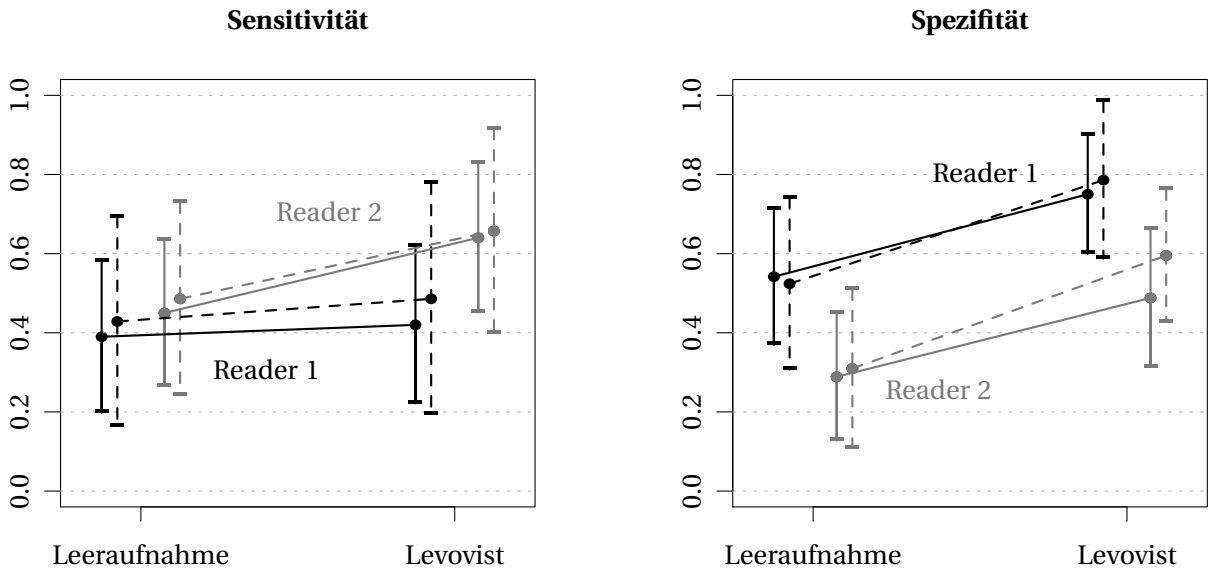

Abbildung 7.3: Schätzer von Sensitivität und Spezifität mit 95%-Konfidenzintervallen der Levovist-Daten: Die durchgezogenen Linien zeigen die ungewichtete Schätzung, die gestrichelten die gewichtete.

Tabelle 7.2: Ergebnisse der Analyse von Sensitivität und Spezifität des Levovist-Datensatzes

	Sensitivität		Spezifität	
	p-Werte ungewichtet	p-Werte gewichtet	p-Werte ungewichtet	p-Werte gewichtet
Methode	0.0345	0.0575	0.0017	0.0056
Reader	0.0460	0.0504	0.0002	0.0022
Interaktion	0.0532	0.1035	0.9410	0.8144

7.2 Brustkrebsdiagnostik mit und ohne CAD[2]

Brustkrebs ist die bei über 40-jährigen Frauen am häufigsten auftretende Krebsform (siehe zum Beispiel Mettin, 1999). Diagnosemethoden zur (Früh-)Erkennung von Brustkrebs sind somit ein wichtiges Forschungsgebiet im Bereich der bildgebenden Diagnostik. Die vorliegende Studie dient zur Evaluation des klinischen Nutzens eines CAD-Systems (Computer-Assisted Diagnosis) bei der Interpretation von Mammographiebildern. Im Rahmen dieser Studie untersuchen drei Reader jeweils 100 maligne Fälle von Brustkrebs (histologisch gesichert) sowie 103 Kontrollen (aus einer Screening-Population). Wie in der täglichen Routine üblich werden hierbei sowohl die linke als auch die recht Brust untersucht, wodurch an der gleichen Patientin mehrere Beobachtungseinheiten vorliegen und somit Clusterdaten auftreten. Details der Studienplanung und Durchführung finden sich bei Sohns u. a. (2010). Die Befundung der einzelnen Brust erfolgt dabei anhand der sogenannten BI-RADS[3] Kategorien 1 bis 5:

[2] Die Daten dieser Studie wurden durch die Abteilung Diagnostische Radiologie der Universität Göttingen erhoben und freundlicherweise von PD Dr. Silvia Obenauer zur Verfügung gestellt.

[3] Breast Imaging Reporting and Data System

1 Normalbefund
2 sicher benigner / gutartiger Befund
3 vermutlich gutartiger Befund
4 suspekter / verdächtiger Befund
5 Befund mit sehr hoher Wahrscheinlichkeit für Malignität

Die drei auswertenden Reader weisen hierbei einen unterschiedlichen Erfahrungshorizont auf: Reader 3 ist Facharzt mit mehrjähriger Mammographieerfahrung, Reader 1 ein Assistenzarzt mit etwa sechsmonatiger Mammographieerfahrung und Reader 2 ein Student ohne Erfahrung. Neben der Frage nach der Effektivität des CAD-Systems an sich stellt sich somit ebenfalls die Frage, ob das CAD-System allen Readern in gleichem Ausmaß behilflich sein kann oder ob die Güte der computergestützen Diagnose von der Erfahrung des Arztes abhängt. Da alle Patientinnen mit und ohne CAD-System befundet wurden, liegt auch in diesem Fall das erste faktorielle Design vor. Abbildung 7.4 zeigt die ROC-Kurven der drei Reader mit und ohne CAD.

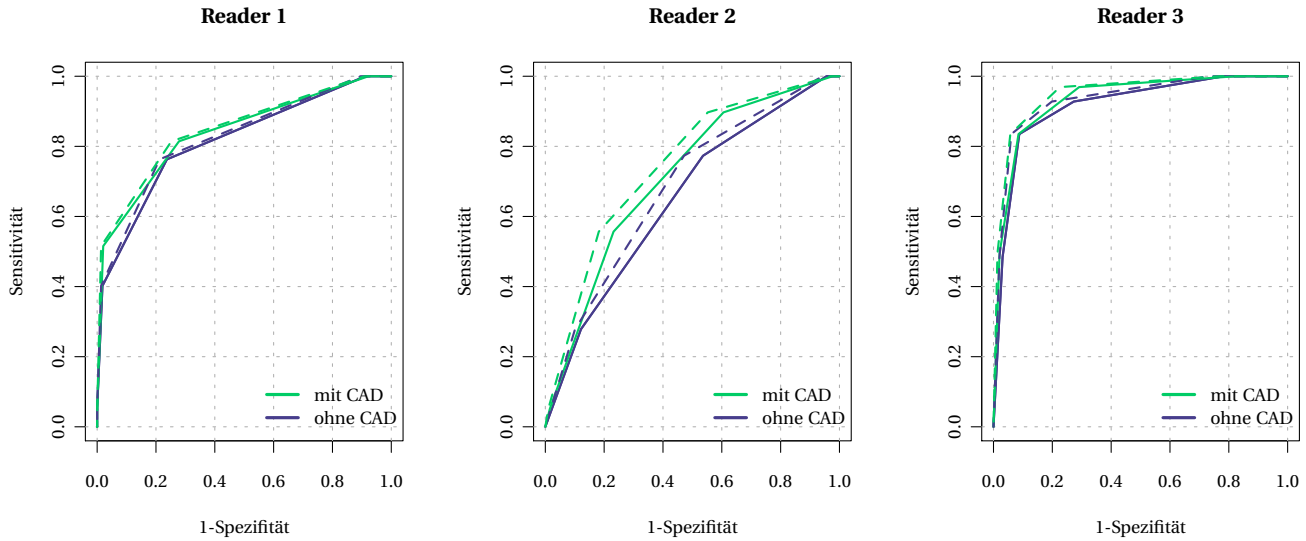

Abbildung 7.4: ROC-Kurven des CAD Datensatzes: Die durchgezogenen Linien zeigen die ungewichtete Schätzung, die gestrichelten die gewichtete.

Aus diesen ROC-Kurven ergeben sich schließlich die AUCs aus Abbildung 7.5 (siehe nächste Seite), sowie die in Tabelle 7.3 dargestellten Ergebnisse des statistischen Tests.
Wie bereits auf Grund der unterschiedlichen Erfahrungswerte der Reader zu vermuten war, lässt sich sowohl im gewichteten als auch im ungewichteten Fall ein deutlicher Unterschied in der diagnostischen Fähigkeit der drei Reader erkennen. Des Weiteren ist eine Zunahme der diagnostischen Güte durch das CAD-System festzustellen, eine signifikante Wechselwirkung zwischen Reader und Methode kann jedoch nicht nachgewiesen werden.

Tabelle 7.3: Ergebnisse der AUC-Analyse des CAD-Datensatzes

	p-Werte ungewichtet	p-Werte gewichtet
Methode	0.0023	0.0014
Reader	<0.0001	<0.0001
Interaktion	0.1244	0.0612

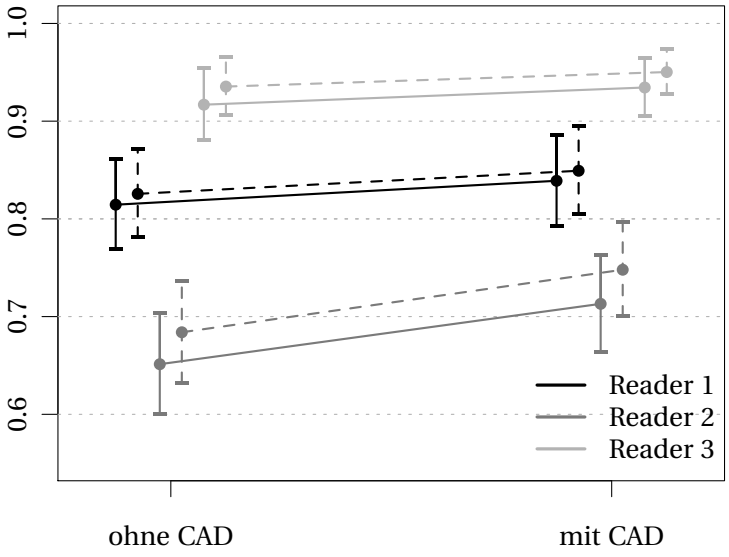

Abbildung 7.5: AUC-Schätzer mit 95%-Konfidenzintervallen der CAD-Daten: Die durchgezogenen Linien zeigen die ungewichtete Schätzung, die gestrichelten die gewichtete.

In einem weiteren Schritt wird die diagnostische Güte nun mittels Sensitivität und Spezifität evaluiert. Zu diesem Zweck wird im Folgenden $\gamma = 3.5$ als Cut-Off gewählt. Befunde der BI-RADS-Kategorien 4 oder 5 werden somit als positiv eingestuft, die übrigen als negativ. Abbildung 7.6 zeigt die hieraus resultierenden Schätzer für Sensitivität und Spezifität, sowie die zugehörigen 95%- Konfidenzintervalle.

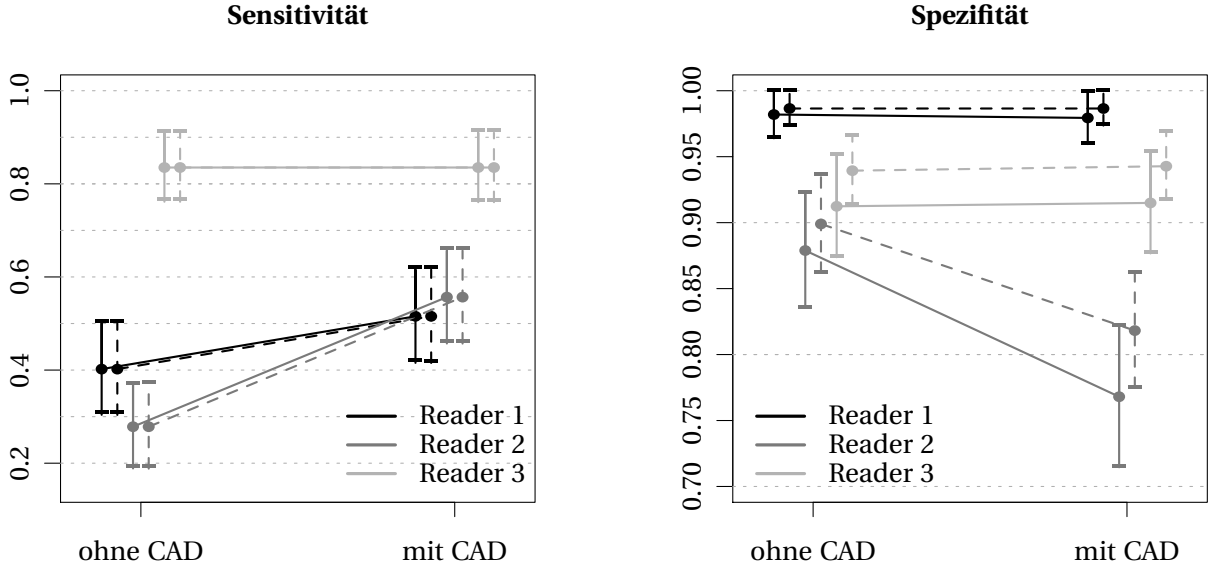

Abbildung 7.6: Schätzer von Sensitivität und Spezifität mit 95%-Konfidenzintervallen der CAD-Daten: Die durchgezogenen Linien zeigen die ungewichtete Schätzung, die gestrichelten die gewichtete.

Im Gegensatz zur AUC-Analyse sind bei der Evaluation von Sensitivität und Spezifität Wechselwirkungen zwischen Reader und Methode zu erkennen: Während für den Facharzt mit mehrjähriger Berufserfahrung

die Verwendung des CAD-Systems weder zu einer Verbesserung von Sensitivität oder Spezifität zu führen scheint, führt das CAD-System zu einer Verbesserung der Sensitivität für den Assistenzarzt und den Studenten. Bei der Analyse der Spezifität zeigt sich für den Assistenz- und Facharzt keine Verbesserung der Diagnostik durch das CAD-System. Für den Studenten hingegen verschlechtert sich die Spezifität: mit CAD-System stellt dieser mehr falsch positive Diagnosen als ohne. Diese Interpretationen der Ergebnisse spiegeln sich auch in den Resultaten in Tabelle 7.4 wieder.

