Ann-Luise
Müller-Sarnowski

Inter Beobachter Variabilität in der pädiatrischen Bronchoskopie

Titelbild: National Cancer Institute der USA 1978, AV-7800-3633; unknown Photographer in the public domain for free reuse at https://visuals.nci.nih.gov/details.cfm?imageid=1950 & https://commons.wikimedia.org/wiki/File:Bronchoscopy_nci-vol-1950-300.jpg

Aus der
Kinderklinik und Kinderpoliklinik
im
Dr. von Haunerschen Kinderspital
der
Ludwig-Maximilians-Universität München
Vorstand: Prof. Dr. med. Dr. sci. nat. Christoph Klein

Inter-Beobachter-Variabilität in der pädiatrischen Bronchoskopie

Dissertation
zum Erwerb des Doktorgrades der Medizin
an der Medizinischen Fakultät der
Ludwig–Maximilians-Universität zu München

vorgelegt von
Ann-Luise Müller-Sarnowski
aus Freital
2017

Mit Genehmigung der Medizinischen Fakultät
der Universität München

Berichterstatter:	Prof. Dr. med. Thomas Nicolai
Mitberichterstatter:	Prof. Dr. med. Andrea Koch
Mitbetreuung durch den promovierten Mitarbeiter:	Dr. med. Felix Müller-Sarnowski (statistische Beratung & Programmierung)
Dekan:	Prof. Dr. med. dent. Reinhard Hickel
Tage der mündlichen Prüfung:	23.11.2017

INHALTSVERZEICHNIS

Danksagung

Besonderer Dank gilt Herrn Dr. Gerstlauer aus der Kinderpneumologie des Klinikums Augsburg, der im Rahmen einer Hospitation an der Dr. von Haunerschen Kinderklinik wesentlich an Mitschnitt und Zusammenstellung der Videobibliothek beteiligt war. Meinem Ehemann, Dr. Felix Müller-Sarnowski, danke ich für die umfangreiche statistische Auswertung, die erst mit maßgeschneidertem Programmcode in der Statistiksprache R möglich wurde.

Abkürzungen und Akronyme

Abkürzung	Bedeutung
a	Kranke (engl. affected)
acc	Genauigkeit (engl. accuracy)
ASGE	American Society for Gastointestinal Endoscopy
AUC	Area Under Curve
B	Bangdiwala
Bhp	Bhapkar-Test
c	Gesundheitszustand (engl. condition)
CART	Classification and Regression Trees
DICOM	Digital Imaging and Communications in Medicine
ESGE	European Society for Gastointestinal Endoscopy
exp	Erwartungswert
f	Falsche
fdr	false discovery rate
fn	Falsch Negative (engl. false negative)
for	false omission rate
fp	Falsch Positive (engl. false positive)
fpr	false positive rate = Alpha
iac	Ungenauigkeit
kC	Kappa Cohen
kF	Kappa Fleiss
LR-	negative Likelihood ratio
LR+	positive Likelihood ratio
McN	MvNemar-Test
mpa	durchschnittliche positive Übereinstimmung
MST	Minimal Standard Terminology
n	Negative
NA	not applicable / not available
nag	negative Übereinstimmung
npv	negativer prädiktiver Wert
odd	odds
OMED	World Organisation of Digestive Endoscopy
OR	odds ratio
p	Positive
pag	positive Übereinstimmung
pca	Prozentuale Übereinstimmung (engl. percentual agreement)
pcn	Präzision (engl. precision)
ppv	positiver prädiktiver Wert
pre	Prävalenz
Q	Yule Q
ROC	Receiver Operator Characteristic
s	Stichprobe (engl. sample)
sen	Sensitivität
spe	Spezifität
t	Richtige
tn	Richtig Negative (engl. true negative)
tnr	true negative rate
tpr	true positive rate
tpr	Richtig Positive (eng, ture positive)
tpr	true positive rare = Beta
Y	Yule Y

1 Einführung

Kapitelverzeichnis

1.1 Fragestellung

Die Bronchoskopie der Atemwege ist als diagnostisches wie auch interventionelles Verfahren ein wichtiger Bestandteil der pädiatrischen Pneumologie. Wie alle endoskopischen Methoden ist die Bronchoskopie von Beobachtung und Geschick des durchführenden Arztes abhängig. Im Gegensatz zu technischen und laborchemischen diagnostischen Hilfsmitteln, ist über die Verlässlichkeit bronchoskopischer Befunde wenig bekannt. Testmetrische Kennwerte wie Sensitivität, Spezifität, positive und negative prädiktive Werte, Likelihood- und odds ratios (OR) fehlen in der Literatur. Die Ausbildungscurricula der pädiatrischen Bronchoskopie basieren auf Expertenmeinungen. Empirische Untersuchungen, welche Faktoren für eine gute Ausbildung tatsächlich ausschlaggebend sein könnten wurden bislang nicht durchgeführt. Diese Studie möchte einen Beitrag zu Beantwortung dieser offenen Fragen leisten.

1.2 Ziel dieser Studie

Ziel dieser Arbeit ist die Untersuchung der Inter-Beobachter-Variabilität in der pädiatrischen Bronchoskopie und ihre Abhängigkeit von Faktoren wie Ausbildung und klinischer Erfahrung. Es sollen die diagnostischen Stärken wie auch Schwächen der pädiatrischen Bronchoskopie aufgezeigt und die Verlässlichkeit von bronchoskopischen Befunden abgeschätzt werden. Diese Daten sind Grundlage für die Entwicklung eines einheitlichen, zentrenübegreifenden Befundschemas. Ein klar definiertes, gleichzeitig mensch- und maschinenlesbares Befundsystem („literate programming“ (Knuth, 1984))mit einheitlicher Terminologie ist die Grundlage für

- die elektronische Archivierung in standardisierten Formaten (z. B. DICOM SR (Hussein u. a., 2004a, 2004b)),
- Qualitätsmanagement (Häussinger u. a., 2004),
- eine evidenzbasierte Ausbildung,
- Epidemiologische Untersuchungen,
- telemedizinische Anwendungen inklusive
- Expertensysteme (engl. clinical decision support, CDS) und
- multizentrische klinische Studien.

Im Unterschied zur Laborforschung sind klinische Zielvariablen meist nicht objektiv messbar, sondern beruhen auf subjektiven Beurteilungen von Ärzten. Die Abschätzung der Inter-Beobachter-Variabilität klinischer Befunde und ein darauf aufbauendes definiertes Befundsystem leisten daher auch einen wesentlichen Beitrag zur translationalen Forschung: denn erst Informationen zur Inter-Beobachter-Variabilität ermöglichen die Auswahl verlässlicher klinischer Zielgrößen, die zur Evaluation von Erkenntnissen aus der Grundlagenforschung herangezogen werden können. Daten zur Inter-Beobachter-Variabilität ermöglichen auch den Vergleich der Effektivität verschiedener diagnostischer Verfahren. Dabei stellt sich entgegen den Erwartungen nicht selten heraus, dass klinische Verfahren trotz aller Subjektivität den vermeintlich objektiveren technischen Verfahren überlegen sein können. Beispielsweise wurde gezeigt, dass drei einfache klinische Tests die in der Notfallversorgung wichtige Differentialdiagnose zwischen zentralem und peripherem Schwindel sensitiver beantworten, als eine Magnetresonanztomographie (Kattah u. a., 2009).

1.3 Vorstellung der Studie

Für diese Studie zur Inter-Beobachter-Variabilität in der pädiatrischen Bronchoskopie wurden 50 Mitglieder der Arbeitsgruppe Bronchoskopie der Gesellschaft für pädiatrische Pneumologie

(GPP) gebeten, eine Videobibliothek von 42 Bronchoskopiemitschnitten mit einem standardisierten Befundbogen zu beurteilen. Die Videoaufzeichnungen wurden dem Bronchoskopiearchiv des Dr. v. Haunerschen Kinderspitals München im Zeitraum von 1999 bis 2001 entnommen und so zusammengestellt, dass das Befundspektrum der pädiatrischen Bronchoskopie weitgehend abgedeckt ist. Es wurden weder begleitende klinische noch demographische Angaben der Patienten übermittelt. Alle Befunde entstammen dem normalen Klinikalltag und wurden ohne spezielle Aufnahmetechnik mitgeschnitten. Der Indikation entsprechend wurden wechselnd starre oder flexible Bronchoskopien in unterschiedlicher Länge und bei variabler Bildqualität durchgeführt. Wie im klinischen Alltag überwiegt der Anteil flexibler Bronchoskopien insgesamt deutlich. Zwanzig Mitgliedern der Arbeitsgruppe Bronchoskopie sandten ausgefüllte Befundbögen zurück.

1.4 Historische Entwicklung der Bronchoskopie

Der HNO-Arzt Gustav Killian gilt als Wegbegründer der Bronchoskopie. Er entfernte 1897 erstmals mittels eines Bronchoskops einen Fremdkörper aus der Trachea. Er experimentierte seit 1896 an der Universität Freiburg mit neuen endoskopischen Techniken und publizierte seine Ergebnisse 1898 in Heidelberg. In seiner Veröffentlichung „Ueber die directe Bronchoskopie“ (Killian, 1898) beschrieb er die Biegsamkeit und Verschieblichkeit der Atemwege und dokumentierte sein Vorgehen. Er benutzte die bereits für die Oesophagoskopie vorhandenen Röhren und führte sie unter Vorziehen des Kehldeckels bis in die Hauptbronchien ein. Damit erweiterte er die durch Kirstein begründete direkte obere Tracheoskopie um die Einsicht in tiefer gelegene Atemwege. Kilian beschränkte sich keineswegs auf die „Einsicht“ der Trachea, sondern unternahm unter Kokainanästhesie der Atemwege auch erste Interventionen, wie Dilatationen der Trachea und Stentimplantationen (Übersicht in Becker, 2010; Killian, 1898; Reprint in Kropp, 2005).

Ikeda, Thoraxchirurg am National Cancer Center Hospital in Tokyo, gelang es durch Weiterentwicklung flexibler Bronchoskope neue diagnostische und therapeutische Verfahren zu etablieren. Er führte die, ebenfalls bereits in der gastrointestinalen Endoskopie verwendeten, fiberoptischen Geräte ein und entwickelte sie weiter. Im Jahr 1966 präsentierte er in Kopenhagen ein flexibles Bronchoskop mit einem Durchmesser von 6 mm und einem Fiberglasfaserbündel mit 15000 Glasfasern (Becker, 2010). Die Anwendung der flexiblen Bronchoskopie beschränkte sich in der ersten Zeit auf die Endoskopie von Erwachsenen, da die Größe der Instrumente die Untersuchung der kindlichen Atemwege nicht zuließ. Erst im Jahre 1978 wurde ein flexibles Bronchoskop verfügbar, das klein genug für die kindlichen Atemwege war. Dieses Instrument hatte einen Durchmesser von 3,5 mm mit einem Saugkanal von 1,2 mm und wurde in den 1980er Jahren kommerzialisiert (Wood, 2001: S. 311–312).

Weitere technische Fortschritte in Größe und Qualität der Endoskope sowie die Entwicklung kleiner Videobronchoskope verbesserten die Betrachtungsmöglichkeit und Therapieoptionen der kindlichen Atemwege. Neben Kaltlichtquellen wurden auch Saugkanäle in kindliche Bronchoskope integriert, was das Spektrum der Interventionsmöglichkeiten erheblich erweiterte. Im Jahr 1999 wurde ein im Durchmesser 2,8 mm großes Endoskop mit Saugkanal und Fiberoptik entwickelt (Wood, 2001: S. 311). Damit ist die pädiatrische Bronchoskopie heutzutage auch schon bei Kleinstkindern im Früh- und Neugeborenenalter möglich.

1.5 Besonderheiten pädiatrischer Bronchoskopie

Die Besonderheit der pädiatrischen Bronchoskopie ergibt sich aus den anatomischen Besonderheiten des kindlichen Respirationstraktes. Im Vergleich zum Erwachsenen ist dieser nicht nur viel kleiner, sondern reagiert auch empfindlicher auf Reize und mit größeren Auswirkungen auf die

Atemfunktion. Der Durchmesser der Trachea eines Frühgeborenen weist, im Gegensatz zu dem eines Erwachsenen, ein Verhältnis von 1 : 5 auf (Mantel u. a., 1995: S. 9).

Anhaltender Stridor und obstruktive Dyspnoe, die meist altersspezifisch auftreten, sind die wichtigste Indikation für eine pädiatrische Bronchoskopie. Während im frühen Neugeborenenalter angeborene Fehlbildungen und Verengungen der Atemwege ursächlich sind, treten mit zunehmenden Alter Erkrankungen wie zum Beispiel Infektionen der Atemwege, Tumore, Fremdkörperaspirationen, Stenosen nach Intubation, Herzfehler mit Links-Rechts-Shunt und Gefäßkompressionen der Trachea in den Vordergrund. (Mantel u. a., 1995: S. 9; Nicolai, Griese, 2010: S. 6)

Die Wahl zwischen starrer und flexibler Bronchoskopie hängt von Faktoren wie Alter, Beschwerdebild und etwaigen Interventionen ab. Während die starre Bronchoskopie lange als Standardmethode der kindlichen Bronchoskopie galt (Mantel u. a., 1995: S. 10), tritt heute durch den Fortschritt von Anästhesie und Technik die flexible Bronchoskopie in den Vordergrund.

Die flexible Bronchoskopie erleichtert die Einsicht in schwer zugängliche Bereiche, wie den Lungenoberlappen und distale Verzweigungen des Bronchialbaums. Sie ist ein für das Kind verhältnismäßig schonendes und zeitsparendes Verfahren. Im Gegensatz dazu ist die starre Bronchoskopie nur in Sedierung und lokaler Anästhesie durchführbar. Letztere gilt jedoch als unverzichtbar zur genauen Darstellung anatomischer Veränderungen bei unklaren Atemwegsobstruktionen, Stentimplantationen, Lasertherapie, Ballondilatationen und bei Interventionen wie z. B. Fremdkörperaspirationen und der diagnostischen bronchoalveolärer Lavage (zur Erregergewinnung z. B. bei Tuberkulose, Immundefekten).

Eine spezielle Methode zur Darstellung des Larynx und der Trachea ist die Stützautoskopie, bei der die Atemwege ohne Intubation mit einer dünnen, starren Optik bei fest eingestelltem Larynxspatel dargestellt werden. Dabei wird das Laryngoskop an einem Stützrohr arretiert und dieses höhenverstellbar und schwenkbar an einem Metallbänkchen am Untersuchungstisch fixiert. Der Untersucher hat damit einen sicheren und kontinuierlichen Überblick über den Kehlkopfeingang. Er kann beide Hände zu Eingriffen z. B. Intubieren oder Absaugen nutzen (Mantel u. a., 1995: S. 10).

1.6 Telemedizin

Zentrales Anliegen der Telemedizin ist die Verbesserung und Sicherung der medizinischen Versorgung durch Anwendung von Telekommunikation. In der Endoskopie kann dies beispielsweise durch den Transfer von Expertenwissen aus Zentren in die Peripherie erreicht werden (Wildi u. a., 2004). Obwohl die technischen Voraussetzungen für Telemedizin in der Bronchoskopie spätestens seit der Jahrtausendwende gegeben sind (Kim u. a., 2000) und Telemedizin in anderen Bereichen der Pädiatrie bereits etabliert ist (Burke u. a., 2015), sind telemedizinische Anwendungen in der Bronchoskopie derzeit (2016) noch eine Ausnahme. Aus Untersuchungen in der Endoskopie lässt sich analog schließen, dass mittels Kompressionsverfahren bewegte Bilder in Echtzeit über größere geographische Entfernung hinweg zur Mitbeurteilung durch Spezialisten in ausreichender Qualität übermittelt werden können (Ohashi u. a., 2008: S. 165). Damit ist es möglich Expertenmeinungen aus spezialisierten Zentren in die Peripherie zu transferieren. Zusätzlich bietet die Telematik eine Möglichkeit zur Verbesserung der Ausbildung.

1.7 Befunddokumentation

Für die Dokumentation bronchoskopischer Befunde bestehen bislang weder Leitlinien, noch liegt eine international anerkannte Terminologie vor (Ernst, Becker, 2001). In der Literatur finden sich nur sehr vereinzelt Vorschläge für einheitliche Befundschemata, die jedoch offenbar nie auf

größere Resonanz stießen (Caliebe, 1968; Deutsche Gesellschaft für Endoskopie, 1974; Ernst, Becker, 2001; Nakhosteen, Zavala, 1983; Oho u. a., 1986). Damit steht die Bronchoskopie zurück hinter verwandten diagnostischen Verfahren, wie der gastrointestinalen Endoskopie. Der europäischen Gesellschaft für gastrointestinale Endoskopie (ESGE) gelang es 1994 mit der sogenannten „Minimal Standard Terminology“ (MST) eine weltweit anerkannte Nomenklatur für die gastroenterologische Endoskopie zu etablieren (Fujino u. a., 2006). Die MST hielt auch einer prospektiven multizentrischen Validierung (Delvaux u. a., 2000) stand. Die MST, die inzwischen von der World Organisation of Digestive Endoscopy (OMED) gepflegt wird, liegt inzwischen in Version 3.0 vor, die das Vokabular auf endoskopischen Ultraschall, Enteroskopie und Komplikationen ausgeweitet hat (Aabakken u. a., 2009). Mit der Integration (Korman, Bidgood, 1997; Korman u. a., 1998) in das DICOM[1] Format (Graham, Perriss, und Scarsbrook 2005) steht auch ein einheitliches Format zur Archivierung bereit. In den wenigen Publikationen zur Befunddokumentation in der Bronchoskopie besteht dahingehend Konsens, dass der schriftliche Befund um Bildmaterial ergänzt werden sollte. Mitschnitte der Untersuchung gelten allgemein als hilfreich für Befunddemonstration, Qualitätsmanagement und Ausbildung. Eine Übersichtsarbeit des ASGE[2] Technology Committee stellt handelsübliche Systeme für die medizinische Bilddokumentation in der Endoskopie vor (Murad u. a., 2014).

Eine auf die praktische Anwendung in der Bronchoskopie zugeschnittene anatomische Nomenklatur der Bronchien wurde bereits 1943 vorgeschlagen (Jackson, Huber, 1943) und mit zunehmender Reichweite der immer dünneren Bronchoskope weiter verfeinert (Mortensen u. a., 1983). Heute gelingt bereits die weitgehend automatisierte anatomische Klassifizierung des Bronchialbaumes (Tschirren u. a., 2005).

Um bronchoskopische Befunde für epidemiologische Untersuchungen, wissenschaftliche Studien, Qualitätsmanagement und Expertensysteme (engl. clinical decision support systems) zu erschließen, ist eine Verschlüsselung der Befunde in ein maschinenlesbares Format notwendig. Bereits 1968 wurde in Kiel ein entsprechender Befundbogen im Lochkartenformat vorgeschlagen (Caliebe, 1968) kurz darauf auch eine umfangreichere Version in Dresden (Wunderlich, 1969). Wenige Jahre später wurden die Vorschläge in einem dritten Befundbogen erneut aufgegriffen (Deutsche Gesellschaft für Endoskopie, 1974). Seither finden sich in der Literatur jedoch kaum Ansätze zu einer Standardisierung der Befunddokumentation in der Bronchoskopie. Die japanische Gesellschaft für Lungenkrebs etablierte in den 1980er Jahren ein Befundschema für Lungenmalignome (Oho u. a., 1986). Ende der 1980er Jahre rief Schindel im Rahmen einer Übersichtsarbeit erneut zu einer einheitlichen Befunddokumentation in der Bronchoskopie auf (Schindl u. a., 1988). Prakash und Edell gaben Mitte der 1990er Jahre allgemeine Empfehlungen zur Bilddokumentation (Prakash, Edell, 1996). Erst zur Jahrtausendwende schlugen Ernst und Becker ein weiteres standardisiertes Bronchoskopieformular vor (Ernst, Becker, 2001). Das Thema einer einheitlichen Befunddokumentation wurde vorwiegend von deutschsprachigen Autoren aufgegriffen.

Bei der Verschlüsselung sollen einerseits wenig Information verloren gehen, gleichzeitig muss der Befundbogen übersichtlich, intuitiv verständlich und leicht auswertbar sein — idealerweise automatisch. Um die Inter-Beobachter-Variabilität untersuchen zu können, wurde für diese Studie ein einfacher Befundfragebogen im Multiple-Choice-Format entwickelt, der sich an den genannten Vorarbeiten orientiert, sich dabei aber auf die wichtigsten Befunde in der pädiatrischen Bronchoskopie beschränkt.

[1] *Digital Imaging and Communications in Medicine (DICOM), ist ein offener Standard für Speicherung und Austausch medizinischer Bildgebung*

[2] *American Society for Gastrointestinal Endoscopy (ASGE)*

1.8 Ausbildung und Qualifikation

Auf Antrag des Vorstandes der Gesellschaft für pädiatrische Pneumologie (GPP) bei der Bundesärztekammer wurde die Zusatzweiterbildung Pädiatrische Pneumologie 2003 vom Bundesärztetag in die Musterweiterbildungsordnung aufgenommen. Seit 2004 ist Kinderpneumologie ein von der Ärztekammer anerkannter Schwerpunkt der Kinderheilkunde. Um die Bezeichnung Kinder-Pneumologie führen zu dürfen, müssen Kinderärzte über die Facharztausbildung zum Kinderarzt hinaus eine 3-jährige Weiterbildung in einem Zentrum für lungenkranke Kinder durchlaufen, wobei ein Kriterienkatalog zu erfüllen ist. Die Ausbildung schließt mit einer Prüfung in der Landesärztekammer ab.

Die für die fakultative Weiterbildung angelegten Maßstäbe orientieren sich an internationalen Kriterien, wie sie bereits in der Schweiz und in den USA (American Thoracic Society, 1997) existieren. Dabei handelt es sich um ein von Experten zusammengestelltes Curriculum. In dieser Studie wurden Anhaltspunkte für ein evidenzbasiertes Ausbildungscurriculum ermittelt. Hierzu wurde die Befundrichtigkeit im Vergleich zur Referenz mit Informationen zu Ausbildung und Erfahrung des Untersuchers korreliert.

1.9 Technische Bildqualität

Die technische Bildqualität endoskopischer Bildgebung lässt sich im Wesentlichen anhand der

- Schärfe und
- Farbdynamik

abschätzen. Die zeitliche Auflösung spielt bei Bildfrequenzen von ≧ 25 Hz, wie sie heute verfahrensunabhängig allgemein üblich sind, in der Bronchoskopie keine wesentliche Rolle, denn die Lunge ist ein vergleichsweise statisches Organ. Die Bildqualität kann sich in der Bildverarbeitungskette erheblich verändern. Insbesondere Speicherung und Fernübertragung der Bildinformation können zum Flaschenhals der Bildqualität werden. So gehen beispielsweise bei der Aufzeichnung auf VHS-Bänder große Anteile der ursprünglichen Bildinformation verloren (siehe Tabelle 1.1 Seite 7). Eine Nachbefundung kann hierdurch eventuell beeinträchtigt werden.

Die digitale Speicherung auf DVD oder Blu-Ray Disc verspricht einen Fortschritt, ist aber wohl noch keine endgültige Lösung. Bereits eine gewöhnliche PAL-Auflösung verursacht bei 24 Bit Farbtiefe und 25 Vollbildern pro Sekunde unkomprimiert einen Datenstrom von 31 MB/s entsprechend 18 Gigabyte/min (Grund, Salm, 2007: S. 364). Im Bereich der Bronchoskopie mit Untersuchungszeiten von meist nur wenigen Minuten ist damit theoretisch eine nahezu verlustfreie Dokumentation des PAL-Signals denkbar. Heute sind jedoch meist Endoskope im Einsatz, die wesentlich höhere Auflösungen als PAL akquirieren und damit erneut den Rahmen der verfügbaren portablen Speichermedien sprengen.

In der Praxis wird der Datenstrom daher mit verlustbehafteten Algorithmen wie MPEG-4 komprimiert. Diese für konventionelle Filmaufnahmen entwickelten Algorithmen verursachen Artefakte, die im Verdacht stehen, die Befunderhebung zu behindern. Untersuchungen aus der Echokardiografie (Gopalswamy u. a., 1997) und Dermatologie (Seidenari u. a., 2004) konnten jedoch keine signifikante Beeinträchtigung der Befundung durch die Kompressions-Algorithmen MPEG-4 und JPEG feststellen. Die Ergebnisse der vorliegenden Studie legen nahe, dass die Bildqualität für die Beurteilung der meisten Krankheitsbilder selbst bei Aufzeichnung auf VHS immer noch ausreichend ist.

1.9.1 Bildschärfe

Die Schärfe ist ein Maß der Detailtreue eines Bildes und wird oft vereinfachend mit dem Begriff Auflösung vermengt. Die Auflösung, die seit dem Siegeszug digitaler Bilder meist in Pixeln angegeben wird, wird intuitiv als wichtigstes Maß der Bildqualität angesehen. Sie ist jedoch nur einer mehrerer Faktoren – wie z. B. Helligkeit, Kontrast, Rauschen – welche die Bildschärfe beeinflussen.

Da es in der endoskopischen Diagnostik vorwiegend um die Beurteilung von Oberflächenstrukturen geht, kommt der Schärfe des Bildmaterials eine entscheidende Bedeutung zu. Tabelle 1.1 gibt einen Überblick über in der Bronchoskopie eingesetzte Normen für bewegte Bilder. Im Sinne einer anschaulichen Darstellung ist, stellvertretend für viele andere Parameter der Bildqualität, jeweils die Auflösung in Pixeln angegeben[3].

Tabelle 1.1: Gängige Normen bewegter Bilder in der Bronchoskopie

	Endoskop		**Monitor**					**Speicherung**			
	Technik	Pixel	Norm	Bildqualität		~Pixel	H:B	Norm	Bildqualität		~Pixel
				Zeilen[4] /Hz/MHz	Pixel				Linien[5]/MHz	Pixel	
analog	Glasfaser	15-40k	PAL-I[6]	625/50/55	~520[7]×576[8]	300k	4:3	VHS-PAL	230/34	~3107×576	178k
			NTSC	525/60/42	~440[7]×486[9]	210k	4:3	VHS-NTSC		~3207×480	153k
								U-Matic SP	330/5	~440×480	211k
								SVHS	400/54	~520×576	300k
digital	Video Sensor	40-80k	PAL[10]		720×576[11]	415k	5:4	DVD		720×576	415k
	HR-Video Sensor	400-800k	SXGA medical HD		1280×1024	13M	5:4				
			HDTV 720p[12]		1280×720	921k	16:9	Blu-ray 720p12		1280×720	921k
			HDTV 1080i[13]		1920×1080	2M	16:9	Blu-ray 1080i13		1920×1080	2M

Modifiziert und ergänzt nach einer Darstellung von Grund (Grund, Salm, 2007) und enzyklopädischen Angaben.

Analoge PAL-Signale (z. B. klassisches Fernsehen) konnten erstmals mit der Einführung von S-VHS, digitale PAL-Signale erst mit der DVD weitgehend verlustfrei gespeichert werden. Heute sind Endoskope verbreitet, deren Auflösung deutlich über den PAL-Standard hinaus geht. Das macht zur Bilddarstellung hochauflösende Computermonitore erforderlich. Eine verlustfreie Speicherung ist z. B. über ein PACS[14] möglich. Für den Datenaustausch wird zumeist auf Kompressionsverfahren und aktuell (2016) gängige portable Speichermedien wie DVD und Blu-ray zurückgegriffen.

[3] Für analoge Signale kann die Auflösung verfahrensbedingt nur näherungsweise in Pixeln angegeben werden.

[4] Zeilen in vertikaler Richtung (im Unterschied zu horizontalen Linien; siehe weiter rechts in der Tabelle)

[5] Ungefähre horizontale Linienzahl

[6] In Deutschland dominiert wie in den meisten anderen europäischen Ländern PAL.

[7] Die horizontale Auflösung kann bei analogen Verfahren nicht exakt in Bildpunkten angegeben werden. Der hier genannte Wert wurde auf Basis des Nyquist-Shannon-Theorems geschätzt.

[8] Im analogen PAL-Standard werden 625 Zeilen kodiert, wovon aber nur 576 die sichtbare Bildinformation tragen.

[9] Im analogen NTSC-Standard werden insgesamt 525 Zeilen kodiert wovon 486 sichtbare Bildinformation tragen.

[10] Digitales PAL entspricht der Norm CCIR 601 bzw. ITU-R 601. Pixel sind hier im Gegensatz zum analogen PAL nicht rechteckig definiert.

[11] Umrechnung analog zu digital nach der Norm CCIR 601.

[12] p steht hier für „progressive". Das bedeutet in diesem Zusammenhang, dass es sich um ein Vollbildformat handelt.

[13] i steht hier für „interlaced". Das bedeutet in diesem Zusammenhang, dass es sich um ein Halbbildformat handelt.

[14] Picture Archiving and Communication System (PACS)

Da die im Allgemeingebrauch heute üblichen Formate in einem am Kinobild orientierten Seitenverhältnis von 16 : 9 für die runden Bilder der Endoskopie ungünstig sind[15], verwenden viele Hersteller qualitativ vergleichbare Computerauflösungen im annähernd quadratischen Format 5 : 4. In diesem Zusammenhang wird der SXGA-Standard (1280 * 1024 Pixel) als „medical HD“ bezeichnet.

1.9.2 Farbdynamik

Im Vergleich zur Schärfe spielt die Farbdynamik eine untergeordnete Rolle. Die Möglichkeiten handelsüblicher, primär für Heimanwender konzipierter Technologien, reichen auch für den professionellen medizinischen Einsatz meist aus. Der in der digitalen Welt verbreitete TrueColor-Farbraum umfasst bei 24-Bit Farbtiefe ca. 167 Millionen Farben, die den wesentlichen Teil des vom Menschen erfassbaren Farbspektrums abdecken. Der Farbraum des menschlichen Auges wird in der Literatur sehr unterschiedlich angegeben, teilweise auf nur 5500 (British Colour Council, 1946; Gekeler, 2007) bzw. ca. 1 Million (Lee, 2005: S. 359) klar differenzierbare Farben geschätzt. In der digitalen Radiologie werden Graustufen heute verbreitet mit 10 Bit entsprechend 1024 Abstufungen codiert (Freyschmidt u. a., 2002: S. 71). Computerbildschirme sind gewöhnlich auf die Darstellung von 256 Grautönen ausgelegt (TrueColor-Farbraum bei 24 Bit Farbtiefe). Angesichts der in Lehrbüchern der Radiologie verbreiteten Annahme, das menschliche Auge könne unter Alltagsbedingungen nur etwa 20 - 30 Grautöne sicher unterscheiden, ist dies ebenfalls kein limitierender Faktor für die Befundung (Bücheler u. a., 2006: S. 75; Laubenberger, Laubenberger, 1999: S. 228; Schlungbaum u. a., 1993: S. 176). Der im Vergleich zu den auf Bildschirmen darstellbaren 256 Grautönen erheblich größere Umfang an Graustufen in digitalen Bilddateien wird mithilfe der sogenannten „Fensterung“ erschlossen. Bei der Fensterung wird jeweils nur ein Ausschnitt („Fenster“) der insgesamt verfügbaren Graustufen dargestellt.

1.10 Krankheitsbilder

Die in vorliegender Studie untersuchten Erkrankungen der Atemwege lassen sich nach ihrer anatomischen Lage in Pathologien des Larynx, der Trachea und der Bronchien einteilen. Die folgende Kurzübersicht der wichtigsten Krankheitsbilder orientiert sich an der Darstellung in gängigen Lehrbüchern (Mantel u. a., 1995; Nicolai, Griese, 2010).

1.10.1 Erkrankungen des Larynx

Veränderungen im Larynx führen häufig zu einer Verengung der Atemwege und zeigen als Leitsymptom einen inspiratorischen Stridor.

1.10.1.1 Supraglottische Pathologien

Veränderungen im supraglottischen Bereich können z. B. durch orofaciale Fehlbildungen wie Mikrognathie mit zurückfallender Zunge (Pierre-Robin-Sequenz), Makroglossie, Zungengrundstruma, Zysten im Bereich der Zunge, Tonsillen-Teratome, nasopharyngeale Angiofibrome und zystische Hygrome hervorgerufen werden. Zusammenfassend führen diese Veränderungen vorwiegend zu einer Verengung der oberen Atemwege. Bei Zungenzysten, Tonsillen-Teratomen, nasopharyngeale Angiofibromen und zystische Hygromen imponiert meist eine Kompression von außen (Mantel u. a., 1995).

[15]Bei 16 : 9 bleiben bei vollständiger Abbildung des runden Endoskopbildes über 50 % der Bildfläche ungenutzt. Bei bildschirmfüllender Darstellung stellt ein 16 : 9 Format nur 43 % ein 5 : 4 Format immerhin 62 % des runden Endoskopbildes dar.

1.10.1.2 Glottische & Subglottische Pathologien

Zu den häufigsten Veränderungen im kindlichen Larynx gehört die infantile Laryngomalazie, die meist mit einer Tracheomalazie einhergeht. Sie stellt eine strukturelle Besonderheit im Säuglingsalter dar und führt zu einem Zusammenfallen des Larynx im Sinne einer Malazie bis hin zur Ateminsuffizienz. Auch Veränderungen der Stimmbänder stellen einen grossen Anteil der Erkrankungen im Larynx. Sie können angeboren oder erworben sein.

Tabelle 1.2: angeborene und erworbene glottische & subglottische Erkrankungen

angeboren	erworben
Larynxspalten	Stimmbandpolypen
Stimmlippensynechien	Larynxfremdkörper
Defekte von Schild-und Ringknorpel	Intubationsgranulome
Epiglottisdysplasien	Stimmbandparesen • nach Verschluss des PDA • bei zentralen Erkrankungen wie ◦ Meningomyelozelen und der ◦ Arnold-Chiari-Malformation oder auch • nach Traumata (z. B. Geburtstraumata)
Kehlkopfzysten	
Laryngozelen,	
kongenitale Diaphragmen,	
Hämangiome	

Die Tabelle stellt die häufigsten angeborenen den häufigsten erworbenen Erkrankungen gegenüber.

Lähmungen und Bewegungsstörungen der Stimmbänder schränken die Zirkulation der Atemluft ein und können zu Schluckstörungen, Aspirationen und Dyspnoe führen. (Mantel u. a., 1995)

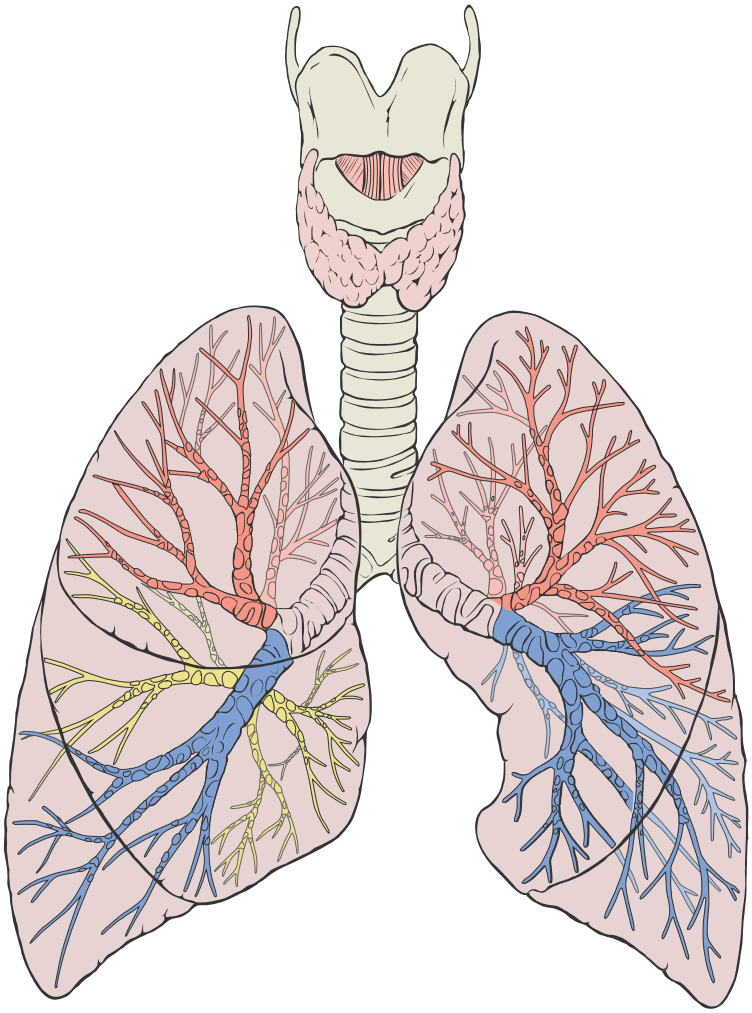
1.10.2 Erkrankungen der Trachea & der Bronchien:

Veränderungen in Trachea gehen ebenfalls mit Stridor und mit Atemnot einher. Wie im Larynx kann man auch hier angeborene und erworbene Veränderungen unterscheiden. Die Tracheomalazie als Besonderheit der kindlichen Atemwege ist ein angeborener oder erworbener Stabilitätsverlust der Trachea mit Erweichung der Knorpelringe. Sie kann segmental oder generalisiert zu einem exspiratorischen Kollaps des betroffenen Trachealsegments führen und in Verbindung mit Laryngo- oder Bronchomalazien auftreten.

Angeborene Veränderungen der Trachea sind tracheale Fisteln (z. B. als tracheoösophageale Fistel bei der Ösophagusatresie Typ IIIb nach Vogt), Hypoplasien, Zysten und Hämangiome, Knorpelfehlbildungen und Schleimhautfalten. Erworbene Veränderungen der trachealen Atemwege treten nach Intubationen, Fremdkörpern oder durch Kompression von außen auf (z. B. Schilddrüsenvergrößerung oder mediastinale Raumforderungen). Als Besonderheit sind Gefäßmissbildungen (Truncus brachiocephalicus, doppelter Aortenboden, rechtsseitige Aortenbogen mit linksseitigem Ductusligament) zu erwähnen, die als pulsierende Kompression auf die Trachea erkennbar sind.

Die Veränderungen im Bronchialsystem haben eine ähnliche Pathogenese wie die der Trachealveränderungen. Symptome sind Dyspnoe, Husten, Hypersekretion, rezidivierende Bronchitiden und Pneumonien. Angeborene Bronchusstenosen sind zumeist in Haupt- und Lappenbronchien lokalisiert und häufig mit anderen Fehlbildungen kombiniert. Erworbene Stenosen werden durch externe Kompressionen bei Gefäßmissbildungen und Tumoren, durch Absaugen bei Intubation, bronchogenen Zysten oder durch Anschwellen der Lymphknoten an der Bronchuscarina hervorgerufen. Bronchomalazien können sowohl angeboren (z. B. Knorpelfehlbildungen) als auch durch Infektionen oder postinfektiös hervorgerufen auftreten. (Mantel u. a., 1995)

Abbildung 1.1: Menschlicher Bronchialbaum



Schema des menschlichen Bronchialbaumes: Oberlappen rot, Mittellappen gelb, Unterlappen blau, Lingula hellblau. Illustration von Patrick J. Lynch, medical illustrator und C. Carl Jaffe, MD, cardiologist unter Creative Commons Lizenz.

1.11 Pathologische Morphologie

Im Sinne einer möglichst objektiven Befunderhebung in der pädiatrischen Bronchoskopie wurden im Befundbogen vorwiegend direkt beobachtbare pathologische Morphologien erfragt. Dazu gehören Grad, Lage und Form von Stenosen sowie deren Spezialformen Malazie, Kompressionen und Pulsationen. Als Schleimhautanomalien wurden Schwellung, Hyperämie und Hypersekretion erfasst. Die Schleimhautbefunde wurden ergänzend zur Einzelanalyse auch im Sinne des Bronchitis Index (Thompson u. a., 1993) als Gesamtbefund für Entzündung untersucht.

1.11.1 Stenosen

Obstruktionen der oberen Atemwege können bei Kindern unter praktischen Gesichtspunkten in akute und chronische einerseits und angeborene bzw. erworbene andererseits unterteilt werden. Stenosen werden durch den

- Stenosegrad ihre
- Ausdehnung (Länge) die
- Konsistenz des das Lumen verlegenden Gewebes und ihre
- Lokalisation

charakterisiert.

1.11.1.1 Stenosegrad als Determinante des Atemwegswiderstandes

Der maximale Stenosegrad einer Lumeneinengung ist als Hauptdeterminante des Atemwegswiderstandes sowohl im Larynx als auch in der Trachea und den Bronchien das wichtigste Kriterium der Befunderhebung. Nach dem Hagen-Poiseuille-Gesetz wird der Atemwegswiderstand angesichts vernachlässigbarer Zähigkeit der Atemluft allein vom maximalen Stenosegrad und der Stenoselänge definiert. Während die Länge der Stenose den Atemwegswiderstand proportional erhöht, wirkt sich eine Verlegung des Lumens (Radius) mit der 4. Potenz aus. Damit kommt der präzisen Einschätzung des Stenosegrades eine entscheidende Bedeutung für das weitere Vorgehen zu.

Exkurs 1: Gesetz nach Hagen-Poiseuille

*Das Hagen-Poiseuille-Gesetz beschreibt die Abhängigkeit des Strömungswiderstands **R** in einer Röhre von deren Länge **l,** dem Innendurchmesser **r** und der dynamischen Viskosität („Zähigkeit“) des darin fließenden Stoffs **η**:*

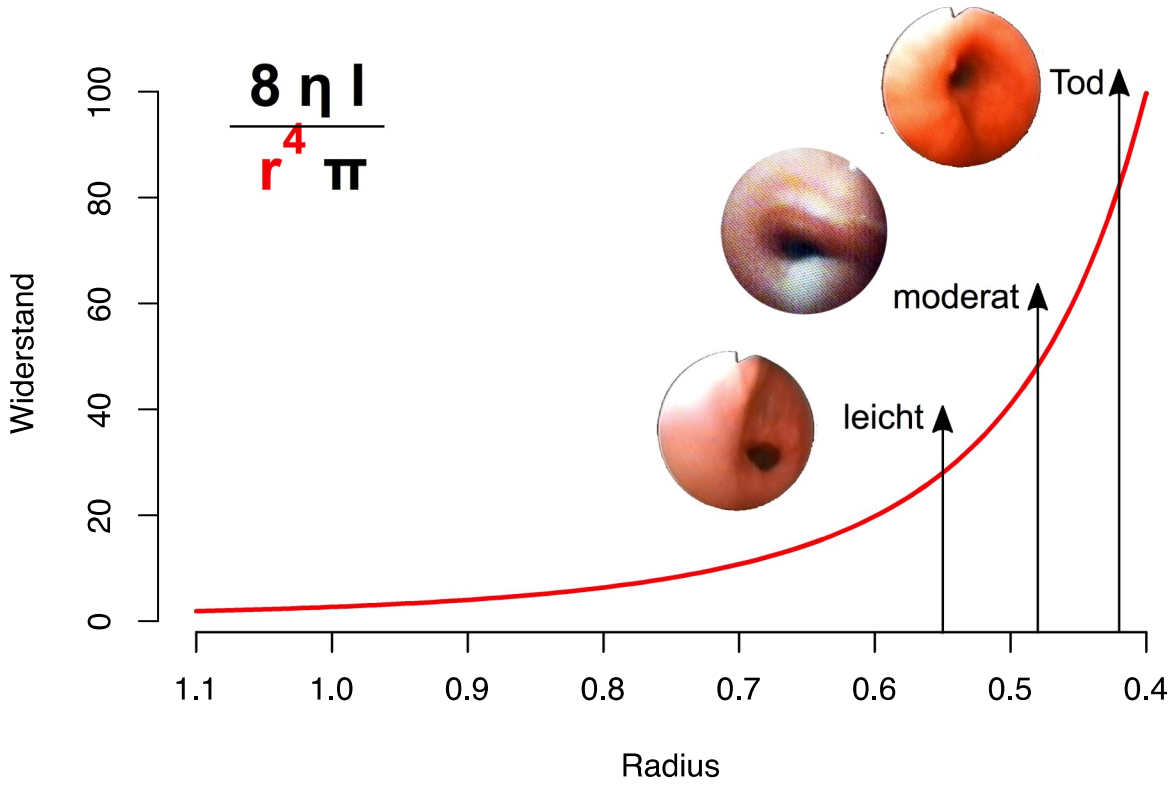
Formel 1.1: Hagen-Poiseuille-Gesetz

$$R = \frac{8\,\eta\,l}{\pi\,r^4}$$

Formel nach Harms und Stöcker (Harms, 1994: S. 72; Stöcker, 1998: S. 190).

Der Widerstand hängt also im Wesentlichen vom Innendurchmesser bzw. Radius ab, der mit der 4. Potenz in die Formel eingeht, wohingegen die anderen Größen sich nur als Faktoren auswirken.