Tabelle 7.4: Ergebnisse der Analyse von Sensitivität und Spezifität des CAD-Datensatzes

	Sensitivität		Spezifität	
	p-Werte ungewichtet	p-Werte gewichtet	p-Werte ungewichtet	p-Werte gewichtet
Methode	<0.0001	<0.0001	0.0006	0.0016
Reader	<0.0001	<0.0001	<0.0001	<0.0001
Interaktion	<0.0001	<0.0001	<0.0001	<0.0001

In einem letzten Analyseschritt werden nun die prädiktiven Werte der beiden diagnostischen Verfahren evaluiert. Da die drei Reader sich deutlich unterscheiden, erfolgt diese Betrachtung nur exemplarisch für Reader 1, den Assistenzarzt. Die Brustkrebsprävalenz nimmt ab einem Alter von 50 Jahren deutlich zu. Auf Grund dessen werden in dieser Arbeit prädiktive Werte für zwei Risikogruppen unterschieden: Frauen im Alter von 40-49 auf der einen Seite und Frauen älter als 50 Jahre auf der anderen Seite. Betrachtet wird hierbei eine normale Screening-Population, für die keine Informationen über weitere Risikofaktoren vorliegen. Die Schätzung der Brustkrebsprävalenz einer solchen Screening-Population befindet sich in Tabelle 7.5 und geht zurück auf Tabar u. a. (1995).[4]

Tabelle 7.5: Brustkrebsprävalenz einer Screening-Population nach Tabar u. a. (1995)

Alter	Anzahl der Erkrankten	Gesamtanzahl	Prävalenz
40-49	418	35448	0.0118
50-74	2049	97677	0.0210

Aus den obigen Prävalenzschätzern erhält man schließlich die in Abbildung 7.7 dargestellten prädiktiven Werte für Reader 1.

Die positiv prädiktiven Werte in einer einfachen Screening-Population sind auf Grund der niedrigen Erkrankungsprävalenz relativ gering, sodass weder mit noch ohne CAD ausschließlich mittels Mammographie diagnostiziert werden sollte. Die BI-RADS-Klassifikation sieht daher im Fall einer positiven Diagnose zunächst eine histologische Sicherung des Befundes durch eine Biopsie vor, bevor weitere Maßnahmen unternommen werden. Die negativ prädiktiven Werte hingegen scheinen relativ hoch, dieses ist jedoch weitestgehend auf die geringe Prävalenz der Erkrankung zurückzuführen, wie Abbildung 7.8 zeigt. Die hier dargestellten prädiktiven Werte als Funktionen der Prävalenz zeigen, dass $p_-(\pi)$ nur wenig größer als $1-\pi$ ist, dass sich die Post-Test-Erkrankungswahrscheinlichkeit nach einer negativen Diagnose also kaum von der Prä-Test-Erkrankungswahrscheinlichkeit unterscheidet. Es sei an dieser Stelle angemerkt, dass die Unterschiede

[4] Da das Studiendesign dieser Studie dem einer Fall-Kontroll-Studie entspricht, kann die Brustkrebsprävalenz keinesfalls aus den Studiendaten geschätzt werden. Zur Berechnung der prädiktiven Werte muss in diesem Fall eine studienexterne Prävalenzschätzung durchgeführt werden.

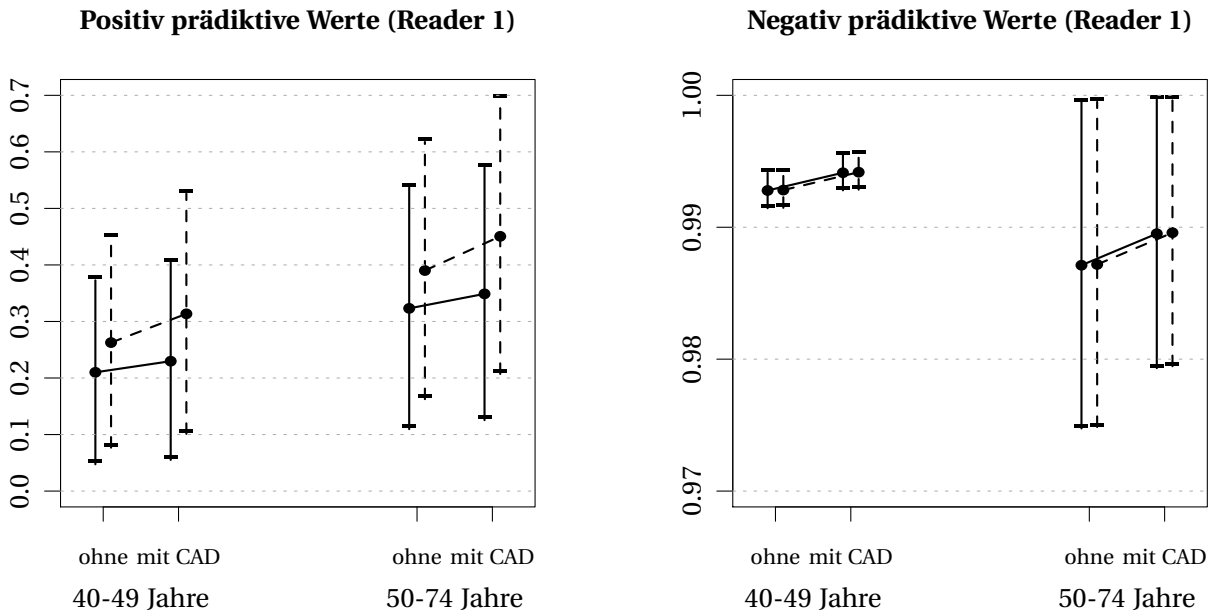

Abbildung 7.7: Schätzer der prädiktiven Werte einer Screening-Population mit 95%-Konfidenzintervallen: Die durchgezogenen Linien zeigen die ungewichtete Schätzung, die gestrichelten die gewichtete.

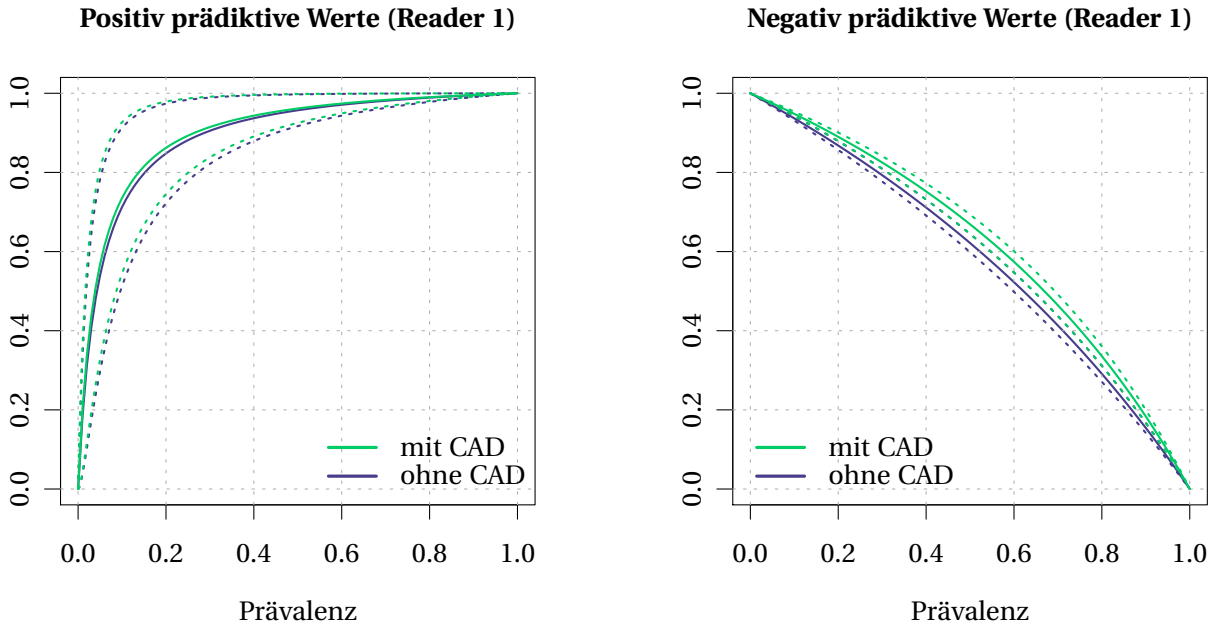

Abbildung 7.8: Darstellung der prädiktiven Werte (des CAD-Systems hier nur die ungewichteten Schätzer) als Funktion der Prävalenz mit den zugehörigen 95%-Konfidenzintervallen

zwischen Prä- und Post-Test-Erkrankunswahrscheinlichkeit am besten durch die sogenannten Likelihoodratios beschrieben werden, die allerdings nicht Thema dieser Arbeit sein sollen. Eine exakte Definition und Interpretation der Likelihoodratios findet man beispielsweise in den Standardwerken von Pepe (2003) oder Zhou u. a. (2002).

8 Zusammenfassung und Ausblick

Der Vielzahl unterschiedlicher analytischer Konzepte zur Evaluation von Diagnosestudien wurde mit dieser Arbeit ein einheitlicher, nichtparametrischer Analyseansatz für AUC, Sensitivität, Spezifität und prädiktive Werte in Studien mit und ohne Clusterdaten gegenübergestellt. Die dargestellte Theorie wurde dabei exemplarisch zur Herleitung einer Analysemethode für die vier Basisdesigns diagnostischer Studien verwendet, um auf diese Art und Weise ein weites Anwendungsgebiet für die präsentierte Methodik zu schaffen. Entwickelt wurden dabei lineare und logistische Modelle.

Likelihoodratios stellen neben den hier präsentierten Gütemaßen eine weitere Möglichkeit zur Charakterisierung eines diagnostischen Verfahrens dar (siehe zum Beispiel Zhou u. a., 2002, Abschnitt 2.7). Sie sind definiert als Quotient aus Sensitivität und 1−Spezifität (positiver Likelihoodratio) beziehungsweise als $\frac{1-se}{sp}$ (negativer Likelihoodratio). Mit dem hier präsentierten Ansatz der δ-Methode zur Analyse prädiktiver Werte lässt sich die dargestellte Methodik gleichsam auf die faktorielle Analyse von Likelihoodratios erweitern.

In diagnostischen Studien können neben den in dieser Arbeit bereits modellierten Faktoren Reader und Methode auch weitere Parameter, wie beispielsweise das Körpergewicht oder das Alter des Patienten Einfluss auf die diagnostische Güte des Verfahrens haben. Eine Erweiterung des hier vorgestellten Analyseansatzes auf Kovariablen wäre daher wünschenswert. Ansätze zur Berücksichtigung von Kovariablen in diagnostischen Studien findet man beispielsweise bei Janes und Pepe (2008) und Janes u. a. (2008), die mittels eines Regressionsansatzes eine auf Kovariablen adjustierte ROC-Kurve schätzen und die zu den adjustierten ROC-Kurven gehörigen AUCs schließlich vergleichen. Zapf (2009) geht das Problem in umgekehrter Art und Weise an, indem sie – basierend auf der Theorie der multivariaten Rangstatistiken – nicht die ROC-Kurve selbst, sondern die AUC direkt auf Kovariablen adjustiert. Dieser Ansatz passt dabei konzeptionell in die in dieser Arbeit präsentierte Methodik, sodass sich durch eine Erweiterung der Ideen von Zapf (2009) auf Ein-Punkt-Verteilungen und Clusterdaten der Anwendungsbereich dieser Arbeit auf Kovariablen weiterentwickeln ließe.

Konietschke (2009) präsentiert einen Ansatz zur Berechnung simultaner Konfidenzintervalle für den nichtparametrischen relativen Effekt. Da dieser gerade der Fläche unter der ROC-Kurve entspricht, wäre eine Erweiterung der Ideen von Konietschke (2009) auf Ein-Punkt-Verteilungen, Clusterdaten und faktorielle Versuchsanlagen erstrebenswert, damit auch im Bereich von faktoriellen Diagnosestudien simultane Konfidenzintervalle für die verschiedenen diagnostischen Gütemaße berechnet werden können.

Da die Datenerhebung in der klinischen Praxis häufig lückenhaft ist, stellen fehlende Werte eine zusätzliche Herausforderung an die statistische Methodik. Zhou und Gatsonis (1996) entwickeln einen Ansatz zur AUC-Evaluation im verbundenen Zwei-Stichproben-Design mit fehlenden Werten und auch Zapf (2009) stellt im Ausblick ihrer Arbeit eine Möglichkeit der Erweiterung ihrer Verfahren auf fehlende Werte vor. Eine derartige Weiterentwicklung der hier präsentierten Methoden wäre ebenfalls wünschenswert, um den Anwendungsbereich der vorgestellten einheitlichen Analysemethode noch ausbauen zu können.

A Definitionen, Sätze und Notationen

A.1 Matrizenrechnung

Matrizen und Vektoren werden stets **fett** geschrieben. Eine Matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ hat dabei die Gestalt:

$$\mathbf{A} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}$$

Die zu $\mathbf{A}$ transponierte Matrix wird mit $\mathbf{A}'$ bezeichnet.

Definition A.1 *[Spezielle Matrizen]*

1. ***Einheitsmatrix***

$$\mathbf{I}_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}$$

2. ***Einservektor***

$$\mathbf{1}_n = (1, ..., 1)'_{1 \times n}$$

3. $\mathbf{n} \times \mathbf{n}$ ***-Einser-Matrix***

$$\mathbf{J}_n = \mathbf{1}_n \mathbf{1}'_n = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}$$

4. ***Zentrierende Matrix***

$$\mathbf{P}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n = \begin{pmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & & \vdots \\ \vdots & & \ddots & -\frac{1}{n} \\ -\frac{1}{n} & \dots & -\frac{1}{n} & 1 - \frac{1}{n} \end{pmatrix}$$

Definition A.2 *[Spur] Für eine quadratische Matrix* $\mathbf{A} \in \mathbb{R}^{n \times n}$ *heißt* $\mathrm{Sp}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$ *die Spur von* $\mathbf{A}$*.*

Definition A.3 *[Rang] Für eine Matrix* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *heißt die Anzahl der linear unabhängigen Zeilenvektoren* $\mathrm{r}(\mathbf{A})$ *der Rang von* $\mathbf{A}$*.*

Definition A.4 *[Kronecker-Summe] 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* $\mathbf{B}$.

Definition A.5 *[Kronecker-Produkt] Für beliebige Matrizen* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *und* $\mathbf{B} \in \mathbb{R}^{p \times q}$ *heißt*

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & & \vdots \\ a_{m1}\mathbf{B} & & a_{mn}\mathbf{B} \end{pmatrix} \in \mathbb{R}^{mp \times nq}$$

Kronecker-Produkt von $\mathbf{A}$ *und* $\mathbf{B}$.

Definition A.6 *[Verallgemeinerte Inverse] Für eine beliebige Matrix* $\mathbf{A}$ *heißt* $\mathbf{A}^-$ *verallgemeinerte Inverse zu* $\mathbf{A}$, *falls* $\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$ *ist. Weiter heißt* $\mathbf{A}$ *reflexive verallgemeinerte Inverse zu* $\mathbf{A}$, *falls* $\mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}^-$ *gilt.*

A.2 Wahrscheinlichkeitstheorie

Definition A.7 *[Verteilungsfunktion] Für eine Zufallsvariable* X *heißt*

$$\begin{aligned} \mathrm{F}^-(x) &= \mathrm{P}(\mathrm{X} < x) \quad \textit{links-stetige} \\ \mathrm{F}^+(x) &= \mathrm{P}(\mathrm{X} \le x) \quad \textit{rechts-stetige} \\ \mathrm{F}(x) &= \frac{1}{2}[\mathrm{F}^+(x) + \mathrm{F}^-(x)] \quad \textit{normalisierte} \end{aligned}$$

Version der Verteilungsfunktion.

Definition A.8 *[Zählfunktion] Die Funktion*

$$\begin{aligned} c^-(x) &= \begin{cases} 0, & x \le 0 \\ 1, & x > 0 \end{cases} \quad \textit{heißt links-stetige} \\ c^+(x) &= \begin{cases} 0, & x < 0 \\ 1, & x \ge 0 \end{cases} \quad \textit{heißt rechts-stetige} \\ c(x) &= \frac{1}{2}[c^+(x) + c^-(x)] \quad \textit{heißt normalisierte} \end{aligned}$$

Version der Zählfunktion.

Definition A.9 *[Mittelränge] Es sei* $c(x)$ *definiert wie in Definition A.8, ferner seien* $x_1, \dots, x_\mathrm{N}$ *beliebige reelle Zahlen, dann heißt*

$$r_i = \frac{1}{2} + \sum_{i=1}^{\mathrm{N}} c(x_i - x_j)$$

der Mittelrang von x_i *unter allen Zahlen* $x_1, \dots, x_\mathrm{N}$.

Definition A.10 *[Asymptotische Äquivalenz] Zwei Folgen von Zufallsvariablen* $\mathrm{X_N}$, $\mathrm{Y_N}$ *heißen asymptotisch äquivalent* ($\mathrm{X_N} \doteqdot \mathrm{Y_N}$), *wenn* $\forall \epsilon > 0$ *gilt* $\lim_{\mathrm{N}\to\infty} \mathrm{P}(|\mathrm{X_N} - \mathrm{Y_N}| > \epsilon) = 0$. *Gilt* $\mathrm{Y_N} \sim \mathrm{F_Y}$ *und* $\mathrm{X_N} \doteqdot \mathrm{Y_N}$ *so wird verkürzend* $\mathrm{X_N} \mathrel{\dot{\sim}} \mathrm{F_Y}$ *geschrieben.*

Satz A.1 (Slutzky) *1. Sei* $\mathbf{X}_N \in \mathbb{R}^k$, $N \geq 1$, *eine Folge von Zufallsvariablen mit* $\mathbf{X}_N \xrightarrow{p} a$, *wobei* $\mathbf{a} \in \mathbb{R}^k$ *konstant ist und sei ferner* $g(\cdot)$ *stetig in* $\mathbf{a}$, *dann gilt*

$$g(\mathbf{X}_N) \xrightarrow{p} g(\mathbf{a})$$

2. Sei $\mathbf{X}_N \in \mathbb{R}^k$, $N \geq 1$, *eine Folge von Zufallsvariablen mit* $\mathbf{X}_N \xrightarrow{\mathscr{L}} \mathbf{X} \sim \mathbf{F}(\mathbf{x})$ *und sei* $g(\cdot)$ *ein* F*-fast-überall stetige Funktion*

$$g(\mathbf{X}_N) \xrightarrow{\mathscr{L}} g(\mathbf{X}) \sim \mathbf{F}_g(\mathbf{x})$$

Beweis: Siehe Slutsky (1925).

Satz A.2 (Cramer) *Die Abbildung* $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^k$ *sei auf einer Umgebung von* $\mu \in \mathbb{R}^d$ *stetig differenzierbar. Es sei* $\mathbf{X}_N$ *eine Folge von d-dimensionalen Zufallsvektoren mit* $\sqrt{N}(\mathbf{X}_N - \mu) \xrightarrow{\mathscr{L}} \mathbf{X}$. *Dann gilt* $\sqrt{N}\left[\mathbf{g}(\mathbf{X_N}) - \mathbf{g}(\mu)\right] \xrightarrow{\mathscr{L}} \mathbf{Dg}(\mu)\mathbf{X}$. *Insbesondere gilt, falls* $\sqrt{N}(\mathbf{X}_N - \mu) \xrightarrow{\mathscr{L}} \mathbf{U} \sim N(\mathbf{0}, \Sigma)$*:*

$$\sqrt{N}\left[\mathbf{g}(\mathbf{X_N}) - \mathbf{g}(\mu)\right] \xrightarrow{\mathscr{L}} \mathbf{V} \sim N(\mathbf{0}, \mathbf{Dg}(\mu)\Sigma\mathbf{Dg}(\mu)')$$

BEWEIS. Siehe zum Beispiel Ferguson (1996, Kapitel 7, Theorem 7). □

Satz A.3 (δ-Satz) *Es sei* $\phi \in \mathbb{R}$ *eine Konstante und* T_N *eine Folge von Zufallsvariablen, sodass für* $N \to \infty$ *gilt* $\sqrt{N}(T_N - \phi) \xrightarrow{\mathscr{L}} T$. *Ferner sei die Funktion* $g(t)$ *an der Stelle* ϕ *stetig differenzierbar. Dann gilt:*

$$\sqrt{N}\left[g(T_N) - g(\phi)\right] \xrightarrow{\mathscr{L}} g'(\phi) \cdot T$$

BEWEIS. Der Beweis ist ein Spezialfall des Satzes von Cramer. □

Satz A.4 (Jensen-Ungleichung) *Sei* X *eine Zufallsvariable mit Verteilungsfunktion* $F(x)$ *und* $E(X) < \infty$ *und sei* $g(x)$ *eine konvexe Funktion. Dann gilt*

$$g\left[E(X)\right] \leq E\left[g(X)\right]$$

oder $$g\left[\int x dF(x)\right] \leq \int g(x) dF(x).$$

BEWEIS. Siehe Jensen (1906). □

Satz A.5 (c_r-Ungleichung) *Für beliebige (abhängige oder unabhängige) Zufallsvariablen* X *und* Y *gilt*

$$E|X+Y|^r \leq c_r\left[E(|X|^r) + E(|Y|^r)\right], \qquad c_r = \begin{cases} 1 & 0 < r \leq 1 \\ 2^{r-1} & r > 1. \end{cases}$$

Insbesondere gilt für $r = 2$ *die Ungleichung* $E(X+Y)^2 \leq 2E(X^2) + 2E(Y^2)$.

BEWEIS. Siehe zum Beispiel Gut (2005, Kapitel 3, Theorem 2.2). □

B Beweise

B.1 Asymptotische Verteilung des gewichteten AUC-Schätzer

B.1.1 Asymptotische Äquivalenz des gewichteten AUC-Schätzer

Satz B.1 *Es sei*

$$\widehat{\mathrm{B}}^{(l)} = \int \mathrm{F}_0^{(l)} d\widehat{\mathrm{F}}_1^{(l)} - \int \mathrm{F}_1^{(l)} d\widehat{\mathrm{F}}_0^{(l)} + 1 - 2 \cdot \int \mathrm{F}_0^{(l)} d\mathrm{F}_1^{(l)},\ l = 1, \ldots, d$$

und es sei $\widehat{\mathbf{B}} = (\widehat{\mathrm{B}}^{(1)}, \ldots, \widehat{\mathrm{B}}^{(d)})'$*, dann gilt:*

$$\left\| \sqrt{\mathrm{N}} \left(\widehat{\mathbf{AUC}} - \mathbf{AUC} \right) - \sqrt{\mathrm{N}} \widehat{\mathbf{B}} \right\|_2 \to \mathbf{0}.$$

BEWEIS. Es genügt zu zeigen, dass $\mathrm{E}\left[\sqrt{\mathrm{N}} \left(\widehat{\mathrm{AUC}}^{(l)} - \mathrm{AUC}^{(l)} \right) - \sqrt{\mathrm{N}} \widehat{\mathrm{B}}^{(l)} \right]^2 \to 0$, $\forall l = 1, \ldots, d$. Es gilt (beispielsweise nachzulesen bei Brunner u. a., 2002, Beweis zu Theorem 3.3.):

$$\begin{aligned} \sqrt{\mathrm{N}} \left(\widehat{\mathrm{AUC}}^{(l)} - \mathrm{AUC}^{(l)} \right) &= \sqrt{\mathrm{N}} \widehat{\mathrm{B}}^{(l)} + \sqrt{\mathrm{N}} \int \left(\widehat{\mathrm{F}}_0^{(l)} - \mathrm{F}_0^{(l)} \right) d \left(\widehat{\mathrm{F}}_1^{(l)} - \mathrm{F}_1^{(l)} \right) \\ &= \sqrt{\mathrm{N}} \widehat{\mathrm{B}}^{(l)} + \sqrt{\mathrm{N}} \widehat{\mathrm{A}}^{(l)}. \end{aligned}$$

Das heißt, es muss $\mathrm{E}(\sqrt{\mathrm{N}} \widehat{\mathrm{A}}^{(l)})^2 \to 0$ gezeigt werden. Nun gilt

$$\begin{aligned} \mathrm{E}(\sqrt{\mathrm{N}} \widehat{\mathrm{A}}^{(l)})^2 &= \mathrm{E} \left[\sqrt{\mathrm{N}} \int \left(\widehat{\mathrm{F}}_0^{(l)} - \mathrm{F}_0^{(l)} \right) d \left(\widehat{\mathrm{F}}_1^{(l)} - \mathrm{F}_1^{(l)} \right) \right]^2 \\ &= \mathrm{E} \left[\sqrt{\mathrm{N}} \sum_{t \in \{u,v\}} \sum_{t' \in \{u,v\}} \frac{n_0^{(t)}}{n_0^{(u)} + n^{(v)}} \frac{n_1^{(t')}}{n_1^{(u)} + n^{(v)}} \int \left(\widehat{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)} \right) d \left(\widehat{\mathrm{F}}_1^{(l|t')} - \mathrm{F}_1^{(l)} \right) \right]^2 \\ &\leq 4 \sum_{t \in \{u,v\}} \sum_{t' \in \{u,v\}} \left(\frac{n_0^{(t)}}{n_0} \frac{n_1^{(t')}}{n_1} \right)^2 \mathrm{E} \left[\sqrt{\mathrm{N}} \int \left(\widehat{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)} \right) d \left(\widehat{\mathrm{F}}_1^{(l|t')} - \mathrm{F}_1^{(l)} \right) \right]^2. \end{aligned} \tag{B.1}$$

Betrachte daher zunächst $\mathrm{E}\left[\sqrt{\mathrm{N}} \int \left(\widehat{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)} \right) d \left(\widehat{\mathrm{F}}_1^{(l|t')} - \mathrm{F}_1^{(l)} \right) \right]^2 = \mathrm{E}\left[\sqrt{\mathrm{N}} \widehat{\mathrm{A}}_{tt'}^{(l)} \right]^2$. Es gilt

$$\begin{aligned} \mathrm{E} \left[\sqrt{\mathrm{N}} \widehat{\mathrm{A}}_{tt'}^{(l)} \right]^2 &= \mathrm{N} \cdot \mathrm{E} \left[\int \left(\widehat{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)} \right) d\widehat{\mathrm{F}}_1^{(l|t')} - \int \left(\widehat{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)} \right) d\mathrm{F}_1^{(l)} \right]^2 \\ &= \mathrm{N} \cdot \mathrm{E} \left[\frac{1}{n_1^{(t')}} \sum_{k=1}^{n_1^{(t')}} \frac{1}{m_{1k}^{(l|t')}} \sum_{s=1}^{m_{1k}^{(l|t')}} \left(\widehat{\mathrm{F}}_0^{(l|t)}(\mathrm{X}_{1ks}^{(l|t')}) - \mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')}) \right) - \int \left(\widehat{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)} \right) d\mathrm{F}_1^{(l)} \right]^2 \end{aligned}$$

$$= \mathrm{N} \cdot \mathrm{E}\left[\frac{1}{n_1^{(t')}} \sum_{k=1}^{n_1^{(t')}} \frac{1}{m_{1k}^{(l|t')}} \sum_{s=1}^{m_{1k}^{(l|t')}} \left(\frac{1}{n_0^{(t)}} \sum_{k'=1}^{n_0^{(t)}} \frac{1}{m_{0k'}^{(l|t)}} \sum_{s'=1}^{m_{0k'}^{(l|t)}} c(\mathrm{X}_{1ks}^{(l|t')} - \mathrm{X}_{0k's'}^{(l|t)}) - \mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')}) \right) \right.$$
$$\left. - \left(\frac{1}{n_0^{(t)}} \sum_{k'=1}^{n_0^{(t)}} \frac{1}{m_{0k'}^{(l|t)}} \sum_{s'=1}^{m_{0k'}^{(l|t)}} \int \left(c(x - \mathrm{X}_{0k's'}^{(l|t)}) - \mathrm{F}_0^{(l)}(x) \right) d\mathrm{F}_1^{(l)}(x) \right) \right]^2.$$

Zur besseren Lesbarkeit bezeichnen im Folgenden

$$\varphi_1\left(\mathrm{X}_{1ks}^{(l|t')}, \mathrm{X}_{0k's'}^{(l|t)}\right) = c(\mathrm{X}_{1ks}^{(l|t')} - \mathrm{X}_{0k's'}^{(l|t)}) - \mathrm{F}_0^{(l)}(\mathrm{X}_{1ks}^{(l|t')})$$
$$\varphi_2\left(\mathrm{X}_{0k's'}^{(l|t)}\right) = \int \left(c(x - \mathrm{X}_{0k's'}^{(l|t)}) - \mathrm{F}_0^{(l)}(x) \right) d\mathrm{F}_1^{(l)}(x).$$