Abbildung 1.2: Abhängigkeit des Atemwegswiderstandes vom Durchmesser



Im Beispiel wurden die dynamische Viskosität η und die Länge l gleich 1 gesetzt. Die vierte Potenz des Radius als wesentliche Determinante ist rot hervorgehoben. Nach einer Darstellung in (Bruce, Rothera, 2009: S. 89)

1.11.1.2 Klassifikation von Atemwegsstenosen

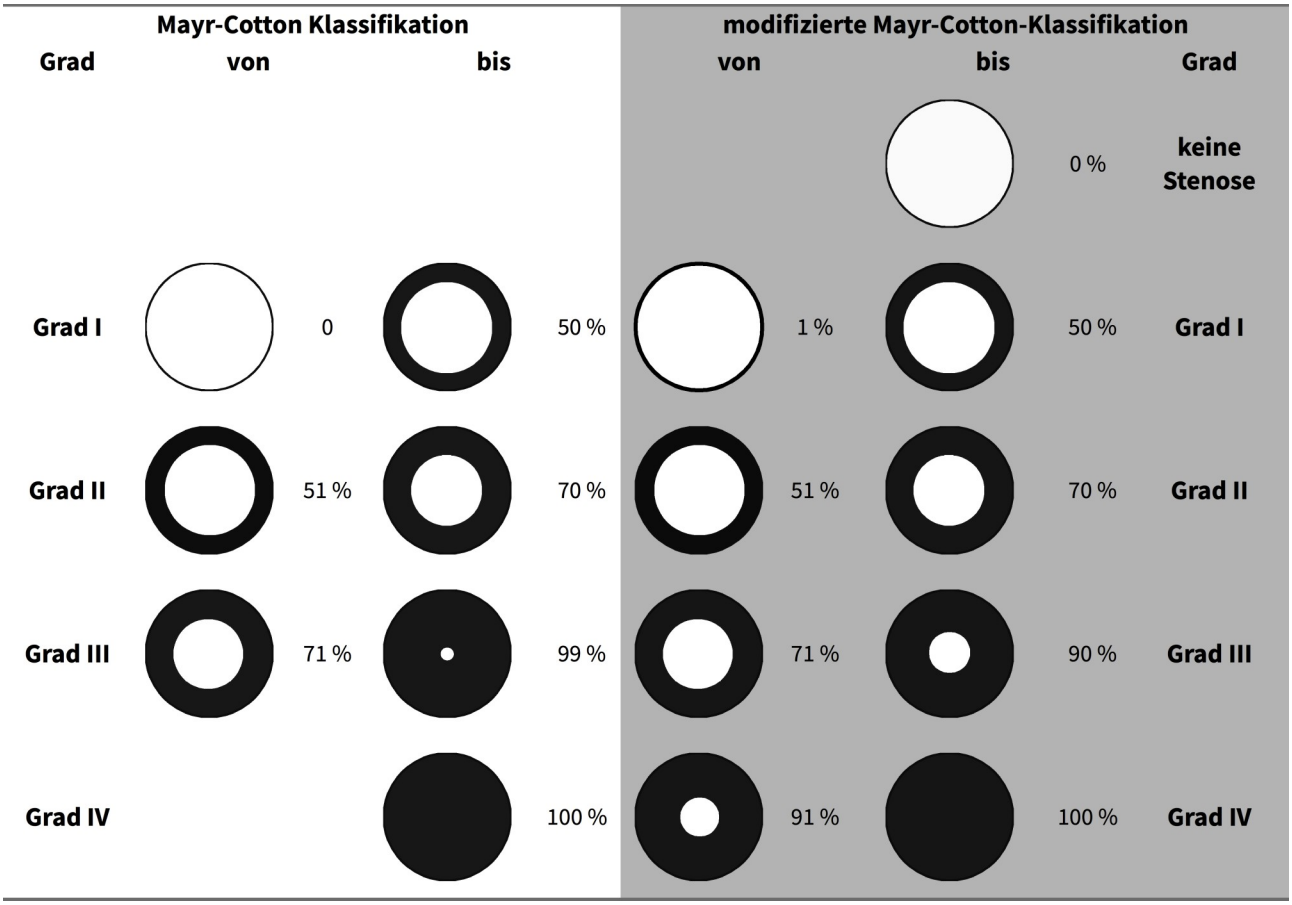
Eine allgemeingültige Klassifikation von Stenosen über den gesamten Atemweg, von Larynx bis Lappenbronchus, ist derzeit (2017) nicht verfügbar. Es wurden jedoch Klassifikationen für laryngotracheale und subglottische Stenosen bei Kindern entwickelt, die sich in den wesentlichen Punkten decken und sich prinzipiell auch auf andere Abschnitte des Atemwegs übertragen las-

sen (Cotton, 1984; Myer u. a., 1994). Ein früheres Klassifikationssystem, das auf Länge, Durchmesser und Konsistenz der Stenose basierte (Grundfast u. a., 1987) konnte sich nicht durchsetzen. Auch dessen Weiterentwicklung bewährte sich in der Praxis nicht (unter dem Akronym FLECS[16] von der American Society of Pediatric Otolaryngologists (ASPO) propagiert; unpublizierte Daten 1988). Dagegen sind die Klassifikationen nach Cotton und Myer inzwischen im Klinikalltag gut etabliert[17].

1.11.1.2.1 Klassifikation für Kinder nach Myer-Cotton

Bei Kindern wurden für solide laryngotracheale Stenosen, die anhand von 100 Fällen entwickelte Cotton-Klassifikation (Cotton, 1984) und für solide subglottische Stenosen die Myer-Cotton Klassifikation (Myer u. a., 1994) vorgeschlagen. Die Cotton-Klassifikation orientiert sich ausschließlich am vom Untersucher subjektiv wahrgenommenen, die Klassifikation nach Myer-Cotton am im Vergleich zu endotrachealen Tuben geschätzten Stenosegrad. Angesichts der besonderen anatomischen Verhältnisse bei Kindern und aufgrund des Hagen-Poiseulle-Gesetzes, ist sie eine für die klinische Anwendung sinnvolle Vereinfachung. Wie eine retrospektive Analyse zeigte, hat die Cotton-Klassifikation prognostischen Wert hinsichtlich der erfolgreichen Dekanülierung (McCaffrey, 1992).

Tabelle 1.3: Myer-Cotton-Klassifikation



Prozentuale Grenzen der Myer-Cotton-Klassifikation (Cotton, 1984; Myer u. a., 1994) *im Vergleich zur modifizierten Klassifikation nach Myer-Cotton. Letztere unterscheidet zwischen Normalbefunden und leichtgradigen Stenosen.*

Die Mayr-Cotton-Klassifikation unterscheidet nicht zwischen der Abwesenheit von Stenosen und Stenosen bis zu 50 %. Grad IV der Klassifikation ist auf eine vollständige Verlegung des Lumens beschränkt. Um zwischen Normalbefunden und leichtgradigen Stenosen einerseits und zwi-

[16] *engl. Function Lumen Extent Consistency Site*

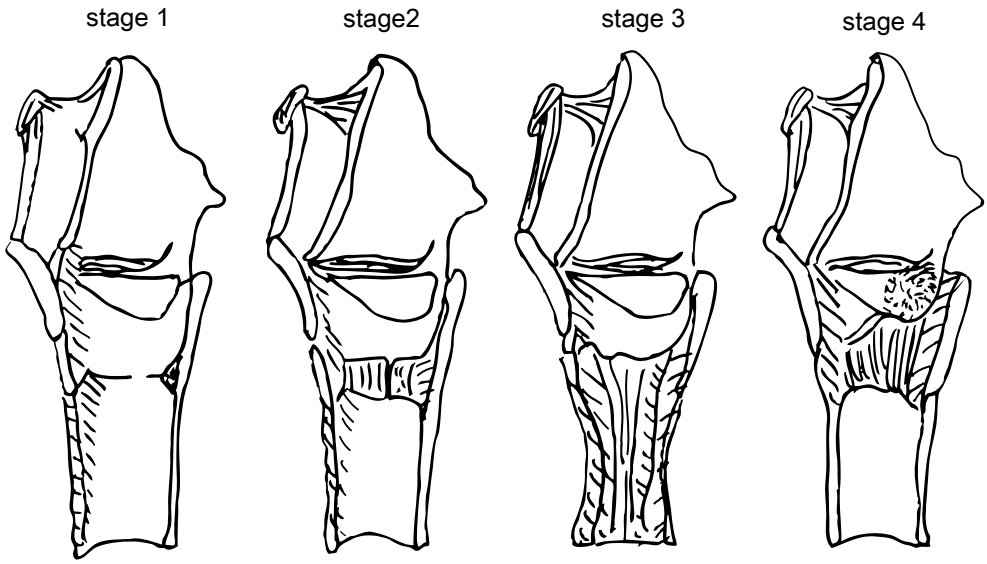
[17] *Angaben gemäß einer historischen Darstellung von Myer (Myer u. a., 1994: S. 319).*

schen hochgradigen Stenosen andererseits differenzieren zu können, wurde die Mayr-Cotton-Klassifikation in dieser Studie an ihren Extremen erweitert: unauffällige Befunde bilden eine eigene Klasse, Grad IV wird auf Kosten von Klasse III mit weiteren Grenzen definiert.

1.11.1.2.2 Klassifikation für Erwachsene nach McCaffrey

Für Erwachsene entwickelte McCaffrey an 72 retrospektiv untersuchten Fällen eine Klassifikation, die eine Prognose hinsichtlich erfolgreicher Dekanülierung und Abwesenheit von Dyspnoe nach Extubation erlaubt (McCaffrey, 1992). Als signifikanter Einflussfaktor kristallisierte sich neben dem Stenosedurchmesser die Lokalisation der Stenose heraus. Gemäß einer Untersuchung McCaffreys (McCaffrey, 1991) ist bei Erwachsenen – anders als bei Kindern – der Stenosegrad offenbar nicht der entscheidende Prädiktor für eine erfolgreiche Dekanülierung. McCaffreys Klassifikation orientiert sich deshalb an der Lokalisation der Stenose. McCaffrey schlägt vier Klassen vor, die basierend auf Stenoselokalisation (subglottisch, tracheal, glottisch) und der Stenoselänge (> 1 cm <) gebildet werden.

Abbildung 1.3: McCaffrey Klassifikation



Nach einer Darstellung von McCaffrey (McCaffrey, 1992).

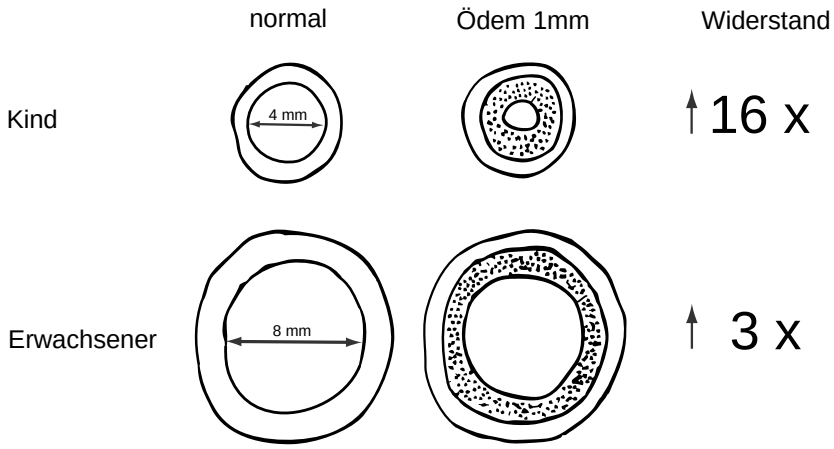
1.11.1.3 Besondere anatomische Verhältnisse bei Kindern

In kindlichen Atemwegen wirken sich Stenosen wegen der besonderen anatomischen Verhältnisse besonders stark aus. Das subglottische Lumen bei Kindern bis 12 Jahren wurde in einer histopathologischen Untersuchung mit durchschnittlich 16,4 mm^2, bei Erwachsenen mit 35,8 mm^2 angegeben, das der Trachea mit 17,1 mm^2 bei Kindern und 43,6 mm^2 bei Erwachsenen (Eckel u. a., 2000). Eine CT/MRT-Studie fand für Kleinkinder zwischen 2 und 6 Jahren einen subglottischen Durchmesser von 7–10 mm (Hudgins u. a., 1997: S. 242 Tabelle 2). Videobronchoskopisch wird das Lumen des linken Hauptbronchus für Kinder zwischen 5 und 10 Jahren auf 27,4 mm^2 beziffert, das des rechten auf 36,5 mm^2 (Masters u. a., 2006: S. 503 Tabelle 1).

Hinzu kommt die erhöhte Reagibilität der Schleimhäute mit rascher Ödembildung. Während sich beim Kind die maximale Engstelle der Atemwege im Bereich des relativ starren Ringknorpels befindet, liegt sie beim Erwachsenen im Bereich der beweglichen Stimmbänder (Mantel u. a., 1995: S. 9). Als weitere anatomische Besonderheit, die u. a. für die Wahl endotrachealer Tuben von Bedeutung ist, wurde lange Zeit angenommen, der kindliche Larynx sei initial trichterförmig und entwickle sich im Laufe der Adoleszenz zur Zylinderform (Eckenhoff, 1951). Jüngere Studien

konnten diese Theorie jedoch nicht stützen (Dalal u. a., 2008, 2009; Eckel u. a., 2000; Litman u. a., 2003; Übersicht bei Motoyama, 2009).

Abbildung 1.4: Auswirkung von Ödemen bei Erwachsenen im Vergleich zu Kindern



Bei Kindern wirken sich Stenosen bzw. Ödeme aufgrund der anatomischen Verhältnisse und des Hagen-Poiseuille-Gesetz stärker aus, als bei Erwachsenen: ein Ödemsaum von 1 mm wirkt sich in einem Lumen mit dem Durchmesser 4 mm im Vergleich zu einem Lumen von 8 mm Durchmesser im Verhältnis 16 : 3 aus (nach Höhne u. a., 2006).

1.11.1.4 Hindernisse der visuellen Bestimmung des Stenosegrades

Die verlässliche Vermessung von Stenosen via Endoskop ist mit zahlreichen Schwierigkeiten behaftet. Dazu zählen

- die optische Distorsion („Verzerrung“) der Endoskoplinsen
- psychophysikalische, durch die Wahrnehmung bedingte, Aspekte und der
- unbekannte Beobachtungsabstand zur Stenose.

1.11.1.4.1 Distorsion

Die Optische Distorsion (Fachbegriff für die Verzerrung der Linse des Endoskops) wird anschaulich als „Fischaugeneffekt“ bezeichnet. Linsen mit Fischaugeneffekt werden in der Endoskopie zu Gunsten eines optimierten Blickfeldes eingesetzt. Bei der Beurteilung von Stenosen ist der sonst wünschenswerte Weitwinkel-Effekt problematisch (Dörffel u. a., 1999, 2003).

1.11.1.4.2 Psychophysik

Da Stenosegrade mit dem aktuell im klinischen Alltag gängigen Instrumentarium nicht objektiv gemessen werden können, ist man auf die subjektive visuelle Schätzung des Untersuchers angewiesen. Damit unterliegt die Beurteilung des Stenosegrades den Gesetzen und Besonderheiten der menschlichen Psychophysik (siehe Exkurs 2).

Die Bronchoskopie bildet beim heutigen Stand der Technik die dreidimensionalen Hohlräume des Bronchialsystems auf plane Bilder ab, die mit Bildschirmen oder Okularen dargestellt werden. Die Schätzung der Stenosegrade reduziert sich somit im Wesentlichen auf die Schätzung des Verhältnisses zweier annähernd kreisförmiger Flächen (dem maximalen und minimalen Lumen) zueinander.

Gerade die Schätzung von Kreisflächen bereitet Menschen besondere Schwierigkeiten. Psychophysikalische Untersuchungen aus der Kartographie (Chang, 1977; Cox, 1976; Ekman, Junge, 1961; Flannery, 1971; Meihoefer, 1973: S. 64) Statistik (Cleveland u. a., 1982; Croxton, 1932) und Psychologie (Teghtsoonian, 1965) zeigten eine tendenzielle Unterschätzung der Verhältnisse von Kreisflächen. Die Verhältnisse von Balken zueinander werden verlässlicher geschätzt, als die von Kreisen (Croxton, 1932: S. 55–56): Kreisflächen mit Faktor 2 werden im Durchschnitt als 1,7 fach

eingeschätzt, Kreisflächen die sich um den Faktor 10 unterscheiden dagegen nur als um Faktor 6 differierend empfunden (Teghtsoonian, 1965: S. 394).

Exkurs 2: Die Psychophysik und ihre 3 klassischen Gesetze

Die Psychophysik beschäftigt sich mit der Beziehung zwischen objektiv messbaren „physikalischen" Reizen und deren subjektiver „psychischen" Wahrnehmung. Für diesen Zusammenhang wurden 3 klassische Gesetze formuliert: das Weber-Gesetz, das Weber-Fechner-Gesetz und die Stevenssche Potenzfunktion.

Webersches Gesetz

Der Anatom und Physiologe Ernst Heinrich Weber (24.06.1795 Wittenberg; † 26.01.1878 Leipzig) entdeckte bei Untersuchungen des Tastsinns, dass der für eine Unterscheidung zweier Reize notwendige Reizunterschied ΔR zur Reizintensität R, unabhängig von deren absoluter Größe, in einem festen Verhältnis steht.*

Formel 1.2: Webersches Gesetz

$$\frac{\Delta R}{R_0} = \frac{R_1 - R_0}{R_0} = k$$

Der kleinste jeweils merkliche Reizunterschied ΔR (engl. JND) ist also proportional zur Reizgröße R. Demnach wird z. B. bei einem Gewicht von 30g eine Veränderung von 1g bemerkt, bei einem Gewicht von 3 kg hingegen erst eine Veränderung von 300g. Heute ist Weber wohl eher aufgrund seines klinischen Tests zur Gehörprüfung bekannt.

Fechnersches Gesetz

Auf Webers Erkenntnissen aufbauend postulierte der Physiker und Mediziner Gustav Theodor Fechner (19.04.1801 Groß Särchen; † 18.11.1887 Leipzig) dass „gleiche Empfindungszuwüchse gleichen relativen Reizzuwüchsen zugehören [...]" bzw. „dass der Empfindungsunterschied derselbe bleibt, wenn das Reizverhältniss dasselbe bleibt". Daraus kann ein logarithmischer Zusammenhang zwischen Empfindung und Reiz abgeleitet werden.*

Formel 1.3: Fechner Gesetz

$$E = k \times \log R + f$$

Fechner erhielt sein Gesetz durch Integration aus dem Weberschen Gesetz. Die Formel kann zur Hervorhebung des in ihr enthaltenen Weber Gesetzes umgeformt werden zu:

Formel 1.4: Weber-Fechner-Gesetz

$$E = k \times \ln \frac{R}{R_0}$$

Fechners Formel gab erstmals ein mathematisches Maß für die subjektive Empfindungsintensität an, weswegen er als Begründer der Psychophysik gilt.

Stevenssche Potenzfunktion

Mitte des 19. Jahrhunderts modellierte Stevens die Beziehung zwischen Reiz und Empfindung mit Potenzfunktionen (Stevens, 1957).

1.11.1.5 Verfahren zur Bestimmung des Stenosegrades

Zahlreiche Verfahren wurden und werden zur Bestimmung des maximalen Stenosegrades herangezogen, um die bekanntermaßen stark fehlerbehaftete, rein subjektive Schätzung des Stenosegrades weit möglichst zu objektivieren.

1.11.1.5.1 Vergleich mit Gegenständen bekannter Größe

In der klinischen Bronchoskopie ist die gängigste Methode zur Bestimmung des Stenosegrades, neben der visuellen Schätzung, bislang der Vergleich mit Gegenständen bekannter Größe, insbesondere endotrachealen Tuben (Myer u. a., 1994). Studien aus der Gastroenterologie deuten darauf hin, dass solche Vergleiche – zum Beispiel mit einem Lineal – die visuelle Einschätzung tatsächlich verbessern können (Gopalswamy u. a., 1997)[18]. Die immer wieder propagierte geöffnete Biopsiezange scheint als Maßstab allerdings wenig geeignet. Sie führte in einer Studie sogar zu größeren Abweichungen als die rein visuelle Beurteilung (Gopalswamy u. a., 1997; Margulies

[18] *Die durchschnittliche Abweichung zur Referenz lag bei Messung mit dem Lineal bei 34 % gegenüber 64 % bei visueller Schätzung ohne Hilfsmittel.*

u. a., 1994, 1994; Vakil u. a., 1994). Der Vergleich mit endotrachealen Kanülen ist zwar weit verbreitet, scheint aber in der Literatur bisher kaum mit anderen Messverfahren evaluiert. Unter anderem schränkt die Distorsion die Genauigkeit dieser Techniken in der Praxis ein (Vakil, 1995).

1.11.1.5.2 Technische Messverfahren

Es wurden zahlreiche Versuche unternommen die Messungen mit technischen Hilfsmitteln zu präzisieren: Dazu gehören die dreidimensionale Vermessung mittels Laryngoskop (Kleinsasser u. a., 1994), elektrisch geladenen Wasserstrahlen (Wakabayashi u. a., 1994) und mittels Lasertechnik („ENDOSCAN“) (Müller, 2004, 2003).

1.11.1.5.3 virtuelle Bronchoskopie

Als nicht invasive Alternative zur fiberoptischen Bronchoskopie wurde versucht, Stenosen über eine 3D-Rekonstruktion von Multidetektor-CT-Daten zu vermessen (De Wever u. a., 2005). Die Virtuelle Bronchoskopie erwies sich zwar als hilfreich bei der Bestimmung der Lage einer Obstruktion, jedoch konnte die Pathologie der Obstruktion meist weniger eindeutig differenziert werden, als bei direkter Beobachtung durch das Bronchoskop. Zudem fehlt der virtuellen Bronchoskopie die Möglichkeit einer oft erforderlichen direkten Intervention, sodass ein bronchoskopischer Eingriff ohnehin nicht vermieden werden kann. Die virtuelle Bronchoskopie kann daher kein Ersatz für die herkömmliche Bronchoskopie sein, sondern lediglich ein ergänzendes diagnostisches Verfahren (Sodhi u. a., 2010).

1.11.1.6 Stenoseform

Stenoseform und Stenoselänge geben Hinweise auf spezifische Krankheitsbilder im Bereich der Atemwege. So imponieren intubationsbedingte Stenosen oft als kurzstreckige ringförmige Stenose im Bereich des Ringknorpels mit lochblendenartiger Morphologie. Auch angeborene Larynxstenosen zeigen sich subglottisch als ringförmige Stenose, sind jedoch langstreckiger als erworbene Stenosen. Bei einer Stimmlippensynechie bzw. einem Larynxdiaphragma – auch als Glottissegel bezeichnet – sieht man eine membranartige Synechie über der anterioren Kommissur im ventralen Stimmlippenabschnitt. Diese Membranbildung kann man vereinzelt auch supra- oder subglottisch beobachten. Segmentäre Trachealstenosen können als bindegewebige Segelbildung kurzstreckig oder als zirkuläre starre sowohl segmental als auch langstreckig erscheinen.

In der klinischen pädiatrischen Bronchoskopie ist die Faustregel verbreitet, Stenosen bis 1 cm als kurzstreckig und über 1 cm als langstreckig zu bezeichnen, wobei natürlich die verschiedenen Altersstufen beachtet werden müssen. (Darstellung nach Mantel u. a., 1995; und Nicolai, Griese, 2010)

1.11.1.7 Spezialformen der Atemwegsobstruktion

Als Spezialformen der Atemwegsobstruktion wurden dynamische Verengungen durch Malazie und Pulsationen sowie sichtbare Kompressionen von außen erfragt. Dabei konnte jeweils die Lage in Larynx, Trachea, Bronchus, in Assoziation zu Stenosen und ein generalisierter Befund angegeben werden.

1.11.1.7.1 Malazie

Malazien sind in Lungenfunktionstests nicht immer zuverlässig erkennbar und können auch mit CT und MRT wegen der eingeschränkten zeitlichen Auflösung nicht sicher dargestellt werden. Die Videobronchoskopie bleibt trotz ihres invasiven Charakters daher der Goldstandard bei der Diagnose von Malazien (Carden u. a., 2005). Voraussetzung zur Beurteilung der dynamischen Veränderungen ist eine erhaltene Spontanatmung, also eine flache Sedierung.

1.11.1.7.2 Pulsationen

Pulsationen werden durch aberrierende Gefäße wie z. B. einen doppelten Aortenbogen, eine Pulmonalisschlinge oder den Truncus brachiocephalicus des Neugeborenen verursacht. Sie imponieren videobronchoskopisch durch pulsatile Bewegungen der Atemwege.

1.11.1.7.3 Kompressionen

Ursache von Kompressionen sind vorwiegend Gefäßaberrationen (Kussman u. a., 2004) gelegentlich aber auch tracheo- oder broncho-ösophageale Fisteln sowie Neoplasien, die das Atemwegslumen von außen einengen. Kompressionen unterscheiden sich bildmorphologisch von gewöhnlichen Stenosen dadurch, dass in der Regel wenig Schleimhautveränderungen zu erkennen sind (wie z. B. bei Stenosen durch Entzündungen) und dass das Lumen evtl. nur einseitig und konvex verlegt ist.

1.11.2 Schleimhaut

Videobronchoskopisch sind die wichtigsten Kriterien der Schleimhautbeurteilung Schwellung bzw. Ödem, Hyperämie und Hypersekretion. Diese Befunde sind zugleich der wichtigste Hinweis auf eine Schleimhautentzündung.

1.11.2.1 Entzündung

Thompson ergänzte die Befunde Schwellung, Hyperämie und Hypersekretion um die Verletzlichkeit der Schleimhäute (engl. friability) und leitete aus ihnen den sogenannten „Bronchitis Index (BI)“ ab. Dazu wurde jedem der Schleimhautbefunde eine dreistufige Skala zugeordnet. Der BI wurde an gesunden Erwachsenen, beschwerdefreien Rauchern und Patienten mit chronischer Bronchitis evaluiert. Dabei ergab sich für Gesunde ein Durchschnitt von 2,3, für Raucher von 8,5 und für Bronchitispatienten von 13,2 (Thompson u. a., 1993).

Tabelle 1.4: Bronchitis Index

Befund	0	1	2	3
Erythem	normal	leicht rot (engl. light red)	rot (engl. red)	fleischig rot (engl. beefy red)
Ödem	Normaler Atemweg (engl. normal airway)	unscharf begrenzte Bifurkationen (engl. blunting of bifurcations)	Atemwegsverengung (engl. airway narrowing)	Atemwegsverschluss (engl. airway occluded)
Sekretion	normal	klare Schleimstraßen (engl. strands of clear mucus)	Schleimpfropfen (engl. globules of mucus)	Atemwegsverschluss (engl. airway occluded)
Verletzlichkeit (engl. friability)	normal	punktuelle submuköse Einblutungen (engl. punctate submucosal hemorrhages)	streifige submuköse Einblutungen (engl. linear submucosal hemorrhages)	offeneBlutung (engl. frank bleeding)

Semiquantitative Skala zur bronchoskopischen Beurteilung von Schleimhautentzündungen nach Thompson (Thompson u. a., 1993)

Im Befundbogen dieser Studie wurde die dreistufige Skala der Schleimhautbefunde auf das Vorhandensein respektive Fehlen des jeweiligen Befundes vereinfacht. Die Verletzlichkeit wurde nicht mit einbezogen, da sie einerseits vom Videomitschnitt nicht sicher erhebbar ist und andererseits eine vernachlässigbare Vorhersagekraft hinsichtlich Entzündungen zu haben scheint (Thompson u. a., 1993: S. 1484 Tabelle 3).

1.11.2.2 Entzündungsbereich

Für die Beurteilung des Entzündungsbereiches wurde das detaillierte anatomische Befundschema auf Ebene der Bronchialsegmente auf die übergeordneten Abschnitte Larynx, Trachea und

Bronchus reduziert. Darüber hinaus konnte eine Assoziation zum Stenosebereich sowie eine generalisierte Befundausbreitung angegeben werden.

Literatur

Aabakken, Lars; Rembacken, Bjorn; LeMoine, Olivier; u. a. (2009): „Minimal standard terminology for gastrointestinal endoscopy - MST 3.0“. In: *Endoscopy*. 41 (8), S. 727–728, DOI: 10.1055/s-0029-1214949.

American Thoracic Society (1997): „The role of the pediatric pulmonary physician in the American health care system. American Thoracic Society“. In: *American journal of respiratory and critical care medicine*. 155 (4), S. 1486–1488, DOI: 10.1164/ajrccm.155.4.9105100.

Becker, Heinrich D (2010): „Bronchoscopy: the past, the present, and the future“. In: *Clinics in Chest Medicine*. 31 (1), S. 1–18, Table of Contents, DOI: 10.1016/j.ccm.2009.11.001.

British Colour Council (1946): *The British Colour Council Dictionary of Colour Standards: A List of Colour Names Referring to the Colours Shown in the Companion Volume*. London: The Council.

Bruce, I. A.; Rothera, M. P. (2009): „Upper airway obstruction in children“. In: *Pediatric Anesthesia*. 19 , S. 88–99.

Bücheler, Von Egon; Lackner, Klaus-Jürgen; Göttsche, Thomas; u. a. (2006): *Einführung in die Radiologie*. Georg Thieme Verlag. — ISBN: 3-13-316011-7

Burke, Bryan L.; Hall, R. W.; Care, the SECTION ON TELEHEALTH (2015): „Telemedicine: Pediatric Applications“. In: *Pediatrics*. 136 (1), S. e293–e308, DOI: 10.1542/peds.2015-1517.

Caliebe, W. (1968): „Dokumentationsgerechte Befunderhebung bei Bronchoskopien“. In: *European Archives of Oto-Rhino-Laryngology*. 191 (2), S. 628–631, DOI: 10.1007/BF00492143.

Carden, Kelly A; Boiselle, Philip M; Waltz, David A; u. a. (2005): „Tracheomalacia and Tracheobronchomalacia in Children and Adults: An In-depth Review“. In: *Chest*. 127 (3), S. 984–1005, DOI: 10.1378/chest.127.3.984.

Chang, KANG-TSUNG (1977): „Visual Estimation of Graduated Circles“. In: *Cartographica: The International Journal for Geographic Information and Geovisualization*. 14 (2), S. 130–138, DOI: 10.3138/V35N-2387-G7M8-P527.

Cleveland, William S.; Harris, Charles S.; McGill, Robert (1982): „Judgments of Circle Sizes on Statistical Maps“. In: *Journal of the American Statistical Association*. 77 (379), S. 541–547.

Cotton, Robin T. (1984): „Pediatric laryngotracheal stenosis“. In: *Journal of Pediatric Surgery*. 19 (6), S. 699–704, DOI: 10.1016/S0022-3468(84)80355-3.

Cox, Carleton W. (1976): „Anchor Effects and the Estimation of Graduated Circles and Squares“. In: *Cartography and Geographic Information Science*. 3 , S. 65–74, DOI: 10.1559/152304076784080195.

Croxton, Frederick E. (1932): „Graphic Comparisons by Bars, Squares, Circles, and Cubes“. In: *Journal of the American Statistical Association*. 27 (177), S. 54–60.

Dalal, Priti G; MURRAY, DAVID; FENG, ANGELA; u. a. (2008): „Upper airway dimensions in children using rigid video-bronchoscopy and a computer software: description of a measurement technique“. In: *Pediatric Anesthesia*. 18 (7), S. 645–653, DOI: 10.1111/j.1460-9592.2008.02533.x.

Dalal, Priti G; Murray, David; Messner, Anna H; u. a. (2009): „Pediatric laryngeal dimensions: an age-based analysis“. In: *Anesthesia and Analgesia*. 108 (5), S. 1475–1479, DOI: 10.1213/ane.0b013e31819d1d99.

De Wever, W; Bogaert, J; Verschakelen, J A (2005): „Virtual bronchoscopy: accuracy and usefulness--an overview“. In: *Seminars in Ultrasound, CT, and MR*. 26 (5), S. 364–373.

Delvaux, M; Crespi, M; Armengol-Miro, J R; u. a. (2000): „Minimal standard terminology for digestive endoscopy: results of prospective testing and validation in the GASTER project“. In: *Endoscopy*. 32 (4), S. 345–55, DOI: 10774976.

Deutsche Gesellschaft für Endoskopie (1974): *Fortschritte der Endoskopie. Verhandlungsbericht; mit 56 Tab. Bd. 5. Bd. 5.* Stuttgart, New York: Schattauer. — ISBN: 978-3-7945-0425-1

Dörffel, W V; Fietze, I; Hentschel, D; u. a. (1999): „A new bronchoscopic method to measure airway size“. In: *The European Respiratory Journal: Official Journal of the European Society for Clinical Respiratory Physiology*. 14 (4), S. 783–788.

Dörffel, W V; Sugano, Y T; Stalling, D; u. a. (2003): „[Laser-based endoscopic measurement of airway dimensions]“. In: *Pneumologie (Stuttgart, Germany)*. 57 (9), S. 503–509, DOI: 10.1055/s-2003-42220.

Eckel, H. E.; Sprinzl, G. M.; Sittel, C.; u. a. (2000): „Zur Anatomie von Glottis und Subglottis beim kindlichen Kehlkopf“. In: *HNO*. 48 (7), S. 501–507, DOI: 10.1007/s001060050606.

Eckenhoff, J E (1951): „Some anatomic considerations of the infant larynx influencing endotracheal anesthesia“. In: *Anesthesiology*. 12 (4), S. 401–410.

Ekman, Gosta; Junge, Kenneth (1961): „Psychological Relations in the Perception of Length, Area, and Volume“. In: *Scandinavian Journal of Psychology*. 2 , S. 1–10, DOI: DOI: 10.1111/j.1467-9450.1961.tb01215.x.

Ernst, Armin; Becker, Heinrich D. (2001): „Documentation in Bronchology“. In: *Clinics in Chest Medicine*. 22 (2), S. 373–379.

Flannery, James John (1971): „The Relative Effectiveness of Some Common Graduated Point Symbols in the Presentation of Quantitative Data“. In: *Cartographica: The International Journal for Geographic Information and Geovisualization*. 8 (2), S. 96–109, DOI: 10.3138/J647-1776-745H-3667.

Freyschmidt, Von Jürgen; Schmidt, Th; Aichinger, Horst; u. a. (2002): *Handbuch diagnostische Radiologie*. o.V. — ISBN: 3-540-41419-3

Fujino, Masayuki A; Bito, Shigeru; Takei, Kazuko; u. a. (2006): „Terminology and global standardization of endoscopic information: Minimal Standard Terminology (MST)“. In: *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*. 1 , S. 2606–9, DOI: 10.1109/IEMBS.2006.259911.

Gekeler, Hans (2007): *Handbuch der Farbe: Systematik, Ästhetik, Praxis*. 6., Aufl. Dumont Buchverlag. — ISBN: 3-8321-7289-0

Gopalswamy, Narasimh; Shenoy, Vishwanath N.; Choudhry, Umesh; u. a. (1997): „Is in vivo measurement of size of polyps during colonoscopy accurate?,“. In: *Gastrointestinal Endoscopy*. 46 (6), S. 497–502, DOI: 10.1016/S0016-5107(97)70003-8.

Graham, R.N.J.; Perriss, R.W.; Scarsbrook, A.F. (2005): „DICOM demystified: A review of digital file formats and their use in radiological practice“. In: *Clinical Radiology*. 60 (11), S. 1133–1140, DOI: 10.1016/j.crad.2005.07.003.

Grund, K.; Salm, R. (2007): „Systeme für die Endoskopie“. In: *Medizintechnik*. o.V., S. 347–366.

Grundfast, K M; Morris, M S; Bernsley, C (1987): „Subglottic stenosis: retrospective analysis and proposal for standard reporting system“. In: *The Annals of Otology, Rhinology, and Laryngology*. 96 (1 Pt 1), S. 101–105.

Harms, Volker (1994): *Physik für Mediziner und Pharmazeuten*. 6. Auflage. Kiel: Harms Volker. — ISBN: 3-86026-027-8

Häussinger, K; Ballin, A; Becker, H D; u. a. (2004): „[Recommendations for quality standards in bronchoscopy]“. In: *Pneumologie (Stuttgart, Germany)*. 58 (5), S. 344–56, DOI: 15162262.

Höhne, C.; Haack, M.; Machotta, A.; u. a. (2006): „Atemwegsmanagement in der Kinderanästhesie“. In: *Der Anaesthesist*. 55 (7), S. 809–820, DOI: 10.1007/s00101-006-1045-0.

Hudgins, P. A.; Siegel, J.; Jacobs, I.; u. a. (1997): „The normal pediatric larynx on CT and MR“. In: *AJNR. American journal of neuroradiology*. 18 (2), S. 239–245.

Hussein, Rada; Engelmann, Uwe; Schroeter, Andre; u. a. (2004a): „DICOM Structured Reporting: Part 1. Overview and Characteristics“. In: *Radiographics*. 24 (3), S. 891–896, DOI: 10.1148/rg.243035710.

Hussein, Rada; Engelmann, Uwe; Schroeter, Andre; u. a. (2004b): „DICOM Structured Reporting: Part 2. Problems and Challenges in Implementation for PACS Workstations“. In: *Radiographics*. 24 (3), S. 897–909, DOI: 10.1148/rg.243035722.

Jackson, Chevalier L.; Huber, John Franklin (1943): „Correlated Applied Anatomy of the Bronchial Tree and Lungs With a System of Nomenclature“. In: *Chest*. 9 (4), S. 319–326, DOI: 10.1378/chest.9.4.319.

Kattah, Jorge C.; Talkad, Arun V.; Wang, David Z.; u. a. (2009): „HINTS to Diagnose Stroke in the Acute Vestibular Syndrome Three-Step Bedside Oculomotor Examination More Sensitive Than Early MRI Diffusion-Weighted Imaging“. In: *Stroke*. 40 (11), S. 3504–3510, DOI: 10.1161/STROKEAHA.109.551234.

Killian, Gustav (1898): „Ueber directe Bronchoscopie“. In: *Münchener Medizinische Wochenschrift*. (27), S. 844–847.

Kim, C. Y.; Etemad, B.; Glenn, T. F.; u. a. (2000): „Remote clinical assessment of gastrointestinal endoscopy (tele-endoscopy): an initial experience.“. In: *Proceedings of the AMIA Symposium*., S. 423–427.

Kleinsasser, N.; Krosdorf, D.; Merkenschlager, A.; u. a. (1994): „Endoscopic, 3-dimensional measurement of neoplasms and stenoses of the larynx and trachea.“. In: *Laryngo-Rhino-Otol*. 73 , S. 428–431.

Knuth, Donald Ervin (1984): „Literate programming“. In: *The Computer Journal*. 27 (2), S. 97–111.

Korman, L Y; Bidgood, W D (1997): „Representation of the Gastrointestinal Endoscopy Minimal Standard Terminology in the SNOMED DICOM microglossary“. In: *Proceedings: A Conference of the American Medical Informatics Association / ... AMIA Annual Fall Symposium. AMIA Fall Symposium*., S. 434–8, DOI: 9357663.

Korman, L.Y; Delvaux, M; Bidgood, Dean (1998): „Structured reporting in gastrointestinal endoscopy:: Integration with DICOM and minimal standard terminology“. In: *International Journal of Medical Informatics*. 48 (1–3), S. 201–206, DOI: 10.1016/S1386-5056(97)00126-3.

Kropp, R (2005): „Die Erfindung der modernen Bronchoskopie“. In: *Pneumologie (Stuttgart, Germany)*. 59 (10), S. 725–729, DOI: 10.1055/s-2005-915549.

Kussman, Barry D.; Geva, Tal; McGowan, Francis X. (2004): „Cardiovascular causes of airway compression“. In: *Paediatric Anaesthesia*. 14 (1), S. 60–74.

Laubenberger, Theodor; Laubenberger, Jörg (1999): *Technik der medizinischen Radiologie*. Deutscher Ärzteverlag. — ISBN: 3-7691-1132-X

Lee, Von Hsien-Che (2005): *Introduction to color imaging science*. Cambridge University Press. — ISBN: 0-521-84388-X

Litman, Ronald S; Weissend, Eric E; Shibata, Dean; u. a. (2003): „Developmental changes of laryngeal dimensions in unparalyzed, sedated children“. In: *Anesthesiology*. 98 (1), S. 41–45.

Mantel, Karl; Nicolai, Th; Merkenschlager, A. (1995): *Kinder-Bronchoskopie-Leitfaden*. Thieme, Stuttgart. — ISBN: 3131156414

Margulies, C; Krevsky, B; Catalano, M F (1994): „How accurate are endoscopic estimates of size?“. In: *Gastrointestinal Endoscopy*. 40 (2 Pt 1), S. 174–177.

Masters, Ian; Ware, Robert; Zimmerman, Paul; u. a. (2006): „Airway sizes and proportions in children quantified by a video-bronchoscopic technique“. In: *BMC Pulmonary Medicine*. 6 (1), S. 5, DOI: 10.1186/1471-2466-6-5.

McCaffrey, T V (1992): „Classification of laryngotracheal stenosis“. In: *The Laryngoscope*. 102 (12 Pt 1), S. 1335–1340, DOI: 10.1288/00005537-199212000-00004.

McCaffrey, T V (1991): „Management of subglottic stenosis in the adult“. In: *The Annals of Otology, Rhinology, and Laryngology*. 100 (2), S. 90–94.

Meihoefer, Hans-Joachim (1973): „The Visual Perception of the Circle in Thematic Maps/Experimental Results“. In: *Cartographica: The International Journal for Geographic Information and Geovisualization*. 10 (1), S. 63–84, DOI: 10.3138/2771-5577-5417-369T.

Mortensen, J. D.; Young, Johanne D.; Stout, Lisa; u. a. (1983): „A numerical identification system for airways in the lung“. In: *The Anatomical Record*. 206 (1), S. 103–114, DOI: 10.1002/ar.1092060112.

Motoyama, Etsuro K (2009): „The shape of the pediatric larynx: cylindrical or funnel shaped?“. In: *Anesthesia and Analgesia*. 108 (5), S. 1379–1381, DOI: 10.1213/ane.0b013e31819ed494.

Müller, A (2004): „[Modern diagnostics of tracheal stenosis]“. In: *Laryngo- Rhino- Otologie*. 83 (6), S. 381–386, DOI: 10.1055/s-2004-814585.

Müller, Andreas (2003): *ENDOSCAN: Entwicklung eines Verfahrens zur endoskopischen Vermessung von Trachealstenosen*. 1. Aufl. Shaker. — ISBN: 3-8322-2250-2

Murad, Faris M.; Banerjee, Subhas; Barth, Bradley A.; u. a. (2014): „Image management systems“. In: *Gastrointestinal Endoscopy*. 79 (1), S. 15–22, DOI: 10.1016/j.gie.2013.07.048.

Myer, C M; O'Connor, D M; Cotton, R T (1994): „Proposed grading system for subglottic stenosis based on endotracheal tube sizes“. In: *The Annals of Otology, Rhinology, and Laryngology*. 103 (4 Pt 1), S. 319–323.

Nakhosteen, John A; Zavala, Donald C (1983): *Atlas und Lehrbuch der flexiblen Bronchoskopie*. o.V. — ISBN: 978-3-662-05898-5

Nicolai, Thomas; Griese, Matthias (2010): *Praktische Pneumologie in der Pädiatrie - Diagnostik: Rationale Differenzialdiagnostik*. 1. Aufl. Thieme, Stuttgart. — ISBN: 3131460814

Ohashi, Kumiko; Sakamoto, N.; Watanabe, M.; u. a. (2008): „Development of a Telediagnosis Endoscopy System over Secure Internet“. In: *Methods of Information in Medicine*., DOI: 10.3414/ME0488.

Oho, Kenkichi; Amemiya, Ryuta; Kaneko, Masahiro; u. a. (1986): „Proposed New Classification of Bronchoscopic Findings in Lung Cancer Cases“. In: *Haigan*. 26 (1), S. 1–10, DOI: 10.2482/haigan.26.1.

Prakash, U B; Edell, E S (1996): „The documentation of thoracic endoscopy“. In: *Chest Surgery Clinics of North America*. 6 (2), S. 193–203, DOI: 8724274.

Schindl, R; Aigner, K; Würtz, J; u. a. (1988): „Bronchoskopiedokumentation“. In: *Praxis Und Klinik Der Pneumologie*. 42 (7), S. 573–575.

Schlungbaum, Von Werner; Flesch, Udo; Stabell, Uwe (1993): *Medizinische Strahlenkunde*. Walter de Gruyter. — ISBN: 3-11-012850-0

Seidenari, Stefania; Pellacani, Giovanni; Righi, Elena; u. a. (2004): „Is JPEG compression of videomicroscopic images compatible with telediagnosis? Comparison between diagnostic performance and pattern recognition on uncompressed TIFF images and JPEG compressed ones“. In: *Telemedicine journal and e-health: the official journal of the American Telemedicine Association*. 10 (3), S. 294–303, DOI: 10.1089/tmj.2004.10.294.

Sodhi, Kushaljit Singh; Aiyappan, Senthil Kumar; Saxena, Akshay Kumar; u. a. (2010): „Utility of multidetector CT and virtual bronchoscopy in tracheobronchial obstruction in children“. In: *Acta Paediatrica (Oslo, Norway: 1992)*. 99 (7), S. 1011–1015, DOI: 10.1111/j.1651-2227.2010.01729.x.

Stevens, Stanley Smith (1957): „On the psychophysical law“. In: *Psychological Review*. 64 (3), S. 153–181.

Stöcker, Horst (1998): *Taschenbuch der Physik: Formeln, Tabellen, Übersichten*. 3. Auflage. Thun und Frankfurt am Main: Deutsch Harri Gmbh. — ISBN: 3-8171-1556-3

Teghtsoonian, Martha (1965): „The Judgment of Size“. In: *The American Journal of Psychology*. 78 (3), S. 392–402.

Thompson, Austin B.; Huerta, Guillermo; Robbins, Richard A.; u. a. (1993): „The Bronchitis Index: A Semiquantitative Visual Scale for the Assessment of Airways Inflammation“. In: *Chest*. 103 (5), S. 1482–1488, DOI: 10.1378/chest.103.5.1482.

Tschirren, Juerg; McLennan, Geoffrey; Palágyi, Kálmán; u. a. (2005): „Matching and anatomical labeling of human airway tree“. In: *IEEE transactions on medical imaging*. 24 (12), S. 1540–1547.

Vakil, N (1995): „Measurement of lesions by endoscopy: an overview“. In: *Endoscopy*. 27 (9), S. 694–697.

Vakil, N; Smith, W; Bourgeois, K; u. a. (1994): „Endoscopic measurement of lesion size: improved accuracy with image processing“. In: *Gastrointestinal Endoscopy*. 40 (2 Pt 1), S. 178–183.

Wakabayashi, T; Nakazawa, S; Yoshino, J; u. a. (1994): „A new method of real-time endoscopic measurement with an electric catheter“. In: *Endoscopy*. 26 (5), S. 466–469.

Wildi, Stephan M; Kim, Christopher Y; Glenn, Tammy F; u. a. (2004): „Tele-endoscopy: a way to provide diagnostic quality for remote populations“. In: *Gastrointestinal endoscopy*. 59 (1), S. 38–43.

Wood, R E (2001): „The emerging role of flexible bronchoscopy in pediatrics“. In: *Clinics in Chest Medicine*. 22 (2), S. 311–317, viii.

Wunderlich, P (1969): „EIn dokumentationsgerechter Befundbericht für die Kinderbronchologie“. In: *Zeitschrift für Erkrankungen der Atmungsorgane mit Folia Bronchologica*. 131 , S. 123–130.