Mit dieser Notation gilt:

$$\mathrm{E}\left[\sqrt{\mathrm{N}}\widehat{\mathrm{A}}_{tt'}^{(l)}\right]^2 = \mathrm{N} \cdot \mathrm{E}\left[\frac{1}{n_1^{(t')}} \sum_{k=1}^{n_1^{(t')}} \frac{1}{m_{1k}^{(l|t')}} \sum_{s=1}^{m_{1k}^{(l|t')}} \frac{1}{n_0^{(t)}} \sum_{k'=1}^{n_0^{(t)}} \frac{1}{m_{0k'}^{(l|t)}} \sum_{s'=1}^{m_{0k'}^{(l|t)}} \varphi_1\left(\mathrm{X}_{1ks}^{(l|t')}, \mathrm{X}_{0k's'}^{(l|t)}\right) - \varphi_2\left(\mathrm{X}_{0k's'}^{(l|t)}\right) \right]^2$$
$$= \frac{\mathrm{N}}{\left(n_0^{(t)} n_1^{(t')}\right)^2} \sum_{k_1=1}^{n_1^{(t')}} \sum_{k_2=1}^{n_1^{(t')}} \frac{1}{m_{1k_1}^{(l|t')} m_{1k_2}^{(l|t')}} \sum_{s_1=1}^{m_{1k_1}^{(l|t')}} \sum_{s_2=1}^{m_{1k_2}^{(l|t')}} \sum_{k_1'=1}^{n_0^{(t)}} \sum_{k_2'=1}^{n_0^{(t)}} \frac{1}{m_{0k_1'}^{(l|t)} m_{0k_2'}^{(l|t)}}$$
$$\sum_{s_1'=1}^{m_{0k_1'}^{(l|t)}} \sum_{s_2'=1}^{m_{0k_2'}^{(l|t)}} \underbrace{\mathrm{E}\left[\prod_{w=1}^{2} \left(\varphi_1\left(\mathrm{X}_{1k_w s_w}^{(l|t')}, \mathrm{X}_{0k_w' s_w'}^{(l|t)}\right) - \varphi_2\left(\mathrm{X}_{0k_w' s_w'}^{(l|t)}\right) \right) \right]}_{=0, \text{ wenn } k_1 \neq k_2 \text{ oder } k_1' \neq k_2'}$$
$$= \frac{\mathrm{N}}{\left(n_0^{(t)} n_1^{(t')}\right)^2} \sum_{k_1=1}^{n_1^{(t')}} \frac{1}{\left(m_{1k_1}^{(l|t')}\right)^2} \sum_{s_1=s_2=1}^{m_{1k_1}^{(l|t')}} \sum_{k_1'=1}^{n_0^{(t)}} \frac{1}{\left(m_{0k_1'}^{(l|t)}\right)^2} \sum_{s_1'=s_2'=1}^{m_{0k_1'}^{(l|t)}} \mathrm{E}\left[\prod_{w=1}^{2} \left(\varphi_1\left(\mathrm{X}_{1k_w s_w}^{(l|t')}, \mathrm{X}_{0k_w' s_w'}^{(l|t)}\right) - \varphi_2\left(\mathrm{X}_{0k_w' s_w'}^{(l|t)}\right) \right) \right]$$
$$\leq \frac{\mathrm{N}}{\left(n_0^{(t)} n_1^{(t')}\right)^2} \sum_{k_1=1}^{n_1^{(t')}} \frac{1}{\left(m_{1k_1}^{(l|t')}\right)^2} \sum_{s_1=s_2=1}^{m_{1k_1}^{(l|t')}} \sum_{k_1'=1}^{n_0^{(t)}} \frac{1}{\left(m_{0k_1'}^{(l|t)}\right)^2} \sum_{s_1'=s_2'=1}^{m_{0k_1'}^{(l|t)}} 4 = \frac{4 \cdot \mathrm{N}}{n_0^{(t)} n_1^{(t')}}.$$

Das Einsetzen dieses Resultats in Gleichung (B.1) liefert schließlich:

$$\mathrm{E}(\sqrt{\mathrm{N}}\widehat{\mathrm{A}}^{(l)})^2 \leq 4 \sum_{t \in \{u,v\}} \sum_{t' \in \{u,v\}} \left(\frac{n_0^{(t)}}{n_0} \frac{n_1^{(t')}}{n_1} \right)^2 \mathrm{E}\left[\sqrt{\mathrm{N}} \int \widehat{\mathrm{F}}_0^{(l|t)} - \mathrm{F}_0^{(l)} d\left(\widehat{\mathrm{F}}_1^{(l|t')} - \mathrm{F}_1^{(l)} \right) \right]^2$$
$$\leq 4 \sum_{t \in \{u,v\}} \sum_{t' \in \{u,v\}} \left(\frac{n_0^{(t)}}{n_0} \frac{n_1^{(t')}}{n_1} \right)^2 \frac{4 \cdot \mathrm{N}}{n_0^{(t)} n_1^{(t')}} = \frac{16\mathrm{N}}{(n_0 n_1)^2} \sum_{t \in \{u,v\}} \sum_{t' \in \{u,v\}} \underbrace{n_0^{(t)} \cdot n_1^{(t')}}_{\leq n_0 \cdot n_1}$$
$$\leq \frac{64\mathrm{N}}{n_0 n_1} = \frac{64(n_0 + n_1 - n^{(v)})}{n_0 n_1} = \frac{64}{n_0} + \frac{64}{n_1} - \frac{64 n^{(v)}}{n_0 n_1} \leq \frac{64}{n_0} + \frac{64}{n_1} \to 0.$$

□

B.1.2 Asymptotische Normalität des gewichteten AUC-Schätzers

Satz B.2 *Unter den Voraussetzungen 4.3 und 4.5 ist die Statistik* $\sqrt{\mathrm{N}}(\widehat{\mathbf{AUC}}-\mathbf{AUC})$ *asymptotisch normalverteilt mit Erwartungswert* $\mathbf{0}$ *und Kovarianzmatrix* $\mathbf{V}_{\widehat{\mathrm{AUC}}}$.

BEWEIS. Um die asymptotische Normalität von $\sqrt{\mathrm{N}}(\widehat{\mathbf{AUC}}-\mathbf{AUC})$ nachzuweisen, genügt es nach Satz 4.22 zu zeigen, dass $\sqrt{\mathrm{N}}\widehat{\mathbf{B}}$ asymptotisch normalverteilt ist. Nun gilt

$$
\begin{aligned}
\widehat{\mathrm{B}}^{(l)} &= \int \mathrm{F}_0^{(l)}\, d\widehat{\mathrm{F}}_1^{(l)} - \int \mathrm{F}_1^{(l)}\, d\widehat{\mathrm{F}}_0^{(l)} + 1 - 2\cdot \mathrm{AUC}^{(l)} \\
&= \frac{n^{(v)}}{n_1}\left(\frac{1}{n^{(v)}}\sum_{k=1}^{n^{(v)}} \overline{\mathrm{Y}}_{1k\cdot}^{(l|v)}\right) + \frac{n_1^{(u)}}{n_1}\left(\frac{1}{n_1^{(u)}}\sum_{k=1}^{n_1^{(u)}} \overline{\mathrm{Y}}_{1k\cdot}^{(l|u)}\right) \\
&\quad - \frac{n^{(v)}}{n_0}\left(\frac{1}{n^{(v)}}\sum_{k=1}^{n^{(v)}} \overline{\mathrm{Y}}_{0k\cdot}^{(l|v)}\right) - \frac{n_0^{(u)}}{n_0}\left(\frac{1}{n_0^{(u)}}\sum_{k=1}^{n_0^{(u)}} \overline{\mathrm{Y}}_{0k\cdot}^{(l|u)}\right) + 1 - 2\cdot \mathrm{AUC}^{(l)} \\
&= \frac{1}{n^{(v)}}\sum_{k=1}^{n^{(v)}}\left(\frac{n^{(v)}}{n_1}\overline{\mathrm{Y}}_{1k\cdot}^{(l|v)} - \frac{n^{(v)}}{n_0}\overline{\mathrm{Y}}_{0k\cdot}^{(l|v)}\right) \\
&\quad + \frac{1}{n_1^{(u)}}\sum_{k=1}^{n_1^{(u)}}\left(\frac{n_1^{(u)}}{n_1}\overline{\mathrm{Y}}_{1k\cdot}^{(l|u)}\right) - \frac{1}{n_0^{(u)}}\sum_{k=1}^{n_0^{(u)}}\left(\frac{n_0^{(u)}}{n_0}\overline{\mathrm{Y}}_{0k\cdot}^{(l|u)}\right) + 1 - 2\cdot \mathrm{AUC}^{(l)} \\
&= \frac{1}{n^{(v)}}\sum_{k=1}^{n^{(v)}}\left(\frac{n^{(v)}}{n_1}\overline{\mathrm{Y}}_{1k\cdot}^{(l|v)} - \frac{n^{(v)}}{n_0}\overline{\mathrm{Y}}_{0k\cdot}^{(l|v)}\right) + \frac{1}{n_1^{(u)}}\sum_{k=1}^{n_1^{(u)}}\left(\frac{n_1^{(u)}}{n_1}\overline{\mathrm{Y}}_{1k\cdot}^{(l|u)}\right) - \frac{1}{n_0^{(u)}}\sum_{k=1}^{n_0^{(u)}}\left(\frac{n_0^{(u)}}{n_0}\overline{\mathrm{Y}}_{0k\cdot}^{(l|u)}\right) + 1 - 2\cdot \mathrm{AUC}^{(l)}. \qquad \text{(B.2)}
\end{aligned}
$$

$\widehat{\mathrm{B}}^{(l)}$ ist damit – abgesehen von einer Konstanten – die Summe der Mittelwerte der unabhängigen Zufallsvariablen $\frac{n^{(v)}}{n_1}\overline{\mathrm{Y}}_{1k\cdot}^{(l|v)} - \frac{n^{(v)}}{n_0}\overline{\mathrm{Y}}_{0k\cdot}^{(l|v)}$, $\frac{n_0^{(u)}}{n_0}\overline{\mathrm{Y}}_{0k\cdot}^{(l|u)}$ und $\frac{n_1^{(u)}}{n_1}\overline{\mathrm{Y}}_{1k\cdot}^{(l|u)}$. Die asymptotische Normalität von $\sqrt{\mathrm{N}}\widehat{\mathrm{B}}^{(l)}$ folgt daher aus dem zentralen Grenzwertsatz nach Lindeberg (Lindeberg, 1922), da diese Zufallsvariablen gleichmäßig beschränkt sind. Die multivariate Normalität von $\sqrt{\mathrm{N}}\widehat{\mathbf{B}}$ lässt sich analog mit Hilfe des Lindeberg'schen zentralen Grenzwertsatzes kanonisch mit der Cramer-Wold-Technik nachweisen. □

B.2 Schätzung der Kovarianzmatrix des gewichteten Schätzers

Satz B.3 *Es seien*

$$\begin{aligned}
&\mathbf{M}_{i\cdot}^{(\cdot)} = diag(m_{i\cdot}^{(1|\cdot)},\ldots,m_{i\cdot}^{(d|\cdot)}) \in \mathbb{R}^{d\times d}, && i = 0,1\\
&\mathbf{M}_{i\cdot}^{(t)} = diag(m_{i\cdot}^{(1|t)},\ldots,m_{i\cdot}^{(d|t)}) \in \mathbb{R}^{d\times d}, && t \in \{u,v\},\ i = 0,1 \text{ und}\\
&\mathbf{M}_{ik}^{(t)} = diag(m_{ik}^{(1|t)},\ldots,m_{ik}^{(d|t)}) \in \mathbb{R}^{d\times d}, && t \in \{u,v\},\ i = 0,1,\ k = 1,\ldots n_i^{(t)}
\end{aligned}$$

Diagonalmatrizen, welche die Gesamtanzahlen an gesunden ($i = 0$) beziehungsweise kranken ($i = 1$) Beobachtungseinheiten der Komponenten $1,\ldots,d$ darstellen. Es seien in Anlehnung an die Definition der asymptotischen Rangtransformationen in Gleichung (4.29) *die beobachtbaren gewichteten Rangtransformationen*

$$\widetilde{\mathrm{Y}}_{iks}^{(l|t)} = \widetilde{\mathrm{F}}_{1-i}^{(l)}\left(\mathrm{X}_{iks}^{(l|t)}\right), \qquad l = 1,\ldots,d,\ t \in \{u,v\},\ i = 0,1,\ k = 1,\ldots,n_i^{(t)},\ s = 1,\ldots,m_{ik}^{(l|t)}$$

definiert. Die Kovarianzmatrix $\mathbf{V}_{\widetilde{\mathrm{AUC}}}$ von $\sqrt{\mathrm{N}}(\widetilde{\mathbf{AUC}} - \mathbf{AUC})$ wird dann durch $\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}}} = \widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}}}^{(v)} + \widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}},0}^{(u)} + \widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}},1}^{(u)}$ konsistent geschätzt, wobei

$$\begin{aligned}
\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}}}^{(v)} = \frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\sum_{k=1}^{n^{(v)}} &\left(\left(\mathbf{M}_{1\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{1k\cdot}^{(v)} - \left(\mathbf{M}_{0\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{0k\cdot}^{(v)} - \left[\mathbf{M}_{1k}^{(v)}\left(\mathbf{M}_{1\cdot}^{(v)}\mathbf{M}_{1\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{1\cdot\cdot}^{(v)} - \mathbf{M}_{0k}^{(v)}\left(\mathbf{M}_{0\cdot}^{(v)}\mathbf{M}_{0\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{0\cdot\cdot}^{(v)}\right]\right)\\
&\cdot\left(\left(\mathbf{M}_{1\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{1k\cdot}^{(v)} - \left(\mathbf{M}_{0\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{0k\cdot}^{(v)} - \left[\mathbf{M}_{1k}^{(v)}\left(\mathbf{M}_{1\cdot}^{(v)}\mathbf{M}_{1\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{1\cdot\cdot}^{(v)} - \mathbf{M}_{0k}^{(v)}\left(\mathbf{M}_{0\cdot}^{(v)}\mathbf{M}_{0\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{0\cdot\cdot}^{(v)}\right]\right)',\\
\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}},i}^{(u)} = \frac{\mathrm{N}n_i^{(u)}}{n_i^{(u)}-1}\sum_{k=1}^{n_i^{(u)}} &\left(\left(\mathbf{M}_{i\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{ik\cdot}^{(u)} - \mathbf{M}_{ik}^{(u)}\left(\mathbf{M}_{i\cdot}^{(u)}\mathbf{M}_{i\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{i\cdot\cdot}^{(u)}\right)\left(\left(\mathbf{M}_{i\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{ik\cdot}^{(u)} - \mathbf{M}_{ik}^{(u)}\left(\mathbf{M}_{i\cdot}^{(u)}\mathbf{M}_{i\cdot}^{(\cdot)}\right)^{-1}\widetilde{\mathbf{Y}}_{i\cdot\cdot}^{(u)}\right)'.
\end{aligned}$$

BEWEIS. Es gilt:

$$\sqrt{\mathrm{N}}\widetilde{\mathbf{B}} = \sqrt{\mathrm{N}}\left(\sum_{k=1}^{n^{(v)}}\left(\mathbf{M}_1^{-1}\mathbf{Y}_{1k\cdot}^{(v)} - \mathbf{M}_0^{-1}\mathbf{Y}_{0k\cdot}^{(v)}\right) + \sum_{k=1}^{n_1^{(u)}}\left(\mathbf{M}_1^{-1}\mathbf{Y}_{1k\cdot}^{(u)}\right) - \sum_{k=1}^{n_0^{(u)}}\left(\mathbf{M}_0^{-1}\mathbf{Y}_{0k\cdot}^{(l)}\right) + \mathbf{1}_d - 2\cdot\mathbf{AUC}\right)$$

Für die Kovarianzmatrix $\mathrm{Cov}\left(\sqrt{\mathrm{N}}\widetilde{\mathbf{B}}\right) = \mathbf{V}_{\widetilde{\mathrm{AUC}}}$ gilt daher:

$$\begin{aligned}
\mathbf{V}_{\widetilde{\mathrm{AUC}}} &= \mathrm{NCov}\left(\sum_{k=1}^{n^{(v)}}\left(\mathbf{M}_1^{-1}\mathbf{Y}_{1k\cdot}^{(v)} - \mathbf{M}_0^{-1}\mathbf{Y}_{0k\cdot}^{(v)}\right)\right) + \mathrm{NCov}\left(\sum_{k=1}^{n_1^{(u)}}\left(\mathbf{M}_1^{-1}\mathbf{Y}_{1k\cdot}^{(u)}\right)\right) + \mathrm{NCov}\left(\sum_{k=1}^{n_0^{(u)}}\left(\mathbf{M}_0^{-1}\mathbf{Y}_{0k\cdot}^{(l)}\right)\right)\\
&= \mathbf{V}_{\widetilde{\mathrm{AUC}}}^{(v)} + \mathbf{V}_{\widetilde{\mathrm{AUC}},1}^{(u)} + \mathbf{V}_{\widetilde{\mathrm{AUC}},0}^{(u)}
\end{aligned}$$

Es gilt nun $\forall\, l,r = 1,\ldots,d$

$$\begin{aligned}
&\mathrm{E}\left(\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}}}[l,r] - \mathbf{V}_{\widetilde{\mathrm{AUC}}}[l,r]\right)^2\\
&\quad= \mathrm{E}\left(\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}}}^{(v)}[l,r] + \widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}},1}^{(u)}[l,r] + \widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}},0}^{(u)}[l,r] - \mathbf{V}_{\widetilde{\mathrm{AUC}}}^{(v)}[l,r] - \mathbf{V}_{\widetilde{\mathrm{AUC}},1}^{(u)}[l,r] - \mathbf{V}_{\widetilde{\mathrm{AUC}},0}^{(u)}[l,r]\right)^2\\
&\quad\le 4\mathrm{E}\left(\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}}}^{(v)}[l,r] - \mathbf{V}_{\widetilde{\mathrm{AUC}}}^{(v)}[l,r]\right)^2 + 4\mathrm{E}\left(\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}},1}^{(u)}[l,r] - \mathbf{V}_{\widetilde{\mathrm{AUC}},1}^{(u)}[l,r]\right)^2 + 4\mathrm{E}\left(\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}},0}^{(u)}[l,r] - \mathbf{V}_{\widetilde{\mathrm{AUC}},0}^{(u)}[l,r]\right)^2. \qquad \text{(B.3)}
\end{aligned}$$

Die Konsistenz von $\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}}}$ kann daher durch einen separaten Nachweis der Konsistenz der einzelnen Summanden in (B.3) gezeigt werden. Die Methodik der Beweisführung wird in dieser Arbeit nur beispielhaft an Hand von $\mathrm{E}(\widetilde{\mathbf{V}}_{\widetilde{\mathrm{AUC}}}^{(v)}[l,r] - \mathbf{V}_{\mathrm{AUC}}^{(v)}[l,r])^2$ illustriert, da sich die Beweistechnik kanonisch auf

$\mathrm{E}(\widetilde{\mathbf{V}}^{(u)}_{\mathrm{AUC},i}[l,r]-\mathbf{V}^{(u)}_{\mathrm{AUC},i}[l,r])^2$, $i=0,1$ übertragen lässt. Definiere hierfür

$$\check{\mathbf{V}}^{(v)}_{\widetilde{\mathrm{AUC}}}=\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\sum_{k=1}^{n^{(v)}}\left(\left(\mathbf{M}^{(\cdot)}_{1\cdot}\right)^{-1}\mathbf{Y}^{(v)}_{1k\cdot}-\left(\mathbf{M}^{(\cdot)}_{0\cdot}\right)^{-1}\mathbf{Y}^{(v)}_{0k\cdot}-\left[\mathbf{M}^{(v)}_{1k}\left(\mathbf{M}^{(v)}_{1\cdot}\mathbf{M}^{(\cdot)}_{1\cdot}\right)^{-1}\mathbf{Y}^{(v)}_{1\cdot\cdot}-\mathbf{M}^{(v)}_{0k}\left(\mathbf{M}^{(v)}_{0\cdot}\mathbf{M}^{(\cdot)}_{0\cdot}\right)^{-1}\mathbf{Y}^{(v)}_{0\cdot\cdot}\right]\right)$$
$$\cdot\left(\left(\mathbf{M}^{(\cdot)}_{1\cdot}\right)^{-1}\mathbf{Y}^{(v)}_{1k\cdot}-\left(\mathbf{M}^{(\cdot)}_{0\cdot}\right)^{-1}\mathbf{Y}^{(v)}_{0k\cdot}-\left[\mathbf{M}^{(v)}_{1k}\left(\mathbf{M}^{(v)}_{1\cdot}\mathbf{M}^{(\cdot)}_{1\cdot}\right)^{-1}\mathbf{Y}^{(v)}_{1\cdot\cdot}-\mathbf{M}^{(v)}_{0k}\left(\mathbf{M}^{(v)}_{0\cdot}\mathbf{M}^{(\cdot)}_{0\cdot}\right)^{-1}\mathbf{Y}^{(v)}_{0\cdot\cdot}\right]\right)'.$$

Es gilt nun $\forall\, l,r=1,\ldots,d$

$$\mathrm{E}\left(\widetilde{\mathbf{V}}^{(v)}_{\widetilde{\mathrm{AUC}}}[l,r]-\mathbf{V}^{(v)}_{\widetilde{\mathrm{AUC}}}[l,r]\right)^2\leq 2\mathrm{E}\left((\widetilde{\mathbf{V}}^{(v)}_{\widetilde{\mathrm{AUC}}}[l,r]-\check{\mathbf{V}}^{(v)}_{\widetilde{\mathrm{AUC}}}[l,r]\right)^2+2\mathrm{E}\left(\check{\mathbf{V}}^{(v)}_{\widetilde{\mathrm{AUC}}}[l,r]-\mathbf{V}^{(v)}_{\widetilde{\mathrm{AUC}}}[l,r]\right)^2. \tag{B.4}$$

Die beiden Summanden werden nun separat betrachtet, wobei der Konsistenzbeweis beider Summanden komponentenweise erfolgt. Hierfür seien $l,r\in\{1,\ldots,d\}$ beliebig, dann gilt:

$$\begin{aligned}
&\mathrm{E}\left[\check{\mathbf{V}}^{(v)}_{\widetilde{\mathrm{AUC}}}[l,r]-\mathbf{V}^{(v)}_{\widetilde{\mathrm{AUC}}}[l,r]\right]^2\\
&=\mathrm{E}\left[\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\sum_{k=1}^{n^{(v)}}\left(\frac{1}{m^{(l|\cdot)}_{1\cdot}}\mathrm{Y}^{(l|v)}_{1k\cdot}-\frac{1}{m^{(l|\cdot)}_{0\cdot}}\mathrm{Y}^{(l|v)}_{0k\cdot}-\left[\frac{m^{(l|v)}_{1k}}{m^{(l|v)}_{1\cdot}m^{(l|\cdot)}_{1\cdot}}\mathrm{Y}^{(l|v)}_{1\cdot\cdot}-\frac{m^{(l|v)}_{0k}}{m^{(l|v)}_{0\cdot}m^{(l|\cdot)}_{0\cdot}}\mathrm{Y}^{(l|v)}_{0\cdot\cdot}\right]\right)\right.\\
&\quad\cdot\left(\frac{1}{m^{(r|\cdot)}_{1\cdot}}\mathrm{Y}^{(r|v)}_{1k\cdot}-\frac{1}{m^{(r|\cdot)}_{0\cdot}}\mathrm{Y}^{(r|v)}_{0k\cdot}-\left[\frac{m^{(r|v)}_{1k}}{m^{(r|v)}_{1\cdot}m^{(r|\cdot)}_{1\cdot}}\mathrm{Y}^{(r|v)}_{1\cdot\cdot}-\frac{m^{(r|v)}_{0k}}{m^{(r|v)}_{0\cdot}m^{(r|\cdot)}_{0\cdot}}\overline{\mathrm{Y}}^{(r|v)}_{0\cdot\cdot}\right]\right)\\
&\quad\left.-\mathrm{NCov}\left(\sum_{k=1}^{n^{(v)}}\frac{1}{m^{(l|\cdot)}_{1\cdot}}\mathrm{Y}^{(l|v)}_{1k\cdot}-\frac{1}{m^{(l|\cdot)}_{0\cdot}}\mathrm{Y}^{(l|v)}_{0k\cdot},\sum_{k=1}^{n^{(v)}}\frac{1}{m^{(r|\cdot)}_{1\cdot}}\mathrm{Y}^{(r|v)}_{1k\cdot}-\frac{1}{m^{(r|\cdot)}_{0\cdot}}\mathrm{Y}^{(r|v)}_{0k\cdot}\right)\right]^2\\
&=\mathrm{E}\left[\sum_{i=0}^{1}\sum_{i'=0}^{1}\frac{-1^{1-i}}{m^{(l|\cdot)}_{i\cdot}}\frac{-1^{1-i'}}{m^{(r|\cdot)}_{i'\cdot}}\left(\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\sum_{k=1}^{n^{(v)}}\left(\mathrm{Y}^{(l|v)}_{ik\cdot}-\frac{m^{(l|v)}_{ik}}{m^{(l|v)}_{i\cdot}}\mathrm{Y}^{(l|v)}_{i\cdot\cdot}\right)\left(\mathrm{Y}^{(r|v)}_{i'k\cdot}-\frac{m^{(r|v)}_{i'k}}{m^{(r|v)}_{i'\cdot}}\mathrm{Y}^{(r|v)}_{i'\cdot\cdot}\right)\right.\right.\\
&\quad\left.\left.-\,\mathrm{N}\sum_{k=1}^{n^{(v)}}\mathrm{Cov}\left(\mathrm{Y}^{(l|v)}_{ik\cdot},\mathrm{Y}^{(r|v)}_{i'k\cdot}\right)\right)\right]^2\\
&\leq 4\left(n^{(v)}\right)^4\sum_{i=0}^{1}\sum_{i'=0}^{1}\left(\frac{1}{m^{(l|\cdot)}_{i\cdot}}\frac{1}{m^{(r|\cdot)}_{i'\cdot}}\right)^2\mathrm{E}\left[\frac{\mathrm{N}}{n^{(v)}\left(n^{(v)}-1\right)}\sum_{k=1}^{n^{(v)}}\left\{\left(\mathrm{Y}^{(l|v)}_{ik\cdot}-\frac{m^{(l|v)}_{ik}}{m^{(l|v)}_{i\cdot}}\mathrm{Y}^{(l|v)}_{i\cdot\cdot}\right)\left(\mathrm{Y}^{(r|v)}_{i'k\cdot}-\frac{m^{(r|v)}_{i'k}}{m^{(r|v)}_{i'\cdot}}\mathrm{Y}^{(r|v)}_{i'\cdot\cdot}\right)\right.\right.\\
&\quad\left.\left.-\,\frac{\mathrm{N}}{n^{(v)}}\mathrm{Cov}\left(\mathrm{Y}^{(l|v)}_{ik\cdot},\mathrm{Y}^{(r|v)}_{i'k\cdot}\right)\right\}\right]^2.
\end{aligned}$$

Da die $\mathrm{Y}^{(l|v)}_{ik\cdot}$ durch $m_{\max}$ gleichmäßig beschränkt sind und $\mathrm{E}(\mathrm{Y}^{(l|v)}_{ik\cdot})=\mathrm{E}(\frac{m^{(l|v)}_{ik}}{m^{(l|v)}_{i\cdot}}\cdot\mathrm{Y}^{(l|v)}_{i\cdot\cdot})$ lässt sich analog zu Werner (2006, Theorem 3.5) zeigen, dass

$$\mathrm{E}\left[\frac{\mathrm{N}}{n^{(v)}\left(n^{(v)}-1\right)}\sum_{k=1}^{n^{(v)}}\left(\left(\mathrm{Y}^{(l|v)}_{ik\cdot}-\frac{m^{(l|v)}_{ik}}{m^{(l|v)}_{i\cdot}}\mathrm{Y}^{(l|v)}_{i\cdot\cdot}\right)\left(\mathrm{Y}^{(r|v)}_{i'k\cdot}-\frac{m^{(r|v)}_{i'k}}{m^{(r|v)}_{i'\cdot}}\mathrm{Y}^{(r|v)}_{i'\cdot\cdot}\right)-\frac{\mathrm{N}}{n^{(v)}}\mathrm{Cov}\left(\mathrm{Y}^{(l|v)}_{ik\cdot},\mathrm{Y}^{(r|v)}_{i'k\cdot}\right)\right)\right]^2=\mathrm{O}\left(\frac{1}{n^{(v)}}\right)$$

gilt. Daher folgt:

$$\begin{aligned}
\mathrm{E}\left[\check{\mathbf{V}}^{(v)}_{\widetilde{\mathrm{AUC}}}[l,r]-\mathbf{V}^{(v)}_{\widetilde{\mathrm{AUC}}}[l,r]\right]^2&\leq 4\cdot\left(n^{(v)}\right)^4\sum_{i=0}^{1}\sum_{i'=0}^{1}\left(\frac{1}{m^{(l|\cdot)}_{i\cdot}}\frac{1}{m^{(r|\cdot)}_{i'\cdot}}\right)^2\mathrm{O}\left(\frac{1}{n^{(v)}}\right)\\
&\leq\mathrm{O}\left(\frac{1}{m^{(l|\cdot)}_{0\cdot}}+\frac{1}{m^{(l|\cdot)}_{1\cdot}}+\frac{1}{m^{(r|\cdot)}_{0\cdot}}+\frac{1}{m^{(r|\cdot)}_{1\cdot}}\right)\to 0,
\end{aligned}$$

da $\frac{n^{(v)}}{m_{i\cdot}^{(l|\cdot)}} \leq 1$, $i = 0,1$, $l = 1,\ldots,d$. Die Konsistenz des zweiten Summanden aus Gleichung (B.4) ist damit bewiesen.