2 Material

Kapitelverzeichnis

2.1 Technische Ausstattung

2.1.1 Diagnostik und Dokumentation

Für die Befunderhebung und deren Dokumentation kamen folgende technischen Geräte zum Einsatz, die zusammen auf einem mobilen Untersuchungswagen arrangiert waren:

2.1.1.1 Bronchoskope & Lichtquelle

Flexible Bronchoskopien wurden mit den Bronchoskopen des Typs BF 3C30 bzw. BF-N2 der Firma Olympus (Shinjuku, Japan) durchgeführt. Die Geräte haben folgende Spezifikationen:

Tabelle 2.1: Spezifikationen Olympus BF 3C30

Sichtfeld	120°
Tiefenschärfe	3-50 mm
Außendurchmesser distales Ende	35 mm
Maximal mögliche Beugung	nach oben 180° nach unten 130°
Außendurchmesser Einführungskanal	36 mm
Arbeitslänge	550 mm
Gesamtlänge	840 mm
Innendurchmesser Instrumentenkanal	12 mm
Minimal sichtbare Distanz der Biopsiezange	3 mm vom distalen Ende

Angaben gemäß dem Herstellerhandbuch.

Tabelle 2.2: Spezifikationen Olympus BF-N2

Sichtfeld	75°
Tiefenschärfe	2-50 mm
Außendurchmesser distales Ende	18 mm
Maximal mögliche Beugung	nach oben 160° nach unten 90°
Außendurchmesser flexibler Schlauch	22 mm
Arbeitslänge	550 mm
Gesamtlänge	770 mm

Angaben gemäß dem Herstellerhandbuch.

Als Lichtquelle wurde ein Richard Wolf (Knittlingen) 108 Auto-TCP-Lichtprojektor eingesetzt.

2.1.1.2 Bildverarbeitung

Flexible und starre Bronchoskopien wurden einheitlich mit einer an das Okular des Bronchoskops andockbaren endoskopischen Kamera aufgenommen und parallel an einen Monitor und einen Videorekorder überspielt. Bei den verwendeten Endoskopen erfolgt die optische Übertragung zwischen distalem Ende und Okular am proximalen Ende des Bronchoskops über Glasfaserkabel nach dem Prinzip der Totalreflexion. Die genaue Auflösung der Endoskopoptiken ließ sich auch im Rahmen einer Anfrage bei den Herstellern nicht eruieren. Gewöhnlich finden Glasfaserkabel mit zwischen 30000 und 50000 Fasern Verwendung, sodass die Faserzahl als limitierender Faktor der Bildqualität ausscheidet. Das Bild der Optik wurde mit einer Endocam 5501 CCD Kamera der Firma Richard Wolf (Knittlingen) abgegriffen. Der CCD-Sensor der Kamera registriert intern Bilder mit einer Auflösung von 681x582 Pixeln, die als analoges PAL (ca. 625x582 Pixel) ausgegeben werden. Die Bronchoskopien wurden mit einem VO-9850 U-Matic Videokassettenrekorder der Firma Sony (Minato, Japan) im SP Modus aufgezeichnet. Das von der Kamera ausge-

gebene Bildmaterial wurde auf diesem Weg nahezu verlustfrei gespeichert. Da den an der Studie beteiligten Untersuchern nur VHS-Videorekorder zugänglich waren, wurden die Mitschnitte des U-Matic Rekorders auf handelsübliche VHS-Videokasetten überspielt.

2.1.2 Datenverarbeitung

2.1.2.1 Hardware

Für die Datenverarbeitung wurden ein Vaio PCG-Z1XEP Laptop der Firma Sony (Minato, Japan), sowie ein MacBook Pro vom Frühjahr 2011 der Firma Apple (Cupertino, USA) verwendet. Festplatte (bzw. SSD) und Arbeitsspeicher wurden bei beiden Geräten durch leistungsstärkere Komponenten ersetzt.

Tabelle 2.3: Computer Hardware

	Vaio PCG-Z1XEP	**MacBook Pro (Frühjahr 2011)**
Prozessor	Intel Pentium M Prozessor 1.50 GHz Centrino	27 GHz Intel Core i7
Speicher	1 GB	16 GB 1333 MHz DDR3
Grafikkarte	ATI Mobility Radeon M6-C16 16MB	Intel HD Graphics 3000 512 MB

Eckdaten der für die Auswertung verwendeten Computerhardware.

2.1.2.2 Software

2.1.2.2.1 Betriebssysteme

Das Vaio Laptop wurde unter den Betriebssytemen Windows XP Professional (Microsoft Redmond USA) und Ubuntu (Canonical Millbank London) in den Versionen 8.04 LTS bzw. 10.04 LTS betrieben. Das MacBook Pro war mit Mac OS X Lion 10.8.3 bis Mac OS X Yosemite 10.10 (Apple, Cupertino, USA) und Windows 7 Ultimate (Microsoft, Redmond, USA) ausgestattet.

2.1.2.2.2 Textverarbeitung und Schriften

Texte wurden in OpenOffice (Sun Microsystems, Santa Clara, USA; später Oracle, Redwood City, USA und zuletzt Apache Software Foundation, Delaware, USA) bzw. LibreOffice Writer (The Document Foundation, Berlin) verfasst. Dabei wurden die quelloffenen Schriftfamilien Source Sans Pro (Hunt, 2015; Adobe, San José, USA) und Vegur (Sagano, 2015) eingesetzt.

2.1.2.2.3 Literaturverwaltung und Recherche

Zur Literaturrecherche kam Firefox (Mozilla, Mountain View, USA) mit der Erweiterung Zotero (Roy Rosenzweig Center for History and New Media, George Mason University, USA) zum Einsatz. Referenzen wurden mithilfe des Zotero-Plugins für LibreOffice in den Text übernommen.

2.1.2.2.4 Tabellenkalkulation

Die Daten des Befund- und Arztfragebogens wurden mit den Programmen **Excel** (Microsoft, Redmond, USA) bzw. LibreOffice **Calc** (The Document Foundation, Berlin) in Tabellen übertragen und anschließend für die Berechnungen umformatiert sowie um Dummy-Variablen ergänzt.

2.1.2.2.5 Datenbank

Um komplexere Analysen zu ermöglichen, wurden die Tabellen im Verlauf aus der Tabellenkalkulation in die SQL-Datenbank MySQL (Sun Microsystems, Santa Clara, USA; später Oracle, Redwood City, USA) übertragen. Virtuelle Variablen wurden als SQL-Views realisiert. Datenbankabfragen wurden mit dem MySQL Query Browser entworfen und das Ergebnis über die Bibliothek RMySQL in die Statistiksprache R importiert.

2.1.2.2.6 Statistische Analyse

Sämtliche Berechnungen wurden mithilfe von Tabellenkalkulationen und der auf maßgeschneiderte statistische Analysen spezialisierten Programmiersprache R durchgeführt.

2.1.2.2.6.1 Excel Kalkulationsvorlagen

Explorative und deskriptive Analysen sowie die Berechnung grundlegender Maßzahlen wurde zunächst mittels Vorlagen für die Tabellenkalkulation Excel durchgeführt. Neben einer frei verfügbaren Vorlage (Mackinnon, 2000) kam dabei auch das kommerziell erhältliche AgreeStat (Gwet, 2001) zum Einsatz.

2.1.2.2.6.2 R-Projekt

Komplexere statistische Analysen wurden mithilfe der Programmiersprache R (R Core Team, 2016) umgesetzt, einer quelloffenen Implementierung der Sprache S (Becker, Chambers, 1984). Die Funktionalität des Basissystems wurde durch zahlreiche Bibliotheken erweitert. Eine Übersicht über die wichtigsten der verwendeten Bibliotheken gibt die folgende Tabelle.

Tabelle 2.4: Verwendete R-Bibliotheken

Bibliothek	Autoren / Referenz	wichtige Funktionen	Anwendung
caret	(Kuhn, 2008)	ConfusionMatrix agreementplot	Kontingenztafeln mit Kennwerten Übereinstimmungs-Diagramm
DiagnosisMed	(Brasil, 2010)	LRgraph	Likelihood ratio graphs
Irr	(Gamer u. a., 2012)		Berechnung von Kappa nach Fleiss
multiclasstesting	(Nardini, Liu, 2014)		sen und ppv in Kontingenztafeln
odfWeave	(Kuhn u. a., 2014; o. A., o. J.)		Formatierung von Ergebnissen
plotmo	(Milborrow, 2016)		graphische Variableninteraktion
stringr	(Wickham, 2016)		String-Operationen
foreach	(Analytics, Weston, 2015)		Schleifen
gvlma	(Pena, Slate, 2014)	gvlma.lm	Überprüfung Annahmen linearer Modelle
relaimpo	(Groemping, 2006)		Variablenwichtigkeit im linearen Modell
randomForest	(Liaw, Wiener, 2002)		Imputation, Variablenwichtigkeit
regr0	(Stahel, 2013)	regr	Multiple lineare Regression
reshape	(Wickham, Hadley, 2007)		Reformatierung von Daten zur Analyse
RMySQL	(Ooms u. a., 2016)		Datenbankimport
ROCR	(Sing u. a., 2005)		Berechnung diverser Kenngrößen
rpart	(Therneau u. a., 2015)		Rekursive Partitionierung
sqldf	(Grothendieck, 2014)		Reformatierung von Datensätzen
vcd	(Meyer u. a., 2006)	assoc	Assoziationsdiagramme
vcdextra	(Friendly, 2016)	expand.dft	Expansion von Tabellen

Übersicht über die wichtigsten R-Bibliotheken, die für die statistische Analyse eingesetzt wurden.

Als Editoren für R-Code wurden Emacs (Stallman, 2015, 1981) mit der Erweiterung (major mode) Emacs Speaks Statistics (ESS; Rossini A.J.[1] u. a., 2004) sowie R-Studio (RStudio Team, 2015) verwendet. Der Quellcode wurde im Versionierungssystem git (Torvalds, 2005) gepflegt, wobei R-Studio, gitlab (Saparoschez, 2011) und Tower (Günther, 2010) als graphische Benutzeroberflächen (GUI; engl. Graphical User Interface) genutzt wurden.

2.2 Videomaterial

Die Videobibliothek dieser Studie beinhaltet 42 Mitschnitte von Bronchoskopien, die zu 6 Themen- bzw. Diagnosegruppen zusammengestellt wurden.

Tabelle 2.5: Übersicht der 6 Diagnosegruppen der Videobibliothek

Gruppe	Thema	Anzahl der Mitschnitte	Bilder ab Seite
I	Tracheomalazie	5	27
II	Trachealeinengung	8	28
III	Verhältnisse der Hauptbronchien	5	31
IV	Stimmbandbeweglichkeit	10	33
V	Kompressionen der Trachea und der Bronchien	6	36
VI	Larynxanomalien	8	38

Themengruppen der Videobibliothek

Um einen Eindruck der Videobibliothek zu vermitteln, werden im folgenden Abschnitt ausgewählte Standbilder der einzelnen Mitschnitte demonstriert. Die Standbilder sind in der Regel gemäß ihrer zeitlichen Abfolge bzw. von proximal nach distal angeordnet. In jeder Gruppe sind in einer Übersichtstabelle zu den einzelnen Videos das Alter des Patienten, die Diagnose des Goldstandards und die Dauer des Mitschnitts angegeben.

2.2.1 Gruppe I: Tracheomalazie

Tabelle 2.6: Übersicht Videomitschnitte Gruppe I – Tracheomalazie

Video	Alter	Diagnosen	Dauer
I-1	2 Monate	Tracheobronchomalazie	1 Minute 59 Sekunden
I-2	7 Monate	Laryngotracheobronchomalazie mit 90 % Trachealstenose	1 Minute 51 Sekunden
I-3	6 Jahre	Langstreckige proximale Trachealstenose 70 % durch Malazie	2 Minuten 19 Sekunden
I-4	1 Jahr	Laryngomalazie; narbige Cricoidstenose 70 %; Tracheostomagranulom Tracheobronchitis narbige Stimmlippenveränderung	1 Minute 28 Sekunden
I-5	1 Jahr	Tracheomalazie Stomagranulom Retrogenie	31 Sekunden

Abbildung 2.1: **Video I-1** Tracheobronchomalazie

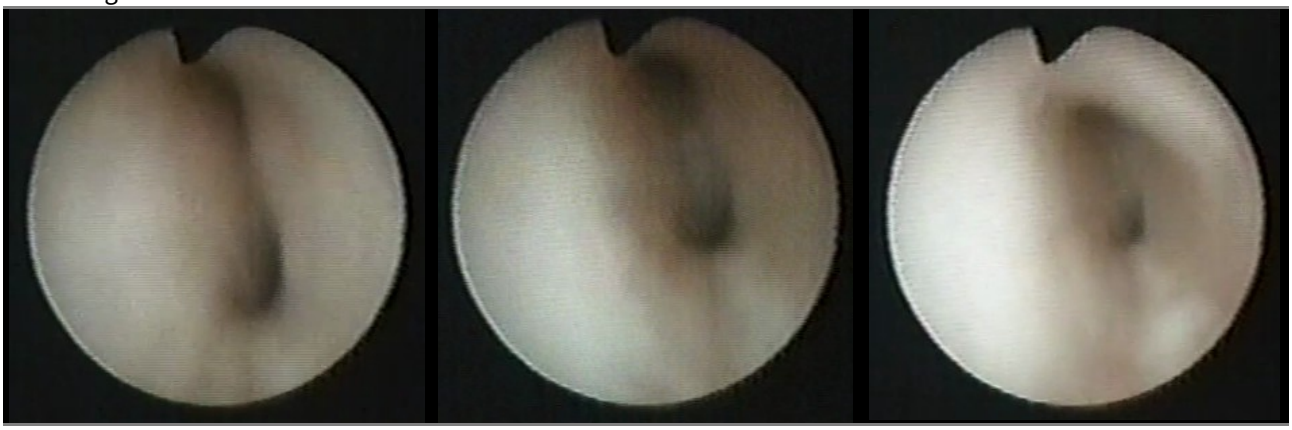


Abbildung 2.2: **Video I-2** Laryngotracheobronchomalazie mit Trachealstenose

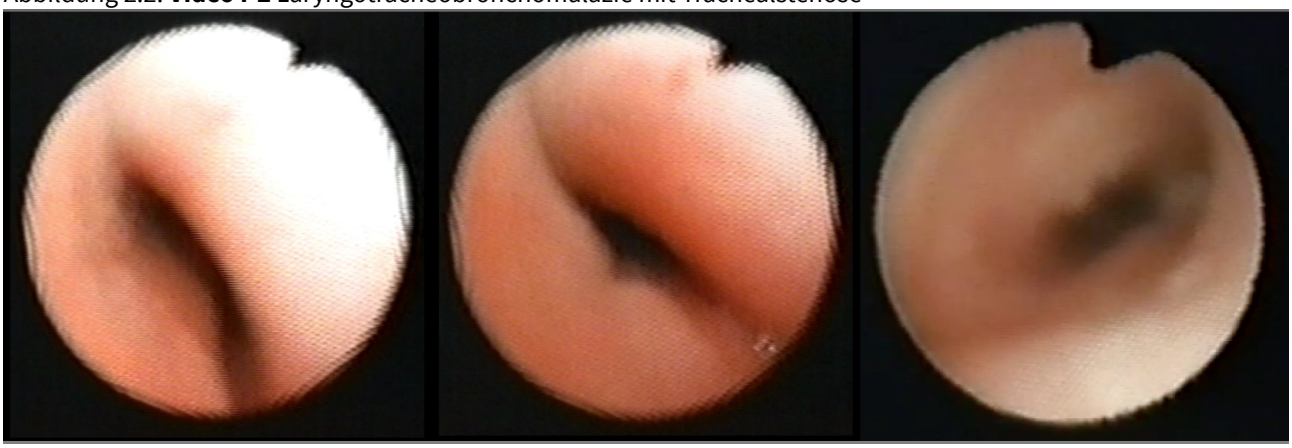


Abbildung 2.3: **Video I-3** Langstreckige proximale Trachealstenose durch Malazie

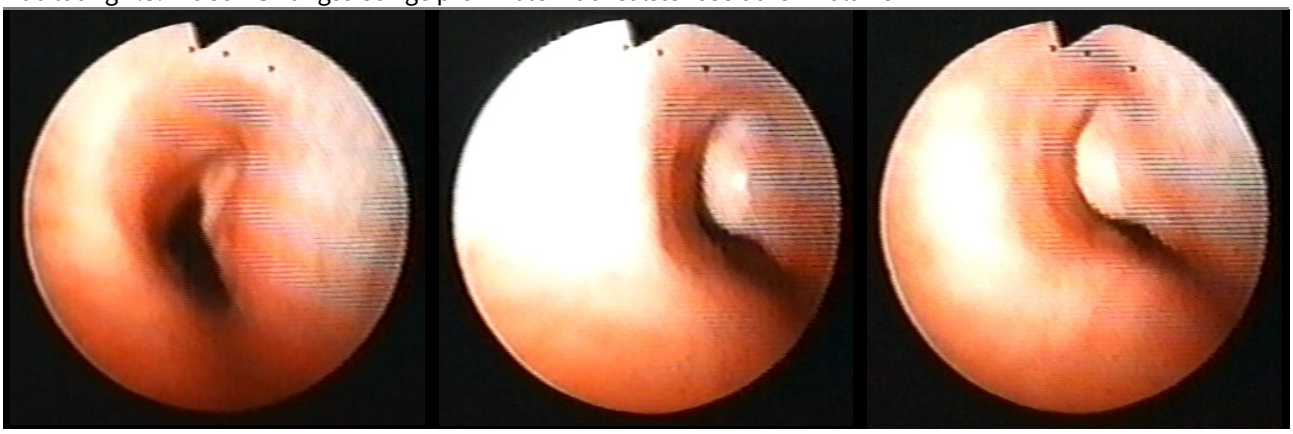


Abbildung 2.4: **Video I-4** Laryngomalazie, Cricoidstenose, Tracheostomagranulom, Bronchitis

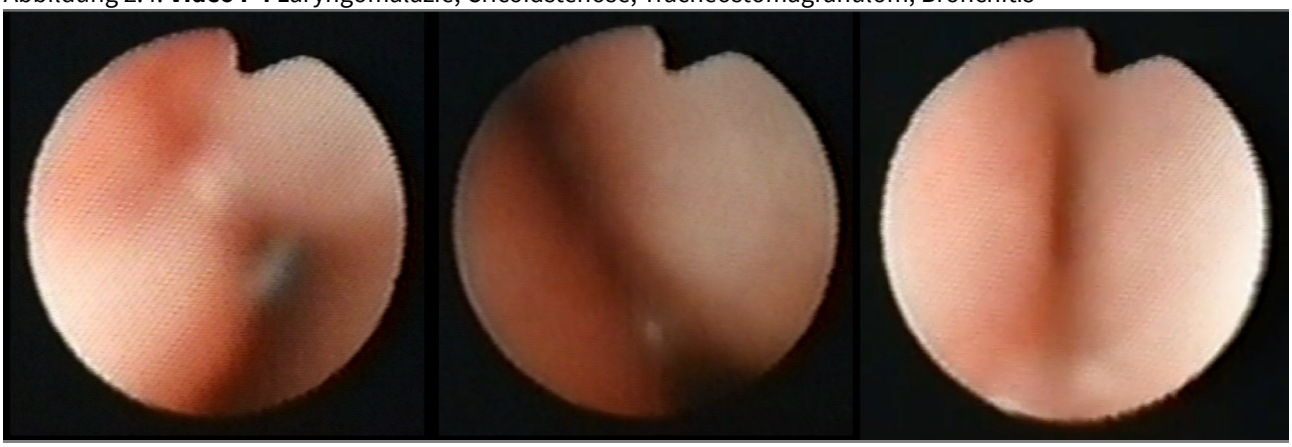
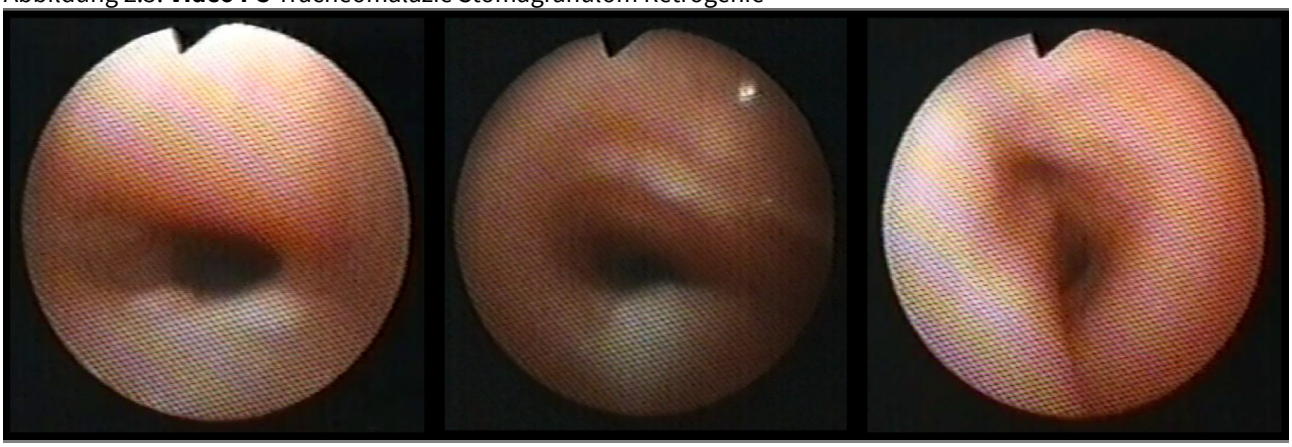


Abbildung 2.5: **Video I-5** Tracheomalazie Stomagranulom Retrogenie



2.2.2 Gruppe II: Trachealeinengung

Tabelle 2.7: Übersicht Videomitschnitte Gruppe II - Trachealeinengungen

Video	Alter	Diagnose	Dauer
II-1	6 Monate	Entzündliche glottische und langstreckige subglottische Larynxstenose von 70 %	25 Sekunden
II-2	9 Jahre 8 Monate	Langstreckige Trachealstenose der mittleren und distalen Trachea von 30 % durch Knorpelringfehlbildung und narbige Veränderungen	1 Minute 14 Sekunden
II-3	3 Jahre 5 Monate	Subglottische Larynxstenose von 80 %, Cricoidmalformation	2 Minuten 50 Sekunden
II-4	11 Monate	Larynxstenose von 50 % Stimmlippenpolster	1 Minute 31 Sekunden
II-5	4 Jahre 8 Monate	Cricoidstenose von 70 % entzündliche glottisch und subglottische Larynxstenose	23 Sekunden
II-6	3 Jahre 3 Monate	Schwere Tracheobronchitis	1 Minute 22 Sekunden
II-7	6 Monate	Cricoidstenose von 95 %	1 Minute 1 Sekunde
II-8	9 Jahre 9 Monate	Distale Trachealstenose von 90 %	32 Sekunden

Abbildung 2.6: **Video II-1** Entzündliche glottische & langstreckige subglottische Larynxstenose

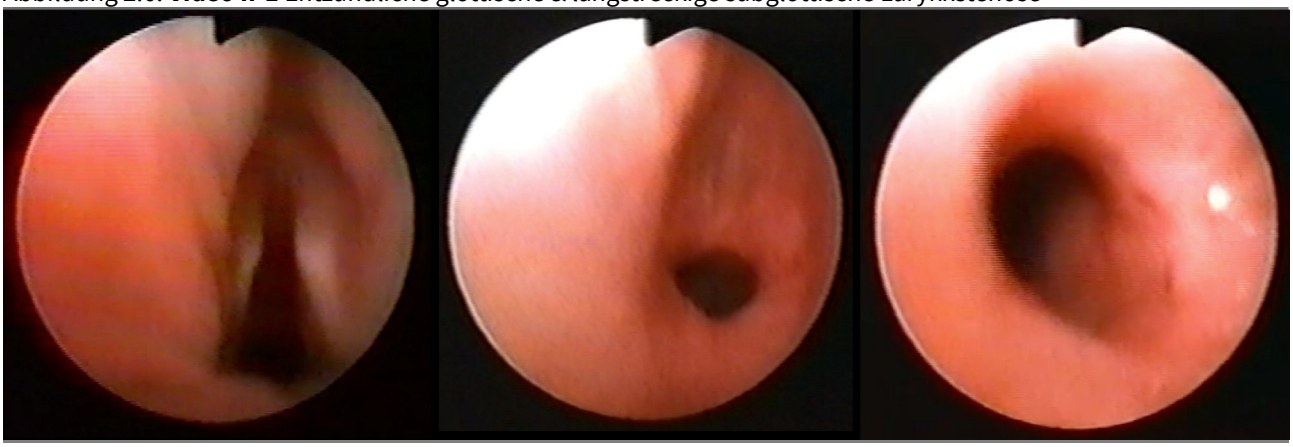


Abbildung 2.7: **Video II-2** Langstreckige Trachealstenose durch Knorpelringfehlbildung

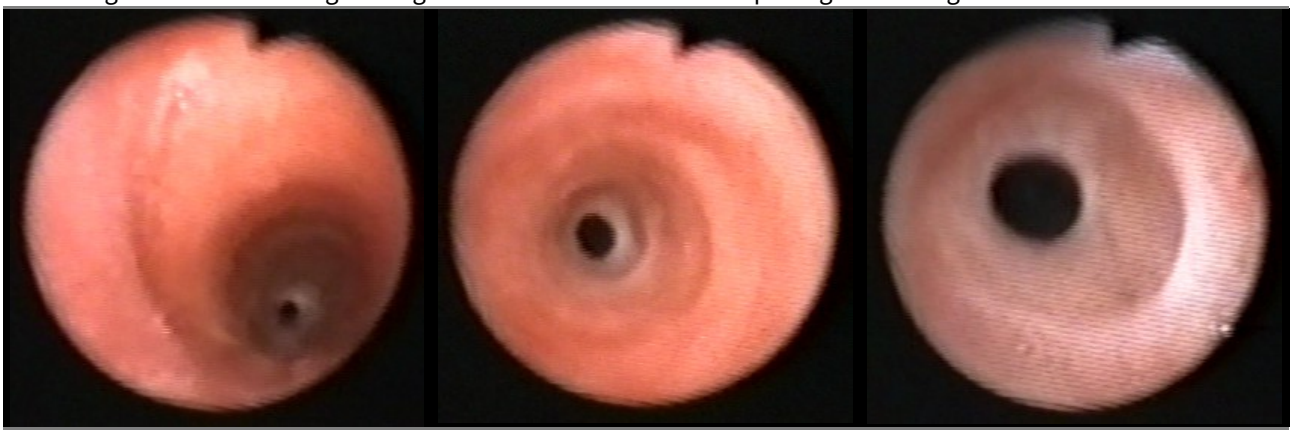


Abbildung 2.8: **Video II-3** Subglottische Larynxstenose, Cricoidmalformation

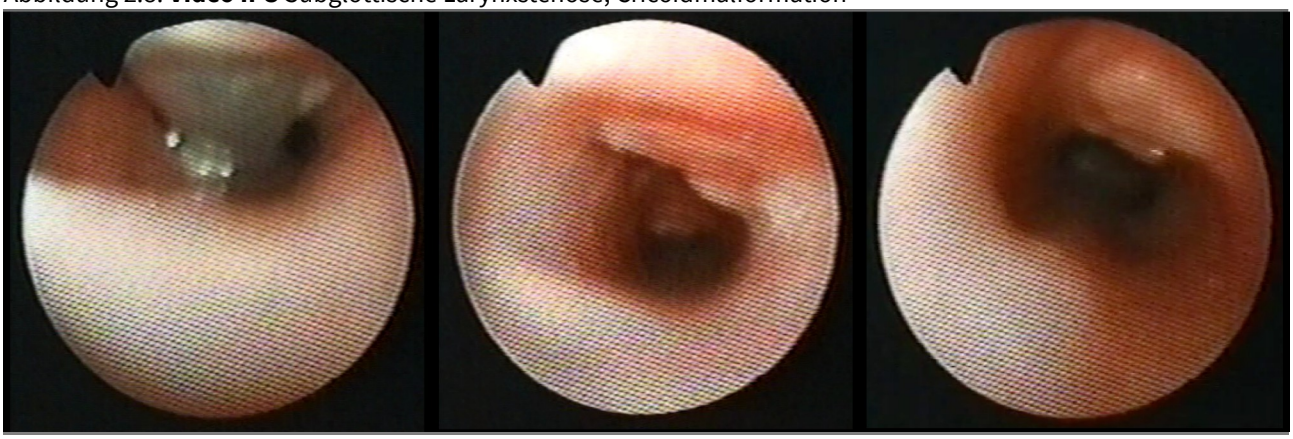


Abbildung 2.9: **Video II-4** Larynxstenose, Stimmlippenpolster

Abbildung 2.10: **Video II-5** Cricoidstenose, entzündliche glottisch & subglottische Larynxstenose

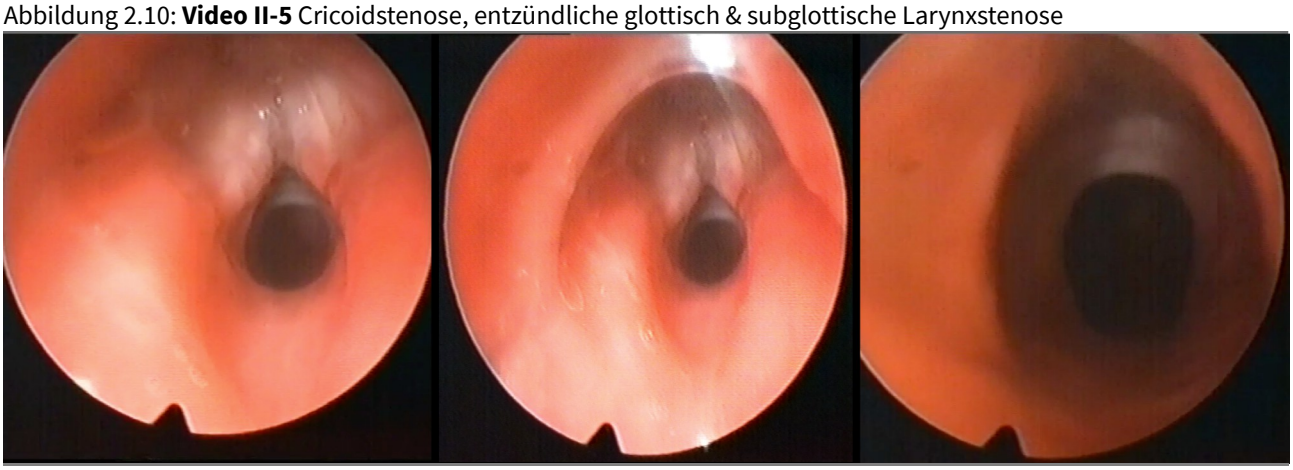


Abbildung 2.11: **Video II-6** Schwere Tracheobronchitis

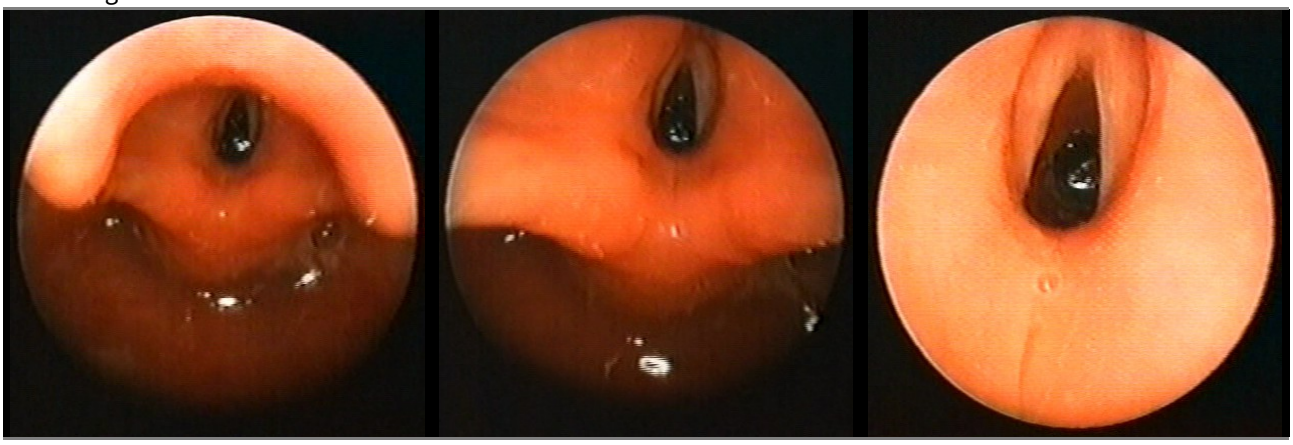


Abbildung 2.12: **Video II-7** Cricoidstenose

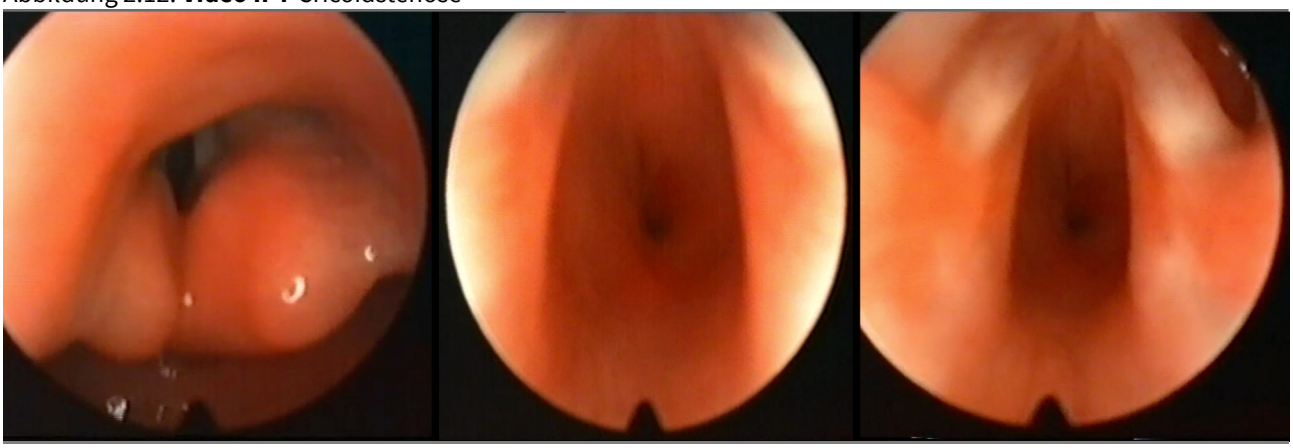
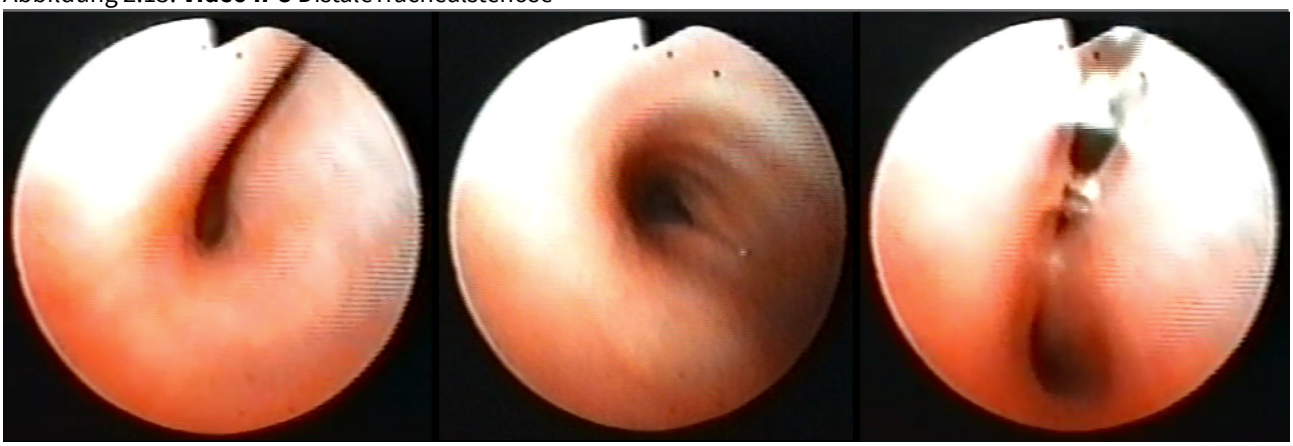


Abbildung 2.13: **Video II-8** DistaleTrachealstenose



2.2.3 Gruppe III: Verhältnisse der Hauptbronchien

Tabelle 2.8: Übersicht Gruppe III Verhältnisse der Hauptbronchien

Video	Alter	Diagnose	Dauer
III-1	1 Monat	Kompression des rechten Hauptbronchus von 95 %-100 % Kompression des linken Hauptbronchus von 70 % V.a. Pulmonalisschlinge	59 Sekunden
III-2	2 Monate	Stenose des linken Hauptbronchus von 80 % Entzündung im Glottisbereich	55 Sekunden
III-3	6 Jahre 3 Monate	Stenose im linken Unterlappenbronchus pulsierende Kompression	59 Sekunden
III-4	4 Jahre 8 Monate	Stenose im rechten Mittellappen von 40 % geringe Bronchitis	1 Minute 54 Sekunden
III-5	22 Tage	Laryngomalazie mit geringer subglottischer Entzündung	2 Minuten 11 Sekunden

Abbildung 2.14: **Video III-1 Kompression linker & rechter Hauptbronchus, V.a. Pulmonalisschlinge**

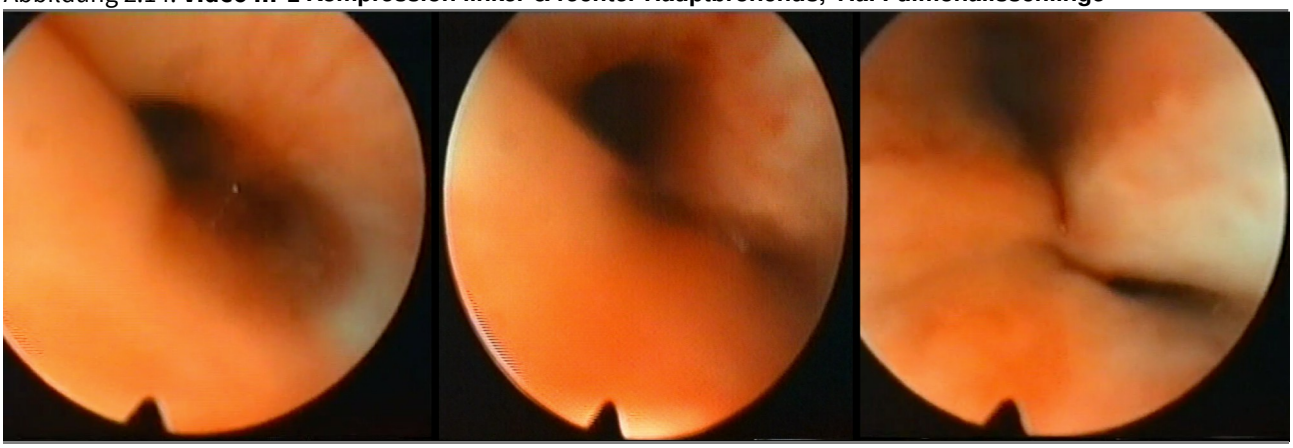


Abbildung 2.15: **Video III-2** Stenose des linken Hauptbronchus, Entzündung im Glottisbereich

Abbildung 2.16: **Video III-3** Stenose linker Unterlappenbronchus pulsierende, Kompression

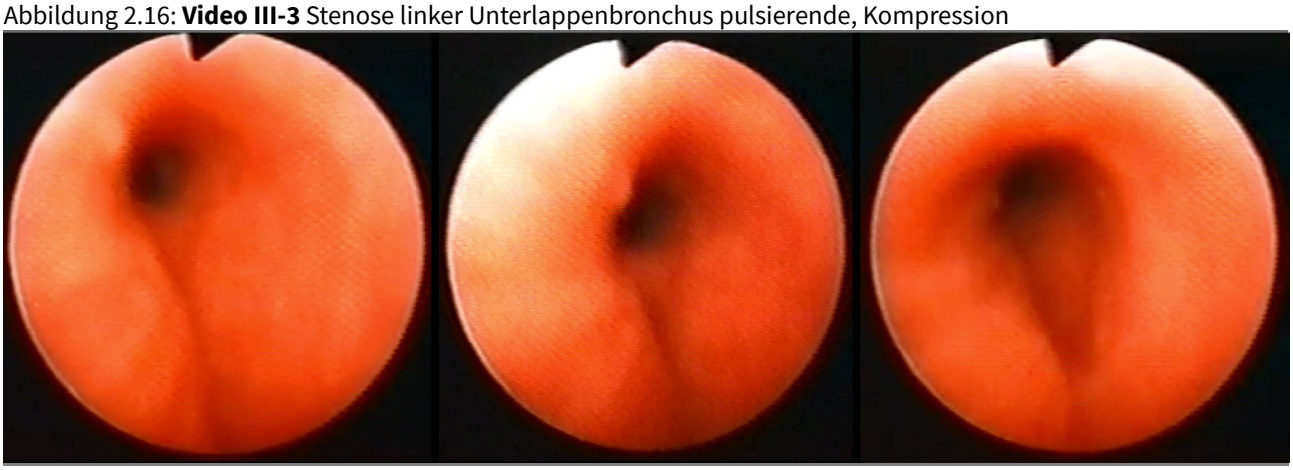
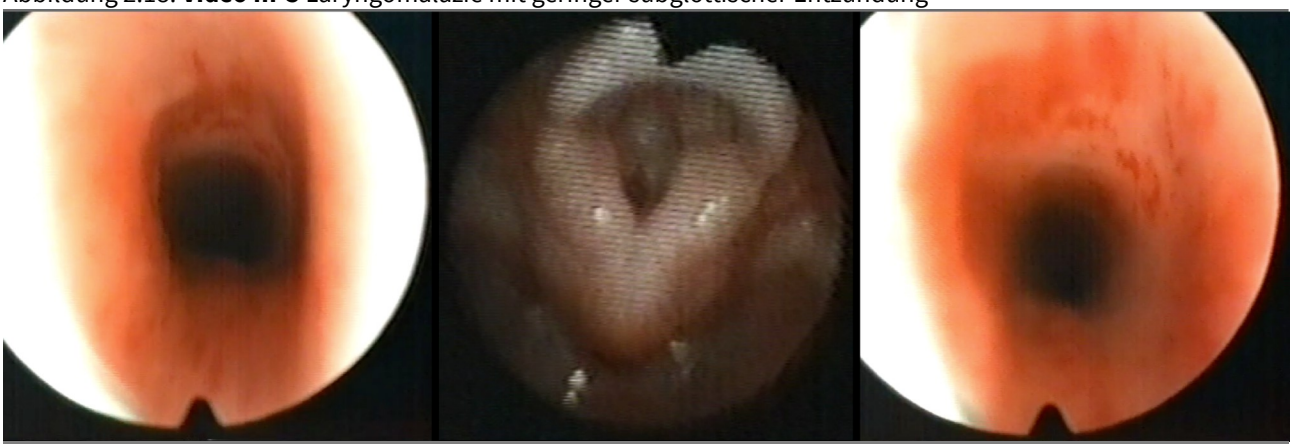


Abbildung 2.17: **Video III-4** Stenose rechter Mittellappen, geringe Bronchitis

Abbildung 2.18: **Video III-5** Laryngomalazie mit geringer subglottischer Entzündung



2.2.4 Gruppe IV Stimmbandbeweglichkeit

Tabelle 2.9: Übersicht Videomittschnitte Gruppe IV - Stimmbandbeweglichkeit

Video	Alter	Diagnose	Dauer
IV-1	10 Jahre 1 Monat	V.a. Stimmlippendysfunktion normaler Larynx	1 Minute 5 Sekunden
IV-2	3 Monate	Stimmlippenparese beidseits	1 Minute 25 Sekunden
IV-3	14 Jahre 9 Monate	Narbige Stimmlippenfixierung subglottische Ringstenose im Ringknorpelbereich	4 Minuten 8 Sekunden
IV-4	4 Jahre 11 Monate	Bilaterale Stimmlippenparese	2 Minuten 45 Sekunden
IV-5	6 Jahre 3 Monate	Narbige Larynxstenose von 80 % narbige Stimmlippenverwachsung	1 Minute 2 Sekunden
IV-6	4 Jahre 11 Monate	Recurrensparese beidseits	2 Minuten 48 Sekunden
IV-7	8 Monate	Stimmlippenparese beidseits	3 Minuten 2 Sekunden
IV-8	11 Jahre 6 Monate	V.a. Stimmlippendysfunktion normaler Larynx	49 Sekunden
IV-9	10 Monate	Abduktionshemmung der Stimmlippen mit Stridor	3 Minuten 43 Sekunden
IV-10	1 Jahr 3 Monate	Larynxspalte 1. Grades	2 Minuten 31 Sekunden

Abbildung 2.19: **Video IV-1** V.a. Stimmlippendysfunktion, normaler Larynx

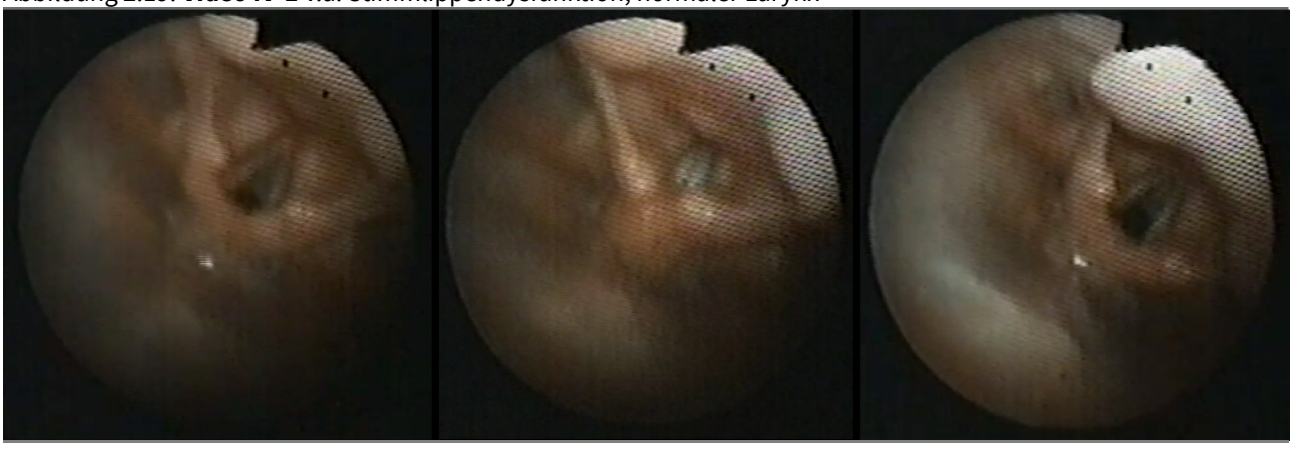


Abbildung 2.20: **Video IV-2** Stimmlippenparese beidseits

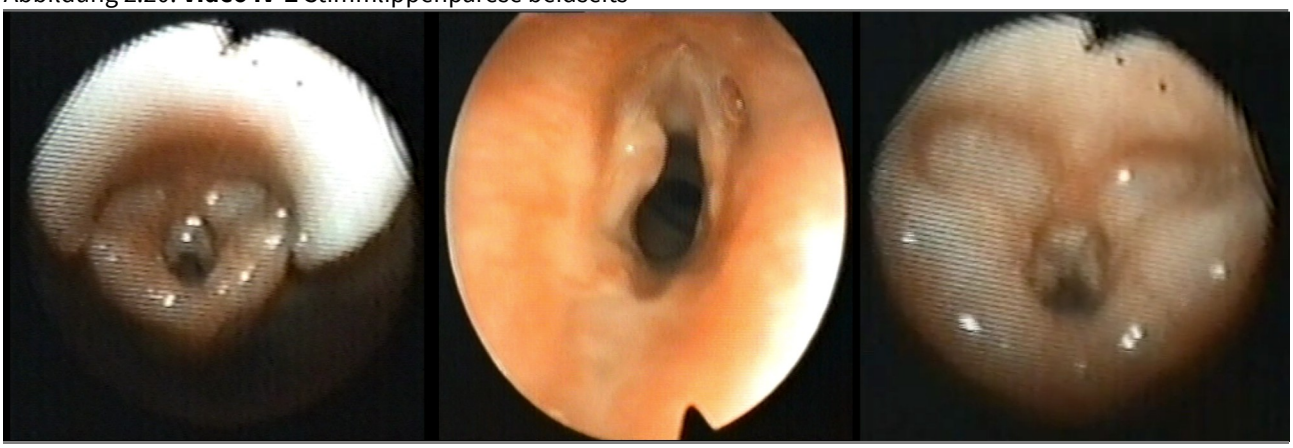


Abbildung 2.21: **Video IV-3** Narbige Stimmlippenfixierung, subglottische Ringstenose

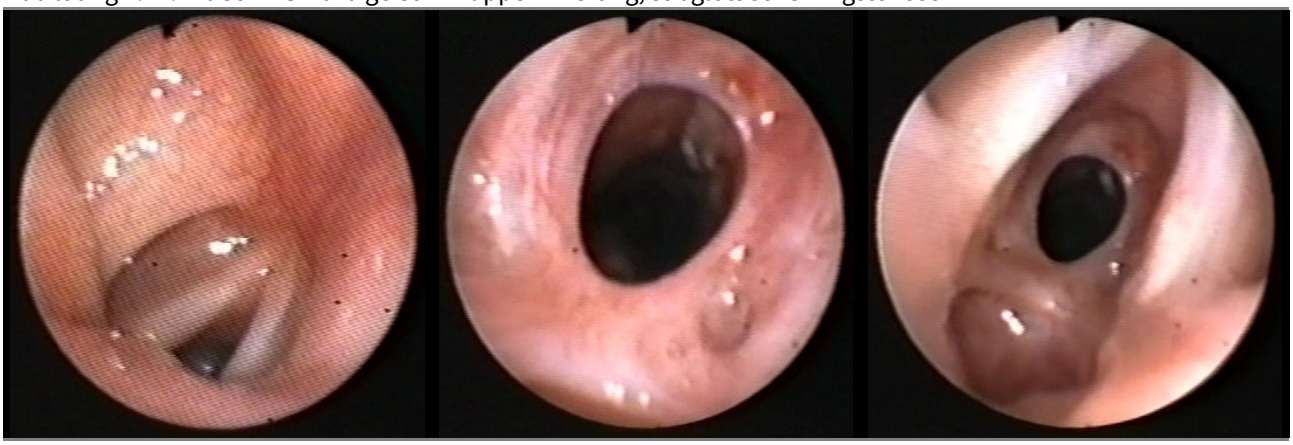


Abbildung 2.22: **Video IV-4** Bilaterale Stimmlippenparese

Abbildung 2.23: **Video IV-5** Narbige Larynxstenose, narbige Stimmlippenverwachsung

Abbildung 2.24: **Video IV-6** Recurrensparese beidseits

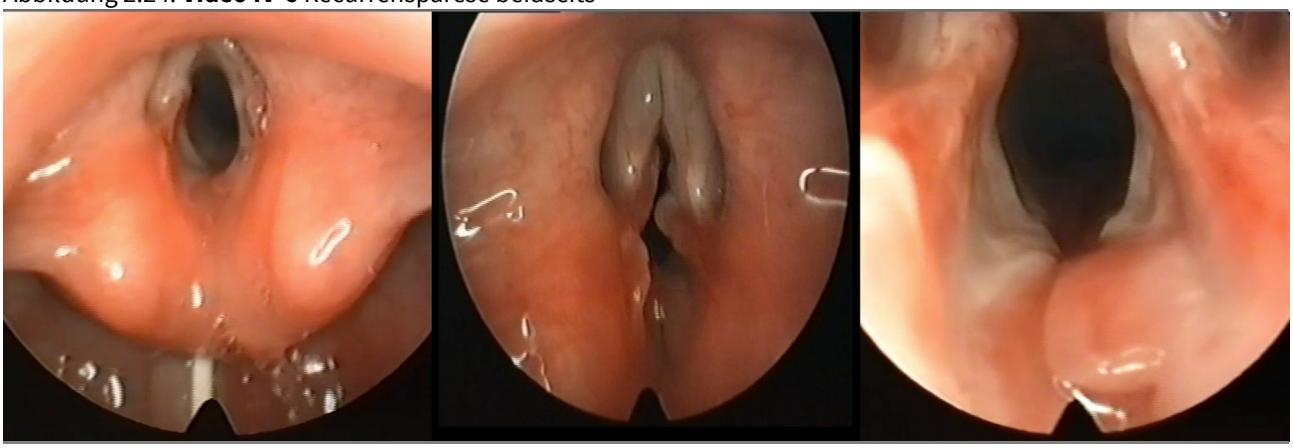


Abbildung 2.25: **Video IV-7** Stimmlippenparese beidseits

Abbildung 2.26: **Video IV-8** V.a. Stimmlippendysfunktion normaler Larynx

Abbildung 2.27: **Video IV-9** Abduktionshemmung der Stimmlippen mit Stridor

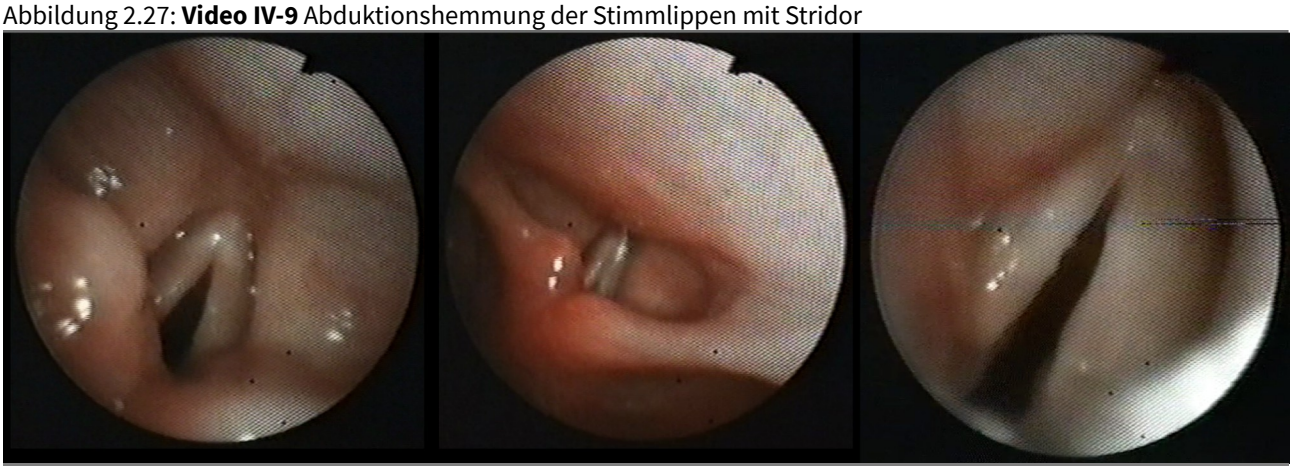
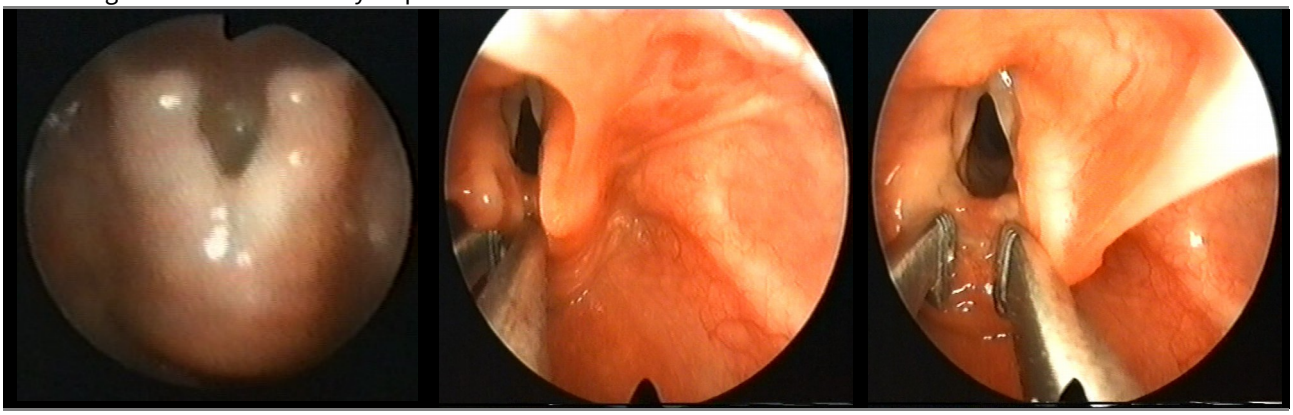


Abbildung 2.28: **Video IV-10** Larynxspalte 1. Grades



2.2.5 Gruppe V Kompression der Trachea und der Bronchien

Tabelle 2.10: Übersicht Videomitschnitte Gruppe VI - Kompression Trachea & Bronchien

Video	Alter	Diagnose	Dauer
V-1	5 Monate	Subglottisches Hämangiom mit 80 % Stenose	1 Minute 57 Sekunden
V-2	6 Monate	Trachealstenose von 90 % bei V.a. Truncuskompression Tracheobronchitis	4 Minuten 39 Sekunden
V-3	9 Monate	Subglottisches Hämangiom mit 90 % Stenose	12 Sekunden
V-4	3 Jahr 2 Monate	Bronchogene Zyste mit Stenose von 70 % des linken Haupt- & Lappenbronchus	3 Minuten 3 Sekunden
V-5	1 Jahr 3 Monate	Chronischer Ösophagusfremdkörper mit 95 % Tracheakompression	2 Minuten 12 Sekunden
V-6	12 Jahre 5 Monate	V.a. doppelten Aortenbogen mit Tracheakompression von 40 % und Stammbronchuskompression von 80 %	4 Minuten 28 Sekunden

Abbildung 2.29: **Video V-1** Subglottisches Hämangiom mit Stenose

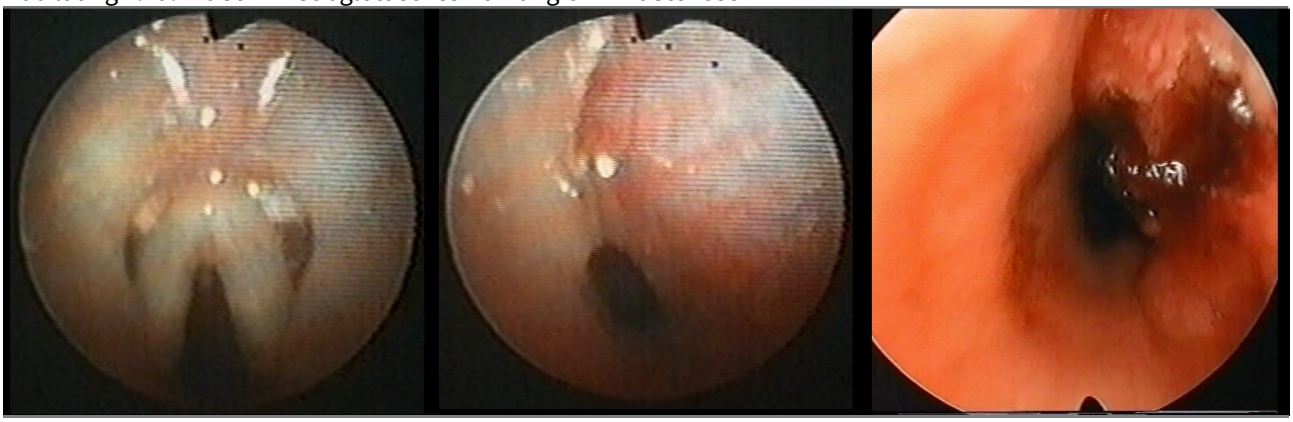


Abbildung 2.30: **Video V-2** Trachealstenose bei V.a. Truncuskompression, Tracheobronchitis

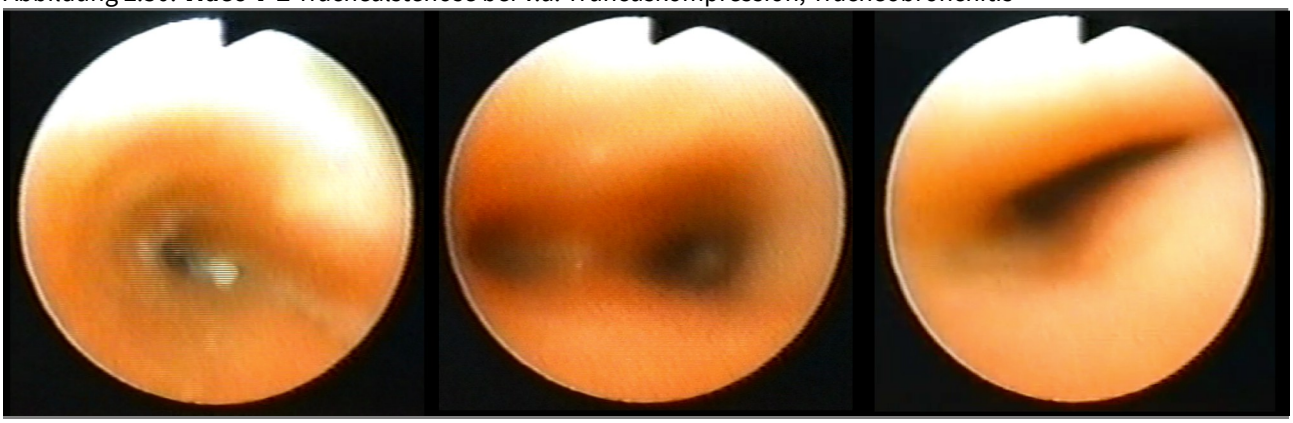


Abbildung 2.31: **Video V-3** Subglottisches Hämangiom mit 90 % Stenose

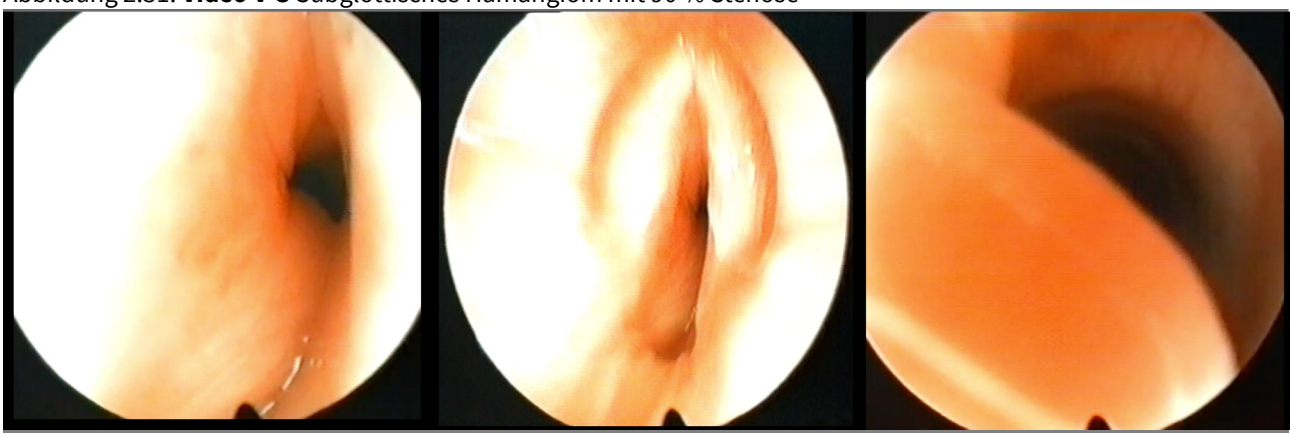


Abbildung 2.32: **Video V-4** Bronchogene Zyste mit Stenose des linken Haupt- & Lappenbronchus

Abbildung 2.33: **Video V-5** Chronischer Ösophagusfremdkörper mit Tracheakompression

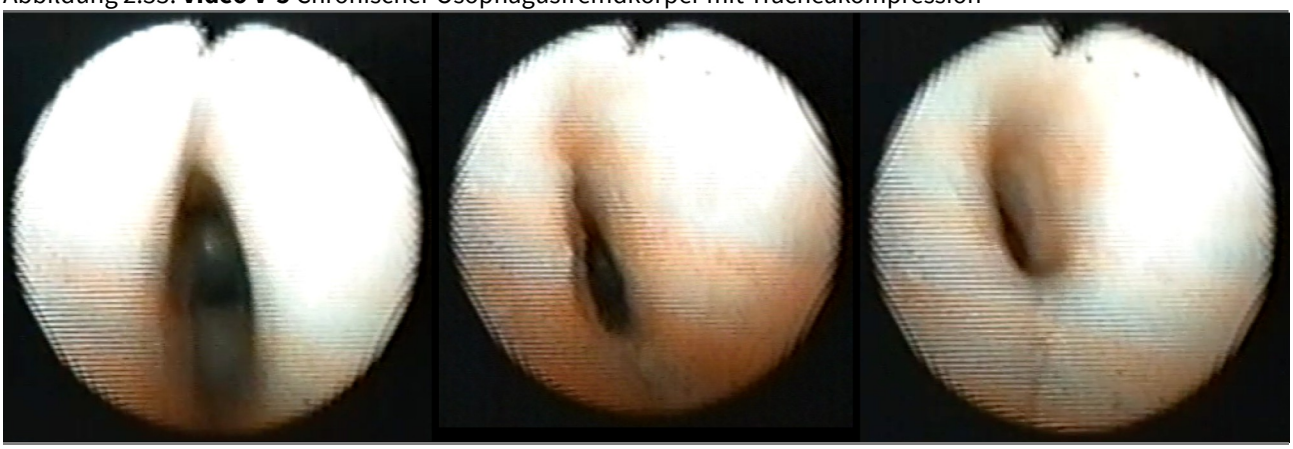


Abbildung 2.34: **Video V-6** V.a. doppelten Aortenbogen mit Kompressionen

2.2.6 Gruppe VI Larynxanomalien

Tabelle 2.11: Übersicht Videomitschnitte Gruppe VI - Larynxanomalien

Video	Alter	Diagnose	Dauer
VI-1	18 Tage	Larynxspalte	3 Minuten 28 Sekunden
VI-2	10 Monate	Puderaspiration	45 Sekunden
VI-3	2 Jahre 10 Monate	Larynxstenose von 70 % bei Larynxpapillomatose	1 Minute 23 Sekunden
VI-4	1 Jahr 5 Monate	Fibrinöse Laryngo-Tracheobronchitis	3 Minuten 5 Sekunden
VI-5	6 Tage	Larynxzyste links	1 Minute 13 Sekunden
VI-6	3 Monate	Infantiler Larynx	1 Minute 5 Sekunden
VI-7	3 Jahre 7 Monate	Larynxstenose von 70 % bei Larynxpapillomatose	1 Minute 2 Sekunden
VI-8	13 Jahre 6 Monate	Trachearuptur subglottisch	4 Minuten 15 Sekunden

Abbildung 2.35: **Video VI-1** Larynxspalte

Abbildung 2.36: **Video VI-2** Puderaspiration

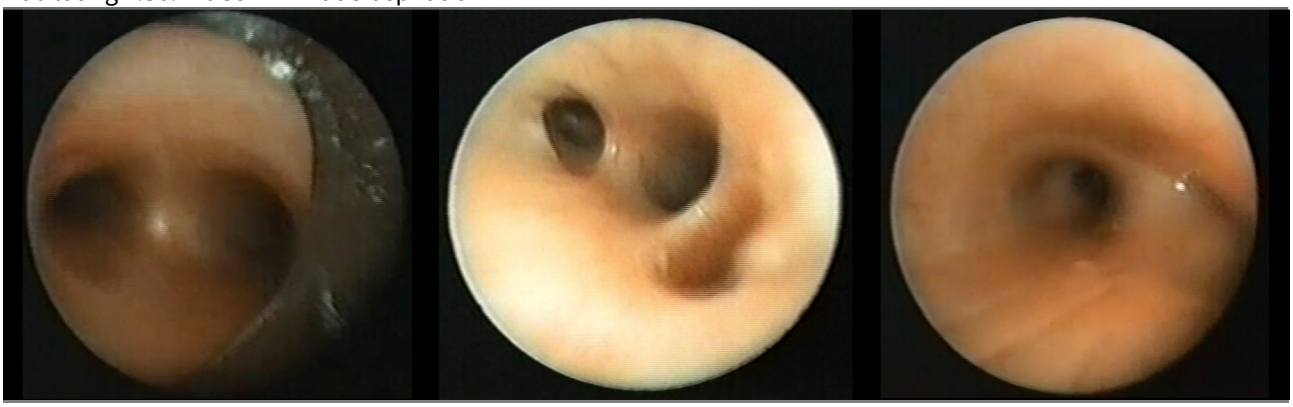


Abbildung 2.37: **Video VI-3** Larynxstenose bei Larynxpapillomatose

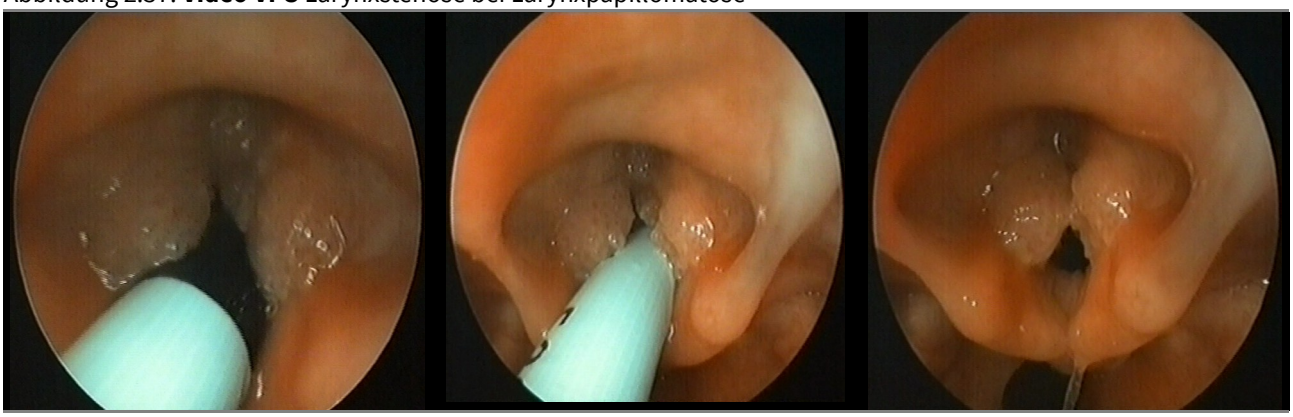


Abbildung 2.38: **Video VI-4** Fibrinöse Laryngo-Tracheobronchitis

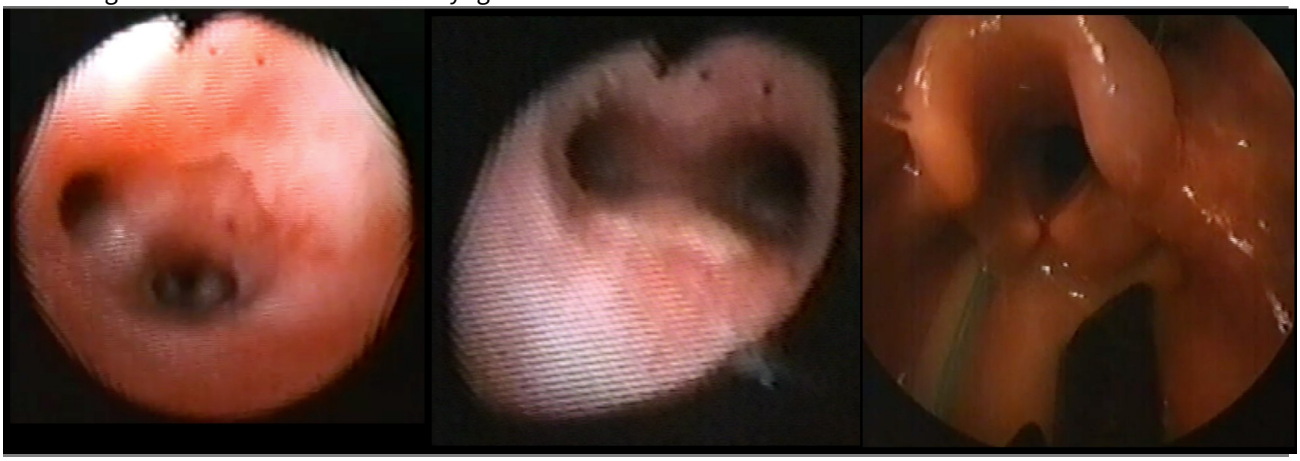


Abbildung 2.39: **Video VI-5** Larynxzyste links

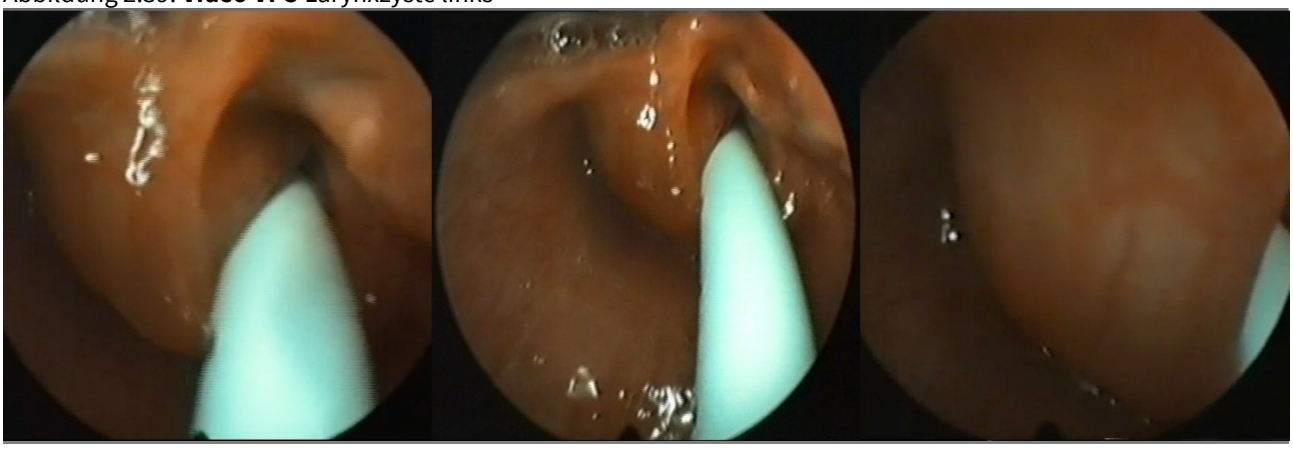


Abbildung 2.40: **Video VI-6** Infantiler Larynx

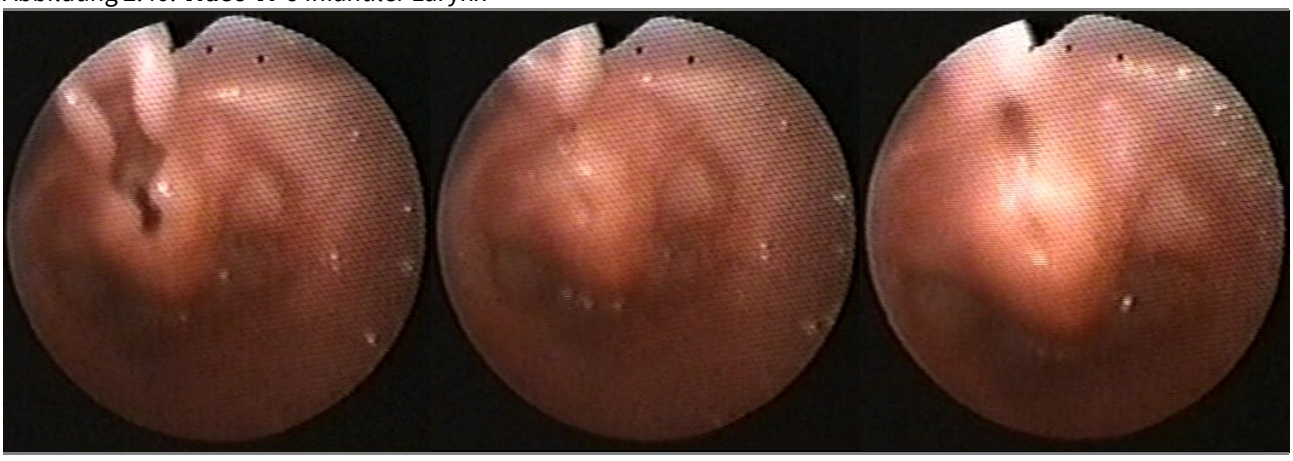


Abbildung 2.41: **Video VI-7** Larynxstenose bei Larynxpapillomatose

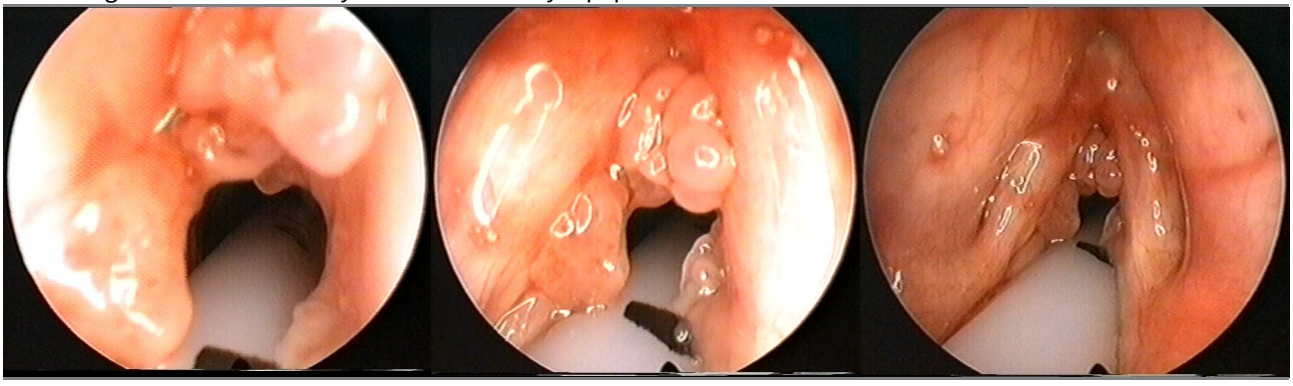
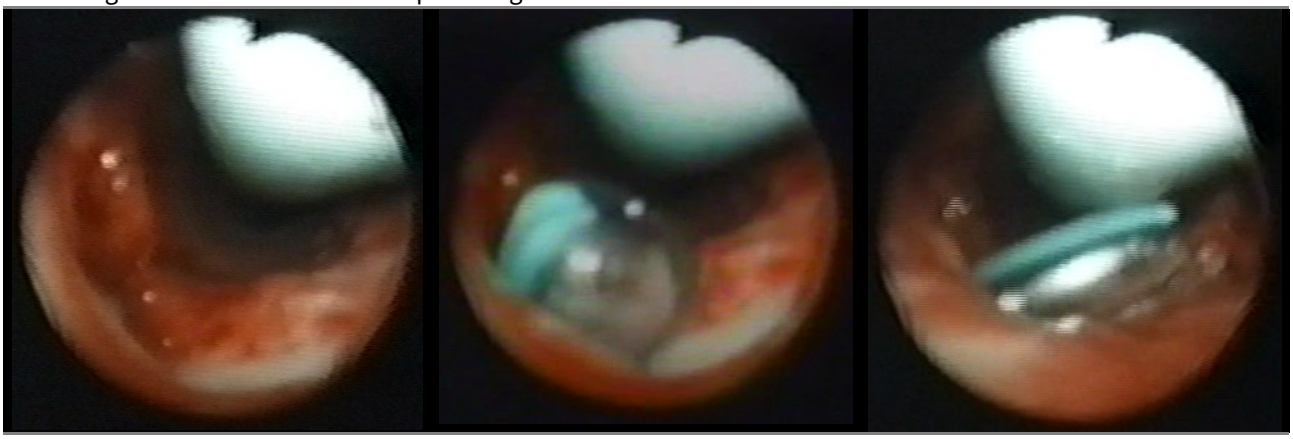


Abbildung 2.42: **Video VI-8** Trachearuptur subglottisch



2.3 Fragebögen

2.3.1 Befundfragebogen

Die Befunde wurden mit einem einheitlichen Befundbogen erhoben, der abgesehen von wenigen Freitextfeldern (z. B. für die Hauptdiagnose) im multiple-choice-Verfahren konstruiert ist (siehe Anhang).

2.3.1.1 Videoqualität

Der Begriff der Videoqualität definiert sich in dem für diese Studie entworfenen Modell aus den Parametern Bildqualität, Aufnahmedauer und Aufnahmesituation.

2.3.1.1.1 Bildqualität

Die technisch maximal mögliche Bildqualität war im Rahmen dieser Studie durch den VHS-Standard begrenzt, da zur Dokumentation der bronchoskopischen Untersuchungen am Ende der Verarbeitungskette VHS-Videobänder eingesetzt wurden. VHS (engl. Video Home System) ist ein analoges Aufzeichnungsverfahren, dessen effektive Bildauflösung umgerechnet vertikal bei ca. 576, horizontal in etwa bei 310 Pixel liegt.

VHS bot sich im Rahmen dieser Studie als Medium der Wahl an, da das Bildmaterial auf einem U-Matic Videokassettenrekorder (Sony V05850-P) dokumentiert wurde, das aufgrund der geringen Verbreitung von U-Matic nicht direkt weiter gegeben werden konnte. Ein direkter Transfer des Bildmaterials auf DVD war mit der zur Verfügung stehenden Gerätschaft leider nicht möglich.

Da nicht alle Untersucher über einheitliche Abspielgerät und Bildschirme verfügten, war die objektive technische Bildqualität nicht bestimmbar. Aufgrund von Studien aus verwandten Anwendungsbereichen (Barbier u. a., 2007; Seidenari u. a., 2004; Spencer u. a., 2000) ist jedoch kein wesentlicher Einfluss der objektiven technischen Bildqualität auf die Befundung zu erwarten.

Um dennoch einen möglichen Einfluss der Bildqualität auf die Befundung abschätzen zu können, wurde über eine Skala mit den Abstufungen „schlecht“ „ausreichend“ und „gut“ eine einfache subjektive Einschätzung der Bildqualität erfasst.

2.3.1.1.2 Aufnahmedauer & Aufnahmesituation

Die Aufnahmedauer konnte gemäß dem subjektiven Empfinden des Untersuchers als „zu kurz“, „ausreichend“ und „gut“ bewertet werden. Für die Aufnahmesituation wurde ebenfalls der subjektive Eindruck des Befunders als „schlecht“ oder „gut“ erfragt.

2.3.1.2 Hauptdiagnose

Die Hauptdiagnose wurde zunächst als Freitext erhoben und im Rahmen der Auswertung nach Vergleich mit dem Befund des Goldstandards in die Kategorien „keine Angabe“, „andere Diagnose“, „ähnliche Diagnose“, „gleiche bzw. synonyme Diagnose“ klassifiziert.

2.3.1.3 Stenosen

Für Stenosen wurden die drei Befundqualitäten Stenosegrad, Stenoselokalisation und Stenoseform erfasst. Die Auswertung des Stenosegrades für alle Abschnitte des Atemwegs erfolgte nach der Myer-Cotton Klassifikation. Der maximale Stenosegrad wurde zunächst als freie Prozentangabe erhoben und anschließend gemäß der Myer-Cotton-Klassifikation in eine der 4 Klassen verschlüsselt. Die Stenoselokalisation wird im Fragebogen für den Larynx, die Trachea, die Haupt- und Lappenbronchien getrennt angegeben. Für den Larynx wurde die Stenoselokalisation in supraglottisch, glottisch und subglottisch eingeteilt, die Lokalisation in den Hauptbronchien in rechte oder linke Lage. Die Lappenbronchien wurden in rechten und linken Oberlappen, Mittellappen und Lingula, sowie in Unterlappen eingeteilt. Die Merkmale der Stenoseform konnten als kurzstreckig, langstreckig, membranös und ringförmig charakterisiert werden.

2.3.1.4 Schleimhaut

In Anlehnung an den Bronchitis-Index (BI) (Thompson u. a., 1993) wurde die Schleimhaut nach den drei Kriterien

- Schwellung,
- Hyperämie und
- Hypersekretion

beurteilt. Die Verletzlichkeit der Schleimhaut wurde wegen der geringen prognostischen Kraft einerseits und der fehlenden Prüfung dieses Kriteriums in vielen der vorgelegten Videos andererseits nicht erhoben. Im multiple-choice-Verfahren wurde nach Anwesenheit bzw. Abwesenheit dieser Kriterien gefragt. Dabei musste auch das Fehlen eines Kriteriums aktiv durch setzen eines Kreuzes angegeben werden. Auf diese Weise sollte sichergestellt werden, dass auch tatsächlich eine Bewertung stattgefunden hat und die Beurteilung nicht einfach übergangen wurde. Wie nicht anders zu erwarten, entstand trotzdem ein erheblicher Anteil von Fehlwerten durch nicht abgegebene Urteile. In die eigentliche Auswertung wurden jedoch nur die Datensätze mit einer aktiven Bewertung (vorhanden/nicht vorhanden) einbezogen. Fehlwerte wurden dabei als eine eigene Kategorie behandelt. Zum Vergleich wurden die Daten auch unter der Prämisse ausgewertet, dass es sich bei nicht abgegebenen Befunden um unauffällige Befunde handelt.

2.3.1.5 Entzündung und Entzündungslokalisation

Entzündungen wurden mit dem binären Klassifikator „ja“ / „nein“ erfasst. Falls der Untersucher eine entzündete Schleimhaut erkannt hat, sollte hier die Lokalisation erfasst werden. Als mögliche Angaben standen der Stenosebereich, der Larynx, der Bronchus und die generalisierte Lokalisation zur Verfügung. Der pauschale Befund „Entzündung“ wurde im Rahmen der Auswertung mit den Schleimhautbefunden korreliert um das Konzept des Bronchitis-Index als Schleimhautsyndrom der Entzündung zu evaluieren.

2.3.1.6 Malazie, Pulsationen, Kompressionen

Malazie, Pulsationen und Kompressionen wurden als Sonderformen von Stenosen aufgefasst. Wie bei Entzündungen wurde die anatomische Lokalisation auf ein grobes Schema der anatomischen Abschnitte (Larynx, Trachea, Bronchus) sowie die Assoziation zu Stenosen beschränkt.

2.3.2 Arztfragebogen

Mit dem Untersucherfragebogen wurde potentielle Einflussgrößen der Befundqualität erfasst. Die erhobenen Daten lassen sich in Angaben

- zur **Person**
- zur **Ausbildung**
- zur **Erfahrung**

gliedern und wurden vorwiegend im multiple-choice Format – teils auch durch Freitextfelder für Zahlenangaben und Anmerkungen – erhoben.

2.4 Daten

Die Fragebögen wurden von Hand in eine Tabellenkalkulation übertragen und dabei numerisch codiert. Aus der Tabellenkalkulation wurden die Daten via Textdatei (csv-Format; engl. comma separated value) in eine SQL Datenbank eingepflegt die als zentraler Datenspeicher für alle weiteren Analysen diente. Über die Schnittstelle RMySQL wurden die Daten in die Arbeitsumgebung der Statistiksprache R übernommen.

2.4.1 Virtuelle Variablen

Die real im Fragebogen erhobenen Variablen wurden um virtuelle Variablen ergänzt, die nach Rechenvorschriften aus den realen Variablen erzeugt wurden. So wurde z. B. die Gesamterfahrung in Bronchoskopie als Gesamtzahl aller während Ausbildung und klinischer Tätigkeit absolvierten Bronchoskopien definiert.

2.4.2 Rekodierung

Im vereinfachten Lokalisationsschema, das bei Entzündungen und den speziellen Stenosen Malazie, Kompressionen und Pulsationen zur Auswahl stand, war die Kategorie „generalisiert“ enthalten. Diese Kategorie wurde vor der Auswertung rekodiert: wurde „gesamt“ gewählt, wurden die übrigen Kategorien positiv gesetzt. Die Kategorie „gesamt“ selbst wurde entfernt.

2.4.3 Fehlwerte

Wie nicht anders zu erwarten, wurden die Fragebögen nicht immer vollständig beantwortet, sodass Fehlwerte entstanden. Da diese Fehlwerte einige statistische Verfahren unmöglich machen, wurden sie in einem zweistufigen Verfahren via informierter Schätzung (engl. informed guess) und Imputation mittels random forests zu einem lückenlosen Datensatz komplettiert. Unter „informed guess“ wird die Ergänzung von Fehlwerten durch Logik anhand der übrigen vorhandenen Informationen bzw. durch plausible Annahmen verstanden. Imputation mit random forests schätzt Fehlwerte auf Basis der „Verwandtheit“ bzw. Ähnlichkeit von Variablen (engl. proximity).

2.4.4 Datenabdeckung

Bedingt durch Rekodierung, Fehlwerte und fehlende Überschneidung der gewählten Befundklassen zwischen Untersuchern und dem Goldstandard konnten Teile des Datensatzes in manche Berechnungen nicht einbezogen werden. Der Anteil der verwendbaren Daten am Gesamtdatensatz wurde jeweils als sogenannte „Datenabdeckung“ angegeben.

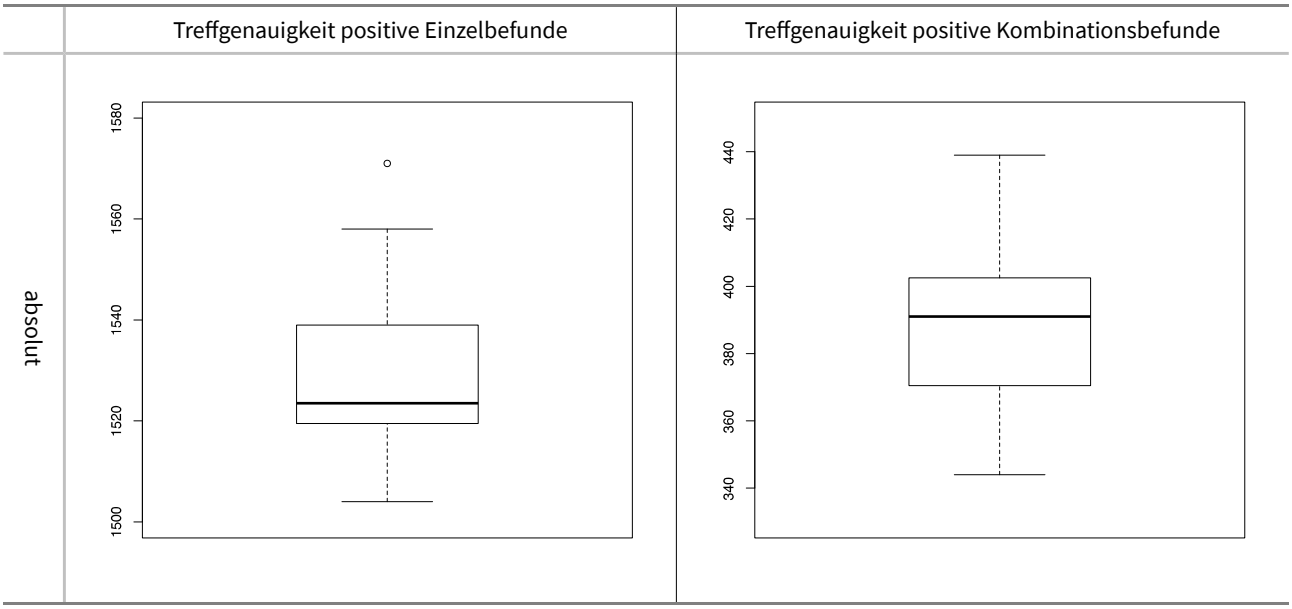
2.4.5 Definition von Zielvariablen

Die Befundrichtigkeit, also die Übereinstimmung mit dem Goldstandard (Genauigkeit; engl. accuracy), ist eine numerisch gut erfassbares Maßzahl der Befundqualität, in die sowohl die Übereinstimmung in negativen Befunden (Normalbefunde; Spezifität), als auch die Übereinstimmung in positiven Befunden (pathologischen Befunde; Sensitivität), eingeht. Die Richtigkeit wurde deshalb als Zielvariable ausgewählt um Faktoren in Ausbildung und Erfahrung zu identifizieren, die möglicherweise Einfluss auf die Befundqualität nehmen.