Im nächsten Schritt soll nun $\mathrm{E}\left(\widetilde{\mathbf{V}}^{(v)}_{\widehat{\mathrm{AUC}}}[l,r] - \check{\mathbf{V}}^{(v)}_{\widehat{\mathrm{AUC}}}[l,r]\right)^2$ abgeschätzt werden. Zur Vereinfachung der Notation seien hierfür im Folgenden

$$\mathrm{A}_k^{(l|v)} = \frac{1}{m_{1\cdot}^{(l|\cdot)}}\mathrm{Y}_{1k\cdot}^{(l|v)} - \frac{1}{m_{0\cdot}^{(l|\cdot)}}\mathrm{Y}_{0k\cdot}^{(l|v)}, \qquad \mathrm{B}_k^{(l|v)} = \frac{m_{1k}^{(l|v)}}{m_{1\cdot}^{(l|v)}}\frac{1}{m_{1\cdot}^{(l|\cdot)}}\mathrm{Y}_{1\cdot\cdot}^{(l|v)} - \frac{m_{0k}^{(l|v)}}{m_{0\cdot}^{(l|v)}}\frac{1}{m_{0\cdot}^{(l|\cdot)}}\mathrm{Y}_{0\cdot\cdot}^{(l|v)}, \text{ sowie}$$

$$\widetilde{\mathrm{A}}_k^{(l|v)} = \frac{1}{m_{1\cdot}^{(l|\cdot)}}\widetilde{\mathrm{Y}}_{1k\cdot}^{(l|v)} - \frac{1}{m_{0\cdot}^{(l|\cdot)}}\widetilde{\mathrm{Y}}_{0k\cdot}^{(l|v)}, \qquad \widetilde{\mathrm{B}}_k^{(l|v)} = \frac{m_{1k}^{(l|v)}}{m_{1\cdot}^{(l|v)}}\frac{1}{m_{1\cdot}^{(l|\cdot)}}\widetilde{\mathrm{Y}}_{1\cdot\cdot}^{(l|v)} - \frac{m_{0k}^{(0|v)}}{m_{0\cdot}^{(l|v)}}\frac{1}{m_{0\cdot}^{(l|\cdot)}}\widetilde{\mathrm{Y}}_{0\cdot\cdot}^{(l|v)}.$$

Dann gilt $\forall r,l = 1,\ldots,d$ mit der Jensen- und der c_r-Ungleichung:

$$\begin{aligned}
&\mathrm{E}\left(\widetilde{\mathbf{V}}^{(v)}_{\widehat{\mathrm{AUC}}}[l,r] - \check{\mathbf{V}}^{(v)}_{\widehat{\mathrm{AUC}}}[l,r]\right)^2 \\
&= \mathrm{E}\left(\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\sum_{k=1}^{n^{(v)}}\left(\widetilde{\mathrm{A}}_k^{(l|v)} - \widetilde{\mathrm{B}}_k^{(l|v)}\right)\left(\widetilde{\mathrm{A}}_k^{(r|v)} - \widetilde{\mathrm{B}}_k^{(r|v)}\right) - \frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\sum_{k=1}^{n^{(v)}}\left(\mathrm{A}_k^{(l|v)} - \mathrm{B}_k^{(l|v)}\right)\left(\mathrm{A}_k^{(r|v)} - \mathrm{B}_k^{(r|v)}\right)\right)^2 \\
&= \left(\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\right)^2 \cdot \left(n^{(v)}\right)^2 \mathrm{E}\left(\frac{1}{n^{(v)}}\sum_{k=1}^{n^{(v)}}\left(\widetilde{\mathrm{A}}_k^{(l|v)} - \widetilde{\mathrm{B}}_k^{(l|v)}\right)\left(\widetilde{\mathrm{A}}_k^{(r|v)} - \widetilde{\mathrm{B}}_k^{(r|v)}\right) - \left(\mathrm{A}_k^{(l|v)} - \mathrm{B}_k^{(l|v)}\right)\left(\mathrm{A}_k^{(r|v)} - \mathrm{B}_k^{(r|v)}\right)\right)^2 \\
&\leq \left(\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\right)^2 \cdot n^{(v)}\sum_{k=1}^{n^{(v)}}\mathrm{E}\left(\left(\widetilde{\mathrm{A}}_k^{(l|v)} - \widetilde{\mathrm{B}}_k^{(l|v)}\right)\left(\widetilde{\mathrm{A}}_k^{(r|v)} - \widetilde{\mathrm{B}}_k^{(r|v)}\right) - \left(\mathrm{A}_k^{(l|v)} - \mathrm{B}_k^{(l|v)}\right)\left(\mathrm{A}_k^{(r|v)} - \mathrm{B}_k^{(r|v)}\right)\right)^2 \\
&= \left(\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\right)^2 \cdot n^{(v)}\sum_{k=1}^{n^{(v)}}\mathrm{E}\Big(\left[\widetilde{\mathrm{A}}_k^{(l|v)} - \widetilde{\mathrm{B}}_k^{(l|v)}\right]\left[\left(\widetilde{\mathrm{A}}_k^{(r|v)} - \widetilde{\mathrm{B}}_k^{(r|v)}\right) - \left(\mathrm{A}_k^{(r|v)} - \mathrm{B}_k^{(r|v)}\right)\right] \\
&\qquad\qquad + \left[\mathrm{A}_k^{(r|v)} - \mathrm{B}_k^{(r|v)}\right]\left[\left(\widetilde{\mathrm{A}}_k^{(l|v)} - \widetilde{\mathrm{B}}_k^{(l|v)}\right) - \left(\mathrm{A}_k^{(l|v)} - \mathrm{B}_k^{(l|v)}\right)\right]\Big)^2 \\
&\leq 2\left(\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\right)^2 \cdot n^{(v)}\Bigg[\sum_{k=1}^{n^{(v)}}\mathrm{E}\Big(\left[\widetilde{\mathrm{A}}_k^{(l|v)} - \widetilde{\mathrm{B}}_k^{(l|v)}\right]\left[\left(\widetilde{\mathrm{A}}_k^{(r|v)} - \widetilde{\mathrm{B}}_k^{(r|v)}\right) - \left(\mathrm{A}_k^{(r|v)} - \mathrm{B}_k^{(r|v)}\right)\right]\Big)^2 \\
&\qquad\qquad + \mathrm{E}\Big(\left[\mathrm{A}_k^{(r|v)} - \mathrm{B}_k^{(r|v)}\right]\left[\left(\widetilde{\mathrm{A}}_k^{(l|v)} - \widetilde{\mathrm{B}}_k^{(l|v)}\right) - \left(\mathrm{A}_k^{(l|v)} - \mathrm{B}_k^{(l|v)}\right)\right]\Big)^2\Bigg].
\end{aligned}$$

Nun gilt:

$$\begin{aligned}
\left|\mathrm{A}_k^{(r|v)} - \mathrm{B}_k^{(r|v)}\right| &= \left|\frac{1}{m_{1\cdot}^{(r|\cdot)}}\mathrm{Y}_{1k\cdot}^{(r|v)} - \frac{1}{m_{0\cdot}^{(r|\cdot)}}\mathrm{Y}_{0k\cdot}^{(r|v)} - \frac{m_{1k}^{(r|v)}}{m_{1\cdot}^{(r|v)}}\frac{1}{m_{1\cdot}^{(r|\cdot)}}\mathrm{Y}_{1\cdot\cdot}^{(r|v)} + \frac{m_{0k}^{(r|v)}}{m_{0\cdot}^{(r|v)}}\frac{1}{m_{0\cdot}^{(r|\cdot)}}\mathrm{Y}_{0\cdot\cdot}^{(r|v)}\right| \\
&= \left|\frac{1}{m_{1\cdot}^{(r|\cdot)}}\left(\mathrm{Y}_{1k\cdot}^{(r|v)} - \frac{m_{1k}^{(r|v)}}{m_{1\cdot}^{(r|v)}}\mathrm{Y}_{1\cdot\cdot}^{(r|v)}\right) - \frac{1}{m_{0\cdot}^{(r|\cdot)}}\left(\mathrm{Y}_{0k\cdot}^{(r|v)} - \frac{m_{0k}^{(r|v)}}{m_{0\cdot}^{(r|v)}}\mathrm{Y}_{0\cdot\cdot}^{(r|v)}\right)\right| \\
&\leq \left|\frac{m_{\max}}{m_{1\cdot}^{(r|\cdot)}}\right| + \left|\frac{m_{\max}}{m_{0\cdot}^{(r|\cdot)}}\right| = \frac{m_{\max}}{m_{1\cdot}^{(r|\cdot)}} + \frac{m_{\max}}{m_{0\cdot}^{(r|\cdot)}}.
\end{aligned}$$

Analog zeigt man $|\widetilde{\mathrm{A}}_1^{(l|v)} - \widetilde{\mathrm{B}}_k^{(l|v)}| \leq \frac{m_{\max}}{m_{1\cdot}^{(l|\cdot)}} + \frac{m_{\max}}{m_{0\cdot}^{(l|\cdot)}}$. Somit folgt:

$$
\begin{aligned}
&\mathrm{E}\left(\widetilde{\mathbf{V}}^{(v)}_{\widetilde{\mathrm{AUC}}}[l,r] - \check{\mathbf{V}}^{(v)}_{\widetilde{\mathrm{AUC}}}[l,r]\right)^2 \\
&\leq 2\left(\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\right)^2 \cdot n^{(v)} \left[\sum_{k=1}^{n^{(v)}} \mathrm{E}\left(\left[\frac{m_{\max}}{m_{1\cdot}^{(r|\cdot)}} + \frac{m_{\max}}{m_{0\cdot}^{(r|\cdot)}}\right]\left[\left(\widetilde{\mathrm{A}}_k^{(r|v)} - \widetilde{\mathrm{B}}_k^{(r|v)}\right) - \left(\mathrm{A}_k^{(r|v)} - \mathrm{B}_k^{(r|v)}\right)\right]\right)^2\right. \\
&\qquad\qquad \left. + \mathrm{E}\left(\left[\frac{m_{\max}}{m_{1\cdot}^{(l|\cdot)}} + \frac{m_{\max}}{m_{0\cdot}^{(l|\cdot)}}\right]\left[\left(\widetilde{\mathrm{A}}_k^{(l|v)} - \widetilde{\mathrm{B}}_k^{(l|v)}\right) - \left(\mathrm{A}_k^{(l|v)} - \mathrm{B}_k^{(l|v)}\right)\right]\right)^2\right] \\
&\leq 2\left(\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\right)^2 n^{(v)} \left[\frac{m_{\max}}{m_{1\cdot}^{(r|\cdot)}} + \frac{m_{\max}}{m_{0\cdot}^{(r|\cdot)}}\right]^2 \sum_{k=1}^{n^{(v)}} \mathrm{E}\left(\left[\left(\widetilde{\mathrm{A}}_k^{(r|v)} - \widetilde{\mathrm{B}}_k^{(r|v)}\right) - \left(\mathrm{A}_k^{(r|v)} - \mathrm{B}_k^{(r|v)}\right)\right]\right)^2 \\
&\quad + 2\left(\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\right)^2 n^{(v)} \left[\frac{m_{\max}}{m_{1\cdot}^{(l|\cdot)}} + \frac{m_{\max}}{m_{0\cdot}^{(l|\cdot)}}\right]^2 \sum_{k=1}^{n^{(v)}} \mathrm{E}\left(\left[\left(\widetilde{\mathrm{A}}_k^{(l|v)} - \widetilde{\mathrm{B}}_k^{(l|v)}\right) - \left(\mathrm{A}_k^{(l|v)} - \mathrm{B}_k^{(l|v)}\right)\right]\right)^2 \\
&\leq 4\left(\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\right)^2 n^{(v)} \left[\frac{m_{\max}}{m_{1\cdot}^{(r|\cdot)}} + \frac{m_{\max}}{m_{0\cdot}^{(r|\cdot)}}\right]^2 \sum_{k=1}^{n^{(v)}} \left(\mathrm{E}\left[\widetilde{\mathrm{A}}_k^{(r|v)} - \mathrm{A}_k^{(r|v)}\right]^2 + \mathrm{E}\left[\widetilde{\mathrm{B}}_k^{(r|v)} - \mathrm{B}_k^{(r|v)}\right]^2\right) \\
&\quad + 4\left(\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\right)^2 n^{(v)} \left[\frac{m_{\max}}{m_{1\cdot}^{(l|\cdot)}} + \frac{m_{\max}}{m_{0\cdot}^{(l|\cdot)}}\right]^2 \sum_{k=1}^{n^{(v)}} \left(\mathrm{E}\left[\widetilde{\mathrm{A}}_k^{(l|v)} - \mathrm{A}_k^{(l|v)}\right]^2 + \mathrm{E}\left[\widetilde{\mathrm{B}}_k^{(l|v)} - \mathrm{B}_k^{(l|v)}\right]^2\right). \qquad \text{(B.5)}
\end{aligned}
$$