Die Befundrichtigkeit wurde auf Ebene der Einzelbefunde („Symptome“) sowie auf Ebene von Kombinationen mehrerer inhaltlich zusammengehöriger Befunde („Syndrome“) berechnet. Berechnungen mit der Befundrichtigkeit von Einzelbefunden als Zielvariable werden im Folgenden als Analyse „auf Symptomebene“ mit der Befundrichtigkeit von Befundkombinationen als Zielvariable analog hierzu als Analyse „auf Syndromebene“ bezeichnet. Auf Symptomebene sind bei

vollständiger Übereinstimmung mit dem Goldstandard maximal 1806 positive Treffer möglich auf Syndromebene maximal 630. Die von den Befundern erreichten Treffer liegen auf Symptomebene zwischen 1504 (83 %) und 1571 (87 %) auf Syndromebene zwischen 344 (55 %) und 439 (70 %). Der Median liegt auf Symptomebene bei 1524 Treffern (84 %) auf Syndromebene bei 391 Treffern (62 %).

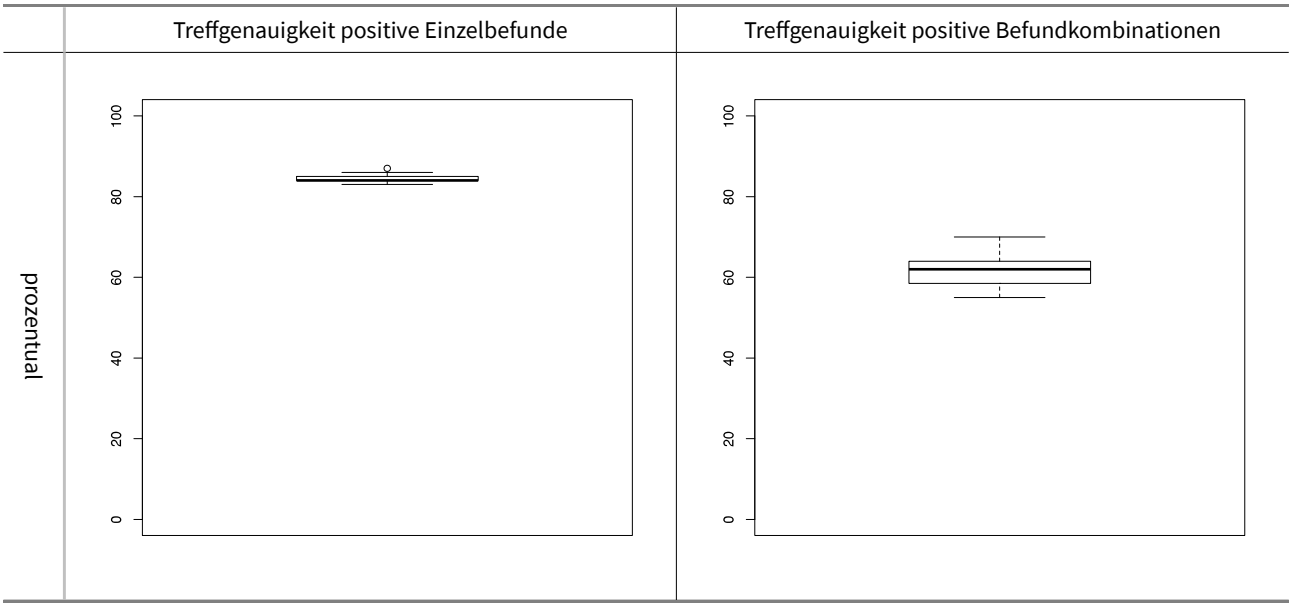
Abbildung 2.43: Absolute Übereinstimmung mit dem Goldstandard



Die Boxplots der Absolutwerte zeigen jeweils nur den Ausschnitt der beobachteten Werte.

Wie die Boxplots der prozentualen Treffgenauigkeit in Abbildung 2.44 illustrieren, liegen – besonders auf Ebene der Einzelbefunde – nur verhältnismäßig geringfügige Unterschiede der Befundgenauigkeit vor. Die prozentuale Spannweite zwischen dem Schlechtesten und dem Besten Befunder liegt auf Ebene der Einzelbefunde bei nur 4 %[19] auf Ebene von Kombinationsbefunden bei immerhin 15 %[20] der jeweils maximal möglichen Trefferzahl.

Abbildung 2.44: Prozentuale Übereinstimmung mit dem Goldstandard



Einen Eindruck der Lage auf der Gesamtskala vermitteln die Boxplots der prozentualen Werte.

[19] *Berechnung aus absoluten Treffern: (1571-1504)/1806 = 3,71 %*

[20] *Berechnung aus absoluten Treffern: (439-344)/630 = 15,08 %*

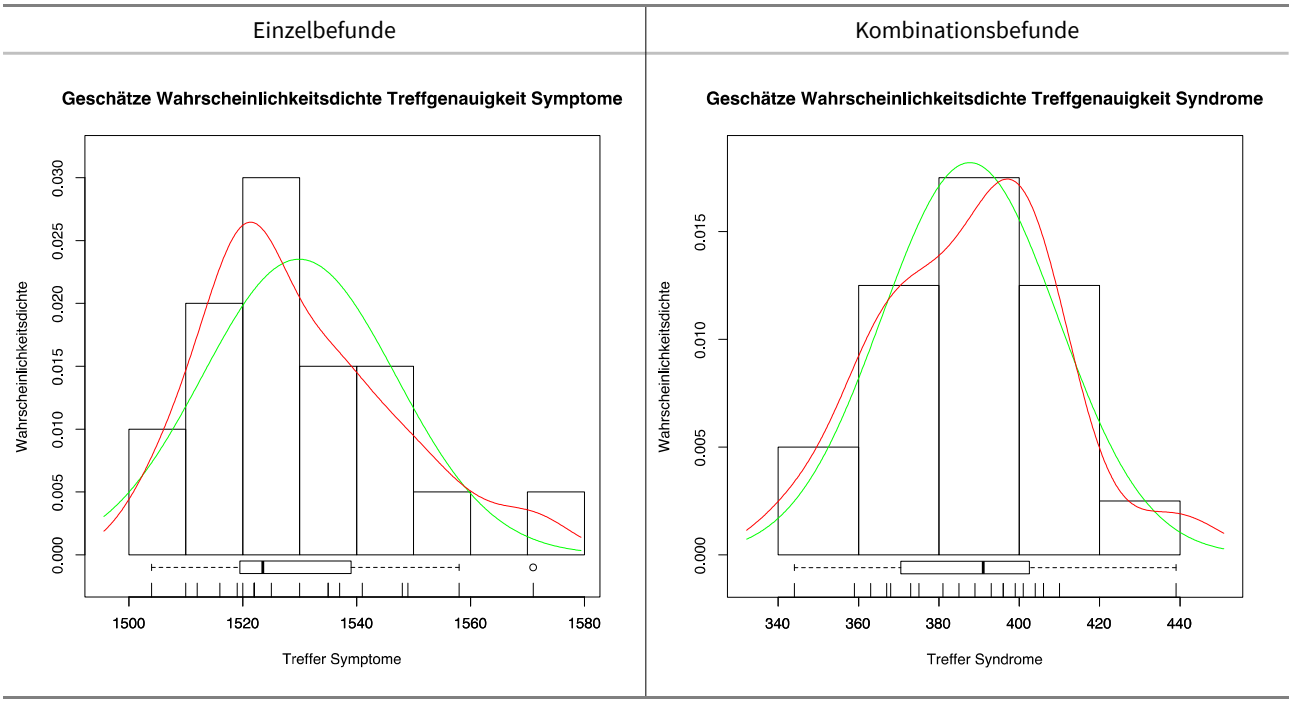
Das beruht auf Ebene der Einzelbefunde darauf, dass 1478 (83,7 %) negative Befunde 286 (16,3 %) positiven Befunden gegenüberstehen. Die Prävalenz der negativen Befunde übersteigt also mehr als das fünffache die der positiven Befunde. Unterschiede bei den wenigen positiven Befunden werden durch die überwiegende Mehrheit negativer Befunde überdeckt.

Dieser Effekt lässt sich bei Betrachtung von Kombinationsbefunden, bei denen 295 (52 %) Normalbefunde 293 (48 %) positiven Befunden gegenüberstehen erheblich abschwächen. Hier ist das Verhältnis zwischen unauffälligen Befunden und pathologischen Befunden nahezu perfekt ausgeglichen, sodass die Unterschiede in der Befundprävalenz als störender Einflussfaktor ausgeschaltet werden. Dadurch treten die Unterschiede in der Befundrichtigkeit deutlicher hervor: Das Feld der insgesamt erreichten Treffer ist bei Kombinationsbefunden im Vergleich zu Einzelbefunden deutlich weiter aufgespreizt (siehe prozentuale Boxplots in Abbildung 2.44). Die prozentuale Distanz der richtigen Befunde zwischen dem schlechtesten und dem besten Befunder ist fast viermal so groß wie auf Syndromebene (15 % gegenüber 4 %).

Abbildung 2.45 zeigt die Verteilungen der Treffer auf Ebene von Einzel- bzw. Kombinationsbefunden in einer kombinierten Darstellung aus Streudiagramm, Histogramm, Boxplot, Kerndichteschätzung (rot) und geschätzter Normalverteilung (grün). Die Verteilung der Treffer der Einzelbefunde ist nach links verschoben, die rechte Flanke der Kurve läuft flacher aus, als die linke Flanke. Im Vergleich hierzu sind die Treffer der Befundkombinationen annähernd symmetrisch verteilt. Der rote Kurvenverlauf der richtigen Befundkombinationen folgt der grün dargestellten Normalverteilung enger als derjenige der richtigen Einzelbefunde. Beide Verteilungen weisen bei hohen Trefferzahlen wegen eines Ausreißers einen Knick auf.

Gemäß dem optischen Eindruck kann zumindest hinsichtlich der richtigen Befundkombinationen als Zielvariable eine Normalverteilung angenommen werden, womit eine wichtige Voraussetzung für die Anwendung eines linearen Modells gegeben ist. Der eingehenden numerischen Analyse der Voraussetzungen für lineare Modelle unter Einbeziehung der unabhängigen Variablen ist Abschnitt 2.4.6.5 gewidmet.

Abbildung 2.45: Verteilungen der Treffgenauigkeit



Vergleich der Verteilungen von Einzelbefunden und Kombinationsbefunden.

2.4.6 Aufbereitung des Datensatzes

Für die Anwendung linearer Modelle sollten insbesondere folgende Voraussetzungen erfüllt sein:

- Normalität der Daten
- Unabhängigkeit der Variablen
- Linearität
- Varianzhomogenität (Homoskedastizität)

Darüber hinaus sollte auch Multikollinearität zwischen den erklärenden Variablen vermieden werden. In einem schrittweisen Verfahren wurde der Datensatz aufbereitet, um diesen Anforderungen der multiplen linearen Regression zu genügen.

2.4.6.1 Komplettierung durch Imputation

Im ersten Schritt wurden Fehlwerte beseitigt, da die derzeit verfügbaren Rechenverfahren der multiplen linearen Regression nur bedingt auf lückenhafte Datensätze anwendbar sind. Dabei wurde zunächst versucht Fehlwerte aus der in anderen Variablen enthaltenen Information zu ergänzen („informed guess“). Die verbleibenden Fehlwerte wurden mithilfe von random forests imputiert und anschließend auf Ihre Plausibilität überprüft.

2.4.6.2 Geringer Informationsgehalt

Die Variable „AusbildungFlexible“ enthält keine Information, da sämtliche Untersucher sie mit ja beantwortet haben. Die Frage nach der Qualifikation wurde – bei einem Fehlwert – durchgehend mit „Facharzt“ beantwortet, sodass sie sich ebenfalls nicht als Prädiktor eignet. Beide Variablen wurden deshalb aus dem Datensatz entfernt.

2.4.6.3 Dummy Variablen

In einem weiteren Schritt wurden Dummy-Variablen entfernt. Dabei handelt es sich um Variablen, die nachträglich hinzugefügt wurden, um Informationen anderer Variablen auf einem anderen Skalenniveau aufzubereiten. Die Variablen Hospitationen, Kursteilnahme, AusbildungStarre, AusbildungInterventionelle, ErfahrungFlexible, ErfahrungStarre, ErfahrungInterventionelle, enthalten keinerlei eigenständige Information, sondern geben lediglich als boolsche Variablen (Ja/nein) an, ob die zugehörigen numerischen Variablen Informationen enthalten.

2.4.6.3.1 Abgeleitete Variablen

Aus anderen Variablen aufsummierte Variablen wurden wegen der dadurch erzeugten Kollinearität mit den ursprünglichen Variablen entfernt. Dazu zählen die Variablen ErfahrungFlexibleGesamtAnzahl, ErfahrungStarreGesamtAnzahl und ErfahrungInterventionelleGesamtAnzahl.

2.4.6.4 Multikollinearität

Korrelieren zwei Variablen zu stark untereinander, können sie die Schätzung eines korrekten Regressionsmodells behindern. Eine stark korrelierenden Variablen muss daher vor der Schätzung aus dem Modell entfernt werden.

2.4.6.4.1 Variance Inflation Factor (VIF)

Die Multikollinearität zwischen den verbleibenden Variablen wurde mithilfe des Variance Inflation Factors (VIF) untersucht. Die Berechnung des VIF ist nur für numerische Variablen möglich, weswegen die kategorialen Variablen HospitationenKlinikart, AusbidungKlinikArt und ErfahrungKlinikart nicht mit einbezogen werden konnten. Der mit Abstand höchste VIF findet sich für die Anzahl und Tage der Hospitationen. Da die Dauer in Tagen den Einfluss von Hospitationen genauer repräsentiert, als deren Anzahl, wurde die Variable HospitationenTage gestrichen.

Tabelle 2.12: Variance Inflation Factor - 1. Durchgang

Variables	VIF
Alter	1.270314
HospitationenAnzahl	25.400842
HospitationenTage	43.589964
KursteilnahmeAnzahl	1.317422
KursteilnahmeTage	3.945810
AusbildungJahre	1.297552
AusbildungFlexibleAnzahl	1.760727
AusbildungStarreAnzahl	1.400745
AusbildungInterventionelleAnzahl	8.780942
ErfahrungFlexibleAnzahl	2.158548
ErfahrungStarreAnzahl	3.621099
ErfahrungInterventionelleAnzahl	4.052032

Prüfung der Multikollinearität anhand des Variance Inflation Factor (VIF)

Nach erneuter Berechnung des VIF ergibt sich ein homogeneres Bild. Kein Wert überschreitet den üblicherweise angesetzten Grenzwert eines VIF von 10. Legt man einen strengeren Schwellenwert von 5 bzw. sqrt(VIF)>2 als Maßstab an, ist die Anzahl der interventionellen Bronchoskopien während der Ausbildung als hinsichtlich der Kollinearität problematisch einzustufen. Im Sinne einer ausgewogenen Zusammenstellung von Einflussgrößen wurde diese grenzwertige Variable jedoch beibehalten. Anders als in Analogie zu Anzahl und Tagen der Hospitationen zu erwarten wäre, liegt bei Anzahl und Tagen der Kursteilnahme zumindest rein rechnerisch keine signifikante Kollinearität vor, obgleich sich hier immerhin der zweithöchste VIF-Wert findet. Um den Einfluss einzelner Größen nicht schon durch die Anzahl der repräsentierenden Variablen zu verzerren, sowie wegen der rein logisch zu erwartenden Kollinearität wurde abschließend auch die Variable KursteilnahmeAnzahl entfernt.

Tabelle 2.13: Variance Inflation Factor - 2. Durchgang

	Variables	VIF
1	Alter	1.255279
2	HospitationenTage	2.183166
3	KursteilnahmeAnzahl	1.293677
4	KursteilnahmeTage	3.345849
5	AusbildungJahre	1.295901
6	AusbildungFlexibleAnzahl	1.687336
7	AusbildungStarreAnzahl	1.307428
8	AusbildungInterventionelleAnzahl	4.995339
9	ErfahrungFlexibleAnzahl	1.981710
10	ErfahrungStarreAnzahl	2.343306
11	ErfahrungInterventionelleAnzahl	2.762599

Prüfung der Multikollinearität anhand des Variance Inflation Factor (VIF)

Im Gegensatz zu den vorherigen Schritten ergibt sich aus dem Entfernen der Variable „Anzahl der Kursteilnahmen“ keine wesentliche Homogenisierung der VIF-Werte. Zusammen mit den kategorialen Variablen der Klinikart der Hospitationen, der Ausbildung und der Erfahrung bilden die verbliebenden 10 Variablen aus Tabelle 2.14 den Datensatz für die Multiple lineare Regression.

Tabelle 2.14: Variance Inflation Factor – 3. Durchgang

	Variables	VIF
1	Alter	1.213290
2	HospitationenTage	2.169464
3	KursteilnahmeTage	3.313180
4	AusbildungJahre	1.290554
5	AusbildungFlexibleAnzahl	1.451064
6	AusbildungStarreAnzahl	1.300248
7	AusbildungInterventionelleAnzahl	4.819313
8	ErfahrungFlexibleAnzahl	1.959089
9	ErfahrungStarreAnzahl	2.314798
10	ErfahrungInterventionelleAnzahl	2.508496

Prüfung der Multikollinearität anhand des Variance Inflation Factor (VIF)

2.4.6.5 Testbatterie der Voraussetzungen linearer Modelle

Die Normalität der Zielvariablen wurde bereits in Abschnitt 2.4.5 (Seite 43) untersucht. Hier werden die erklärenden Variablen einer Testbatterie unterzogen (Peña, Slate, 2006), die neben den genannten Annahmen linearer Regression[21] auch auf Schiefe und Kurtosis (Wölbung) prüft.

Tabelle 2.15: Test auf Voraussetzungen lineares Modell Symptomebene

	Value	p-value	Decision
Global Stat	5.013496	0.28592	Assumptions acceptable.
Skewness	1.060134	0.30318	Assumptions acceptable.
Kurtosis	0.119507	0.72957	Assumptions acceptable.
Link Function	3.829910	0.05035	Assumptions acceptable.
Heteroscedasticity	0.003945	0.94992	Assumptions acceptable.

Prüfung der Normalität mithilfe einer Testbatterie.

Sowohl für richtige Einzelbefunde, als auch für richtige Kombinationsbefunde wurden die Voraussetzungen für die Anwendung linearer Modelle auf dem üblichen Signifikanzniveau von p=0,05 erfüllt. Datengrundlage war der via random forests komplettierte Datensatz.

Tabelle 2.16: Test auf Voraussetzungen lineares Model Syndromebene

	Value	p-value	Decision
Global Stat	2.352e+00	0.6713	Assumptions acceptable.
Skewness	1.758e-01	0.6750	Assumptions acceptable.
Kurtosis	3.631e-06	0.9985	Assumptions acceptable.
Link Function	1.913e+00	0.1667	Assumptions acceptable.
Heteroscedasticity	2.635e-01	0.6077	Assumptions acceptable.

Prüfung der Normalität mit Hilfe einer Testbatterie.

2.4.6.6 Variablenauswahl in den Modellen

Tabelle gibt einen Überblick über die Verwendung der Variablen des Befundfragebogens in den Modellen des Ergebnisteils. Lineare Modelle wurden mit „lm Auswahl“ berechnet.

[21] *Linearität, Homoskedastizität (Varianzhomogenität), Unabhängigkeit und Normalität*

Tabelle 2.17: Modelle mit den jeweils untersuchten Variablen

Variable	lm	lm vif	lm Auswahl	tree
Alter	+	+	+	+
Qualifikation	kaum Information			+
Hospitationen	dummy			
HospitationenKlinikart	+	+	+	+
HospitationenAnzahl	+	vif		+
HospitationenTage	+	+	+	+
Kursteilnahme	dummy			
KursteilnahmeAnzahl	+	+	gestrichen	+
KursteilnahmeTage	+	+	+	+
AusbildungJahre	+	+	+	+
AusbildungKlinikart	+	+	+	+
AusbildungFlexible	keine Information (alle Befunder gleich)			
AusbildungFlexibleAnzahl	+	+	+	+
AusbildungStarre	dummy			
AusbildungStarreAnzahl	+	+	+	+
AusbildungInterventionelle	dummy			
AusbildungInterventionelleAnzahl	+	vif		+
ErfahrungKlinikart	+	+	+	+
ErfahrungFlexible	dummy			
ErfahrungFlexibleAnzahl	+	+	+	+
ErfahrungFlexibleGesamtAnzahl	konstruiert			+
ErfahrungStarre	dummy			
ErfahrungStarreAnzahl	+	+	+	+
ErfahrungStarreGesamtAnzahl	konstruiert			+
ErfahrungInterventionelle	dummy			
ErfahrungInterventionelleAnzahl	+	+	+	+
ErfahrungInterventionelleGesamtAnzahl	konstruiert			+
Anzahl im Modell berücksichtigter Variablen	15	13	12	19

Übersicht der für verschiedene Modelle gewählten Variablen.

Literatur

Analytics, Revolution; Weston, Steve (2015): *foreach: Provides Foreach Looping Construct for R*. o.V.

Barbier, Paolo; Alimento, Marina; Berna, Giovanni; u. a. (2007): „High-grade Video Compression of Echocardiographic Studies: A Multicenter Validation Study of Selected Motion Pictures Expert Groups (MPEG)-4 Algorithms“. In: *Journal of the American Society of Echocardiography*. 20 (5), S. 527–536, DOI: 10.1016/j.echo.2006.10.006.

Becker, Richard A.; Chambers, John M. (1984): *S: an interactive environment for data analysis and graphics*. Belmont, Calif: Wadsworth Advanced Book Program (The Wadsworth statistics/probability series). — ISBN: 978-0-534-03313-2

Brasil, Pedro (2010): *DiagnosisMed: Diagnostic test accuracy evaluation for medical professionals.* o.V.

Friendly, Michael (2016): *vcdExtra: „vcd“ Extensions and Additions*. o.V.

Gamer, Matthias; Lemon, Jim; <puspendra.pusp22@gmail.com>, Ian Fellows Puspendra Singh (2012): *irr: Various Coefficients of Interrater Reliability and Agreement*. o.V.

Groemping, Ulrike (2006): „Relative Importance for Linear Regression in R: The Package relaimpo“. In: *Journal of Statistical Software*. 17 (1), S. 1–27.

Grothendieck, G. (2014): *sqldf: Perform SQL Selects on R Data Frames*. o.V.

Günther, Tobias (2010): *Tower*. Stuttgart: fournova.

Gwet, Kilem (2001): *Statistical Tables for Inter-Rater Reliability Testing*. STATAXIS Publishing Company. — ISBN: 0-9708062-1-3

Hunt, Paul (2015): „adobe/source-sans-pro“. *GitHub*. Abgerufen am 09.08.2015 von https://github.com/adobe/source-sans-pro.

Kuhn, Max (2008): „Building Predictive Models in R Using the caret Package“. In: *Journal of Statistical Software*. 28 (5), S. 1–26.

Kuhn, Max; Coulter, Nathan; Lenon, Patrick; u. a. (2014): *odfWeave: Sweave processing of Open Document Format (ODF) files*. o.V.

Liaw, Andy; Wiener, Matthew (2002): „Classification and Regression by randomForest“. In: *R News*. 2 (3), S. 18–22.

Mackinnon, A (2000): „A spreadsheet for the calculation of comprehensive statistics for the assessment of diagnostic tests and inter-rater agreement“. In: *Computers in Biology and Medicine*. 30 (3), S. 127–134.

Meyer, David; Zeileis, Achim; Hornik, Kurt (2006): „The Strucplot Framework: Visualizing Multi-way Contingency Tables with vcd“. In: *Journal of Statistical Software*. 17 (3), S. 1–48.

Milborrow, Stephen (2016): *plotmo: Plot a Model's Response and Residuals*. o.V.

Nardini, Christine; Liu, Yuanhua (2014): *Multiclasstesting: Performance of N-ary classification testing*. o.V.

o. A. (o. J.): „Sweave and the Open Document Format – The odfWeave Package“. In: *RNews*.

Ooms, Jeroen; James, David; DebRoy, Saikat; u. a. (2016): *RMySQL: Database Interface and „MySQL“ Driver for R*. o.V.

Peña, Edsel A.; Slate, Elizabeth H. (2006): „Global Validation of Linear Model Assumptions“. In: *Journal of the American Statistical Association*. 101 (473), S. 341, DOI: 10.1198/016214505000000637.

Pena, Edsel A.; Slate, Elizabeth H. (2014): *gvlma: Global Validation of Linear Models Assumptions*. o.V.

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

Rossini A.J.[1]; Heiberger R.M.[2]; Sparapani R.A.[3]; u. a. (2004): „Emacs Speaks Statistics: A Multiplatform, Multipackage Development Environment for Statistical Analysis“. In: *Journal of Computational & Graphical Statistics*. 13 , S. 247–261, DOI: doi:10.1198/1061860042985.

RStudio Team (2015): *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc.

Sagano, Sora (2015): „ドットコロン - フォント: Vegur“. Abgerufen am 05.10.2012 von http://www.dotcolon.net/font/?id=vegur.

Saparoschez, Dimitri (2011): *GitLab*. Utrecht.

Seidenari, Stefania; Pellacani, Giovanni; Righi, Elena; u. a. (2004): „Is JPEG compression of videomicroscopic images compatible with telediagnosis? Comparison between diagnostic performance and pattern recognition on uncompressed TIFF images and JPEG compressed ones“. In: *Telemedicine journal and e-health: the official journal of the American Telemedicine Association*. 10 (3), S. 294–303, DOI: 10.1089/tmj.2004.10.294.

Sing, Tobias; Sander, Oliver; Beerenwinkel, Niko; u. a. (2005): „ROCR: visualizing classifier performance in R“. In: *Bioinformatics*. 21 (20), S. 3940–3941, DOI: 10.1093/bioinformatics/bti623.

Spencer, K; Solomon, L; Mor-Avi, V; u. a. (2000): „Effects of MPEG compression on the quality and diagnostic accuracy of digital echocardiography studies“. In: *Journal of the American Society of Echocardiography: Official Publication of the American Society of Echocardiography*. 13 (1), S. 51–7.

Stahel, Werner (2013): *The R-Function regr and Package regr0 for an Augmented Regression Analysis*. ETH Zürich.

Stallman, Richard (2015): *Emacs*. Boston: Free Software Foundation (FSF).

Stallman, Richard M. (1981): „EMACS the Extensible, Customizable Self-documenting Display Editor“. In: *Proceedings of the ACM SIGPLAN SIGOA Symposium on Text Manipulation*. New York, NY, USA: ACM, S. 147–156, DOI: 10.1145/800209.806466. — ISBN: 978-0-89791-050-7

Therneau, Terry; Atkinson, Beth; Ripley, Brian (2015): *rpart: Recursive Partitioning and Regression Trees*. o.V.

Thompson, Austin B.; Huerta, Guillermo; Robbins, Richard A.; u. a. (1993): „The Bronchitis Index: A Semiquantitative Visual Scale for the Assessment of Airways Inflammation“. In: *Chest*. 103 (5), S. 1482–1488, DOI: 10.1378/chest.103.5.1482.

Torvalds, Linus (2005): *Git*. o.V.

Wickham; Hadley (2007): „Reshaping data with the reshape package“. In: *Journal of Statistical Software*. 21 (12).

Wickham, Hadley (2016): *stringr: Simple, Consistent Wrappers for Common String Operations*. o.V.

3 Methoden

KAPITELVERZEICHNIS

Grundlage der hier zusammengestellten Methoden ist eine umfangreiche Literaturrecherche. Für allgemeine Statistik wurden Nachschlagewerke und einführende Lehrbücher für Mediziner konsultiert (Bortz u. a., 2008; Bortz, Lienert, 2008; Bortz, Weber, 2005; Hilgers u. a., 2007; Sachs, Hedderich, 2006; Weiß, 2013). Zur Inter-Beobachter-Variabilität wurden Monographien (Gwet, 2014; Wirtz, Caspar, 2002), Übersichtsarbeiten in der Form von Fachartikeln (unter anderem Chmura Kraemer u. a., 2002; Feuerman, Miller, 2008; Hallgren, 2012; Krummenauer, 2003; Viera, Garrett, 2005; Wirtz, Kutschmann, 2007) und Internetseiten (Uebersax, 2010) herangezogen. Für die medizinische Testtheorie bzw. die Analyse kategorialer Daten wurde auf bekannte Referenzwerke zurückgegriffen (Agresti, 2002; Andersen, 1997; Everitt, 1986; Fienberg, 2007; Fleiss, 2003; Friendly, 2000; Kateri, 2014; Kraemer, 1992; Rudas, 1998; Zhou, 2011). Bei der Umsetzung der Berechnungen in der Statistiksprache R wurden wertvolle Hinweise aus der umfangreichen, frei verfügbaren Dokumentation (Handl, Kuhlenkasper, 2016; R-Project, 2016), aus kommerziellen Lehrbüchern (Kabacoff, 2011; Spector, 2008), von Webseiten (Kabacoff, 2014) und von der R-Mailingliste bezogen. Ausgangspunkt der Berechnungen mit Algorithmen des maschinellen Lernens waren anwendungsorientierte Lehrbücher zum Thema (Hastie u. a., 2008; James u. a., 2013; Lantz, 2013).

Die Analyse der Inter-Beobachter-Variabilität wurde unter drei Gesichtspunkten durchgeführt:

- **Beschreibung der Verteilung** der erhobenen Befunde (**Deskription**)
- **relativer Vergleich** der Untersucher **untereinander** (**Präzision**)
- **absoluter Vergleich** der Untersucher **mit der Referenz** (**Richtigkeit**).

Die Dreiteilung in **Deskription**, **Präzision** und **Richtigkeit** gliedert die Darstellung der Befundübereinstimmung in dieser Studie. Jeder dieser Aspekte wird durch Kenngrößen beschrieben, die in diesem Kapitel vorgestellt werden. Die Begriffe der Testtheorie werden teilweise variabel gebraucht. In Fachgebieten wie Statistik, Epidemiologie, Labormedizin und Maschinenlernen etablierten sich jeweils eigene Terminologien. Im Interesse einer möglichst konsistenten Begrifflichkeit orientiert sich diese Studie an der dritten Ausgabe des international vocabulary of metrology (VIM) der Working Group 2 des Joint Committee for Guides in Metrology (JCGM/WG 2) (Brinkmann, Deutsches Institut für Normung, 2012; Joint Committee for Guides in Metrology (JCGM), 2012), sowie an ausgewählten Monographien, insbesondere denen von Everitt (1986) und Kraemer (1992). Unter dem Überbegriff **Deskription** werden Häufigkeitsverteilungen von Befunden dargestellt. Tabellarische Aufstellungen zeigen, wie oft ein Befund erhoben wurde, von wie vielen verschiedenen Befundern und in wie vielen verschiedenen Videomitschnitten. Dabei wird die Gruppe der Befunder dem Goldstandard gegenüber gestellt. Die **Präzision** untersucht die Übereinstimmung der Befunder untereinander. Als führende Maßzahlen hierfür wurden die durchschnittliche positive Übereinstimmung und das Kappa nach Fleiss gewählt. Die **Richtigkeit** vergleicht die Befunde der Befunder mit denen des Goldstandards. Die wichtigsten Maßzahlen der Richtigkeit sind die Sensitivität (Wahrscheinlichkeit einen Kranken als krank zu erkennen), der positive prädiktive Wert (Wahrscheinlichkeit von Krankheit bei positivem Befund) und das Kappa nach Cohen, eine für zufällige Übereinstimmung korrigierte Genauigkeit.

Tabelle 3.1: Allgemeine Tabellenstruktur im Ergebnisteil

	Deskription des Antwortverhaltens von Referenz und Befundern	**Präzision** Übereinstimmung der Befunder untereinander	**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz
Vergleich	ohne Bezug	relativ	absolut
Gruppe	Befunder & Referenz	Befunder untereinander	Befunder versus Referenz
Kenngrößen (Beispiele)	• Häufigkeiten • Verteilungen	• positive Übereinstimmung • Kappa Fleiss	• Sensitivität • positiver prädiktiver Wert • Kappa Cohen

Die Tabellen des Ergebnisteils sind nach den Kriterien Deskription, Präzision und Richtigkeit gegliedert.

3.1 Definitionen aus der Epidemiologie

Die in der Studie verwendeten Maßzahlen basieren auf den epidemiologischen Begriffen Grundgesamtheit, Stichprobe und Prävalenz, die im Folgenden kurz eingeführt werden.

3.1.1 Grundgesamtheit

Die Grundgesamtheit (kurz **p** nach engl. population) ist die Gesamtheit aller Individuen, über die eine Aussage getroffen werden soll. In dieser Studie sind das alle pädiatrischen Patienten, die einer Bronchoskopie zugeführt werden, bzw. alle Ärzte, die pädiatrische Bronchoskopien durchführen.

3.1.2 Stichprobe

Da die Grundgesamtheit selbst meist nicht untersucht werden kann, wird sie durch eine Stichprobe (kurz **s** nach engl. sample) geschätzt. Die Schätzung der Grundgesamtheit durch die Stichprobe fällt um so genauer aus, je größer der Stichprobenumfang ist. Diese Studie repräsentiert die Grundgesamtheit pädiatrischer Bronchoskopien mit einer Stichprobe von 42 Videomitschnitten, die so zusammengestellt ist, dass sie das Spektrum bronchoskopischer Befunde in der Pädiatrie weitgehend abdeckt. Die Grundgesamtheit der behandelnden Ärzte wird durch 20 in der Pädiatrie tätige Ärzte aus der Arbeitsgruppe Bronchoskopie der Gesellschaft für pädiatrische Pulmonologie (GPP) vertreten.

3.1.3 Prävalenz

Unter Prävalenz versteht man den Anteil der Kranken an der Grundgesamtheit. Da die Grundgesamtheit selbst meist nicht untersucht werden kann, wird sie in der Praxis über den Anteil der Kranken in einer (möglichst repräsentativen) Stichprobe geschätzt. Aus der Häufigkeitsverteilung in der Stichprobe wird die Wahrscheinlichkeit, an einer Krankheit zu leiden, abgeleitet.

Formel 3.1: Prävalenz

englisch: prevalence Abkürzung: pre

Definition: Anteil der Kranken an der Grundgesamtheit.

Formel:
$$Prävalenz = \frac{Kranke(a)}{Stichprobe(s)} = \frac{tp + fn}{tn + fn + fp + tp}$$

Literatur: (Hedderich, Sachs, 2016: S. 191; Hilgers u. a., 2007: S. 82)

Zu den Bezeichnungen der Kategorien der Vier-Felder-Tafel (tp, tn, fn, fp) siehe Kapitel 3.2.6 Seite 58.

Die Epidemiologie unterscheidet genauer zwischen der Prävalenz zu einem bestimmten Zeitpunkt (Punktprävalenz) und der in einem bestimmten Zeitraum (Intervallprävalenz). Auch der Kontext, in dem die Prävalenz erhoben wird, ist von Bedeutung. Die Prävalenz von Kopfschmerzen in der Allgemeinbevölkerung unterscheidet sich erheblich von der in einer Hausarztpraxis, die wiederum verschieden zu der in einer Notaufnahme ist. Auch geographische (z. B. föhnbedingter Kopfschmerz am Alpenrand), demographische und ethnische Faktoren müssen berücksichtigt werden. Die Prävalenz in einer Stichprobe muss daher möglichst ähnlich zu der in der Grundgesamtheit sein, über die eine Aussage getroffen werden soll. Die Prävalenz zählt zu den testunabhängigen Eingangsgrößen. Sie heißt deshalb auch „a priori Wahrscheinlichkeit“ bzw. Prätestwahrscheinlichkeit. Die Prävalenz ist eine zentrale Größe, die nahezu alle Kennzahlen medizinischer Tests entscheidend beeinflusst.

3.2 Definitionen aus der Statistik

Neben den im letzten Abschnitt eingeführten epidemiologischen Grundbegriffen sind im Kontext medizinischer Tests auch die folgenden Fachtermini aus der Statistik relevant.

3.2.1 Empirische Wahrscheinlichkeit

Bis heute existieren unterschiedliche Auffassungen darüber, was unter Wahrscheinlichkeit zu verstehen ist. Für die Anwendung in der Medizin ist es üblich, sich Wahrscheinlichkeit als Maß für die Gewissheit unterschiedlicher Ausgänge komplexer bzw. zufälliger Prozesse vorzustellen. Die Wahrscheinlichkeit selbst ist nicht direkt beobachtbar oder messbar, wohl aber die aus ihr resultierenden Häufigkeiten von Ereignissen. Die Häufigkeit von Ereignissen erlaubt Rückschlüsse auf die ihnen zugrunde liegende Wahrscheinlichkeit. Und zwar um so genauer, je mehr Ereignisse beobachtet werden (Gesetz der großen Zahl). In diesem Sinne werden die relativen Häufigkeiten innerhalb der untersuchten Stichprobe als Näherungen bzw. Schätzungen für die Wahrscheinlichkeiten in der Grundgesamtheit herangezogen.

3.2.2 Klassen, Klassifikation, Klassierung bzw. Klassifizierung

Die Ausprägungen diskreter Variablen bezeichnet man als Kategorien oder Klassen. Zum Beispiel kann sich die Variable Gesundheitszustand in die Kategorien gesund und krank ausprägen, die Variable Testergebnis in die Klassen negativ und positiv. Ein System kategorialer Variablen mit ihren dazugehörigen Klassen heißt Klassifikation. Die Einordnung eines Variablenwertes in eine Klasse nennt man Klassierung bzw. Klassifizierung. Ärztliche Diagnosen sind also eine Klassifizierung gemäß der Variable Krankheit. Während es für Variablen wie Bluthochdruck oder Laborwerte Grenzwerte gibt, anhand denen die Klassifizierung erfolgt, basiert die Klassifizierung klinischer Untersuchungsbefunde auf letztlich rein subjektiven Kriterien. Die Klassifizierung unterliegt daher unvermeidlich einer gewissen Variabilität. Die Variabilität bei der subjektiven Klassifizierung von Beobachtungen in der pädiatrischen Bronchoskopie ist Gegenstand dieser Studie.

Medizinische diagnostische Tests lassen sich durch die Variable Gesundheitszustand (kurz **c** nach engl. condition) mit den Klassen gesund (kurz **u** nach engl. unaffected) bzw. krank (kurz **a** nach engl. affected) und die Variable Testergebnis (kurz **r** nach engl. result) mit den Klassen positiv (kurz **p** nach engl. positive) bzw. negativ (kurz **n** nach engl. negative) beschreiben. In diesem einfachsten Fall von zwei Variablen mit jeweils 2 Klassen, kann die Beziehung zwischen den beiden Variablen und ihren Klassen in einer sogenannten Vier-Felder-Tafel dargestellt werden. Die Vier-Felder-Tafel zeigt die gemeinsame Verteilung zweier Variablen. Im Falle medizinischer Tests ist das die gemeinsame Verteilung zwischen Gesundheitszustand und Testergebnis.

Mehrdimensionale Klassifikationen können in binäre Klassifikationen zerlegt werden. Die Variable „Stenoselokalisation Larynx“ des Befundbogens setzt sich z. B. aus den Klassen „supraglottisch“, „glottisch“ und „subglottisch“ zusammen. Anstatt die Variable „Stenoselokalisation Larynx“ mit ihren drei Ausprägungen zu betrachten, können die drei Ausprägungen auch als einzelne binäre Variablen mit den Klassen „vorhanden“ bzw. „nicht vorhanden“ angesehen werden. Im Rahmen dieser Studie werden binäre Variablen als Einzelbefunde bzw. „Symptome“ bezeichnet, aus mehreren Einzelbefunden resultierende mehrdimensionale Variablen als Befundkombination bzw. „Syndrome“.

3.2.3 Präzision

Die Präzision beschreibt die Stabilität einer Messung, sagt aber nichts über das Verhältnis der gemessenen Werte zum wahren Wert aus. Sie wird deshalb auch als "Wiederholgenauigkeit" be-

zeichnet. Sie misst die Reproduzierbarkeit eines Messwertes – oder anders formuliert – seine Streuung.

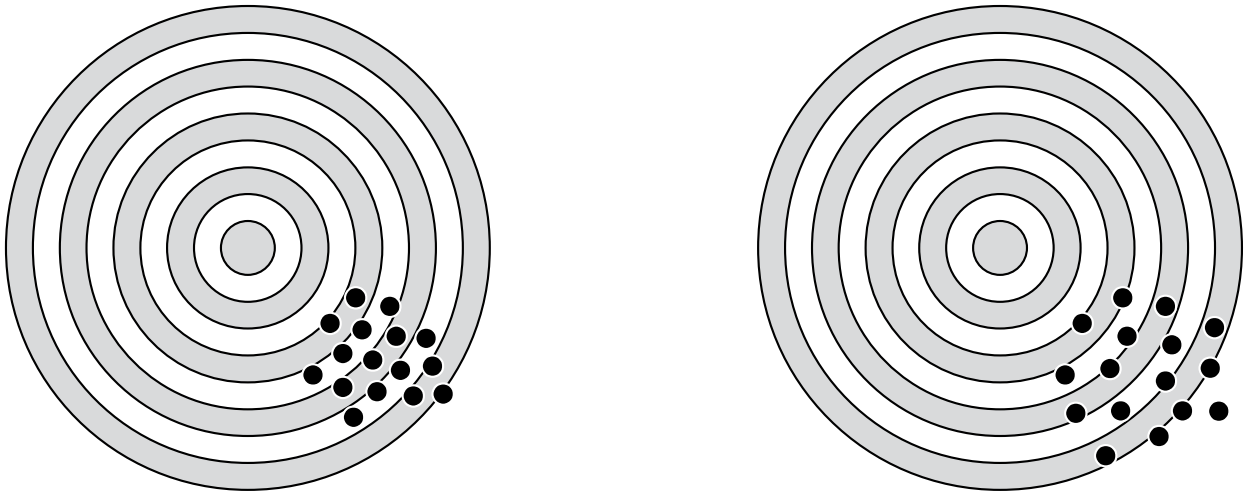
Definition 3.1: Präzision

englisch:	precision	Abkürzung:	pcn
Definition:	„Grad der Übereinstimmung zwischen […] Messgrößen, die durch wiederholte Messungen an den gleichen oder ähnlichen Gegenständen unter festgelegten Bedingungen erhalten wurden." (freie deutsche Übersetzung nach: Joint Committee for Guides in Metrology (JCGM), 2012 Abschnitt 2.15)		

Definition der Präzision gemäß dem internationalen Wörterbuch der Metrologie.

Eine klassische graphische Darstellung der Präzision und ihrer Beziehung zur Richtigkeit ist die sogenannte Zielscheibenanalogie (Abbildung 3.1). Die Präzision misst die Streuung der einzelnen Schüsse untereinander, sagt aber nichts über deren Lage zum Zentrum der Scheibe aus.

Abbildung 3.1: graphische Darstellung der Präzision



links: hohe Präzision geringe Richtigkeit. Rechts: geringe Präzision, geringe Richtigkeit.

Im Kontext der Inter-Beobachter-Variabilität ist die Präzision ein Maß der Übereinstimmung der Befunder untereinander und wird auch als „Konkordanz" bezeichnet. Dabei wird die Präzision im Einklang mit der Definition des International Vocabulary of Metrology (Joint Committee for Guides in Metrology (JCGM), 2012) nicht als Maßzahl, sondern als abstrakter Obergriff verstanden, der anhand diverser Kenngrößen diskutiert werden kann. In diese Studie wurden die durchschnittliche positive Übereinstimmung (Kapitel 3.3.2.1 Seite 65) und das Kappa nach Fleiss (Kapitel 3.3.2.2 Seite 66) als Kennwerte der Präzision herangezogen.

3.2.4 Richtigkeit

Die Richtigkeit vergleicht die Lage von Messwerten zu einer Referenz. Diese Referenz stellt in dieser Studie der Goldstandard dar, dessen Befunde mit dem wahren Befund gleichgesetzt werden. Wie die Zielscheibenanalogie (Abbildung 3.2) illustriert, ist für die Richtigkeit alleine die absolute Lage zu einem Referenzpunkt ausschlaggebend, die Streuung der Schüsse (Präzision) spielt hingegen keine Rolle. Analog zur Präzision, ist die Richtigkeit als abstrakte Größe aufzufassen, die sich durch mehrere Maßzahlen beschreiben lässt.

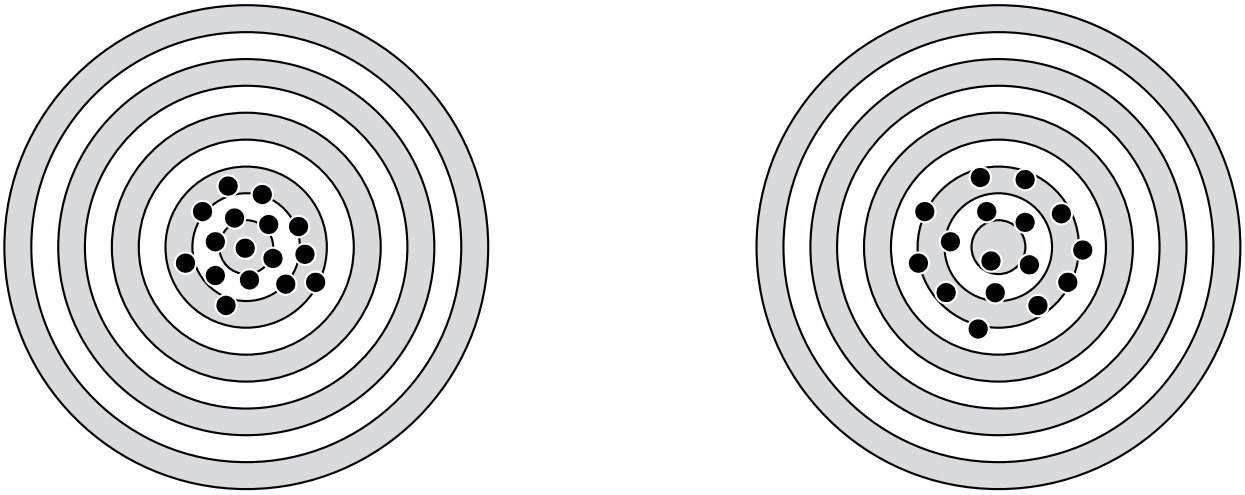
Definition 3.2: Richtigkeit

englisch:	accuracy	Abkürzung:	acc
Definition:	„Grad der Übereinstimmung zwischen einem gemessenem Wert und dem wahren Wert einer Messgröße." (freie deutsche Übersetzung nach: Joint Committee for Guides in Metrology (JCGM), 2012 Abschnitt 2.13)		

Definition der Richtigkeit gemäß dem internationalen Wörterbuch der Metrologie.

Als Kenngrößen der Richtigkeit werden in dieser Studie insbesondere die Sensitivität (Kapitel 3.3.3.3.1 Seite 69), der positive prädiktive Wert (Kapitel 3.3.3.3.3 Seite 71) und das Kappa nach Cohen (Kapitel 3.3.3.6 Seite 78) untersucht. Ergänzend werden weitere verbreitete Maßzahlen wie die Likelihood ratios mit der Area under Curve (AUC) und die Odds ratio diskutiert. Likelihood ratios und Odds ratio sind, im Gegensatz zu Sensitivität, positiv prädiktivem Wert und Kappa Cohen, unabhängig von der Prävalenz und daher besser zum Vergleich unterschiedlicher Befunde geeignet.

Abbildung 3.2: graphische Darstellung der Richtigkeit



Links: hohe Richtigkeit, hohe Präzision. Rechts: hohe Richtigkeit, geringere Präzision.

3.2.5 Goldstandard

Die Berechnung der Richtigkeit erfordert die Kenntnis der wahren Werte, mit dem die Vorhersagen des Testes bzw. die Beurteilungen der Untersucher überprüft werden können. Oft sind die wahren Werte jedoch nicht bekannt und müssen deshalb durch eine Referenz ersetzt werden, die dem wahren Wert möglichst nahe kommt. Eine solche Referenz bezeichnet man als Goldstandard. Im Rahmen dieser Studie wurde die Befundung des Leiters der Arbeitsgruppe pädiatrische Bronchoskopie der Gesellschaft für pädiatrische Pulmonologie als Goldstandard verwendet. Rationale des Goldstandards ist neben der großen Erfahrung in pädiatrischer Bronchoskopie die Kenntnis sämtlicher klinischer Hintergrundinformationen zu den Videomitschnitten der Bronchoskopien. Über den Verlauf wurden die in der jeweiligen Bronchoskopie gestellten Diagnosen bestätigt. Der Referenzbefund wurde nach mehrfacher Durchsicht der Mitschnitte gezielt im Hinblick auf die für die Diagnose relevanten Befunde erstellt.

3.2.6 Kontingenztafel

Der Begriff Kontingenztafel (engl. contingency table) wurde von Karl Pearson geprägt (Pearson, 1904). Kontingenztafeln stellen die gemeinsame Häufigkeitsverteilung kategorialer Merkmale dar, vergleichen also verschiedene Klassifikationen miteinander. Im Falle eines medizinischen Tests wird die Klassifikation des Zustandes (krank oder gesund) mit der Klassifikation des Tests (positiv oder negativ) verglichen. Werden nur zwei Klassifikationen miteinander verglichen, ergibt sich eine einfache Tabelle, in der eine Klassifikation horizontal (Zeilen; z. B. Testergebnis) die andere vertikal (Spalten; z. B. Zustand) aufgetragen wird. Eine solche zweidimensionale Kontingenztafel wird auch Kreuztabelle (engl. cross tab), im Bereich des maschinellen Lernens Konfusionsmatrix (engl. confusion matrix) genannt. Horizontal bzw. vertikal zeigen die Randsummen die

Häufigkeitsverteilung jeweils einer Klassifikation[22]. Die zwischen diesen „Achsen“ aufgespannte Tabelle illustriert die gemeinsame Häufigkeitsverteilung der beiden Klassifikationen. Werden zwei binäre Klassifikationen miteinander verglichen, erhält man als einfachste Form einer Kreuztabelle die sogenannte Vier-Felder-Tafel. Kontingenztafeln sind aber weder auf binäre Klassifikationen noch auf zwei Dimensionen beschränkt. Multidimensionale Kontingenztafeln sind sogar eher die Regel. Höherdimensionale Kontingenztafeln werden zur Analyse jedoch häufig in ein zweidimensionales Format und dieses wiederum in eine binäre Klassifikation kollabiert. Kontingenztafeln sind die Grundlage für fast alle Kenngrößen zur Bewertung diagnostischer und statistischer Tests.

3.2.6.1 Die Felder der Vier-Felder-Tafel

In der Vier-Felder-Tafel (Tabelle 3.2) repräsentieren üblicherweise die Spalten die Realität (wahre Werte) und die Zeilen die Testbefunde (Vorhersagen, Prognosen). Hinsichtlich der Reihenfolge der Klassen sind in der Vier-Felder-Tafel die Anordnungen 0/1 und 1/0 möglich, wobei letztere das Gebräuchlichste ist. Für diese Studie wurde jedoch aus praktischen Gründen die „aufsteigende“ Anordnung 0/1 gewählt. Wie Tabelle 3.2 zeigt, werden die Felder bei den beiden Anordnungen jeweils relativ zueinander diagonal, die Randsummen horizontal bzw. vertikal vertauscht.

Tabelle 3.2: Mögliche Anordnungen der Klassen in der Vier-Felder-Tafel

Anordnung der Klassen 1/0					**Anordnung der Klassen 0/1**				
		Referenz/Realität		Randsumme			Referenz/Realität		Randsumme
		1	0				0	1	
Vorhersage	1	**a** tp	**b** fp	a+b tp+fp=p	Vorhersage	0	**a** tn	**b** fn	a+b tn+fn=n
	0	**c** fn	**d** tn	c+d fn+tn=n		1	**c** fp	**d** tp	c+d fp+tp=p
Randsumme		a+c tp+fn=a	b+c fp+tn=u	a+b+c+d tp+fp+fn+tn=s	Randsumme		b+c fp+tn=u	a+c tp+fn=a	a+b+c+d tp+fp+fn+tn=s

Die Tabelle zeigt die Vier-Felder-Tafel bei unterschiedlicher Anordnung der Klassen. Die Bezeichnung mit Buchstabenpaaren anstelle einzelner Buchstaben identifiziert die einzelnen Felder eindeutig.

Um die Darstellung von der letztlich willkürlichen Reihenfolge der Klassen zu lösen und damit Verwechslungen auszuschließen, werden die Felder der Vier-Felder-Tafel in dieser Arbeit nicht mit Kleinbuchstaben benannt, sondern mit Buchstabenpaaren, die die Felder der Vier-Felder-Tafel, unabhängig von der Anordnung der Klassen, eindeutig identifizieren. Die **richtig Negativen** sind alle Testpersonen, die vom Test korrekt als gesund eingestuft wurden (kurz **tn** nach engl. „true negative“). Analog hierzu sind die **richtig Positiven** alle zu Recht als krank eingestuften Testpersonen (kurz **tp** nach engl. „true positive“). Gesunde Testpersonen bei denen der Test aber positiv ausfällt, heißen **falsch Positive** (kurz **fp** nach engl. „false positive“). Kranke Testpersonen mit einem negativen Testergebnis sind **falsch Negative** (kurz **fn** nach engl. „false negative“). Mit diesen vier Feldern werden alle möglichen Beziehungen zwischen Testergebnis und wahrem Zustand erschöpfend beschrieben. Die vier Elemente der Vier-Felder-Tafel bilden die gemeinsame Verteilung zwischen dem horizontal aufgetragenen Zustand und dem vertikal aufgetragenen Testergebnis. Sie addieren sich deshalb zur Stichprobe (kurz **s** nach engl. sample) der Gesamtheit aller mit dem Test untersuchten Personen.

[22] *Im Beispiel der Vierfeldertafel auf Seite 64 ist horizontal der tatsächliche Zustand mit den Klassen krank/gesund und vertikal die Vorhersage des Testergebnisses (testpositiv/testnegativ) aufgetragen. Dem entsprechend zeigen die horizontalen Randsummen die Verteilung des realen Zustandes c, die vertikalen Randsummen des Testergebnisses r.*

3.2.6.2 Randsummen der Kontingenztafel

Die Randsummen (engl. marginal totals) der Vier-Felder-Tafel fassen die vier Felder tn, tp, fn und fp horizontal (Zeilen) und vertikal (Spalten) zu Gruppen mit einer anschaulichen Bedeutung zusammen. ***Vertikal*** bilden die **richtig Negativen** (kurz **tn**) und die **falsch Positiven** (kurz **fp**) die **Gesunden** (kurz **u** nach engl. unaffected), die **falsch Negativen** (kurz **fn**) und die **richtig Positiven** (kurz **tp**) die **Kranken** (kurz **a** nach engl. affected). ***Horizontal*** finden sich die **richtig Negativen** (kurz **tn**) und die **falsch Negativen** (kurz **fn**) zu den (Test)**Negativen** (kurz **n** nach engl. negative) zusammen, die **falsch Positiven** (kurz **fp**) und die **richtig Positiven** (kurz **tp**) zu den (Test)**Positiven** (kurz **p** nach engl. positive). Auch ***diagonal*** bilden die Felder der Vier-Felder-Tafel wichtige Gruppen: Die **falsch Negativen** (kurz fn) und die **falsch Positiven** (kurz fp) formen die **Falschen** (kurz **f** nach engl. false), die **richtig Negativen** (kurz **tn**) und die **richtig Positiven** (kurz **tp**) die **Richtigen** (kurz **t** nach engl. true).

Tabelle 3.3: Gruppen der Vier-Felder-Tafel mit abgeleiteten Kennwerten

Orientierung		Gruppe	Beispiele abgeleiteter Kennwerte	Formel
Zustand (condition c)	vertikal (Spalten)	Gesunde (u) = condition negative (cn)	Spezifität (= true negative rate) alpha (= false positive rate = p-value)	$tn + fp$
		Kranke (a) = condition positive (cp)	Sensitivität (= true positive rate = power) beta (= false negative rate)	$fn + tp$
(Test)Ergebnis (result r)	horizontal (Zeilen)	Negative (n) = result negative (rn)	negativer prädiktiver Wert false omission rate	$tn + fn$
		Positive (p) = result positive (rp)	positiver prädiktiver Wert (= precision) false discovery rate (= q-value)	$tp + fp$
(Test)Wahrheit (verity v)	diagonal	Richtige (t):	positive Übereinstimmung	$tn + tp$
		Falsche (f):	negative Übereinstimmung	$fp + fn$

Definition wichtiger Gruppen der Vier-Felder-Tafel und aus ihnen abgeleitete Kennwerte.

3.2.6.2.1 Erwartungswerte

Den Randsummen der Kontingenztafel kommt u. a. deshalb eine besondere Bedeutung zu, weil sich aus ihnen die Erwartungswerte der Felder bei zufälliger Verteilung bzw. Unabhängigkeit der in der Kontingenztafel aufgetragenen Merkmale abgeleitet werden können. Diese Erwartungswerte spielen bei der Berechnung von Kappa Cohen und statistischen Tests, wie dem χ-Quadrat-Test, eine zentrale Rolle. Die allgemeine Formel zur Bildung der Erwartungswerte lautet:

Formel 3.2: Allgemeine Formel für Erwartungswerte in Kontingenztafeln

englisch: expected value — Abkürzung: exp

Formel:
$$Erwartungswert = \frac{Zeilensumme \times Spaltensumme}{Stichprobe}$$

Literatur: (Backhaus u. a., 2016: S. 368; Bortz u. a., 2008: S. 105; Kuckartz u. a., 2013: S. 93; Steland, 2004: S. 230; Weiß, 2013: S. 202)

Allgemeine Definition von Erwartungswerten in Kontingenztafeln unter der Annahme unabhängiger Merkmale.

Daraus ergeben sich die Erwartungswerte in der Vier-Felder-Tafel medizinischer Tests:

Tabelle 3.4: Erwartungswerte der Vier-Felder-Tafel

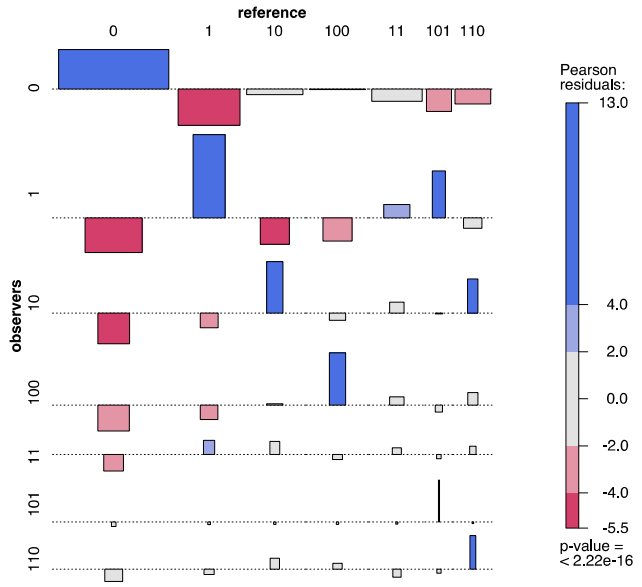
Erwartungswert für	Zeilensumme	Spaltensumme	Formulierung über Prävalenz
tn („etn“)	$= \frac{Negative \times Gesunde}{Stichprobe}$		$= (1 - Prävalenz) \times Negative$
fn („efn“)	$= \frac{Negative \times Kranke}{Stichprobe}$		$= Prävalenz \times Negative$
fp („efp“)	$= \frac{Positive \times Gesunde}{Stichprobe}$		$= (1 - Prävalenz) \times Positive$
tp („etp“)	$= \frac{Positive \times Kranke}{Stichprobe}$		$= Prävalenz \times Positive$

Formeln zur Berechnung der Erwartungswerte in der Vier-Felder-Tafel (Bortz u. a., 2008: S. 104 f.).

Zur Unterscheidung zwischen den beobachteten Werten und den Erwartungswerten werden die Kürzel der Erwartungswerte um „e“ für engl. expected (also etn, efn, efp und etp) ergänzt. Der Quotient $\frac{Kranke}{Stichprobe}$ ist äquivalent zur Prävalenz, der Quotient $\frac{Gesunde}{Stichprobe}$ entspricht 1-Prävalenz. Die Erwartungswerte können daher als Vielfache der Prävalenz bzw. von 1-Prävalenz umformuliert werden (Tabelle 3.4).

Als Visualisierung der Beziehung zwischen beobachteten und erwarteten Häufigkeiten in Kontingenztafeln schlug Cohen sogenannte Assoziationsdiagramme vor (Cohen, 1980), die später von Friendly weiter entwickelt wurden (Friendly, 1992). Meyer implementierte Assoziationsdiagramme im strucplot-framework (Meyer u. a., 2006), das in R über die Bibliothek vcd verfügbar ist.

Abbildung 3.3: Beispiel eines Assoziationsdiagrammes



Das Beispiel eines Assoziationsdiagramms links zeigt die Verhältnisse zwischen den beobachteten und erwarteten Häufigkeiten in der Kontingenztafel der Larynxstenosen. Jede Zelle der Kontingenztafel wird durch ein Rechteck repräsentiert, dessen Fläche proportional zur Differenz zwischen beobachteter und erwarteter Häufigkeit ist. Die Rechtecke sind dabei relativ zu einer Grundlinie positioniert, die Unabhängigkeit zwischen Beobachter und Referenz bzw. den Erwartungswert bei Zufälligkeit repräsentiert. Wenn die beobachtete Häufigkeit die erwartete Häufigkeit übertrifft, ist das Rechteck über der Grundlinie aufgetragen, anderenfalls darunter. Abweichungen sind zusätzlich farblich kodiert: positive Abweichungen entsprechen Blautönen, negative Rottönen. Im Kontext der Übereinstimmung zwischen Befundern sind über den Erwartungen liegende Häufigkeiten (blau) bei den diagonal angeordneten richtigen Befunden wünschenswert. Befunde abseits der Befundübereinstimmung signalisierenden Diagonalen, sollten hingegen möglichst unterhalb der Erwartungswerte liegen (rot).

Das Assoziationsdiagramm illustriert das Verhältnis zwischen Beobachtungswerten und Erwartungswerten in der Kontingenztafel (Friendly, 1992; Meyer u. a., 2003, 2006).

Im Assoziationsdiagramm repräsentiert für jede Zelle einer Kontingenztafel ein Rechteck die Differenz zum jeweiligen Erwartungswert. Die Fläche des Rechtecks ist dabei proportional zur Abweichung und relativ zu einer den Erwartungswert signalisierenden Grundlinie positioniert sowie farblich kodiert: über dem Erwartungswert liegende Häufigkeiten sind über der Grundlinie aufgetragen und in Blautönen gehalten, darunter liegende Häufigkeiten sind unterhalb der Grundlinie angeordnet und rot schattiert. Eine gute Übereinstimmung zwischen den Befundern

und dem Goldstandard manifestiert sich im Assoziationsdiagramm in diagonal gelegenen Häufigkeiten über dem Erwartungswert (blau) – also in häufigen kongruenten Befunden. Befunde abseits der Übereinstimmung signalisierenden Diagonalen sollten hingegen möglichst selten auftreten (grau bis rot), also unterhalb der Erwartungswerte liegen. Numerisch können die Differenzen zu den Erwartungswerten in sogenannten Pearson residuals gemessen werden, die in der Legende korrespondierend zu den Farbschattierungen angegeben sind.

3.2.6.2.2 Homogenität der Randsummen

Um die Diagnosen zweier Befunder miteinander vergleichen zu können, müssen die Randsummen der Befunde ähnlich verteilt sein. Trifft das nicht zu, kann es zu systematischen Abweichungen kommen, die die Interpretierbarkeit vieler Maßzahlen beeinträchtigen. Das ist anhand eines Beispiels nachvollziehbar: Während ein gutmütiger Lehrer seine Schüler stets eine Note besser bewertet, als es der tatsächlichen Leistung entspricht, vergibt sein strenger Kollege immer um eine Stufe schlechtere Noten. Fordert man beide Lehrer auf, dieselbe Schulklasse zu beurteilen, ist nur eine geringe Übereinstimmung feststellbar, obwohl die Bewertung der beiden Lehrer bis auf die Verschiebung um 2 Schulnoten eigentlich einheitlich ist. De facto liegt jedoch nur bei den Schulnoten 3 und 4 eine (scheinbare) Überschneidung der Beurteilung vor. Die Randsummen der zugehörigen Kontingenztafel unterscheiden sich aufgrund der Verschiebung erheblich. Die unterschiedlichen Verteilungen der Randsummen entsprechen differierenden Prävalenzen in den einzelnen Klassen (im Beispiel Schulnoten). **Die Forderung nach Homogenität der Randsummen entspricht also der Forderung nach homogenen Prävalenzen in den Befundkategorien**. Da die meisten Kennzahlen der Übereinstimmung wie z. B. prädiktive Werte und Kappa wesentlich von der Prävalenz beeinflusst sind (Gwet, 2002), können diese Maßzahlen bei unterschiedlichen Randverteilungen nur eingeschränkt interpretiert werden. Ein möglicher Ausweg ist das Heranziehen von weitgehend prävalenzunabhängigen Kennwerten wie z. B. der odds ratio.

Tabelle 3.5: Beispiel divergierende Randsummen

Schulnote	objektive Leistung der Schüler	Bewertung (=Randverteilung) gutmütiger Lehrer	Bewertung (=Randverteilung) strenger Lehrer
1		2	
2	2	4	
3	4	3	2
4	3	1	4
5	1		3
6			1

Durch systematische Über- bzw. Unterbewertung der Leistungen verschieben sich die Randsummen (=Randverteilungen) der Kontingenztafel. Obwohl sich die Bewertung der beiden Lehrer eigentlich nur in einem systematischen Fehler von zwei Notenstufen unterscheidet, scheinen sich die Bewertungen der beiden Lehrer nur bei den Notenstufen 3 und 4 zu überschneiden.

3.2.6.2.2.1 McNemar Test

Der McNemar-Test prüft Vier-Felder-Tafeln auf signifikante Unterschiede in den Randsummen. Bei gleichen Randsummen gilt in der Vier-Felder-Tafel Gesunde = Negative bzw. tn+fp = tn+fn und Kranke = Positive bzw. fn+tp = fp+tp. Somit kürzt sich bei gleichen Randsummen in den Gleichungen tn respektive tp heraus, so dass fp = fn als determinierende Gleichung für übereinstimmende Randsummen übrig bleibt (Darstellung nach Uebersax, 2006). Darauf baut die Teststatistik des McNemar-Testes auf, der nur in der Vier-Felder-Tafel anwendbar ist. In dieser Studie wird das Ergebnis des McNemar-Test bevorzugt in der linken oberen Ecke von Kontingenztafeln dar-

gestellt. Ein signifikanter McNemar-Test – gleichbedeutend mit signifikanten Unterschieden in den Randsummen – wird durch einen schwarzen Zellenhintergrund hervorgehoben.

Formel 3.3: Teststatistik McNemar-Test

englisch:	McNemar test	Abkürzung: McN
Definition:	Test zum Vergleich der Randsummen einer Vier-Felder-Tafel.	
Formel:	Teststatistik McNemar-Test in der Vier-Felder-Tafel= $\frac{(fn-fp)^2}{(fn+fp)}$	
Literatur:	(Fay, 2011; McNemar, 1947)	

Definition der Teststatistik nach Mc Nemar zum Vergleich der Homogenität von Randsummen.

3.2.6.2.2.2 Stuart-Maxwell- und Bhapkar Test

Der Stuart-Maxwell-Test (Everitt, 1986; Maxwell, 1970; Stuart, 1955) deckt, als eine generalisierte Form des McNemar-Test, Randsummeninhomogenitäten in Kontingenztafeln mit mehr als 2 Klassen auf. Der Bhapkar-Test (Bhapkar, 1966) ist eine weitere Variante eines generalisierten McNemar-Testes. Beide Tests sind asymptotisch äquivalent (Keefe, 1982). Der Bhapkar-Test wird im Vergleich zum Stuart-Maxwell-Test als mächtiger beschrieben (Gamer u. a., 2012; Uebersax, 2006) und fand daher im Rahmen dieser Studie bevorzugt Anwendung. Analog zum McNemar Test wird er in der linken oberen Ecke von Kontingenztafeln angegeben und im Falle eines signifikanten Testergebnisses durch einen schwarzen Zellenhintergrund markiert.

Tabelle 3.6: Die Vier-Felder-Tafel mit ihren wichtigsten Maßzahlen

Kontingenztafel **Test** versus **Zustand** „Vier-Felder-Tafel“		**Zustand** (Klassifizierung durch Referenz) wahre Werte bzw. Realität: negativ (0) *gesund* H_0^- falsch	positiv (1) *krank* H_0^+ richtig	Randsummen Σ **Testergebnis** (r)	Verhältnisse Felder zu Randsummen **„rates“**: richtige Befunde	falsche Befunde	Verhältnisse der Felder: Chancen **„odds“**
Test (Klassifizierung durch Befunder) *Vorhersage bzw. Testergebnis*	negativ (0) H_0^{rej} verworfen	**tn** richtig negativ	**fn** falsch negativ → *Fehler 2. Art*	(Test)**Negative** tn+fn=**rn=n**	**negativer prädiktiver Wert** (npv) $\frac{tn}{tn+fn}$	**false omission rate** (for) $\frac{fn}{tn+fn}$	Chance krank vs. Gesund bei Test neg. (oan) $\frac{fn}{tn}$
	positiv (1) H_0^{acc} akzeptiert	**fp** falsch positiv → *Fehler 1. Art*	**tp** richtig positiv	(Test)**Positive** fp+tp=**rp=p**	**positiver prädiktiver Wert** (ppv) precision $\frac{tp}{fp+tp}$	**false discovery rate** (fdr) q-value $\frac{fp}{fp+tp}$	Chance krank vs. Gesund bei Test pos. (oap) $\frac{tp}{fp}$
Zustand (c) Randsummen Σ		**Gesunde** tn+fp=**cn=u**	**Kranke** fn+tp=**cp=a**	**Stichprobenumfang (s)** tn+fn+fp+tp			**odds ratio** (OR) $\frac{\frac{tp}{fp}}{\frac{fn}{tn}} = \frac{\frac{ppv}{fdr}}{\frac{for}{npv}} = \frac{ppv}{\frac{(1-ppv)}{(1-npv)}} = \frac{ppv}{npv}$ =
Verhältnisse zu Randsummen **Raten (engl. „rates“)** Raten mit gleichem Nenner ergänzen sich zu 1 Raten sind bedingte P zweier Randsummen	richtige Befunde	**Spezifität** true negative rate (tnr) 1 - α $\frac{tn}{tn+fp}$	**Sensitivität** true positive rate (tpr) Teststärke bzw. Macht (power; 1 - β) recall $\frac{tp}{fn+tp}$	**Genauigkeit** (acc) accuracy $\frac{tn+tp}{tn+fn+fp+tp}$			
	falsche Befunde	**alpha** α Fehler 1. Art false positive rate (fpr) p-value; 1 - tnr $\frac{fp}{tn+fp}$	**beta** β Fehler 2. Art false negative rate (fnr) 1 - tpr $\frac{fn}{fn+tp}$	**Un**genauigkeit (iac) $\frac{fn+fp}{tn+fn+fp+tp}$			
Verhältnisse der Felder untereinander **Chancen (engl. „odds“)**	im Hinblick auf Testergebnis bei gegebenem Zustand	Chance Test positiv vs. negativ bei Gesundheit (opu) $\frac{fp}{tn} = \frac{fpr}{tnr}$	Chance Test positiv vs. negativ bei Krankheit (opa) $\frac{tp}{fn} = \frac{tpr}{fnr}$	**odds ratio** (OR bzw. odr) $\frac{\frac{tp}{fn}}{\frac{fp}{tn}} = \frac{\frac{tpr}{fnr}}{\frac{fpr}{tnr}} = \frac{tpr}{\frac{(1-tpr)}{tnr}} = \frac{LR+}{LR-}$			Kreuzprodukt $\frac{tp \times tn}{fn \times fp}$
Verhältnisse der Raten **Quotienten (engl. „ratios“)** (of rates)		Likelihood ratio neg. Test (LR- bzw. lrn) $\frac{tpr}{fpr} = \frac{\frac{tp}{fn+tp}}{\frac{fp}{tn+fp}}$ = $\frac{(1-tpr)}{tnr} = \frac{fnr}{tnr}$	Likelihood ratio pos. Test (LR+ bzw. lrp) $\frac{fnr}{tnr} = \frac{\frac{tp}{fn+tp}}{\frac{tn}{tn+fp}}$ = $\frac{tpr}{(1-tnr)} = \frac{tpr}{fpr}$	Area under curve (AUC) $\frac{1}{2} + \frac{1}{2}\frac{tp \times tn - fp \times fn}{(tp+fp)(fn+tn)}$ = $\frac{(Sensitivität + Spezifität)}{2}$			

Modifiziert und erweitert nach einer Darstellung der englischen Wikipedia zum positiv prädiktiven Wert (Wikipedia, 2016). Die Referenzen zu den Formeln sind bei den folgenden Einzeldarstellungen der Maßzahlen angegeben. Zusammenfassende Darstellungen mehrerer Kenngrößen finden sich u. a. bei Powers und Fawcett (Fawcett, 2006; Powers, 2011).

3.3 Kennzahlen

Als Maß der Befundübereinstimmung dienen Kenngrößen. Welche Kenngrößen sich jeweils anbieten, hängt zuvorderst vom Skalenniveau der zu untersuchenden Variable ab, also davon, ob es sich um ein kontinuierliches oder kategoriales Merkmal handelt, wobei innerhalb kategorialer Merkmale nach nominalen und ordinalen Merkmalen unterschieden wird. Ferner lassen sich die Kenngrößen in Maße der Präzision und Maße der Richtigkeit gliedern. Jede Kenngröße beleuchtet dabei immer nur einen Teilaspekt von Präzision oder Richtigkeit. Ein einziges Maß, das sämtliche Vorzüge in sich vereint (engl. omnibus index) gibt es nicht. Um ein möglichst vollständiges Bild des Sachverhaltes zu erhalten, müssen daher verschiedene Kenngrößen betrachtet werden. So sind Kenngrößen etwa auf positive Befunde fokussiert (z. B. Sensitivität), andere auf negative Befunde (z. B. Spezifität), korrigieren für zufällige Übereinstimmung (Kappa) oder sind vom Einfluss der Randsummen bzw. Prävalenz weitgehend unabhängig (z. B. odds ratio). Einige Maßzahlen der Vier-Felder-Tafel lassen sich nicht ohne Weiteres auf mehrdimensionale Arrays verallgemeinern. Die wichtigsten Kenngrößen werden im Folgenden vorgestellt. Für alle Kenngrößen gilt: Die unbekannten Wahrscheinlichkeiten, respektive Häufigkeiten in der Grundgesamtheit, werden gemäß einem empirischen Verständnis von Wahrscheinlichkeit anhand der beobachteten Häufigkeiten in der Stichprobe, die in der Kontingenztafel dargestellt sind, geschätzt.

3.3.1 Skalenniveau und Gliederung

Kontinuierliche Merkmale kommen in der Medizin zwar häufig vor (z. B. Labormesswerte, Blutdruck etc.), werden jedoch zur Interpretation meistens umgehend in kategoriale Größen transformiert (z. B. Normalbefund/pathologischer Befund bei Labormesswerten und Hypotonie / Normotonie/Hypertonie beim Blutdruck). Die meisten Variablen liegen somit auf einer Nominal- bzw. Ordinalskala. Im Befundfragebogen wurde mit dem maximalen Stenosegrad nur ein einziges Merkmal auf einer kontinuierlichen Skala erhoben, das im Rahmen der Auswertung aber auf die ordinale Myer-Cotton Klassifikation abgebildet wurde. Kennzahlen zur Beurteilung der Übereinstimmung kontinuierlicher Merkmale, wie der Intra-Klassen-Korrelationskoeffizient (ICC), fanden im Rahmen dieser Studie daher keine Anwendung. Kennzahlen zur Beurteilung der Übereinstimmung kategorialer Merkmale werden fast ausnahmslos aus Feldern der Vier-Felder-Tafel konstruiert. Nach der Art ihrer Konstruktion kann man diese Kennzahlen gliedern in

- Verhältnisse der Felder zu den Randsummen: **Raten** (engl. **rates**)
- Verhältnisse der Felder untereinander: **Chancen** (engl. **odds**)
- Verhältnisse der Raten zueinander: **Quotienten** (engl. **ratios**)

Die im Rahmen dieser Studie angewendeten Kenngrößen werden im kommenden Abschnitt gemäß dieser Gliederung eingeführt.

3.3.2 Kennzahlen der Präzision

Kennzahlen der Präzision werden unabhängig von einer Referenz (hier dem Goldstandard) berechnet. Sie beschreiben die relative Übereinstimmung der Werte untereinander bzw. ihre Streuung, sagen aber nichts über deren absolute Lage aus (siehe Zielscheibenanalogie Seite 56). Im Gegensatz zu den Kennzahlen der Richtigkeit werden sie unabhängig von der Vier-Felder-Tafel bzw. Kontingenztafeln konstruiert.

3.3.2.1 Durchschnittliche positive Übereinstimmung

In der Literatur ist derzeit keine etablierte, referenzunabhängige Maßzahl für die Übereinstimmung zwischen mehreren Untersuchern bekannt. Im Rahmen dieser Arbeit wurde daher das

Konzept der „durchschnittlichen positiven Übereinstimmung“ (kurz **mag** nach engl. „mean positive agreement“) entwickelt. Dabei wird zunächst eine Tabelle konstruiert in der in den Zeilen die beurteilten Videos und in den Spalten die Untersucher aufgetragen sind (siehe Beispiel Tabelle 3.7). Für jede Klasse und jedes Video wird dann aus dem Verhältnis der in der jeweiligen Klasse übereinstimmenden Untersucher zu allen Untersuchern die prozentuale Übereinstimmung (kurz pca nach engl. „percentual agreement“) bestimmt. Bei Befunden, die von nur einem Untersucher erhobene wurden, wurde die pca für die Berechnung der durchschnittlichen positiven Übereinstimmung gleich 0 gesetzt, da der Begriff Übereinstimmung erst ab 2 Untersuchern sinnvoll definiert werden kann.

Tabelle 3.7 erklärt die durchschnittliche positive Übereinstimmung anhand eines Beispiels mit 4 Untersuchern, 5 Videos und einer binären Klassifikation in positiv (1) und negativ (0). Für die Klasse positiv (1) ergibt sich in Video 1, angesichts fehlender positiver Befunde, eine Übereinstimmung von 0 %. In Video 3 errechnet sich eine prozentuale Übereinstimmung von 25 %. Da es sich um den Befund eines einzigen Untersuchers handelt, liegt jedoch keine Übereinstimmung vor, sodass in die Berechnung der durchschnittliche positiven Übereinstimmung eine 0 einfließt. In den Videos 3, 4 und 5 liegt die prozentuale Übereinstimmung bei 50 %, 75 % respektive 100 %. Die durchschnittliche Übereinstimmung über alle Videos liegt somit bei 75 %[23]. Das Beispiel der mittleren positiven Übereinstimmung für negative Befunde ist komplementär hierzu. Dieser Algorithmus zur Berechnung der positiven Übereinstimmung wurde in einer R-Funktion umgesetzt.

Tabelle 3.7: Beispiel Berechnung mittlere positive Übereinstimmung

Video	Untersucher 1	Untersucher 2	Untersucher 3	Untersucher 4	pca (1)	pca (0)
1	0	0	0	0	0	100
2	1	0	1	0	50	50
3	1	0	0	0	(25) 0	75
4	1	1	1	0	75	(25) 0
5	1	1	1	1	100	0
				mpa	75	75

Vier Untersucher bewerten 5 Videos als positiv (1) oder negativ (0). Die mittlere positive Übereinstimmung (mpa) beträgt im Beispiel für positive wie auch negative Befunde 75 %.

Die mittlere positive Übereinstimmung berücksichtigt nur positive Fälle. Würden im Beispiel der Tabelle 3.7 Videos mit durchgehend negativer Befundung hinzugefügt, hätte das auf die mittlere positive Übereinstimmung keinen Einfluss. Die Prävalenz einheitlich negativer Befunde ist somit für die mittlere positive Übereinstimmung irrelevant. Für die Berechnung der mittleren positiven Übereinstimmung über mehrere positive Befundkategorien (z. B. kombinierte Stenosen) wurde der Durchschnitt der mittleren positiven Übereinstimmung der einzelnen Kategorien gebildet, wobei deren relative Häufigkeit berücksichtigt wurde. Dabei wurden die Kategorien negative Befunde und Fehlwerte nicht berücksichtigt (analog zur Berechnung in den einzelnen Kategorien).

3.3.2.2 Kappa nach Fleiss

Koeffizienten, die Übereinstimmung unter Berücksichtigung zufälliger Übereinstimmung messen, werden mit dem griechischen Buchstaben Kappa bezeichnet. Unter den zahlreichen Varianten von Kappa ist das Kappa nach Fleiss (Fleiss, 1971) eine Möglichkeit, Übereinstimmung unabhängig von einer Referenz zu bestimmen. In R ist das Kappa nach Fleiss u. a. in der Funktion kappam.fleiss aus der Bibliothek irr implementiert (Gamer u. a., 2007, 2012). Alternative Kappa

[23](50 %+75 %+100 %)/3

Indices für mehrere Untersucher sind u. a. das Kappa nach Light (Light, 1971) und das Kappa nach Conger (Conger, 1980).

3.3.3 Kennzahlen der Richtigkeit

Kennzahlen der Richtigkeit prüfen die absolute Übereinstimmung zu einer Referenz (hier dem Goldstandard). Sie beruhen sämtlich auf den Größen der Vier-Felder-Tafel (Seite 64).

3.3.3.1 Genauigkeit

Die Genauigkeit (engl. accuracy, kurz **acc**) ist die wahrscheinlich intuitivste und bekannteste Messgröße der Richtigkeit. Sie bezeichnet die Wahrscheinlichkeit, einen Patienten richtig zu klassifizieren, also entweder als richtig gesund oder richtig krank und entspricht somit dem Verhältnis der richtigen Befunde (siehe Tabelle 3.4 Seite 61) im Vergleich zur Gesamtheit aller Befunde der Stichprobe. Die Genauigkeit berücksichtigt gleichberechtigt richtig positive und richtig negative Befunde. Das Kappa nach Cohen kann als eine Form der Genauigkeit angesehen werden, die für die gemäß den Randsummen zu erwartende, „zufällige“ Wahrscheinlichkeit korrigiert.

Formel 3.4: Genauigkeit

englisch: accuracy — Abkürzung: acc

Definition: Anteil der richtig klassifizierten Personen an der Stichprobe.

Formeln:

$$Genauigkeit = \frac{tn+tp}{tn+tp+fn+fp} = \frac{Richtige(t)}{Stichprobe(s)}$$

$$= \left(\frac{tp+tn}{s}\right)\left(\frac{tp}{tp+tn}\right) + \left(\frac{fp+tn}{s}\right)\left(\frac{tn}{fp+tn}\right)$$

$$= Prävalenz \times Sensitivität + (1 - Prävalenz) \times Spezifität$$

Literatur: (Alberg u. a., 2004; Uebersax, 2014)

Rechnerische Definition der Genauigkeit als Maßzahl der Richtigkeit in der Vier-Felder-Tafel.

Man kann die Genauigkeit als gewichtetes Mittel von Sensitivität und Spezifität eines Testes auffassen, wobei die Sensitivität mit der Prävalenz gewichtet wird, die Spezifität mit dem Komplement der Prävalenz (Formel 3.4). Die Genauigkeit ist als Maßzahl der Übereinstimmung irreführend je größer die Differenz zwischen Sensitivität und Spezifität ist und/oder je mehr die Prävalenz von 50 % abweicht (Alberg u. a., 2004). Um diese Irreführung zu vermeiden wird die Genauigkeit verbreitet durch Maße wie Kappa ersetzt bzw. ergänzt, die für zufällige Übereinstimmung korrigieren.

3.3.3.2 Positive und negative Übereinstimmung

Im Gegensatz zur Genauigkeit, als Gesamtmaß für richtig positive und richtig negative Befunde fokussiert sich die positive Übereinstimmung auf die Richtigkeit positiver Befunde.

Formel 3.5: positive Übereinstimmung

englisch: positive agreement, proportion of specific agreement — Abkürzung: pag

Definition: Maß für die Übereinstimmung in positiven Befunden der Vier-Felder-Tafel.

Formel:

$$positive\ Übereinstimmung = \frac{2\,tp}{s + (tp - tn)} = \frac{tp}{\frac{2\,tp + fn + fp}{2}} = \frac{2\,tp}{2\,tp + fn + fp}$$

Literatur: (Cicchetti, Feinstein, 1990; Cunningham, 2009; Hripcsak, Heitjan, 2002; Uebersax, 2014)

Rechnerische Definition der positiven Übereinstimmung in der Vier-Felder-Tafel.

Fleiss verbalisiert die positive Übereinstimmung als bedingte Wahrscheinlichkeit eines positiven Befundes des einen Beobachters, unter der Bedingung eines positiven Befundes des anderen Be-

obachters, wobei der Beobachter zufällig ausgewählt wird (Fleiss, 1981; Graham, Bull, 1998). Die positive Übereinstimmung betrachtet beide Beobachter als gleichwertig. Die Entsprechung der positiven Übereinstimmung für negative Befunde ist die negative Übereinstimmung.

Formel 3.6: negative Übereinstimmung

englisch:	negative agreement, proportions of specific agreement	Abkürzung: nag
Definition:	Maß für die Übereinstimmung in negativen Befunden der Vierfeldertafel.	
Formel:	$Negative\ Übereinstimmung = \frac{2\,tn}{2\,tn + fn + fp}$	
Literatur:	(Cicchetti, Feinstein, 1990; Cunningham, 2009; Hripcsak, Heitjan, 2002; Uebersax, 2014)	

Rechnerische Definition der negativen Übereinstimmung in der Vierfeldertafel.

Zufallskorrigierte Varianten der positiven und negativen Übereinstimmung sind in beiden Fällen identisch mit dem Kappa nach Cohen.

3.3.3.3 Raten (Verhältnisse der Felder zu den Randsummen)

Raten (engl. rates) sind Verhältnisse, die zu Randsummen der Vier-Felder-Tafel gebildet werden. Prominenteste Vertreter sind die Sensitivität und Spezifität. Raten können auch als bedingte Wahrscheinlichkeiten zweier Randsummen aufgefasst werden. Sie beschreiben die Wahrscheinlichkeit für das Eintreten der einen Randsumme unter der Bedingung der anderen Randsumme: p(Randsumme 1 | Randsumme 2). Die Sensitivität ist zum Beispiel die Wahrscheinlichkeit eines positiven Tests (Randsumme 1) unter der Bedingung von Krankheit (Randsumme 2). Raten machen also entweder eine Aussage über die Wahrscheinlichkeit des Testergebnisses bei gegebenem Zustand, oder die Wahrscheinlichkeit des Zustandes bei gegebenem Testergebnis. Die Raten dieser beiden Gruppen sind insofern spiegelbildlich zueinander, als die Ereignisse der bedingten Wahrscheinlichkeit vertauscht sind (P(A|B) versus P(B|A)).

Tabelle 3.8: Raten der Felder der Vier-Felder-Tafel

Wahrscheinlichkeiten der Testergebnisse bei gegebenem Zustand P(Test | Zustand)	**Wahrscheinlichkeiten der Zustände bei gegebenem Testergebnis P(Zustand | Test)**
Spezifität = true negative rate (tnr) **P(Testnegativ|Gesundheit)**	Negativ prädiktiver Wert (npv) **P(Gesundheit|Testnegativ)**
Sensitivität = true positive rate (tpr) **P(Testpositiv|Krankheit)**	Positiv prädiktiver Wert (ppv) **P(Krankheit|Testpositiv)**
Alpha = false positive rate (fpr) **P(Testpositiv|Gesundheit)**	False discovery rate (fdr) **P(Gesundheit|Testpositiv)**
Beta = false negative rate (fnr) **P(Testnegativ|Krankheit)**	False omission rate (for) **P(Krankheit|Testnegativ)**

Modifiziert nach Tabelle 1 im Supplement zu Westover (2011).

Raten mit der gleichen Randsumme im Nenner ergänzen sich zu 1 und können deshalb als komplementäre Raten bezeichnet werden.

Formel 3.7: Raten mit gleichem Nenner ergänzen sich zu 1

Randsumme im Nenner	**Komplementäre Raten**			**Summe**
Gesunde	Spezifität	+	Alpha	
Kranke	Sensitivität	+	Beta	**=1**
Negative	Negativer prädiktiver Wert	+	False omission rate	
Positive	Positive prädiktiver Wert	+	False discovery rate	

Komplementäre Raten der Vier-Felder-Tafel.

Auch die Verhältnisse der Randsummen zum Stichprobenumfang (ebenfalls eine Randsumme) können als Rate bezeichnet werden. Diese Raten geben den Anteil der Gesunden, Kranken,

(Test)Negativen und (Test)Positiven an der Stichprobe an. Die relativen Anteile der Randsummen am Stichprobenumfang entsprechen dem Erwartungswert der Raten der Felder bei Zufälligkeit.

Formel 3.8: Raten der Randsummen der Vier-Felder-Tafel

$$Gesundrate = \frac{tn+fp}{tn+fn+fp+tp} = \frac{u}{s} = \frac{Gesunde}{Stichprobenumfang} = zufällig\ zu\ erwartender\ npv$$

$$Krankrate = \frac{fn+tp}{tn+fn+fp+tp} = \frac{a}{s} = \frac{Kranke}{Stichprobenumfang} = zufällig\ zu\ erwartender\ ppv$$

$$Negativrate = \frac{tn+fn}{tn+fn+fp+tp} = \frac{n}{s} = \frac{Negative}{Stichprobenumfang} = zufällig\ zu\ erwartende\ Spezifität$$

$$Positivrate = \frac{fp+tp}{tn+fn+fp+tp} = \frac{p}{s} = \frac{Positive}{Stichprobenumfang} = zufällig\ zu\ erwartende\ Sensitivität$$

Raten der Randsummen und Ihre Beziehung zu den zufällig zu erwartenden Raten der Felder der Vier-Felder-Tafel.

3.3.3.3.1 Sensitivität, Spezifität und Youden J

Die Sensitivität ist die Wahrscheinlichkeit, einen Kranken mit einem Test als krank zu erkennen. Sie ist die bedingte Wahrscheinlichkeit, dass der Test im Falle von Krankheit positiv ausfällt.

Formel 3.9: Sensitivität

englisch: sensitivity, true positive rate, power, recall — Abkürzung: sen, tpr

Definition: Wahrscheinlichkeit einen Kranken als krank zu klassifizieren.

Formel:
$$Sensitivität = \frac{tp}{tp+fn} = \frac{richtig\ Testpositive}{Kranke(a)} = 1-\beta$$

Literatur: (Altman, Bland, 1994a; Hilgers u. a., 2007: S. 83; Kraemer, 1992: S. x; Sachs, Hedderich, 2006: S. 133; Spitalnic, 2004b)

Rechnerische Definition der Sensitivität in der Vier-Felder-Tafel.

In der Vier-Felder-Tafel wird die Sensitivität über den Anteil der **richtig als krank erkannten Testpersonen** (tp) an der Gesamtheit der **tatsächlich kranken Testpersonen** (fn+tp) geschätzt. Sie ist eine Maßzahl innerhalb der Gruppe der Kranken. Die Sensitivität wird auch als Falsch-Positive-Rate (engl. **true positive rate**; kurz tpr) bezeichnet und entspricht in der Nomenklatur statistischer Tests der Teststärke bzw. Macht (engl. power). Sie ist ein Maß der „Empfindlichkeit" bzw. Trennschärfe eines Tests. Die Gesamt-Sensitivität über mehrere Klassen wurde gemäß der Definition in der R-Bibliothek multiclasstesting (Nardini, Liu, 2014) bestimmt. Analog zum Kappa nach Cohen (Formel 3.28 Seite 79) wurde eine Variante der Sensitivität vorgeschlagen, die die zufällig zu erwartende Übereinstimmung berücksichtigt.