Es gilt nun $\forall k = 1, \ldots, n^{(v)}$:

$$
\begin{aligned}
\mathrm{E}\left[\widetilde{\mathrm{A}}_k^{(l|v)} - \mathrm{A}_k^{(l|v)}\right]^2 &= \mathrm{E}\left[\frac{1}{m_{1\cdot}^{(l|\cdot)}}\widetilde{\mathrm{Y}}_{1k\cdot}^{(l|v)} - \frac{1}{m_{0\cdot}^{(l|\cdot)}}\widetilde{\mathrm{Y}}_{0k\cdot}^{(0|v)} - \frac{1}{m_{1\cdot}^{(l|\cdot)}}\mathrm{Y}_{1k\cdot}^{(l|v)} + \frac{1}{m_{0\cdot}^{(l|\cdot)}}\mathrm{Y}_{0k\cdot}^{(0|v)}\right]^2 \\
&\leq 2\left(\frac{1}{m_{1\cdot}^{(l|\cdot)}}\right)^2 \mathrm{E}\left[\widetilde{\mathrm{Y}}_{1k\cdot}^{(l|v)} - \mathrm{Y}_{1k\cdot}^{(l|v)}\right]^2 + 2\left(\frac{1}{m_{0\cdot}^{(l|\cdot)}}\right)^2 \mathrm{E}\left[\widetilde{\mathrm{Y}}_{0k\cdot}^{(l|v)} - \mathrm{Y}_{0k\cdot}^{(l|v)}\right]^2 \\
&= 2\left(\frac{m_{1k}^{(l|v)}}{m_{1\cdot}^{(l|\cdot)}}\right)^2 \mathrm{E}\left[\frac{1}{m_{1k}^{(l|v)}}\sum_{k=1}^{m_{1k}^{(l|v)}}\left(\widetilde{\mathrm{Y}}_{1ks}^{(l|v)} - \mathrm{Y}_{1ks}^{(l|v)}\right)\right]^2 + 2\left(\frac{m_{0k}^{(l|v)}}{m_{0\cdot}^{(l|\cdot)}}\right)^2 \mathrm{E}\left[\frac{1}{m_{0k}^{(l|v)}}\sum_{k=1}^{m_{0k}^{(l|v)}}\left(\widetilde{\mathrm{Y}}_{0ks}^{(l|v)} - \mathrm{Y}_{0ks}^{(l|v)}\right)\right]^2 \\
&\leq 2\left(\frac{m_{\max}}{m_{1\cdot}^{(l|\cdot)}}\right)^2 \frac{1}{m_{1k}^{(l|v)}} \sum_{k=1}^{m_{1k}^{(l|v)}} \underbrace{\mathrm{E}\left[\widetilde{\mathrm{Y}}_{1ks}^{(l|v)} - \mathrm{Y}_{1ks}^{(l|v)}\right]^2}_{\leq \frac{4m_{\max}}{m_{0\cdot}^{(l|v)}} \text{ (vergl. Gl. (4.21), S. 38)}} + 2\left(\frac{m_{\max}}{m_{0\cdot}^{(l|\cdot)}}\right)^2 \frac{1}{m_{0k}^{(l|v)}} \sum_{k=1}^{m_{0k}^{(l|v)}} \underbrace{\mathrm{E}\left[\widetilde{\mathrm{Y}}_{0ks}^{(l|v)} - \mathrm{Y}_{0ks}^{(l|v)}\right]^2}_{\leq \frac{4m_{\max}}{m_{1\cdot}^{(l|v)}}} \\
&\leq \mathrm{O}\left(\left[\frac{1}{m_{1\cdot}^{(l|v)}}\right]^2 \cdot \frac{1}{m_{0\cdot}^{(l|v)}} + \left[\frac{1}{m_{0\cdot}^{(l|v)}}\right]^2 \cdot \frac{1}{m_{1\cdot}^{(l|v)}}\right).
\end{aligned}
$$

Analog zeigt man:

$$
\mathrm{E}\left[\widetilde{\mathrm{B}}_k^{(l|v)} - \mathrm{B}_k^{(l|v)}\right]^2 \leq \mathrm{O}\left(\left[\frac{1}{m_{1\cdot}^{(l|v)}}\right]^2 \cdot \frac{1}{m_{0\cdot}^{(l|v)}} + \left[\frac{1}{m_{0\cdot}^{(l|v)}}\right]^2 \cdot \frac{1}{m_{1\cdot}^{(l|v)}}\right).
$$

Durch Einsetzen dieser Resulate in (B.5) erhält man schließlich

$$
\begin{aligned}
&\mathrm{E}\left(\widetilde{\mathbf{V}}^{(v)}_{\widehat{\mathrm{AUC}}}[l,r]-\check{\mathbf{V}}^{(v)}_{\widehat{\mathrm{AUC}}}[l,r]\right)^2 \\
&\quad\leq 4\left(\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\right)^2 n^{(v)}\left[\frac{m_{\max}}{m_{1\cdot}^{(r|\cdot)}}+\frac{m_{\max}}{m_{0\cdot}^{(r|\cdot)}}\right]^2\sum_{k=1}^{n^{(v)}}\mathrm{O}\left(\left[\frac{1}{m_{1\cdot}^{(r|v)}}\right]^2\cdot\frac{1}{m_{0\cdot}^{(r|v)}}+\left[\frac{1}{m_{0\cdot}^{(r|v)}}\right]^2\right) \\
&\qquad+4\left(\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\right)^2 n^{(v)}\left[\frac{m_{\max}}{m_{1\cdot}^{(l|\cdot)}}+\frac{m_{\max}}{m_{0\cdot}^{(l|\cdot)}}\right]^2\sum_{k=1}^{n^{(v)}}\mathrm{O}\left(\left[\frac{1}{m_{1\cdot}^{(l|v)}}\right]^2\cdot\frac{1}{m_{0\cdot}^{(l|v)}}+\left[\frac{1}{m_{0\cdot}^{(l|v)}}\right]^2\right) \\
&\quad\leq 4\left(\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\right)^2 n^{(v)}\left[\frac{m_{\max}}{m_{1\cdot}^{(r|\cdot)}}+\frac{m_{\max}}{m_{0\cdot}^{(r|\cdot)}}\right]^2 n^{(v)}\mathrm{O}\left(\left[\frac{1}{m_{1\cdot}^{(r|v)}}\right]^2\cdot\frac{1}{m_{0\cdot}^{(r|v)}}+\left[\frac{1}{m_{0\cdot}^{(r|v)}}\right]^2\right) \\
&\qquad+4\left(\frac{\mathrm{N}n^{(v)}}{n^{(v)}-1}\right)^2 n^{(v)}\left[\frac{m_{\max}}{m_{1\cdot}^{(l|\cdot)}}+\frac{m_{\max}}{m_{0\cdot}^{(l|\cdot)}}\right]^2 n^{(v)}\mathrm{O}\left(\left[\frac{1}{m_{1\cdot}^{(l|v)}}\right]^2\cdot\frac{1}{m_{0\cdot}^{(l|v)}}+\left[\frac{1}{m_{0\cdot}^{(l|v)}}\right]^2\right) \\
&\quad\leq \mathrm{O}\left(\frac{1}{m_{0\cdot}^{(l|\cdot)}}+\frac{1}{m_{1\cdot}^{(l|\cdot)}}+\frac{1}{m_{0\cdot}^{(r|\cdot)}}+\frac{1}{m_{1\cdot}^{(r|\cdot)}}\right)\to 0,
\end{aligned}
$$

da $\frac{n^{(v)}}{m_{i\cdot}^{(l|\cdot)}}\leq 1$ und $\frac{\mathrm{N}}{m_{i\cdot}^{(l|\cdot)}}\leq\frac{\mathrm{N}}{n_i}\leq \mathrm{N}_0$.
Analog zeigt man:

$$
\mathrm{E}\left(\widetilde{\mathbf{V}}^{(u)}_{\mathrm{AUC},i}[l,r]-\mathbf{V}^{(u)}_{\mathrm{AUC},i}[l,r]\right)^2\leq \mathrm{O}\left(\frac{1}{m_{i\cdot}^{(l|\cdot)}}+\frac{1}{m_{i\cdot}^{(r|\cdot)}}\right)\to 0,\ i=0,1,
$$

und erhält somit die Behauptung. □

C Zusätzliche Simulationsergebnisse

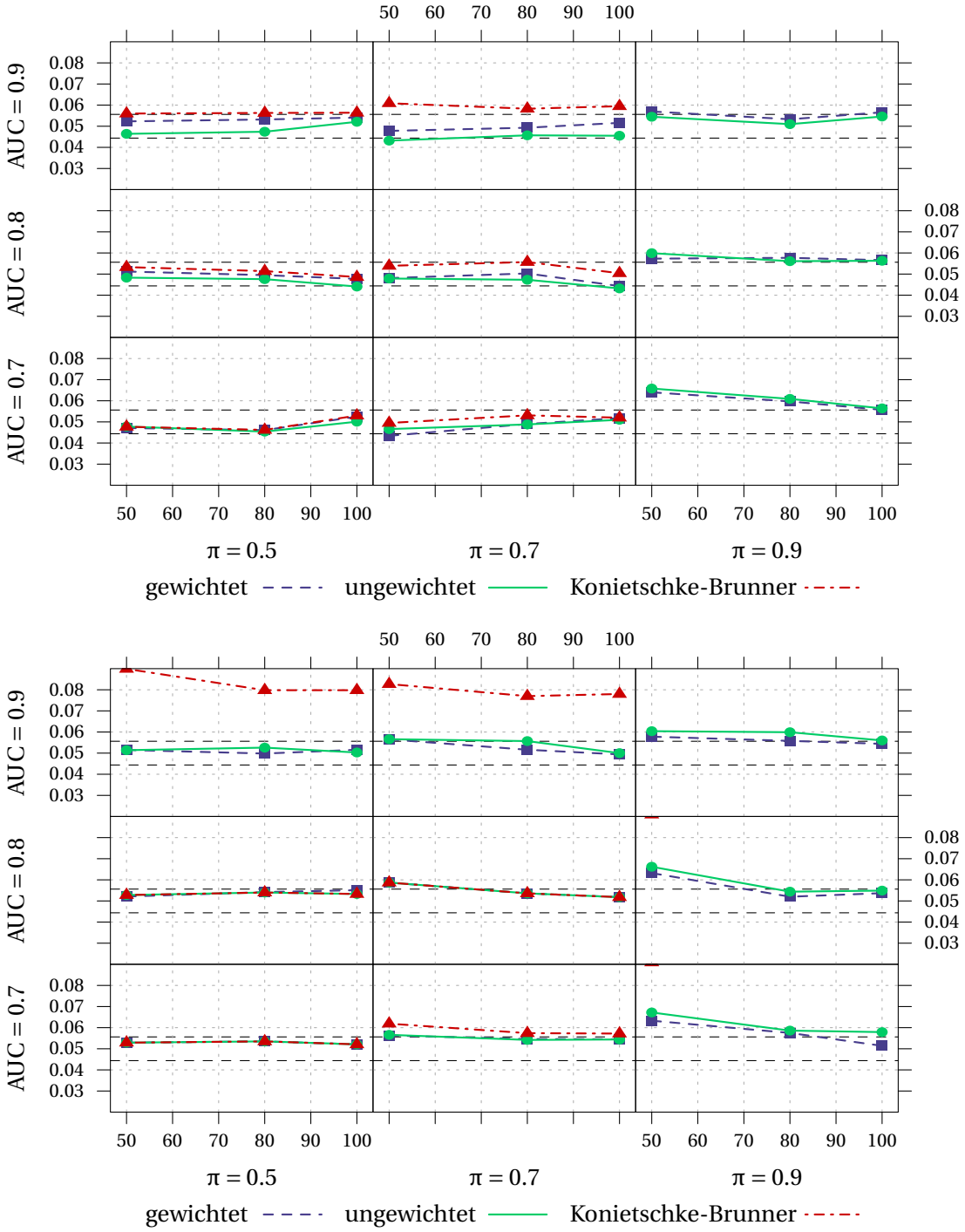

Abbildung C.1: Niveausimulation des Methodeneffekts der AUC III: Simulationen bei 5% nominellem Niveau für 2 Methoden und 3 Reader bei einer Clusterkorrelation von 0.5 / 0.75 und einer Clustergröße von 2 Beobachtungen und bei einer Clusterkorrelation von 0.25 / 0.5 und einer Clustergröße von 5 Beobachtungen; bei $\pi = 0.9$ ist die Methode nach Konietschke-Brunner stark liberal, sodass die zugehörigen Simulationsergebnisse zu Gunsten einer besseren Übersichtlichkeit nicht dargestellt sind.

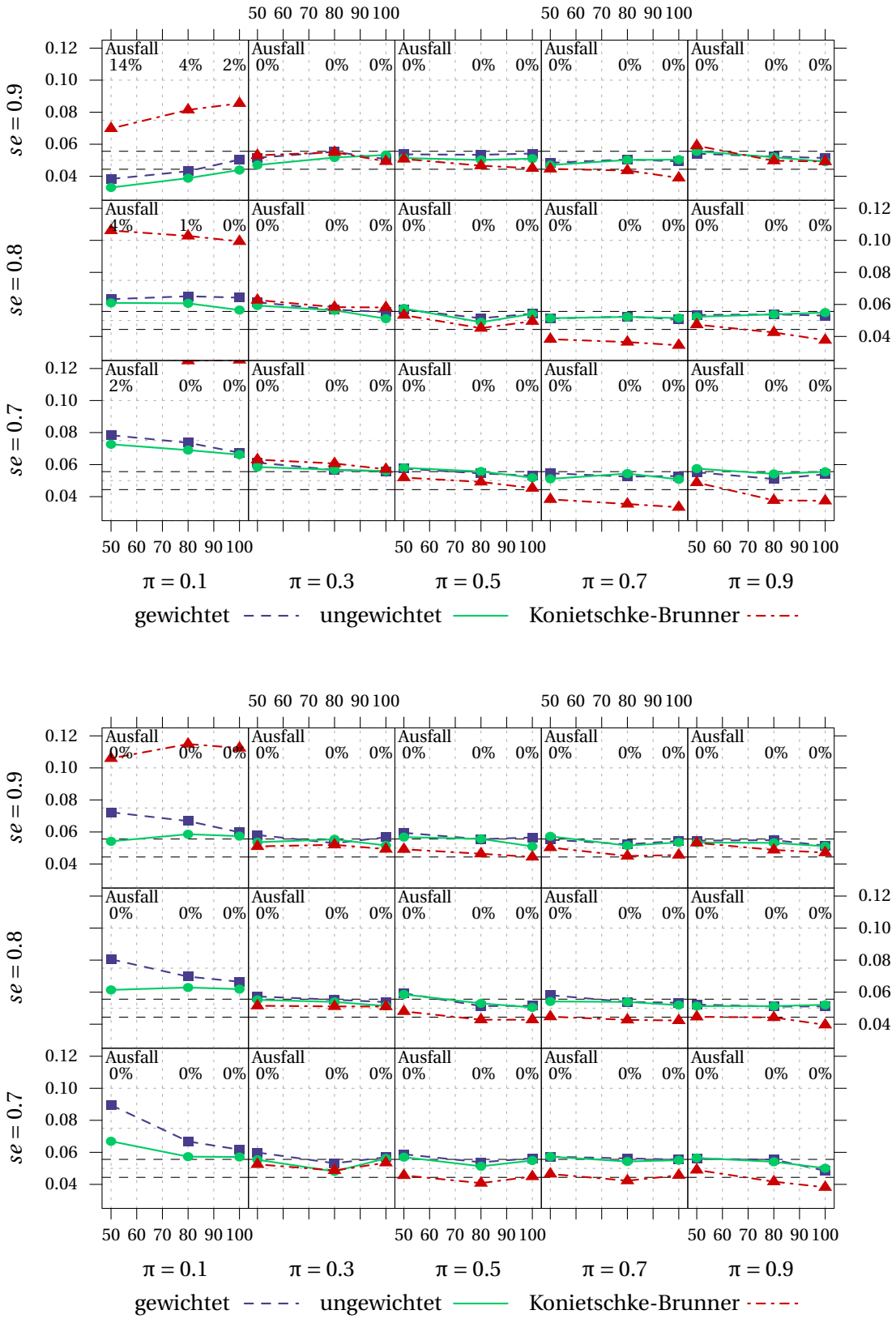

Abbildung C.2: Niveausimulation des Methodeneffekts der Sensitivität III: Simulationen bei 5% nominellem Niveau für 2 Methoden und 3 Reader bei einer Clusterkorrelation von 0.5 / 0.75 und einer Clustergröße von 2 Beobachtungen und bei einer Clusterkorrelation von 0.25 / 0.5 und einer Clustergröße von 5 Beobachtungen