Formel 3.10: korrigierte Sensitivität nach Kraemer, Coughlin und Jamart

englisch: chance corrected sensitivity, rescaled sensitivity — Abkürzungen: kse, λ_{se}, $\kappa(1,0)$

Definition: Zufallskorrigierte Wahrscheinlichkeit einen Kranken mit einem Test als krank zu klassifizieren.

Formeln:
$$kse = \frac{\frac{tp}{fn+tp} - \frac{fp+tp}{tn+fn+fp+tp}}{\frac{tn+fn}{tn+fn+fp+tp}} = \frac{tpr - \frac{p}{s}}{\frac{n}{s}} = \frac{Sensitivität - Positivrate}{Negativrate} = \frac{npv - Negativrate}{Positivrate}$$
$$= \frac{Sensitivität - zufällig\ erwartete\ Sensitivität}{1 - zufällig\ erwartete\ Sensitivität} = \frac{sen - [sen \times pre + (1-spe) \times (1-pre)]}{spe \times (1-pre) + (1-sen) \times pre}$$

Literatur: (Coughlin, Pickle, 1992; Jamart, 1992; Kraemer, 1985, 1992: S. 40)

Rechnerische Definition der korrigierten Sensitivität in der Vier-Felder-Tafel.

Die Spezifität ist die Wahrscheinlichkeit, einen Gesunden als gesund zu klassifizieren. Sie wird in der Vier-Felder-Tafel (Tabelle 3.6 Seite 64) über den Anteil der **richtig als gesund erkannten**

Testpersonen (tn) an der Gesamtheit der tatsächlich gesunden Testpersonen (tn+fp) in der Stichprobe geschätzt. Sie ist eine Maßzahl innerhalb der Gruppe der Gesunden.

Formel 3.11: Spezifität

englisch: specificity, true negative rate — Abkürzung: spe, tnr

Definition: Wahrscheinlichkeit, einen Gesunden mit einem Test als gesund zu klassifizieren.

Formeln: $$Spezifität = \frac{tn}{tn + fp} = \frac{richtig\ Testnegative}{Gesunde} = 1 - \alpha$$

Literatur: (Altman, Bland, 1994a; Kraemer, 1992: S. x; Sachs, Hedderich, 2006: S. 132–133; Spitalnic, 2004b)

Rechnerische Definition der Spezifität in der Vier-Felder-Tafel.

Wie bei der Sensitivität kann auch für die Spezifität in Anlehnung an das Kappa nach Cohen (Formel 3.28 Seite 79) eine Variante berechnet werden, die die zufällig zu erwartende Übereinstimmung mit einbezieht.

Formel 3.12: korrigierte Spezifität nach Kraemer, Coughlin und Jamart

englisch: chance corrected specificity — Abkürzung: ksp, $\lambda_{sp}, \kappa(0,0)$

Definition: Zufallskorrigierte Wahrscheinlichkeit, einen Gesunden mit einem Test als gesund zu klassifizieren.

Formeln:
$$ksp = \frac{\frac{tn}{tn + fp} - \frac{tn + fn}{tn + fn + fp + tp}}{\frac{fp + tp}{tn + fn + fp + tp}} = \frac{tnr - \frac{n}{s}}{\frac{p}{s}} = \frac{Spezifität - Negativrate}{Positivrate} = \frac{ppv - Krankrate}{Gesundrate}$$
$$= \frac{Spezifität - zufällig\ erwartete\ Spezifität}{1 - zufällig\ erwartete\ Spezifität} = \frac{spe - [spe\,(1 - pre) + (1 - sen) \times pre]}{sen \times pre + (1 - spe)(1 - pre)}$$

Literatur: (Coughlin, Pickle, 1992; Jamart, 1998; Kraemer, 1992: S. 41)

Rechnerische Definition der zufallskorrigierten Spezifität in der Vier-Felder-Tafel.

Aus der zufallskorrigierten Sensitivität und Spezifität nach Kraemer, Coughlin und Jamart kann im Verein mit dem Stichprobenumfang die χ^2-Teststatistik abgeleitet werden.

Formel 3.13: Chi-Quadrat-Teststatistik über korrigierte Sensitivität & Spezifität formuliert

englisch: chi-square-statistic — Abkürzung: χ^2

Formeln: $$\chi^2 = s \times kse \times ksp = Stichprobenumfang \times korrigierte\ Sensitivität \times korrigierte\ Spezifität$$

Literatur: (Jamart, 1992: S. 1036; Kraemer, 1992: S. 41)

Die Chi-Quadrat-Statistik kann aus dem Stichprobenumfang und der korrigierten Sensitivität und Spezifität nach Kraemer, Coughlin und Jamart (Formeln 3.10 und 3.12) abgeleitet werden.

Aus den Formeln 3.10 und 3.12 wird ersichtlich, dass zufallskorrigierte Sensitivität und Spezifität alleine aus Sensitivität, Spezifität und Prävalenz konstruierbar sind. Neben der hier vorgestellten zufallskorrigierten Sensitivität und Spezifität nach Kraemer, Coughlin und Jamart (Coughlin, Pickle, 1992; Jamart, 1998; Kraemer, 1992: S. 40–41) wurde von Brenner und Gefeller ein anderer Berechnungsmodus vorgeschlagen, der unabhängig von der Prävalenz ist. Diese Variante ist eng mit den Likelihood ratios und dem Youden J verwandt (Brenner, Gefeller, 1994; Gefeller, Brenner, 1994). Beide Varianten der Sensitivität und Spezifität sind umstritten (Holle, Windeler, 1997).

Das sogenannte Youden J, das auch als Youden Index bezeichnet wird, fasst Sensitivität und Spezifität zu einer gemeinsamen Größe zusammen. Mithilfe des Youden Index kann der optimale Schwellenwert eines diagnostischen Tests bestimmt werden.

Formel 3.14: Youden J

englisch:	Youden J	Abkürzung:	J
Formel:	$\frac{1}{2}(\frac{tp-fn}{tp+fn}+\frac{tn-fp}{fp+tn})=\frac{tp\times tn-fn\times fp}{(tp+fn)(fp+tn)}=Sensitivität+Spezifität-1$		
Literatur:	(Hedderich, Sachs, 2016: S. 181; Youden, 1950)		

Rechnerische Definition des Youden J in der Vier-Felder-Tafel.

3.3.3.3.2 Alpha und Beta

Alpha gibt die Wahrscheinlichkeit eines (falsch) positiven Tests bei Gesunden an und wird auch als falsch-positiv-Rate (engl. false positive rate; kurz fpr) bezeichnet.

Formel 3.15: alpha

englisch:	Alpha, false positive rate, type I error	Abkürzung:	α, fpr
Definition:	Die Wahrscheinlichkeit für einen positiven Test bei Gesunden.		
Formel:	$\frac{fp}{tn+fp}$		
Literatur:	(Hedderich, Sachs, 2016: S. 427; Sachs, Hedderich, 2006: S. 308 f.)		

Rechnerische Definition von Alpha in der Vier-Felder-Tafel.

Im Kontext des statistischen Testens spricht man vom Fehler 1. Art oder Risiko I. Alpha entspricht der Irrtumswahrscheinlichkeit, eine falsche Hypothese als richtig einzustufen. Alpha markiert somit das Signifikanzniveau eines statistischen Tests. Im medizinischen Umfeld werden für alpha üblicherweise Irrtumswahrscheinlichkeiten von 0,05 („signifikant“) oder 0,01 („hoch signifikant“) in Kauf genommen.

Formel 3.16: beta

englisch:	beta, false negative rate, type II error	Abkürzung:	β, fnr
Definition:	Wahrscheinlichkeit eines negativen Testergebnisses bei Krankheit		
Formeln:	$\frac{fn}{fn+tp}=\frac{falsch\ Negative}{Kranke}=1-Sensitivität$		
Literatur:	(Hedderich, Sachs, 2016: S. 427; Sachs, Hedderich, 2006: S. 308 f.)		

Rechnerische Definition von Beta in der Vier-Felder-Tafel.

Beta gibt die Wahrscheinlichkeit eines (falsch) negativen Tests bei Kranken an und wird auch als falsch-negativ-Rate (engl. false negative rate; kurz fnr) bezeichnet. Im Zusammenhang mit statistischen Tests heißt Beta Fehler 2. Art oder Risiko II. Der Fehler 2. Art gibt die Wahrscheinlichkeit an, eine richtige Hypothese als falsch zu verwerfen. Im Gegensatz zum bewusst gewählten Alpha ist Beta ein schwer zu kontrollierender Parameter. Beta ist mit Alpha invers korreliert: Je kleiner α desto größer β und desto kleiner die Teststärke. Alpha und Beta nehmen beide mit zunehmendem Stichprobenumfang (s) ab. Neben α und s ist die Effektgröße die für β bestimmende Variable. (Theorie: Bortz u. a., 2008: S. 36–43; interaktive Graphik: Jeske, 2001: Abschn. Fehler 1. und 2. Art; Praxis in R: Kabacoff, 2011: Kap. 10). Alpha und Beta sind bedingte Wahrscheinlichkeiten: Alpha gilt für den Fall, dass H0 falsch ist (P|H0- → nur 1. Spalte), Beta für den Fall, dass H0 richtig ist (P|H0+ → nur 2. Spalte). Alpha+Spezifität sowie Beta+Sensitivität ergänzen sich deshalb jeweils (in ihrer Spalte) zu 1 (Sachs, Hedderich, 2006).

3.3.3.3.3 positiver und negativer prädiktiver Wert

Der positive prädiktive Wert gibt die Wahrscheinlichkeit an, tatsächlich krank zu sein, wenn das Testresultat positiv ausgefallen ist. Es handelt sich also um die **bedingte** Wahrscheinlichkeit für Krankheit unter der Bedingung eines positiven Testresultates. Der positive prädiktive Wert heißt

auch ***a posteriori* Wahrscheinlichkeit** für Krankheit bzw. Posttestwahrscheinlichkeit - im Gegensatz zur Prävalenz als ***a priori* Wahrscheinlichkeit** für Krankheit. Der positive prädiktive Wert wird in der Vier-Felder-Tafel über das Verhältnis der tatsächlich Kranken an den Testpositiven geschätzt und lässt sich auch durch Prävalenz, Sensitivität und Spezifität beschreiben. Er ist eine Maßzahl innerhalb der Gruppe der Testpositiven. Der gemeinsame Prädiktive Wert für mehrere Kategorien wurde gemäß der Definition und Implementierung in der R-Bibliothek Multiclasstesting (Nardini, Liu, 2014) bestimmt.

Formel 3.17: positiver Vorhersagewert, positiver prädiktiver Wert

englisch:	positive predictive value	Abkürzung: ppv
Definition:	Wahrscheinlichkeit, einen Kranken im Test als krank zu klassifizieren.	
Formeln:	$\frac{tp}{tp+fp}=\frac{richtig\ Testpositive}{Testpositive}=\frac{(Prävalenz\times Sensitivität)}{Prävalenz\times Sensitivität+(1-Prävalenz)\times(1-Spezifität)}$	
Literatur:	(Altman, Bland, 1994b; Bortz, Lienert, 2008: S. 262; Gallagher, 2005; Hilgers u. a., 2007: S. 84; Kraemer, 1992: S. x; Li u. a., 2007; Sachs, Hedderich, 2006: S. 132–133; Weiß, 2013: S. 213)	

Rechnerische Definition des positiven prädiktiven Wertes in der Vier-Felder-Tafel.

Analog zum positiven prädiktiven Wert gibt der negative prädiktive Wert die Wahrscheinlichkeit für Gesundheit bei einem negativen Testresultat an. Der negative prädiktive Wert wird in der Vier-Felder-Tafel über das Verhältnis der tatsächlich Gesunden an den Testnegativen geschätzt. Wie der positive prädiktive Wert kann auch der negative Prädiktive Wert durch Prävalenz, Spezifität und Sensitivität ausgedrückt werden. Das macht deutlich, dass die prädiktiven Werte erheblich von der Prävalenz beeinflusst sind.

Formel 3.18: negativer Vorhersagewert, negativer prädiktiver Wert

englisch:	negative predictive value	Abkürzung: npv
Definition:	Wahrscheinlichkeit, einen Gesunden im Test als gesund zu klassifizieren.	
Formel:	$\frac{tn}{tn+fn}=\frac{richtig\ Testnegative}{Testnegative}=\frac{Spezifität\times(1-Prävalenz)}{Spezifität\times(1-Prävalenz)+(1-Sensitivität)\times Prävalenz}$ $= 1-fdr$	
Literatur:	(Altman, Bland, 1994b; Bortz, Lienert, 2008: S. 263; Gallagher, 2005; Hilgers u. a., 2007: S. 84; Kraemer, 1992: S. x; Li u. a., 2007; Sachs, Hedderich, 2006: S. 132–133)	

Rechnerische Definition des negativen prädiktiven Wertes in der Vier-Felder-Tafel.

3.3.3.3.4 false omission rate und false discovery rate

Der Anteil negativer Testergebnisse, die falsch sind, wird als engl. „false omision rate“ (kurz for), der Anteil positiver Testergebnisse, die falsch sind, als engl. „false discovery rate“ (kurz fdr) bezeichnet. Die false omission rate entspricht der Wahrscheinlichkeit trotz eines negativen Testes krank zu sein, die false discovery rate der Wahrscheinlichkeit trotz eines positiven Testes gesund zu sein.

Formel 3.19: false omission rate

englisch:	false omission rate	Abkürzung: for	false discovery rate	Abkürzung: fdr
Definition:	Anteil falsch Negativer an den Negativen.		Anteil falsch Positiver an den Positiven.	
Formel:	$\frac{fn}{tn+fn}=\frac{Falsch\ Negative}{Negative}=1-npv$		$\frac{fp}{fp+tp}=\frac{Falsch\ Positive}{Positive}=1-ppv$	
Literatur:	(Supplement Westover u. a., 2011: S. 4)			

Rechnerische Definition der false omission rate und false discovery rate in der Vier-Felder-Tafel.

3.3.3.4 Ratios (Verhältnisse der Raten)

Aus den Verhältnissen der Raten zueinander lassen sich weitere Maßzahlen bilden. Die wichtigste von Ihnen ist die sogenannte odds ratio (OR).

3.3.3.4.1 Odds – Chancenverhältnisse

Der Sprachgebrauch von odds im Englischen ist variabel (Fulton u. a., 2012) und kann in etwa mit „Chancenverhältnis“ ins Deutsche übertragen werden. Die odds gibt das Verhältnis zwischen der Wahrscheinlichkeit des Eintretens eines Ereignisses und der des Nicht-Eintretens an. Ein Chancenverhältnis von 1 : 4 bedeutet, dass in einem von 5 Fällen ein positiver Ausgang zu erwarten ist und entspricht somit einer Wahrscheinlichkeit von 0,2 (20 %).

Formel 3.20: Allgemeine Formel für odds

englisch:	odds	Abkürzung:	odd

Definition: Verhältnis zwischen der Wahrscheinlichkeit des Eintretens eines Ereignisses und der des Nicht-Eintretens.

Formel:
$$odds(A) = \frac{P(A)}{1 - P(A)}$$

Literatur: (Hilgers u. a., 2007: S. 235)

Rechnerische Definition der odds in der Vier-Felder-Tafel.

Das Verhältnis zwischen komplementären Raten (z. B. Sensitivität & Beta bzw. Spezifität und Alpha), also solchen, die sich zu 1 ergänzen, entspricht genau der Definition der odds: die den komplementären Raten gemeinsame Randsumme im Nenner kürzt sich im Doppelbruch heraus (Formel 3.21). Die odds der Vier-Felder-Tafel entsprechen somit dem Verhältnis zwischen Feldern, die eine gemeinsame Randsumme bilden. Die odds der Vier-Felder-Tafel werden also gleichzeitig vom Verhältnis zwischen komplementären Raten wie auch vom Verhältnis zwischen zwei Feldern definiert. Dem entsprechend können aus den Feldern der Vier-Felder-Tafel horizontal und vertikal vier unterschiedlich zusammengesetzte odds definiert werden – eine pro Randsumme.

Formel 3.21: Chancenverhältnisse der Vier-Felder-Tafel

	Randsumme	Formeln	In Worten	kurz
bei gegebenem Zustand im Hinblick auf Testergebnis	Kranke	$\frac{tpr}{(1-tpr)} = \frac{Sensitivität}{Beta} = \frac{\frac{tp}{fn+tp}}{\frac{fn}{fn+tp}} = \frac{tp}{fn}$	Wahrscheinlichkeit positiv : negativ bei Krankheit p(p:n|a)	opa
	Gesunde	$\frac{fpr}{(1-fpr)} = \frac{Alpha}{Spezifität} = \frac{\frac{fp}{tn+fp}}{\frac{tn}{tn+fp}} = \frac{fp}{tn}$	Wahrscheinlichkeit positiv : negativ bei Gesundheit p(p:n|u)	opu
bei gegebenem Testergebnis im Hinblick auf Zustand	Negative	$\frac{for}{(1-for)} = \frac{false\ omission\ rate}{negativ\ prädiktiver\ Wert} = \frac{\frac{fn}{tn+fn}}{\frac{tn}{tn+fn}} = \frac{fn}{tn}$	Wahrscheinlichkeit krank : gesund bei negativem Test p(a:u|n)	oan
	Positive	$\frac{ppv}{(1-ppv)} = \frac{positiv\ prädiktiver\ Wert}{false\ discovery\ rate} = \frac{\frac{tp}{fp+tp}}{\frac{fp}{fp+tp}} = \frac{tp}{fp}$	Wahrscheinlichkeit krank : gesund bei positivem Test p(a:u|p)	oap

(Glas u. a., 2003: S. 1130)

Zwei odds formulieren das Chancenverhältnis des Testergebnisses (positiv versus negativ) bei gegebenem Zustand und zwei das Chancenverhältnis des Zustandes (krank versus gesund) bei gegebenem Testergebnis. Die vier odds geben somit die Wahrscheinlichkeit eines positiven Testergebnisses im Vergleich zu einem negativen bei Krankheit und Gesundheit bzw. die Wahrscheinlichkeit von Krankheit im Vergleich zu Gesundheit bei negativem und positivem Test an. Es handelt sich also um bedingte Wahrscheinlichkeiten. Die odds der Vier-Felder-Tafel selbst werden selten diskutiert, sind aber Berechnungsgrundlage für eine der wichtigsten Maßzahlen, die sogenannten odds ratio (OR).

3.3.3.4.2 odds ratio, Yules Q und Yules Y

Wie der Name sagt wird die odds ratio aus dem Verhältnis der odds gebildet. Und zwar entweder aus dem Verhältnis der beiden odds, die vom Zustand ausgehen oder dem Verhältnis der beiden odds, die vom Testergebnis ausgehen (siehe Formel 3.21). Beide Quotienten führen zum gleichen Ergebnis und werden durch Kürzen zum Kreuzprodukt der vier Felder. Die odds ratio zählt zu den prävalenzunabhängigen Maßzahlen. Sie ist ein Maß für den Zusammenhang zwischen Zeilen und Spalten der Kontingenztafel. Diese Eigenschaft macht die odds ratio zur Grundlage der logistischen Regression (Rudas, 1998). Im Falle der Vier-Felder-Tafel medizinischer Tests gibt die odds ratio den Zusammenhang zwischen Zustand und Testergebnis an. Eine odds ratio von 1 bedeutet Unabhängigkeit zwischen Zustand und Testergebnis. Der Test wäre also nutzlos. Je höher die odds ratio, desto größer ist der Zusammenhang zwischen Zustand und Testergebnis und desto besser die Aussagekraft des Testes.

Formel 3.22: odds ratio

englisch: odds ratio — Abkürzung: OR, odr

Definition: Das Verhältnis zwischen komplementären Raten entspricht dem Kreuzprodukt und heißt odds ratio.

Formeln:

$$\frac{tp \times tn}{fn \times fp} = \frac{\text{LR+}}{\text{LR-}} =$$

bei gegebenem Zustand im Hinblick auf das Testergebnis:

$$\frac{\text{opa}}{\text{opu}} = \frac{\frac{tp}{fn}}{\frac{fp}{tn}} = \frac{\frac{tpr}{fnr}}{\frac{fpr}{tnr}} = \frac{\frac{sen}{1-sen}}{\frac{1-spe}{spe}} = \frac{sen \times spe}{(1-sen)\times(1-spe)}$$

bei gegebenem Testergebnis im Hinblick auf den Zustand:

$$\frac{\text{oap}}{\text{oan}} = \frac{\frac{tp}{fp}}{\frac{fn}{tn}} = \frac{\frac{ppv}{fdr}}{\frac{for}{npv}} = \frac{\frac{ppv}{1-ppv}}{\frac{1-npv}{npv}} = \frac{ppv \times npv}{(1-ppv)\times(1-npv)}$$

Literatur: (Glas u. a., 2003: S. 1130; Hilgers u. a., 2007: S. 236; Kraemer, 1992: S. 103)

Rechnerische Definitionen der odds ratio. Die odds ratio kann sowohl von den vom Zustand ausgehenden Größen Sensitivität und Spezifität abgeleitet werden, als auch von den vom Testergebnis ausgehenden Größen positiver und negativer prädiktiver Wert.

Die odds ratio kann Werte zwischen 0 und unendlich annehmen. Für viele Anwendungen ist jedoch eine Verteilung der Werte zwischen 0 und 1 wünschenswert. Yule hat hierfür zwei Transformationen vorgeschlagen, die als Yule Q und Yule Y bekannt sind.

Formel 3.23: Yule Q und Yule Y, Transformationen der Odds ratio

englisch:	Yules Q	Abkürzung: oyq	Yules Y	Abkürzung: oyy
Formel:	$\frac{\text{OR}-1}{\text{OR}+1}$		$\frac{\sqrt{(\text{OR})}-1}{\sqrt{(\text{OR})}+1}$	
Literatur:	(Yule, 1900)		(Walter, 2001; Wirtz, Caspar, 2002: S. 107; Yule, 1912)	

Formeln von Yules Q und Yules Y, die die odds ratio auf Werte zwischen 0 und 1 transformieren.

3.3.3.4.3 Likelihood ratios, ROC und AUC

Als Likelihood ratio bezeichnet man die Wahrscheinlichkeit eines Testergebnisses bei einem Kranken im Verhältnis zur Wahrscheinlichkeit desselben Testergebnisses bei einem Gesunden (McGee, 2002). Die Likelihood ratio kann für ein positives (LR+) wie auch ein negatives (LR-) Testergebnis formuliert werden. Eine positive Likelihood ratio (LR+) >1 zeigt an, dass ein Testergebnis mit dem Vorhandensein der Krankheit assoziiert ist, ein Likelihood ratio < 1 zeigt eine Assoziation mit der Abwesenheit der Krankheit an. Je höher die positive Likelihood ratio über 1, respektive je näher sie an 0 liegt, desto stärker ist der Zusammenhang. In der Praxis kann man bei Likelihood ratios über 10 bzw. unter 0,1 das Vorhandensein bzw. den Ausschluss einer Krankheit annehmen (nach Deeks, Altman, 2004).

Formel 3.24: positive und negative Likelihoodratio

englisch:	positive likelihood ratio Abkürzung: LR+	negative likelihood ratio Abkürzung: LR-
Definition:	Wahrscheinlichkeit (P) eines positiven Tests bei Kranken im Verhältnis zur P eines positiven Tests bei Gesunden.	Wahrscheinlichkeit (P) eines negativen Tests bei Kranken im Verhältnis zur P eines negativen Tests bei Gesunden.
Formel:	$\mathrm{LR+} = \frac{1-\frac{tp}{tp+fn}}{\frac{tn}{tn+fp}} = \frac{1-Sensitivität}{Spezifität} = \frac{fnr}{tnr}$	$\mathrm{LR-} = \frac{1-\frac{tp}{tp+fn}}{\frac{tn}{tn+fp}} = \frac{1-Sensitivität}{Spezifität} = \frac{tpr}{fpr}$
Literatur:	(Akobeng, 2007; Deeks, Altman, 2004; Dujardin u. a., 1994: S. 30; Gallagher, 1998: S. 394; Hayden, Brown, 1999: S. 577; Hedderich, Sachs, 2016: S. 187; Hilgers u. a., 2007: S. 87 & 88; Spitalnic, 2004a)	

Formeln der positiven und negativen Likelihoodratio (LR) in der Vier-Felder-Tafel.

Aus Likelihood ratio und Prävalenz (Prä-Test-Wahrscheinlichkeit) kann über die Prä-Test-odds die Post-Test-odds und aus ihr die Post-Test-Wahrscheinlichkeit berechnet werden. Die Posttestwahrscheinlichkeit gibt die Wahrscheinlichkeit an, dass eine Erkrankung im Falle eines positiven (LR+) bzw. negativen Test (LR-) vorliegt. Diese Berechnung kann auch graphisch mit Hilfe des Fagan-Nomogramms gelöst werden. Dazu wird im Fagan-Nomogramm eine Gerade durch die Skalenwerte der Prä-Test-Wahrscheinlichkeit und der Likelihood ratio gezogen. Die Post-Test-Wahrscheinlichkeit ergibt sich dann aus dem Schnittpunkt dieser Geraden mit der Skala der Post-Test-Wahrscheinlichkeit.

Formel 3.25: Zusammenhang zwischen Likelihoodratio, Prä- und Post-Test-Wahrscheinlichkeit

Rechnerische Bestimmung der post-Test-Wahrscheinlichkeit:

$$prä-Test-odds=\frac{Prävalenz}{(1-Prävalenz)}$$

$$prä-Test-odds\times LR=post-test-odds$$

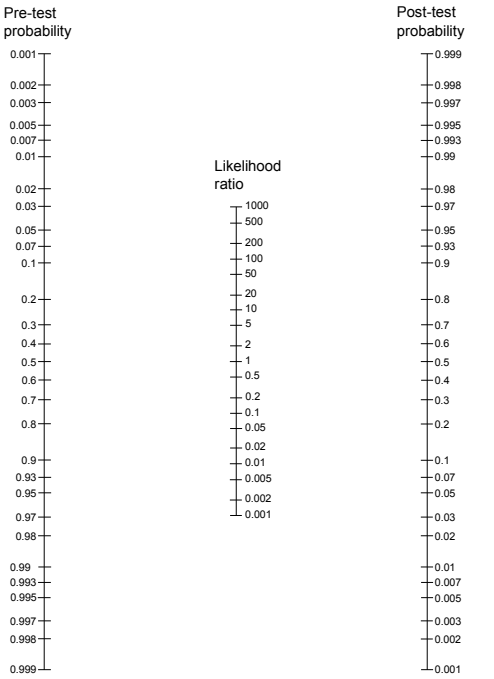
$$post-test-Wahrscheinlichkeit=\frac{post-test-odds}{(post-tests-odds+1)}$$

Tabelle zur Abschätzung des Wahrscheinlichkeitszuwachses:

Likelihood ratio	ungefähre Veränderung der Wahrscheinlichkeit
0,1	-45
0,2	-30
0,3	-25
0,4	-20
0,5	-15
1	0
2	+15
3	+20
4	+25
5	+30
6	+35
7-9	+40
10	+45

Literatur: (Hedderich, Sachs, 2016: S. 188; McGee, 2002)

graphische Lösung Fagan Nomogramm:



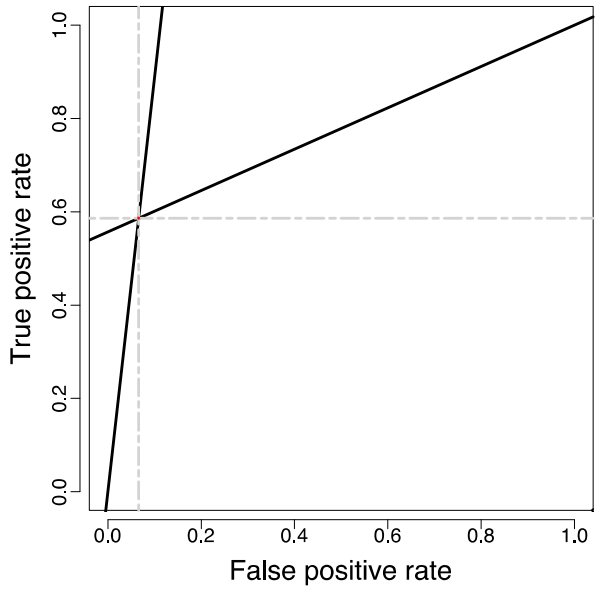
(Fagan, 1975; Herich u. a., 2015)

Mit der Likelihoodratio kann von der Prä-Test- auf die Post-Test-Wahrscheinlichkeit geschlossen werden. Zur schnellen Abschätzung in der Praxis kann die Gleichung mit dem Fagan-Nomogramm graphisch gelöst werden oder eine Tabelle zur groben Abschätzung genutzt werden. Letztere kommt ohne Prävalenz als Eingangsgröße aus.

Aus der Differenz zwischen Prä- und Post-Test-Wahrscheinlichkeit ergibt sich der durch den Test gewonnene Wahrscheinlichkeitszuwachs. Diese durch den Test erzielte Veränderung der Wahrscheinlichkeit kann grob mithilfe einer Tabelle abgeschätzt werden, in die nur die Likelihood ratio eingeht (McGee, 2002), denn die Likelihoodratio ist eine prävalenzunabhängige Größe.

Die Likelihood ratios können als Likelihood-ratio-Graph visualisiert werden (Biggerstaff, 2000). Dabei wird in einem Koordinatensystem vertikal die Sensitivität (synonym Richtig-Positiv-Rate; engl. true positive rate, kurz tpr) und horizontal Alpha (synonym Falsch-Positiv-Rate, engl. false positive rate, kurz fpr) aufgetragen. Die Gerade durch den Ursprung und den von Sensitivität und Alpha des Testes definierten Punkt repräsentiert die positive Likelihood ratio (LR+), denn ihre Steigung ist $\frac{tpr}{fpr}$, was der Definition der LR+ entspricht (Formel 3.24). Analog hierzu ist die Steigung der Geraden durch den von Sensitivität und Alpha des Testes definierten Punkt und den Punkt x=1, y=1 $\frac{fnr}{tnr}$ und damit die negative Likelihood ratio (LR-). Der Likelihood-ratio-Graph entspricht der sogenannten Receiver-Operator-Charakteristik (ROC) für eine binäre Variable. In R sind Likelihood-ratio-Graphen über die Bibliothek DiagnosisMed (Brasil, 2010) verfügbar.

Abbildung 3.4: ROC bzw. Likelihood-ratio-Graph



Der Likelihood-ratio-Graph ist eine Receiver-Operator-Charakteristik (ROC) für binäre Variablen. Im Koordinatensystem aus Sensitivität (engl. true positive rate, kurz tpr) versus Falsch-positiv-Rate (engl. false positive rate, kurz fpr) wird der Punkt y = Sensitivität, x = Falsch Positiv Rate eingetragen. Die gestrichelte Parallele zur x-Achse markiert die Sensitivität, die Parallele zur y-Achse die Falsch Positiv Rate. Die von (0,0) bzw. (1,1) ausgehenden Geraden durch den Punkt (tpr,fpr) haben die Steigung tpr/fpr bzw. fnr/tnr und repräsentieren somit die positive (LR+) respektive negative Likelihood ratio (LR-). Die Trennschärfe des Testes ist um so besser je weiter der Punkt (tpr,fpr) im linken oberen Eck der Graphik liegt, also je steiler LR+ und je flacher LR- verläuft bzw. je größer die Fläche unterhalb der LR+ und LR- ist. Die Fläche unterhalb LR+ und LR- wird als area under curve (kurz AUC) bezeichnet. Sie ist ein Maß der Trennschärfe eines Testes, das LR+ und LR- in sich vereint.

Beispiel einer Receiver-Operator-Charakteristik für den Befund (Vorhandensein bzw. Fehlen) einer subglottischen Larynxstenose (Biggerstaff, 2000).

Aus positiver und negativer Likelihood ratio kann ein gemeinsamer Index gebildet werden, die sogenannte Area Under Curve (AUC). Sie entspricht der gemeinsamen Fläche unter den beiden Geraden im Likelihood-ratio-Graph.

Formel 3.26: Area Under Curve (AUC)

englisch: area under curve Abkürzung: AUC

Formel:
$$\frac{(Sensitivität + Spezifität)}{2} = \frac{1}{2} + \frac{1}{2}\frac{tp \times tn - fp \times fn}{(tp + fp)(fn + tn)}$$

Literatur: (Bangdiwala u. a., 2008: S. 871; Cantor, Kattan, 2000; Hout, 2003)

Formel für die area under curve (AUC) in der Vier-Felder-Tafel.

3.3.3.5 Bangdiwala

Bangdiwala schlägt zur Beurteilung der Befundübereinstimmung ein Diagramm vor (Bangdiwala, 1985; Bangdiwala u. a., 2008; Bangdiwala, Shankar, 2013), das beobachtete Häufigkeiten von Befunden mit den Erwartungswerten vergleicht und dabei Sensitivität und Spezifität, sowie die prädiktiven Werte visualisiert. Aus dem Übereinstimmungsdiagramm nach Bangdiwala kann ein gemeinsamer Kennwert für Sensitivität, Spezifität und prädiktive Werte abgeleitet werden, den Bangdiwala als „B“ bezeichnet.

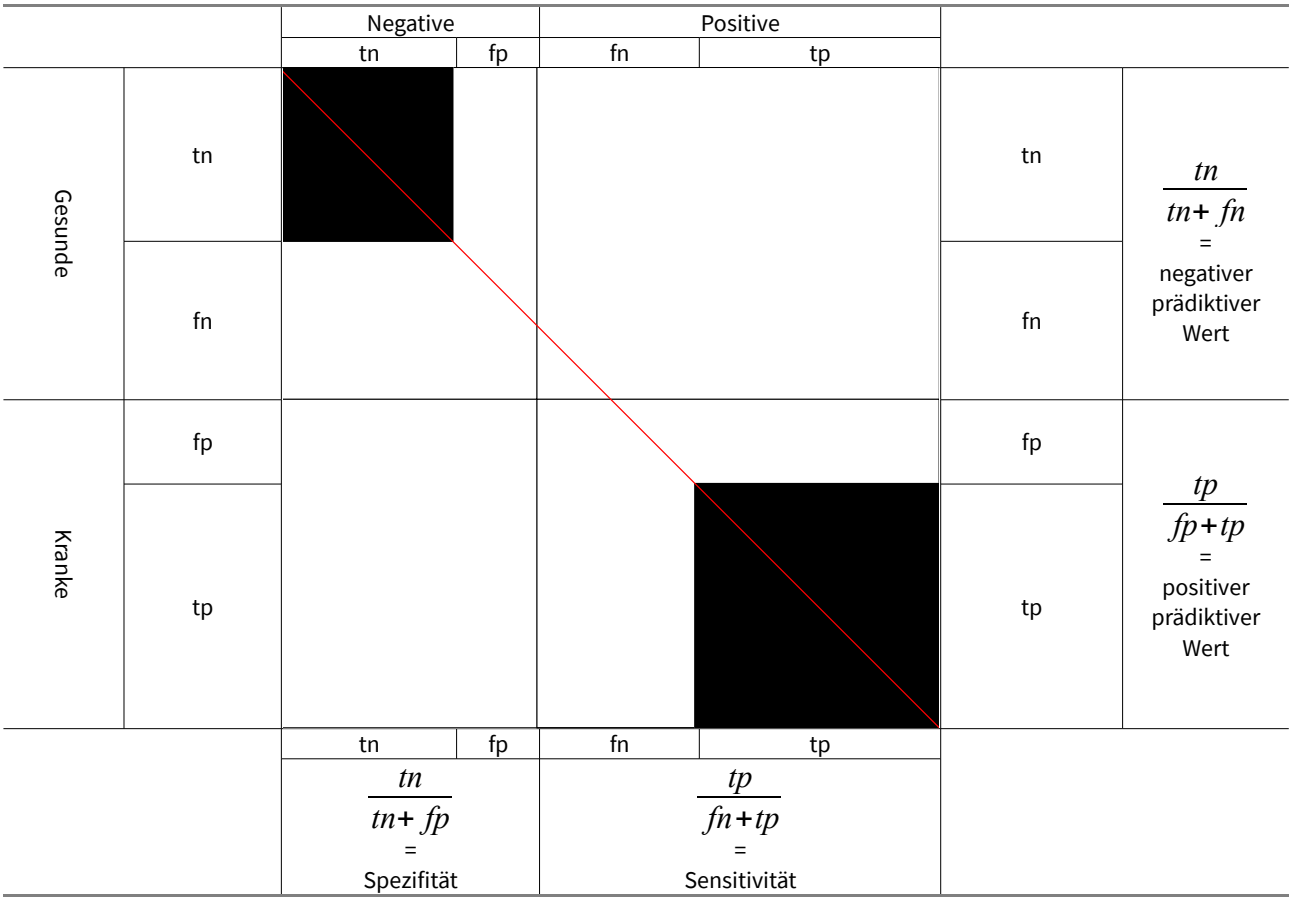
Formel 3.27: Bangdiwala

englisch:	Bangdiwala	Abkürzung: B
Definition:	Verhältnis aus Summe der quadrierten Häufigkeiten richtiger Befunde zur Summe der Erwartungswerte richtiger Befunde. Das entspricht dem Verhältnis der Summe der Flächen beobachteter Häufigkeiten richtiger Befunde im Verhältnis zur Summe der Fläche der maximal möglichen Übereinstimmung richtiger Befunde.	
Formel:	$B = \frac{\sum \text{Richtiger Felder}^2}{\sum Zeilensumme\, richtiger\, Felder \times Spaltensumme\, richtger\, Felder} = \frac{tn^2 + tp^2}{(tn+fn)(tn+fp)+(tp+fp)(tp+fn)}$	
Literatur:	(Bangdiwala, 1985; Bangdiwala u. a., 2008; Bangdiwala, Shankar, 2013; Munoz, Bangdiwala, 1997)	
Implementierung in R:	Funktion agreementplot der Bibliothek vcd	

Definition von Bangdiwalas B.

Die Bedeutung von B kann am besten anhand des Übereinstimmungsdiagramms nach Bangdiwala nachvollzogen werden:

Abbildung 57: Übereinstimmungsdiagramm nach Bangdiwala



Übereinstimmungsdiagramm nach Bangdiwala (nach Bangdiwala u. a., 2008: S. 868 Abbildung 2). *Die Schwarzen Flächen symbolisieren die beobachteten Häufigkeiten richtig negativer (engl. true negative; kurz tn) und richtig positiver (engl. true positive; kurz tp) Befunde, die umgebenden weißen Flächen die jeweiligen Erwartungswerte gemäß den Randsummen. Horizontal entspricht das Verhältnis zwischen der Seitenlänge des schwarzen Quadrates und dem umgebenden Rechteck Spezifität bzw. Sensitivität, vertikal negativem bzw. positivem prädiktivem Wert.*

3.3.3.6 Kappa nach Cohen

Die am besten etablierte Maßzahl zur Beurteilung der Übereinstimmung verschiedener Befunder ist das Kappa nach Cohen. Kappa kann als Variante der Genauigkeit aufgefasst werden, die für die zufällig zu erwartende Übereinstimmung korrigiert.

Formel 3.28: Kappa nach Cohen

englisch:	Kappa Cohen	Abkürzung: kpC, κ

Formel:

$$\kappa = \frac{beobachtete\ Übereinstimmung - zufällige\ Übereinstimmung}{1 - zufällige\ Übereinstimmung}$$

$$\kappa = \frac{2(tp \times tn - fn \times fp)}{(tp + fp)(fp + tn) + (fn + tn)(tp + fn)}$$

Literatur:	(Bortz u. a., 2008: S. 451; Bortz, Lienert, 2008: S. 311; Cohen, 1960; Hedderich, Sachs, 2016: S. 727; Hoehler, 2000; Hripcsak, Heitjan, 2002: S. 101; Langenbucher u. a., 1996: S. 1287)
Implementierung in R:	Funktion kappa2 aus der Bibliothek irr, Funktion Kappa aus der Bibliothek vcd, Funktion confusionMatrix aus der Bibliothek caret, Funktion cohen.kappa aus der Bibliothek psych

Definition des Kappakoeffizienten nach Cohen.

Die Interpretation von Kappa Cohen ist im Vergleich zu verbreiteten Kennwerten wie Sensitivität und Spezifität deutlich weniger intuitiv und anschaulich. Landis hat daher eine Skala zur groben Einschätzung der Bedeutung von Kappa Cohen vorgeschlagen. Für diese Arbeit wurde eine noch etwas weiter untergliederte Skala verwendet.

Tabelle: Interpretation von Kappa Cohen & Bangdiwala

Englisch	Deutsch	Kappa Cohen nach Landis	Kappa Cohen nach Munoz	Bangdiwala nach Munoz	modifizierte Klassifikation dieser Studie	
almost perfect	nahezu perfekt	0,81-1,00	0,75-1	0,65-1	0,81-1,0	perfekt
substantial	stark	0,61-0,80	0,45-0,75	0,35-0,65	0,71-0,80	hervorragend
					0,61-0,70	stark
moderate	moderat	0,41-0,60	0,2-0,45	0,15-0,35	0,51-0,60	gut
					0,41-0,50	moderat
fair	mäßig	0,21-0,40	0-0,2	0,05-0,15	0,31-0,40	mäßig
					0,21-0,30	schwach
slight	gering	0,00-0,20			0,11-0,20	gering
					0,06-0,1	kaum
poor	dürftig	<0	<0	0-0,05	0-0,05	keine

Interpretation von Kappa Cohen und Bangdiwala im Vergleich (Landis, Koch, 1977: S. 165; Munoz, Bangdiwala, 1997).

3.3.3.6.1 Kappa Cohen bei mehreren Beurteilern

Das Kappa nach Cohen ist nur für den Vergleich zwischen zwei Untersuchern definiert. Für die Übereinstimmung mehrerer Befunde wurden spezielle Varianten von Kappa Cohen entwickelt, wie das Kappa nach Fleiss (Fleiss, 1971), das Kappa nach Light (Light, 1971) oder das Kappa nach Conger (Conger, 1980). Diese Kappa-Varianten berechnen jedoch Kappa-Werte innerhalb einer Gruppe gleichberechtigter Befunder, sind also Maßzahlen der Präzision, nicht der Richtigkeit. Das Kappa nach Cohen kann als Maßzahl der Richtigkeit herangezogen werden, wenn einer der beiden verglichenen Untersucher eine Referenz bzw. ein Goldstandard, also den wahren Wert repräsentiert. Um das Konzept von Kappa Cohen als Maßzahl der Richtigkeit auf mehrere Untersucher zu erweitern, gibt es zwei naheliegende Möglichkeiten:

1. Die Bestimmung eines Gesamt-Kappas als Mittelwert paarweiser Kappa-Werte zwischen jedem Befunder und der Referenz.
2. Die Berechnung eines Gesamt-Kappas, indem die Gruppe der Befunder zu einem einzigen virtuellen Befunder vereint wird.

3.3.3.6.1.1 Mittelwert des paarweisen Kappa Cohen

Eine Möglichkeit Kappa Cohen für mehrere Untersucher und eine Referenz zu berechnen besteht darin, den Mittelwert der paarweisen Kappa-Werte zwischen der Referenz und jedem einzelnen Untersucher zu bestimmen. Über ein Histogramm kann die Verteilung der paarweisen Kappa-Werte illustriert werden, was Rückschlüsse auf die Einheitlichkeit der Befunde erlaubt. Das auf diese Weise kalkulierte Kappa Cohen wird im Verlauf als „paarweises Kappa Cohen“ bezeichnet.

3.3.3.6.1.2 Kappa Cohen mit vereinten Befundern

Eine weitere Möglichkeit Kappa Cohen für mehrere Untersucher und eine Referenz zu berechnen ist, alle Untersucher zu einem gemeinsamen virtuellen Befunder zusammen zu fassen. Dabei werden aus 42 Befunden von 20 Untersuchern zu einem Video 840 – sozusagen wiederholte – Befunde eines einzigen virtuell „vereinten“ Untersuchers zu einem Video. Der Befund des aus den 20 Untersuchern vereinten virtuellen Befunders wird zur Berechnung von Kappa Cohen einem in gleicher weise behandelten – also 20 mal wiederholten – Befund des Goldstandards gegenüber gestellt. Daraus kann eine einzige Kontingenztafel konstruiert werden, die sämtliche Befunde der 20 Untersucher mit dem Referenzbefund des Goldstandards vergleicht.

Diese Methode wurde in dieser Studie als bevorzugtes Verfahren gewählt und wird im Weiteren als „vereintes Kappa Cohen“ bezeichnet. Das vereinte Kappa entspricht der Intention einer gemeinsamen Aussage über alle Befunder. Die Analyse sämtlicher Befunde der Untersucher und des Goldstandards in einer gemeinsamen Kontingenztafel ist übersichtlicher als zahlreiche separate Kontingenztafeln und insbesondere hinsichtlich Kombinationsbefunden von Vorteil. Sämtliche Kombinationsbefunde werden beim vereinten Kappa Cohen in einer einzigen Kontingenztafel dargestellt, die somit das gesamte Spektrum aller Befundkategorien abdeckt. Bei paarweiser Berechnung von Kappa Cohen können die Befundkategorien in jeder paarweisen Kontingenztafel variieren. Da nur zwischen Beurteilern und Referenz überlappende Kategorien in die Berechnung mit einbezogen werden können, wirkt sich das auch auf die Fehlwerte aus.

3.3.3.6.1.3 Vergleich der Varianten des globalen Kappa Cohen

Bei den meisten Berechnungen wird neben dem unifizierten Kappa Cohen auch der Mittelwert des paarweisen Kappa Cohen sowie die Verteilung des paarweisen Kappa Cohen angegeben. Das von den paarweisen Mittelwerten abgeleitete Gesamt-Kappa lag meist etwas unter dem unifizierten Gesamt-Kappa, aber in der gleichen Größenordnung wie das unifizierte Kappa.

3.3.3.6.2 Kappa bei differierenden Befundklassen

In dieser Studie wurden inhaltlich zusammengehörige Einzelbefunde („Symptome“) zu Befundkombinationen („Syndromen“) aggregiert. Damit können

- Kappawerte übergeordneter anatomischer Abschnitte berechnet werden und
- es kann zwischen Einfach- und Mehrfachbefunden differenziert werden – z. B. zwischen einfachen und mehrfachen Stenosen.