Literaturverzeichnis

[Bamber 1975] BAMBER, D.: The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Characteristic Graph. In: *Journal of Mathematical Psycology* 12 (1975), S. 387–415

[Bayes 1763] BAYES, T.: An essay towards solving a problem in the doctrine chances. In: *Philosophical Transactions* 53 (1763), S. 370–418

[Beiden u. a. 2000] BEIDEN, S. V. ; WAGNER, R. F. ; CAMPELL, G.: Components-of-variance models and multiple bootstrap experiments. an alternative method for random-effects, receiver operating characteristic analysis. In: *Academic Radiology* 7 (2000), S. 341–349

[Brunner u. a. 1997] BRUNNER, E. ; DETTE, H. ; MUNK, A.: Box-Type Approximations in Nonparametric Factorial Designs. In: *Journal of the American Statistical Association* 92 (1997), S. 1494–1502

[Brunner und Munzel 2002] BRUNNER, E. ; MUNZEL, U.: *Nichtparametrische Datenanalyse*. Springer Verlag, Berlin Heidelberg, 2002

[Brunner u. a. 1999] BRUNNER, E. ; MUNZEL, U. ; PURI, M. L.: Rank-Score Tests in Factorial Designs with Repeated Measures. In: *Journal of Multivariate Analysis* 70 (1999), S. 286–317

[Brunner u. a. 2002] BRUNNER, E. ; MUNZEL, U. ; PURI, M. L.: The multivariate nonparametric Behrens-Fisher problem. In: *Journal of Statistical Planning and Inference* 108 (2002), S. 37–53

[DeLong u. a. 1988] DELONG, E. R. ; DELONG, D. M. ; CLARKE-PEARSON, D. L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. In: *Biometrics* 44 (1988), S. 837–845

[Diamond und Forrester 1979] DIAMOND, G. A. ; FORRESTER, J. S.: Analysis of Probability as an Aid in the Clinical Diagnosis of Coronary-Artery Disease. In: *New England Journal of Medicine* 300 (1979), S. 1350–1358

[Domhof 2001] DOMHOF, S.: *Nichtparametrische relative Effekte*, Georg-August-Universität Göttingen, Dissertation, 2001

[Dorfmann u. a. 1992] DORFMANN, D. D. ; BERBAUM, K. S. ; METZ, C. E.: Receiver operating characteristic rating analysis: generalized to the population of readers and patients with jackknife method. In: *Investigative Radiology* 27 (1992), S. 723–731

[Efron und Tibshirani 1993] EFRON, E ; TIBSHIRANI, R. J.: *An Introduction to Bootstrap*. Chapman & Hall, 1993

[EMEA 2008] EMEA: *Guideline on Clinical Evaluation of Diagnostic Agents (Draft)*, 2008

[EMEA 2009] EMEA: *Guideline on Clinical Evaluation of Diagnostic Agents*, 2009

[FDA 2004] FDA: *Guidance for Industry: Developing Medical Imaging Drug and Biological Products, Part 2: Clinical Indications*, 2004

[Ferguson 1996] FERGUSON, T. S.: *A Course in Large Sample Theory*. Chapman & Hall, 1996

[Gönen u. a. 2001] GÖNEN, M. ; PAMAGEAS, K. S. ; LARSON, S. M.: Statistical Issues in Analysis of Diagnostic Imaging Experiments with Multiple Observations per Patient. In: *Radiology* 221 (2001), S. 763–767

[Gut 2005] GUT, A.: *Probability: a graduate course*. Springer Science+Business Media, Inc., 2005

[Hanley und McNeil 1982] HANLEY, J. A. ; MCNEIL, B. J.: The Meaning and the Use of the Area under a Receiver Operating Characteristic Curve. In: *Radiology* 143 (1982), S. 29–36

[Janes u. a. 2008] JANES, H. ; LONGTON, G. ; PEPE, M.: Accomodating Covariates in ROC Analysis. In: *UW Biostatistics Working Paper Series* (2008)

[Janes und Pepe 2008] JANES, H. ; PEPE, M.: Adjusting for Covariates in Studies of Diagnostic, Screening, or Prognostic Markers: An Old Concept in a New Setting. In: *American Journal of Epidemiology,* 168 (2008), S. 89–97

[Jensen 1906] JENSEN, J. L. W. V.: Sur les fonctions convexes et les inégalités entre les valeurs moyennes. In: *Acta Mathematica* 30 (1906), S. 175–193

[Kaufmann u. a. 2005] KAUFMANN, J. ; WERNER, C. ; BRUNNER, E.: Nonparametric Methods for Analysing the Accuracy of Diagnostic Tests with Multiple Readers. In: *Statistical Methods in Medical Research* 14 (2005), S. 129–146

[Konietschke 2009] KONIETSCHKE, F.: *Simultane Konfidenzintervalle für nichtparametrische relative Kontrasteffekte*, Georg-August-Universität Göttingen, Dissertation, 2009

[Konietschke und Brunner 2009] KONIETSCHKE, F. ; BRUNNER, E.: Nonparametric analysis of clustered data in diagnostic trials: Estimation problems in small sample sizes. In: *Computational Statistics and Data Analysis* 53 (2009), S. 730–741

[Kulle 1999] KULLE, B.: *Nichtparametrisches Behrens-Fisher-Problem im Mehr-Stichprobenfall*, Georg-August-Universität Göttingen, Diplomarbeit, 1999

[Lange 2008] LANGE, K.: *Nichtparametrische Modelle für faktorielle Diagnosestudien*, Georg-August-Universität Göttingen, Diplomarbeit, 2008

[Lange und Brunner 2011] LANGE, K. ; BRUNNER, E.: *Sensitivity, Specificity and ROC-Curves in Multiple Reader Diagnostic Trials – A Unified, Nonparametric Approach.* 2011. – Submitted

[Leisenring u. a. 2000] LEISENRING, W. ; ALONZO, T. ; PEPE, M. S.: Comparisons of Predictive Values of Binary Medical Diagnostic Tests for Paired Designs. In: *Biometrics* 56 (2000), S. 345–351

[Lévy 1925] LÉVY, P.: *Calcul des Probabilités*. Gauthiers-Villars, Paris, 1925

[Li und Zhou 2008] LI, G. ; ZHOU, K.: A unified Approach to Nonparametric Comparison of Receiver Operating Characteristic Curves for Longitudinal and Clustered Data. In: *Journal of the American Statistical Association* 103 (2008), S. 705–713

[Liang und Zeger 1986] LIANG, K. Y. ; ZEGER, S. L.: Longitudinal data analysis using generalized linear models. In: *Biometrika* 73 (1986), S. 13–22

[Lindeberg 1922] LINDEBERG, J. W.: Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. In: *Mathematische Zeitschrift* 15 (1922), S. 211–225

[Liu u. a. 2004] LIU, B. ; METZ, C. E. ; JIANG, Y.: An ROC comparison of four methods of combining information from multiple images of the same patient. In: *Medical Physics* 31 (2004), S. 2552–2563

[Mann und Whitney 1947] MANN, H. D. ; WHITNEY, D. R.: On a test of whether one of two random variables is stochastically larger then the other. In: *Annals of Mathematical Statistics* 18 (1947), S. 50–60

[Mercaldo u. a. 2007] MERCALDO, M. D. ; LAU, F. L. ; ZHOU, X. H.: Confidence Intervals for Predictive Values with an emphasis to case-control studies. In: *Statistics in Medicine* 26 (2007), S. 2170–2183

[Mettin 1999] METTIN, C.: Global breast cancer mortality statistics. In: *CA: A Cancer Journal for Clinicians* 49 (1999), S. 135–137

[Moskowitz und Pepe 2006] MOSKOWITZ, C. S. ; PEPE, M. S.: Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. In: *Clinical Trials* 3 (2006), S. 272–279

[Munzel 1996] MUNZEL, U.: *Multivariate nichtparamatrische Verfahren für feste Faktoren in mehrfaktoriellen Versuchsanlagen*, Georg-August-Universität Göttingen, Dissertation, 1996

[Munzel und Brunner 2000] MUNZEL, U. ; BRUNNER, E.: Nonparametric methods in multivariate factorial designs. In: *Journal of Statistical Planning Inference* 88 (2000), S. 117–132

[Obuchowski 2007] OBUCHOWSKI, N. A.: New Methodological Tools for Multiple-Reader ROC Studies. In: *Radiology* 243 (2007), S. 10–12

[Obuchowski u. a. 2004] OBUCHOWSKI, N. A. ; BEIDEN, S. V. ; BERBAUM, K. S. ; HILLIS, S. L. ; ISHWARAN, H. ; SONG, H. H. ; WAGNER, R. F.: Multireader, Multicase Receiver Operating Characteristic Analysis: An Empirical Comparison of Five Methods. In: *Academic Radiology* 11 (2004), S. 980–995

[Obuchowski 1997] OBUCHOWSKI, N.A.: Nonparametric Analysis of Clustered ROC Curve Data. In: *Biometrics* 53 (1997), S. 170–180

[Pepe 2003] PEPE, M. S.: *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press Inc., New York, 2003

[Puri und Sen 1971] PURI, M. L. ; SEN, P. K.: *Nonparametric Methods in Multivariate Analysis.* Wiley, New York., 1971

[Rutter 2000] RUTTER, C. M.: Bootstrap Estimation of Diagnostic Accuracy with Patient-clustered Data. In: *Academic Radiology* 7 (2000), S. 413–419

[Ruymgaart 1980] RUYMGAART, F. H.: *A unified approach to the asymptotic distribution theory of certain midrank statistics.* Raoult, J. P. (Ed.), Lecture Notes on Mathematics, No. 821, Springer, Berlin, 1980. – S. 1–18

[Sackett u. a. 1996] SACKETT, D. L. ; ROSENBERG, W. M. C. ; GRAY, J. A. M. ; HAYNES, R. B. ; RICHARDSON, W. S.: Evidence based medicine: what it is and what it isn't. In: *BMJ (British Medical Journal)* 312 (1996), S. 71–72

[Slutsky 1925] SLUTSKY, E.: Über stochastische Asymptoten und Grenzwerte. In: *Metron* 5 (1925), S. 3–89

[Smith und Hadgu 1992] SMITH, P. J. ; HADGU, A.: Sensitivity and specificity for correlated observations. In: *Statistics in Medicine* 11 (1992), S. 1503–1509

[Sohns u. a. 2010] SOHNS, C. ; ANGIC, B. ; SOSSALLA, S. ; KONIETSCHKE, F. ; OBENAUER, S.: Computer-assisted Diagnosis in Full Field Digital Mammography –Results in Dependence of Readers Experiences. In: *The Breast Journal* 16 (2010), S. 490–497

[Song 1997] SONG, H. H.: Analysis of Correlated ROC Areas in Diagnostic Testing. In: *Biometrics* 53 (1997), S. 370–382

[Tabar u. a. 1995] TABAR, L. ; FAGERBERG, G. ; CHEN, H. H. ; DUFFY, S. W. ; SMART, C. R. ; GAD, A. ; SMITH, R. A.: Efficacy of breast cancer screening by age: new results from the Swedish Two-County Trial. In: *Cancer* 75 (1995), S. 2507–17

[Werner 2006] WERNER, C.: *Nichtparametrische Analyse von diagnostischen Tests*, Georg-August-Universität Göttingen, Dissertation, 2006

[Werner und Brunner 2007] WERNER, C. ; BRUNNER, E.: Rank methods for the analysis of clustered data in diagnostic trials. In: *Computational Statistics and Data Analysis* 51 (2007), S. 5041–5054

[Zapf 2009] ZAPF, A.: *Multivariates nichtparametrisches Behrens-Fisher-Problem mit Kovariablen*, Georg-August-Universität Göttingen, Dissertation, 2009

[Zhou und Gatsonis 1996] ZHOU, X. H. ; GATSONIS, C. A.: A simple method for comparing correlated ROC curves using incomplete data. In: *Statistics in Medicine* 15 (1996), S. 1687–1693

[Zhou u. a. 2002] ZHOU, X. H. ; OBUCHOWSKI, N. A. ; MCCLISH, D. K.: *Statistical Methods in Diagnostic Medicine*. John Wiley & Sons, Inc., New York, 2002

Lebenslauf

Persönliche Daten

Name	Katharina Lange
Anschrift	Albrecht-von-Haller-Str. 2, 37075 Göttingen
Geburtsdatum	04. August 1984
Geburtsort	Detmold
Familienstand	ledig

Ausbildung

Juni 2003	Allgemeine Hochschulreife
2003-2008	Studium im Diplomstudiengang Mathematik an der Georg-August-Universität Göttingen
Juli 2005	Vordiplom Mathematik
Oktober 2008	Diplom Mathematik Titel der Diplomarbeit: „Nichtparametrische Modelle für faktorielle Diagnosestudien“
April 2009	Aufnahme in den Promotionsstudiengang *„Applied Statistics and Empirical Methods“*

Tätigkeiten

2005-2008	Studentische Hilfskraft am Institut für Mathematische Stochastik, am Lehrstuhl für Wirtschaftstheorie und in der Abteilung Medizinische Statistik im Bereich Forschung und Lehre
seit November 2008	Wissenschaftliche Mitarbeiterin in der Abteilung Medizinische Statistik