Es ist sehr wahrscheinlich, dass von den Untersuchern Befundkombinationen gewählt werden, die vom Goldstandard nicht gewählt wurden und umgekehrt. Anders als bei Betrachtung auf Ebene einzelner Befunde mit der binären Klassifikation vorhanden oder nicht vorhanden, weichen die kombinierten Befundklassen daher zwischen Untersuchern und dem Goldstandard in der Regel voneinander ab. In die Berechnung der Richtigkeit können aber definitionsgemäß nur überlappende Befundklassen einbezogen werden. Befundklassen, die entweder nur von den Befundern oder nur vom Goldstandard gesehen wurden, werden bei der Berechnung der Richtigkeit zu Fehlwerten. Der Anteil der ursprünglich vorhandenen Daten, der in die Berechnung der

Richtigkeit einbezogen werden konnte, ist in der Spalte „Datenabdeckung“ in Prozent angegeben.

3.4 Multiple lineare Regression

Lineare Modelle zählen zu den am häufigsten angewandten statistischen Methoden. Die bei dem gegebenen Datensatz einfachste mögliche Variante eines linearen Modells ist die multiple lineare Regression (ohne Interaktionen), die mit den Funktionen lm und regr in R realisiert wurde. Die Beurteilung der relativen Bedeutung einzelner Variablen auf Grundlage ihrer Koeffizienten ist zwar naheliegend, praktisch aber schwierig, da die Variablen auf verschiedenen Skalen liegen können. Eine direkte Vergleichbarkeit solcher Variablen innerhalb des Models oder gar zwischen verschiedenen Modellen ist nicht gegeben. Um das intuitive Konzept der relativen Wichtigkeit von Variablen zu bedienen wurden eine Reihe von Verfahren entwickelt, die in R u. a. in der Bibliothek relaimpo (Groemping, 2006) implementiert wurden. Berechnet wird bei allen Verfahren der relative Beitrag einer Variable zum Bestimmtheitsmaß R^2. Aus den insgesamt 6 in relaimpo angebotenen Metriken wurden das klassische lmg (Lindeman u. a., 1980) und pmvd (Feldman, 2007) als jüngste Entwicklung für die Analyse ausgewählt.

3.5 Rekursive Partitionierung

Rekursives Partitionieren ist meist unter dem Begriff Entscheidungsbaum (engl. „decision tree“) bekannt. Die Vorteile des rekursiven Partitionierens im Vergleich zu anderen statistischen Modellen liegen u. a. in der

- Anschaulichkeit und somit guten Interpretierbarkeit, der
- Fähigkeit zur Modellierung komplexer Interaktionen, der
- weitgehenden Hypothesenfreiheit (eine theoretische statistische Verteilung der Variablen muss nicht geschätzt werden) und
- der flexiblen Handhabung von Fehlwerten.

Nachdem die relative Bedeutung der einzelnen Variablen mithilfe linearer Modelle untersucht wurde, bietet rekursives Partitionieren eine Möglichkeit, etwaige Wechselwirkungen zwischen den Variablen zu modellieren, die in den vorgestellten linearen Modellen nicht berücksichtigt wurden. Ein weiterer wesentlicher Vorteil ist im Umgang mit Fehlwerten zu sehen. Verfahrensbedingt konnte im Rahmen der multiplen linearen Regression nur ein via Imputation komplettierter Datensatz analysiert werden. Rekursives Partitionieren kann die lückenhaften Originaldaten direkt untersuchen. Im Falle des linearen Modells mit richtigen Einzelbefunden als Zielvariable bestehen Zweifel, ob die Grundvoraussetzungen für ein lineares Modell tatsächlich erfüllt sind. Die Methoden des rekursiven Partitionierens sind als hypothesenfreie Verfahren über derartige Zweifel erhaben. Da von Kollinearität und multiplem Testen keine wesentliche Beeinträchtigung der Baummodelle zu erwarten ist, konnten auch virtuelle Variablen, die sich aus den real erhobenen Variablen ableiten, in die Untersuchung mit einbezogen werden.

3.5.1 CART

Entscheidungsbäume wurden mithilfe der Implementierung rpart (Therneau u. a., 2012) des klassischen CART[24] Algorithmus (Breiman u. a., 1983) in der gleichnamigen R-Bibliothek berechnet. Rpart partitioniert ausschließlich binär und kann sowohl mit nominalskalierten als auch metrischen Zielvariablen umgehen. Variablen mit Fehlwerten werden unter Verwendung sogenann-

[24] *Das Akronym CART steht für Classification and Regression Trees und ist mit dem Titel der Erstbeschreibung (Breiman u. a., 1983) identisch.*

ter Surrogatvariablen[25] in die Analyse mit einbezogen. Hierdurch entstehen vergleichsweise einfache Modelle des originalen Datensatzes. Eine Eliminierung der Fehlwerte ist im Gegensatz zu vielen anderen Algorithmen nicht notwendig.

3.5.2 Random Forests

Sogenannte „random forests“ aggregieren – in der Hoffnung die Unzulänglichkeiten einzelner Bäume herauszumitteln - viele Entscheidungsbäume zu einem Entscheidungswald. Dabei wird in die einzelnen Bäume jeweils nur eine zufällige Auswahl aller Variablen aufgenommen (sogenanntes „bootstrapping“). Der Vorteil der Anschaulichkeit einzelner Bäume geht zwar verloren, dafür sind die Ergebnisse des Entscheidungswaldes robuster. Ähnlich zu linearen Modellen kann in random forest die relative Wichtigkeit von Variablen zueinander untersucht werden. Random forests sind auch in der Lage, Fehlwerte in Datensätzen zu imputieren. Insbesondere die beiden letztgenannten Eigenschaften waren im Rahmen dieser Studie ausschlaggebend für die Anwendung von random forests.

Tabelle 3.9: Überblick über die angewandten Algorithmen des rekursiven Partitionierens

Algorithmus	Implementierung in R	Zielvariablen	Partitionierung	Fehlwerte
CART	rpart library rtree function	nominalskaliert rationalskaliert	binär	Surrogatvariablen
random forest	randomForest library randomForest function	nominalskaliert rationalskaliert	je nachdem mit welchem Modell die einzelnen Bäume generiert werden	Imputation auf Basis der proximity

Vergleich verwendeter Verfahren des rekursiven Partitionierens.

[25] *Stark vereinfacht erklärt werden fehlende Werte durch Werte anderer Variablen die sich möglichst ähnlich verhalten ersetzt.*

Literatur

Agresti, Alan (2002): *Categorical Data Analysis*. 2. Auflage. John Wiley & Sons. — ISBN: 0-471-36093-7

Akobeng, Anthony K. (2007): „Understanding diagnostic tests 2: likelihood ratios, pre- and post-test probabilities and their use in clinical practice“. In: *Acta Pædiatrica*. 96 (4), S. 487–491, DOI: 10.1111/j.1651-2227.2006.00179.x.

Alberg, Anthony J; Park, Ji Wan; Hager, Brant W; u. a. (2004): „The Use of “Overall Accuracy” to Evaluate the Validity of Screening or Diagnostic Tests“. In: *Journal of General Internal Medicine*. 19 (5 Pt 1), S. 460–465, DOI: 10.1111/j.1525-1497.2004.30091.x.

Altman, D. G.; Bland, J. M. (1994a): „Diagnostic tests 1: Sensitivity and specificity.“. In: *BMJ: British Medical Journal*. 308 (6943), S. 1552.

Altman, D. G.; Bland, J. M. (1994b): „Diagnostic tests 2: Predictive values.“. In: *BMJ: British Medical Journal*. 309 (6947), S. 102.

Andersen, Erling B. (1997): *Introduction to the statistical analysis of categorical data*. Berlin ; New York: Springer. — ISBN: 3-540-62399-X

Backhaus, Klaus; Erichson, Bernd; Plinke, Wulff; u. a. (Hrsg.) (2016): *Multivariate Analysemethoden: eine anwendungsorientierte Einführung*. 14., überarb. und aktualisierte Aufl. Berlin: Springer Gabler (Lehrbuch). — ISBN: 978-3-662-46075-7

Bangdiwala, Shrikant I (1985): „A graphical test for observer agreement“. In: *Contributed papers of the 45th and centenary session of the International Statistical Institute*. Amsterdam, S. 307–8.

Bangdiwala, Shrikant I.; Haedo, Ana S.; Natal, Marcela L.; u. a. (2008): „The agreement chart as an alternative to the receiver-operating characteristic curve for diagnostic tests“. In: *Journal of Clinical Epidemiology*. 61 (9), S. 866–874, DOI: 10.1016/j.jclinepi.2008.04.002.

Bangdiwala, Shrikant I; Shankar, Viswanathan (2013): „The agreement chart“. In: *BMC Medical Research Methodology*. 13 (1), DOI: 10.1186/1471-2288-13-97.

Bhapkar, V. P. (1966): „A Note on the Equivalence of Two Test Criteria for Hypotheses in Categorical Data“. In: *Journal of the American Statistical Association*. 61 (313), S. 228–235, DOI: 10.2307/2283057.

Biggerstaff, Brad J. (2000): „Comparing diagnostic tests: a simple graphic using likelihood ratios“. In: *Statistics in Medicine*. 19 (5), S. 649–663, DOI: 10.1002/(SICI)1097-0258(20000315)19:5<649::AID-SIM371>3.0.CO;2-H.

Bortz, Jürgen; Lienert, Gustav A. (2008): *Kurzgefasste Statistik für die klinische Forschung: Leitfaden für die verteilungsfreie Analyse kleiner Stichproben*. 3., aktualisierte u. bearb. Aufl. Springer, Berlin. — ISBN: 3-540-75737-6

Bortz, Jürgen; Lienert, Gustav A.; Boehnke, Klaus (2008): *Verteilungsfreie Methoden in der Biostatistik*. 3., korrigierte Aufl. Heidelberg: Springer. — ISBN: 978-3-540-74706-2

Bortz, Jürgen; Weber, René (2005): *Statistik für Human- und Sozialwissenschaftler: mit 242 Tabellen*. 6., vollst. überarb. und aktualisierte Aufl. Heidelberg: Springer Medizin (Springer-Lehrbuch). — ISBN: 978-3-540-21271-3

Brasil, Pedro (2010): *DiagnosisMed: Diagnostic test accuracy evaluation for medical professionals.* o.V.

Breiman, Leo; Friedman, R. A.; Olshen, R. A.; u. a. (1983): *Classification and Regression Trees*. Wadsworth Publishing Co Inc. — ISBN: 0-534-98053-8

Brenner, Hermann; Gefeller, Olaf (1994): „Chance-corrected measures of the validity of a binary diagnostic test“. In: *Journal of Clinical Epidemiology*. 47 (6), S. 627–633, DOI: 10.1016/0895-4356(94)90210-0.

Brinkmann, Burghart; Deutsches Institut für Normung (Hrsg.) (2012): *Internationales Wörterbuch der Metrologie: grundlegende und allgemeine Begriffe und zugeordnete Benennungen (VIM); deutsch-englische Fassung ISO/IEC-Leitfaden 99:2007 = Vocabulaire international de métrologie = International vocabulary of metrology*. 4. Aufl., Fassung 2012. Berlin: Beuth (Wissen Messwesen). — ISBN: 978-3-410-22472-3

Cantor, Scott B.; Kattan, Michael W. (2000): „Determining the Area under the ROC Curve for a Binary Diagnostic Test“. In: *Medical Decision Making*. 20 (4), S. 468–470, DOI: 10.1177/0272989X0002000410.

Chmura Kraemer, Helena; Periyakoil, Vyjeyanthi S; Noda, Art (2002): „Kappa coefficients in medical research“. In: *Statistics in Medicine*. 21 (14), S. 2109–2129, DOI: 10.1002/sim.1180.

Cicchetti, D V; Feinstein, A R (1990): „High agreement but low kappa: II. Resolving the paradoxes“. In: *Journal of Clinical Epidemiology*. 43 (6), S. 551–558.

Cohen, Ayala (1980): „On the graphical display of the significant components in two-way contingency tables“. In: *Communications in Statistics - Theory and Methods*. 9 (10), S. 1025–1041, DOI: 10.1080/03610928008827940.

Cohen, Jacob (1960): „A Coefficient of Agreement for Nominal Scales“. In: *Educational and Psychological Measurement*. 20 (1), S. 37–46, DOI: 10.1177/001316446002000104.

Conger, Anthony J. (1980): „Integration and generalization of kappas for multiple raters.“. In: *Psychological Bulletin*. 88 (2), S. 322–328.

Coughlin, S. S.; Pickle, L. W. (1992): „Sensitivity and specificity-like measures of the validity of a diagnostic test that are corrected for chance agreement“. In: *Epidemiology (Cambridge, Mass.)*. 3 (2), S. 178–181.

Cunningham, Michael (2009): „More than Just the Kappa Coefficient: A Program to Fully Characterize Inter-Rater Reliability between Two Raters“.

Deeks, Jonathan J.; Altman, Douglas G. (2004): „Statistics Notes: Diagnostic Tests 4: Likelihood Ratios“. In: *BMJ: British Medical Journal*. 329 (7458), S. 168–169, DOI: 10.2307/25468683.

Dujardin, Bruno; Ende, Jef Van den; Gompel, Alfons Van; u. a. (1994): „Likelihood Ratios: A Real Improvement for Clinical Decision Making?“. In: *European Journal of Epidemiology*. 10 (1), S. 29–36, DOI: 10.2307/3520897.

Everitt, Brian (1986): *The analysis of contingency tables*. London; New York: Chapman and Hall (Monographs on Applied Probability and Statistics). — ISBN: 978-0-412-14970-2

Fagan (1975): „Nomogram for Bayes's Theorem“. In: *New England Journal of Medicine*. 293 (5), S. 257–257, DOI: 10.1056/NEJM197507312930513.

Fawcett, Tom (2006): „An introduction to ROC analysis“. In: *Pattern Recognition Letters*. (ROC Analysis in Pattern Recognition) 27 (8), S. 861–874, DOI: 10.1016/j.patrec.2005.10.010.

Fay, Michael P. (2011): „Exact McNemar's Test and Matching Confidence Intervals“.

Feldman, B. (2007): „A theory of attribution“. In: *MPRA paper*. 3349 .

Feuerman, Martin; Miller, Allen R (2008): „Relationships between statistical measures of agreement: sensitivity, specificity and kappa“. In: *Journal of Evaluation in Clinical Practice*. 14 (5), S. 930–933, DOI: 10.1111/j.1365-2753.2008.00984.x.

Fienberg, Stephen E. (2007): *The analysis of cross-classified categorical data*. 2nd ed. New York, NY: Springer. — ISBN: 978-0-387-72824-7

Fleiss, J.L. (1971): „Measuring nominal scale agreement among many raters“. In: *Psychological Bulletin*. (76), S. 378–382.

Fleiss, Joseph L. (1981): *Statistical methods for rates and proportions*. 2d ed. New York: Wiley (Wiley series in probability and mathematical statistics). — ISBN: 978-0-471-06428-2

Fleiss, Joseph L. (2003): *Statistical methods for rates and proportions*. 3rd ed. Hoboken, N.J: J. Wiley (Wiley series in probability and statistics). — ISBN: 978-0-471-52629-2

Friendly, Michael (1992): „Graphical methods for categorical data“. In: *SAS User Group International Conference Proceedings*. Citeseer, S. 190–200.

Friendly, Michael (2000): *Visualizing Categorical Data*. SAS Publishing. — ISBN: 1-58025-660-0

Fulton, Lawrence; Mendez, Francis; Bastian, Nathaniel; u. a. (2012): „Confusion Between Odds and Probability, a Pandemic?“. In: *Journal of Statistics Education*. 20 (3).

Gallagher, E J (1998): „Clinical utility of likelihood ratios“. In: *Annals of emergency medicine*. 31 (3), S. 391–397.

Gallagher, E. John (2005): „Numeric Instability of Predictive Values“. In: *Annals of Emergency Medicine*. 46 (4), S. 311–313, DOI: 10.1016/j.annemergmed.2005.05.011.

Gamer, Matthias; Lemon, Jim; <puspendra.pusp22@gmail.com>, Ian Fellows Puspendra Singh (2012): *irr: Various Coefficients of Interrater Reliability and Agreement*. o.V.

Gamer, Matthias; Lemon, Jim; Fellows, Ian (2007): „The irr Package“.

Gefeller, O.; Brenner, H. (1994): „How to correct for chance agreement in the estimation of sensitivity and specificity of diagnostic tests.“. In: *Methods Archive*. 33 , S. 180–186.

Glas, Afina S; Lijmer, Jeroen G; Prins, Martin H; u. a. (2003): „The diagnostic odds ratio: a single indicator of test performance“. In: *Journal of clinical epidemiology*. 56 (11), S. 1129–1135.

Graham, P; Bull, B (1998): „Approximate standard errors and confidence intervals for indices of positive and negative agreement“. In: *Journal of Clinical Epidemiology*. 51 (9), S. 763–771.

Groemping, Ulrike (2006): „Relative Importance for Linear Regression in R: The Package relaimpo“. In: *Journal of Statistical Software*. 17 (1), S. 1–27.

Gwet, Kilem (2002): „Inter-rater reliability: dependency on trait prevalence and marginal homogeneity“. In: *Statistical Methods for Inter-Rater Reliability Assessment Series*. 2 , S. 1–9.

Gwet, Kilem Li (2014): *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters; [a handbook for researchers, practitioners, teachers & students]*. 4. ed. Gaithersburg, MD: Advanced Analytics, LLC. — ISBN: 978-0-9708062-8-4

Hallgren, Kevin A. (2012): „Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial“. In: *Tutorials in quantitative methods for psychology*. 8 (1), S. 23–34.

Handl, Andreas; Kuhlenkasper, Torben (2016): „Einführung in die Statistik mit R“.

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer-Verlag Gmbh. — ISBN: 0-387-84857-6

Hayden, Stephen R; Brown, Michael D (1999): „Likelihood Ratio: A Powerful Tool for Incorporating the Results of a Diagnostic Test Into Clinical Decisionmaking“. In: *Annals of Emergency Medicine*. 33 (5), S. 575–580, DOI: 10.1016/S0196-0644(99)70346-X.

Hedderich, Jürgen; Sachs, Lothar (2016): *Angewandte Statistik: Methodensammlung mit R*. 15., überarbeitete und erweiterte Auflage. Berlin Heidelberg: Springer Spektrum. — ISBN: 978-3-662-45690-3

Herich, L.; Lehmacher, W.; Hellmich, M. (2015): „Drop the Likelihood Ratio: A Novel Non-electronic Tool for Interpreting Diagnostic Test Results“. In: *Methods of Information in Medicine*. 54 (3), S. 283–287, DOI: 10.3414/ME14-01-0091.

Hilgers, Ralf-Dieter; Schreiber, Viktor; Bauer, Peter (2007): *Einführung in die Medizinische Statistik*. Berlin, Heidelberg: Springer Berlin Heidelberg. — ISBN: 978-3-540-33943-4

Hoehler, Fred K. (2000): „Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity“. In: *Journal of Clinical Epidemiology*. 53 (5), S. 499–503, DOI: 10.1016/S0895-4356(99)00174-2.

Holle, Rolf; Windeler, Jürgen (1997): „Is there a gain from “chance-corrected” measures of diagnostic validity?“. In: *Journal of Clinical Epidemiology*. 50 (1), S. 117–120, DOI: 10.1016/S0895-4356(96)00311-3.

Hout, Wilbert B. van den (2003): „The Area under an ROC Curve with Limited Information“. In: *Medical Decision Making*. 23 (2), S. 160–166, DOI: 10.1177/0272989X03251246.

Hripcsak, George; Heitjan, Daniel F. (2002): „Measuring agreement in medical informatics reliability studies“. In: *Journal of Biomedical Informatics*. 35 (2), S. 99–110, DOI: 10.1016/S1532-0464(02)00500-2.

Jamart, J. (1998): „A comment on Gefeller and Brenner’s chance-corrected sensitivity and specificity“. In: *Methods of Information in Medicine*. 37 (3), S. 307-308-310.

Jamart, J. (1992): „Chance-corrected sensitivity and specificity for three-zone diagnostic tests“. In: *Journal of Clinical Epidemiology*. 45 (9), S. 1035–1039.

James, Gareth; Witten, Daniela; Hastie, Trevor; u. a. (Hrsg.) (2013): *An introduction to statistical learning: with applications in R*. New York: Springer (Springer texts in statistics). — ISBN: 978-1-4614-7137-0

Jeske, Roland (2001): „Online Statistik - Universität Konstanz“. *Online Statistik*. Abgerufen am 24.02.2012 von http://www.uni-konstanz.de/FuF/wiwi/heiler/os/index.html.

Joint Committee for Guides in Metrology (JCGM) (2012): *International Vocabulary of Metrology (VIM) – Basic and general concepts and associated terms*. 3rd edition. Joint Committee for Guides in Metrology (JCGM).

Kabacoff, Robert I. (2014): „Quick-R: Home Page“. *Quick-R - accessing the power of R*. Abgerufen am 20.06.2016 von http://www.statmethods.net/.

Kabacoff, Robert I. (2011): *R in Action: Data Analysis and Graphics with R*. Manning. — ISBN: 1-935182-39-0

Kateri, Maria (2014): *Contingency table analysis: methods and implementation using R*. New York: Springer. — ISBN: 978-0-8176-4810-7

Keefe, Thomas J. (1982): „On the relationship between two tests for homogeneity of the marginal distributions in a two-way classification". In: *Biometrika*. 69 (3), S. 683–684, DOI: 10.1093/biomet/69.3.683.

Kraemer, Helena Chmura (1992): *Evaluating medical tests: objective and quantitative guidelines*. Newbury Park, California: Sage Publications. — ISBN: 0-8039-4611-2

Kraemer, Helena Chmura (1985): „The Robustness of Common Measures of 2 × 2 Association to Bias Due to Misclassifications". In: *The American Statistician*. 39 (4), S. 286–290, DOI: 10.2307/2683705.

Krummenauer, Frank (2003): „VII: Diagnosestudien: Einfache Maße für Validität und Reliabilität". In: *Klin Monatsbl Augenheilkd*. (4), S. 281–283.

Kuckartz, Udo; Rädiker, Stefan; Ebert, Thomas; u. a. (2013): *Statistik: eine verständliche Einführung*. 2., überarbeitete Auflage. Wiesbaden: Springer VS (Lehrbuch). — ISBN: 978-3-531-19889-7

Landis, J. Richard; Koch, Gary G. (1977): „The Measurement of Observer Agreement for Categorical Data". In: *Biometrics*. 33 (1), S. 159–174.

Langenbucher, J.; Labouvie, E.; Morgenstern, J. (1996): „Measuring diagnostic agreement.". In: *Journal of consulting and clinical psychology*. 64 (6), S. 1285.

Lantz, Brett (2013): *Machine learning with R learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications*. Birmingham, UK: Packt Publishing. — ISBN: 978-1-4619-4965-7

Li, Jialiang; Fine, Jason P.; Safdar, Nasia (2007): „Prevalence-dependent diagnostic accuracy measures". In: *Statistics in Medicine*. 26 (17), S. 3258–3273, DOI: 10.1002/sim.2812.

Light, Richard J. (1971): „Measures of response agreement for qualitative data: Some generalizations and alternatives.". In: *Psychological Bulletin*. 76 (5), S. 365–377.

Lindeman, Richard Harold; Merenda, Peter Francis; Gold, Ruth Z (1980): *Introduction to bivariate and multivariate analysis*. Glenview, Ill.: Scott, Foresman. — ISBN: 0-673-15099-2

Maxwell, A. E. (1970): „Comparing the Classification of Subjects by Two Independent Judges". In: *The British Journal of Psychiatry*. 116 (535), S. 651–655, DOI: 10.1192/bjp.116.535.651.

McGee, Steven (2002): „Simplifying Likelihood Ratios". In: *Journal of General Internal Medicine*. 17 (8), S. 647–650, DOI: 10.1046/j.1525-1497.2002.10750.x.

McNemar, Quinn (1947): „Note on the sampling error of the difference between correlated proportions or percentages". In: *Psychometrika*. 12 (2), S. 153–157, DOI: 10.1007/BF02295996.

Meyer, David; Zeileis, Achim; Hornik, Kurt (2006): „The Strucplot Framework: Visualizing Multi-way Contingency Tables with vcd". In: *Journal of Statistical Software*. 17 (3), S. 1–48.

Meyer, David; Zeileis, Achim; Hornik, Kurt (2003): „Visualizing independence using extended association plots". In: *Proceedings of DSC 2003*.

Munoz, Sergio R.; Bangdiwala, Shrikant I. (1997): „Interpretation of Kappa and B statistics measures of agreement". In: *Journal of Applied Statistics*. 24 (1), S. 105–112, DOI: 10.1080/02664769723918.

Nardini, Christine; Liu, Yuanhua (2014): *Multiclasstesting: Performance of N-ary classification testing*. o.V.

Pearson, Karl (1904): *On the theory of contingency and its relation to association and normal correlation.* London: Dulau and Co. (Drapers' Company research memoirs.).

Powers, David Martin (2011): „Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation". In.:

R-Project (2016): „CRAN: Contributed Documentation". *Contributed Documentation*. Abgerufen am 07.03.2011 von http://cran.r-project.org/other-docs.html.

Rudas, Tamás (1998): *Odds ratios in the analysis of contingency tables*. Thousand Oaks: Sage Publications (Sage university papers series). — ISBN: 978-0-7619-0362-8

Sachs, Lothar; Hedderich, Jürgen (2006): *Angewandte Statistik: Methodensammlung mit R*. 12., vollst. neu bearb. Aufl. Berlin; Heidelberg [u.a.]: Springer, Berlin. — ISBN: 978-3-540-32160-6

Spector, Phil (2008): *Data Manipulation with R*. o.V. (Use R!), DOI: 10.1007/978-0-387-74731-6.

Spitalnic, Stuart (2004a): „Test Properties 2: Likelihood Ratios, Bayes' Formula, and Receiver Operating Characteristic Curves". In: *Hospital Physician*. 40 (10), S. 53–58.

Spitalnic, Stuart (2004b): „Test Properties I: Sensitivity, Specificity, and Predictive Values". In: *Hospital Physician*. 40 (9), S. 27–31.

Steland, Ansgar (2004): *Mathematische Grundlagen der empirischen Forschung*. Berlin: Springer (Statistik und ihre Anwendungen). — ISBN: 978-3-540-03700-2

Stuart, Alan (1955): „A Test for Homogeneity of the Marginal Distributions in a Two-Way Classification". In: *Biometrika*. 42 (3–4), S. 412–416, DOI: 10.1093/biomet/42.3-4.412.

Therneau, Terry M.; Atkinson, Beth; Ripley, Brian (2012): *rpart: Recursive Partitioning*. o.V.

Uebersax, John (2006): „McNemar Tests of Marginal Homogeneity". Abgerufen am 10.08.2014 von http://www.john-uebersax.com/stat/mcnemar.htm.

Uebersax, John (2014): „Raw Agreement Indices". Abgerufen am 01.12.2015 von http://www.john-uebersax.com/stat/raw.htm#binspe.

Uebersax, John S. (2010): „Statistical Methods for Rater and Diagnostic Agreement". *Statistical Methods for Rater and Diagnostic Agreement*. Abgerufen am 02.09.2010 von http://www.john-uebersax.com/stat/agree.htm.

Viera, A. J.; Garrett, J. M. (2005): „Understanding interobserver agreement: the kappa statistic". In: *Fam Med*. 37 (5), S. 360–363.

Walter, S. D. (2001): „Hoehler's adjusted kappa is equivalent to Yule's Y". In: *Journal of Clinical Epidemiology*. 54 (10), S. 1072–1072.

Weiß, Christel (2013): *Basiswissen Medizinische Statistik: mit 20 Tabellen; [mit Epidemiologie]*. 6., überarb. Aufl. Berlin: Springer (Springer-Lehrbuch). — ISBN: 978-3-642-34260-8

Westover, M Brandon; Westover, Kenneth D; Bianchi, Matt T (2011): „Significance testing as perverse probabilistic reasoning". In: *BMC Medicine*. 9 , S. 20, DOI: 10.1186/1741-7015-9-20.

Wikipedia (2016): „Positive and negative predictive values". *Wikipedia, the free encyclopedia*.

Wirtz, M.; Kutschmann, M. (2007): „Analyse der Beurteilerübereinstimmung für kategoriale Daten mittels Cohens Kappa und alternativer Maße". In: *Rehabilitation*. 46 (06), S. 370–377, DOI: 10.1055/s-2007-976535.

Wirtz, Markus; Caspar, Franz (2002): *Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Hogrefe-Verlag. — ISBN: 3-8017-1646-5

Youden, W J (1950): „Index for rating diagnostic tests". In: *Cancer*. 3 (1), S. 32–35.

Yule, G. Udny (1900): „On the Association of Attributes in Statistics: With Illustrations from the Material of the Childhood Society, &c". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*. 194 , S. 257–319, DOI: 10.2307/90759.

Yule, G. Udny (1912): „On the Methods of Measuring Association Between Two Attributes". In: *Journal of the Royal Statistical Society*. 75 (6), S. 579–652, DOI: 10.2307/2340126.

Zhou, Xiao-hua (2011): *Statistical methods in diagnostic medicine*. 2nd ed. Hoboken, N.J: Wiley (Wiley series in probability and statistics). — ISBN: 978-0-470-18314-4

4 Ergebnisse

Kapitelverzeichnis

Die Darstellung der Ergebnisse orientiert sich inhaltlich an der Gliederung der Fragebögen.

- Der **Arztfragebogen** erfragte Hintergrundinformationen zur Person des Untersuchers, dessen Ausbildung und Erfahrung.
- Mit dem **Befundfragebogen** wurden die Videomitschnitte der Bronchoskopien in einem einheitlichen Format beurteilt und die subjektiv empfundene Qualität der Videomitschnitte erfasst.

Formal werden die auf den Angaben der Fragebögen basierenden Ergebnisse nach den Aspekten

- **Befundverteilung** (deskriptive Statistik)
- **Konkordanz** bzw. Präzision (relative Befundrichtigkeit innerhalb der Befunder) und
- **Richtigkeit** (absolute Befundrichtigkeit relativ zum Goldstandard)

besprochen. Innerhalb jedes Abschnittes werden zunächst die Befunde einzeln im Sinne unabhängiger Symptome beschrieben, dann als Kombinationsbefund im Sinne eines zusammengehörigen Syndroms. Die Abschnitte sind dabei wie folgt gegliedert:

- **Einzelbefunde**
  - tabellarische Ergebnisse
  - Beschreibung von Befundverteilung, Präzision, Richtigkeit
  - aufbereitete Quelldaten zur Befundrichtigkeit
    - Kontingenztafeln Einzelbefunde
    - Kennwerte der Kontingenztafeln
    - Diagramme zu den Kontingenztafeln: Bangdiwaladiagramm & ROC
- **Kombinationsbefunde**
  - tabellarische Ergebnisse
  - Beschreibung von Befundverteilung, Präzision, Richtigkeit
  - aufbereitete Quelldaten Befundkombinationen
    - Kontingenztafeln überschneidende Befundklassen
    - Kennwerte der Kontingenztafeln
    - Diagramme zu den Kontingenztafeln: Assoziations- & Bangdiwaladiagramm
- **Zusammenfassung**
  kurze Zusammenfassung der wichtigsten Ergebnisse

Der Abschnitt **Einflussgrößen der Befundrichtigkeit** setzt Arzt- und Befundfragebogen zueinander in Relation: Er versucht Unterschiede in Ausbildung und Erfahrung der Untersucher festzustellen, die möglicherweise als Erklärung für die unterschiedlich gute Übereinstimmung mit den Referenzbefunden des Goldstandards herangezogen werden können. Auf diese Weise sollen Anhaltspunkte dafür gesammelt werden, welche Faktoren für eine gute Ausbildung in pädiatrischer Bronchoskopie ausschlaggebend sein könnten.

4.1 Arztfragebogen

Der Arztfragebogen charakterisiert die Gruppe der Untersucher, die an dieser Studie teilgenommen haben. Neben grundlegenden

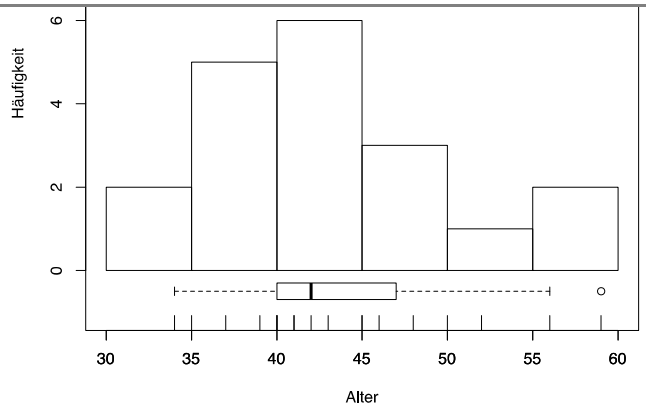
- **demographischen Daten** wurden Informationen zu
- **Ausbildung** und
- **Erfahrung**

eingeholt. Vorrangiges Ziel war dabei die Erfassung möglicher Einflussfaktoren der Befundrichtigkeit.

4.1.1 Alter

Die Untersucher waren zum Zeitpunkt dieser Studie zwischen 34 und 59 Jahre alt. Der Mittelwert betrug 43 Jahre, der Median lag bei 42 Jahren. Mehr als die Hälfte der Untersucher waren zwischen 35 und 45 Jahre alt.

Abbildung 4.1: Alter der Befunder



Altersverteilung der Befunder als Histogramm, Boxplot und Strichverteilung

4.1.2 Qualifikation

Nahezu alle Untersucher (18 von 20) hatten zum Zeitpunkt der Studie bereits die Anerkennung eines Facharztes für Kinderheilkunde erworben, 1 Untersucher hatte noch keine Facharztprüfung abgelegt, 1 Untersucher machte keine Angaben zu seinem Status.

4.1.3 Ausbildung

Neben Dauer und Ausbildungsort erfragt der Arztfragebogen in Bezug auf die Ausbildung

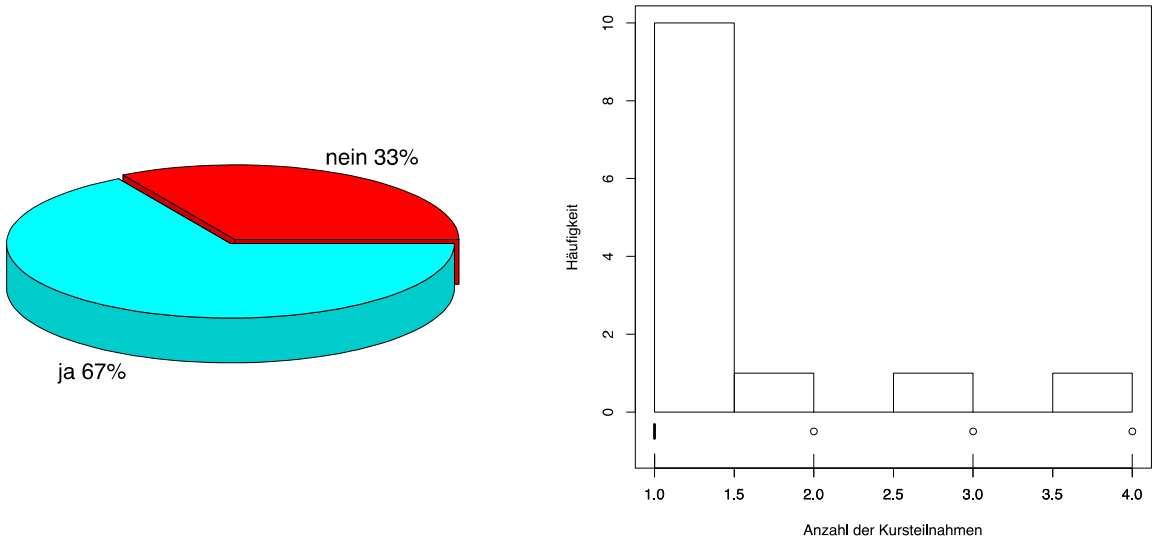
- Kursteilnahmen
- Hospitationen und die
- Anzahl der Untersuchungen

die während der Ausbildung in den einzelnen Disziplinen der Bronchoskopie (flexible, starr, interventionell) absolviert wurden.

4.1.3.1 Kursteilnahme

Zwei Drittel aller Befunder haben an einem Bronchoskopiekurs teilgenommen. Abgesehen von wenigen Ausnahmen wurde nur ein einziger Kurs besucht.

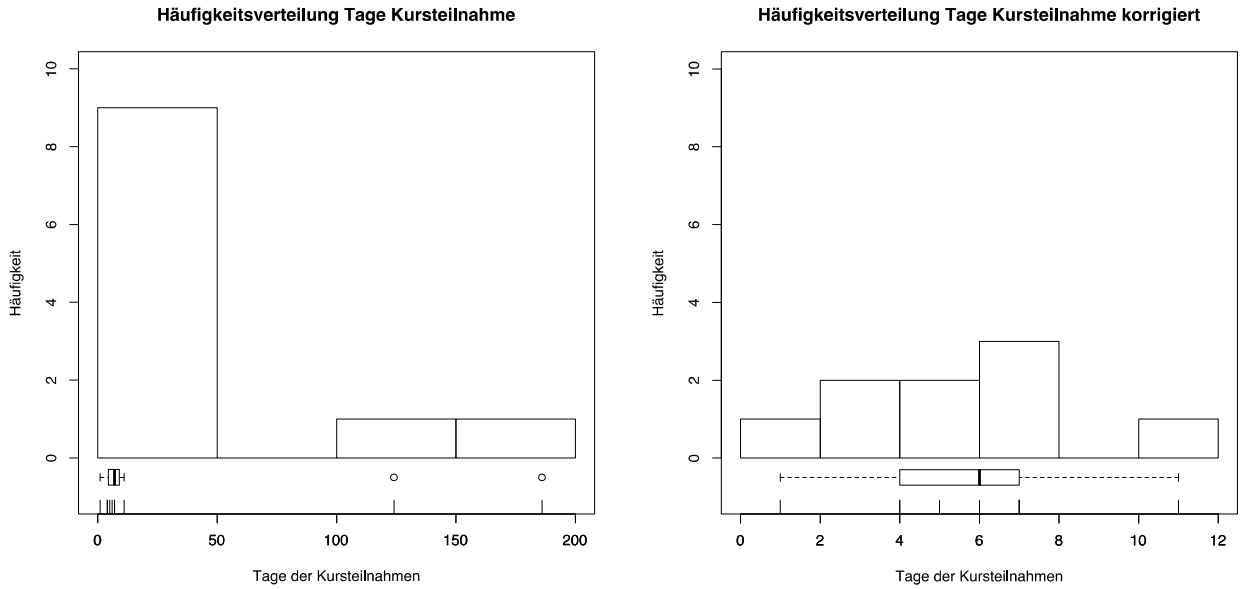
Abbildung 4.2: Kursteilnahme



Links: Anteil der Befunder mit Kursteilnahme. Rechts: Anzahl der Kurse im Balkendiagramm.

Bezieht man zwei mit 124 bzw. 186 Tagen extreme Angaben zur Kursdauer (jeweils nur 1 Kurs!) mit ein, ergibt sich eine mittlere Kursdauer von 33 Tagen bei einem Median von 7 Tagen (Abbildung 4.3 links). Nach Entfernen dieser Ausreißer entsteht ein deutlich ausgewogeneres Bild, bei dem Mittelwert und Median mit 6 Tagen aufeinander treffen (Abbildung 4.3 rechts).

Abbildung 4.3: Dauer der Kursteilnahme



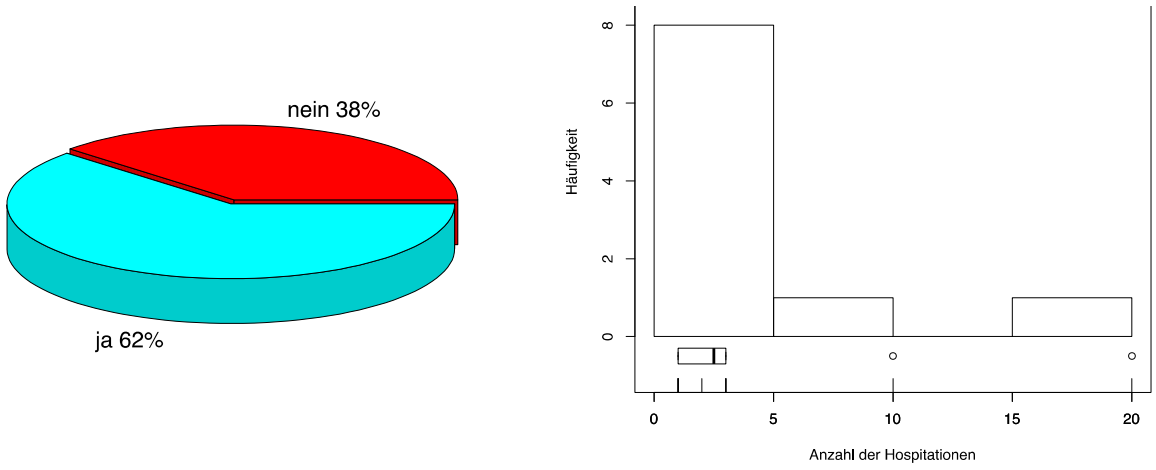
Links: Dauer der Kurse. Rechts: Dauer der Kurse nach Entfernen zweier extremer Angaben (Ausreißer).

4.1.3.2 Hospitationen

Knapp 2/3 der Befunder absolvierte im Rahmen der Ausbildung Hospitationen. Lässt man den Extremwert von 110 Hospitationen unberücksichtigt, haben die Befunder im Median an 2,5 Hospitationen teilgenommen (Mittel 4,5). Dabei setzten sich zwei Befunder mit 10 respektive 20 Hospitationen deutlich vom übrigen Feld ab. Im Median dauerten Hospitationen 28 Tage (Mittel 60). Dabei wurden zwei mit 620 bzw. 730 Tagen extreme Angaben in die Berechnung nicht mit einbezogen. Auch nach Entfernen dieser „Ausreißer“ besteht bei der Hospitationsdauer immer noch eine Variation zwischen 3 und 186 Tagen, wobei sich die meisten Hospitationen insgesamt auf

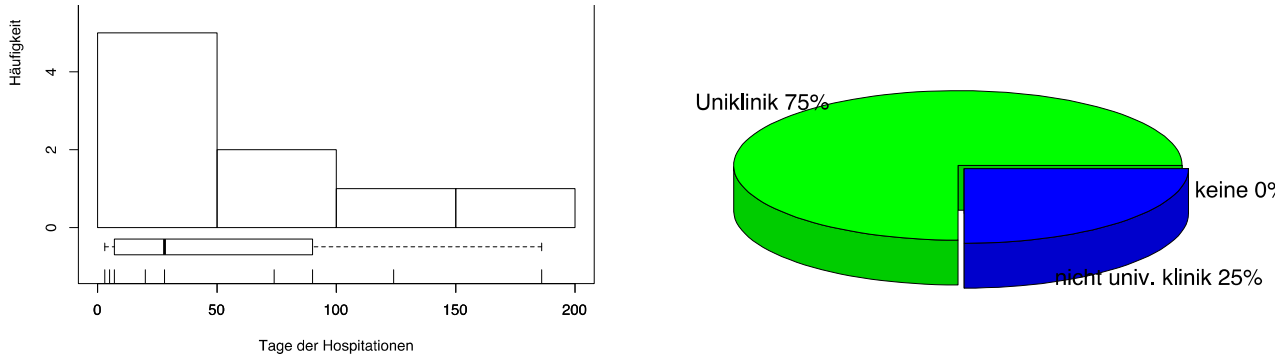
nicht mehr als 50 Tage summieren. Drei Viertel aller Hospitationen fanden an Universitätskliniken statt.

Abbildung 4.4: Hospitationen



Links: Anteil der Befunder mit Hospitationen. Rechts: Anzahl der Hospitationen als Balkendiagramm.

Abbildung 4.5: Hospitationen

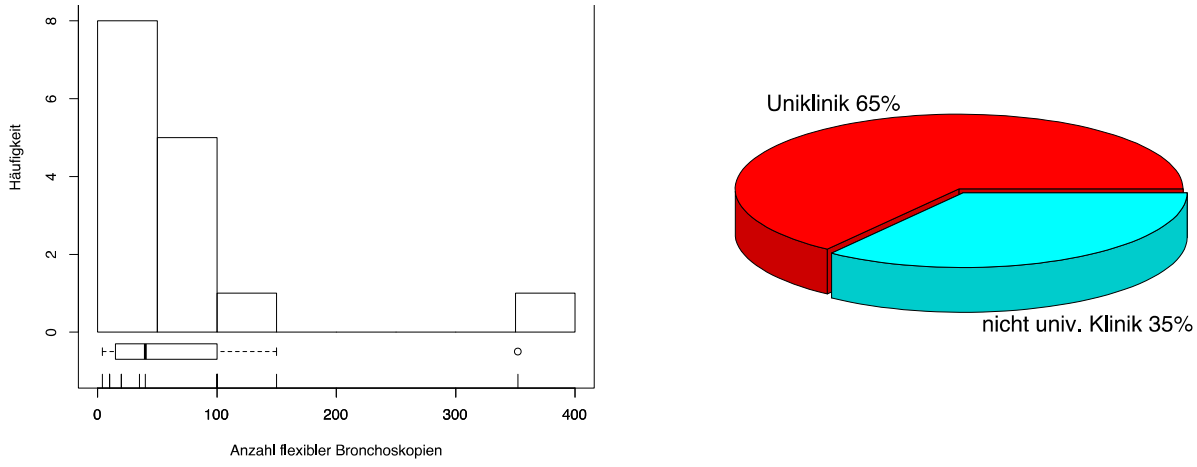


Links: Balkendiagramm der Hospitationsdauer. Rechts: Anteil der Hospitationen an Universitätskliniken.

4.1.3.3 Bronchoskopien in der Ausbildung

Im Median dauerte die Ausbildung in pädiatrischer Bronchoskopie 2 Jahre, im Mittel etwas mehr als 4 Jahre. Zwei Drittel der Teilnehmer an dieser Studie haben ihre Ausbildung an einer Universitätsklinik absolviert.

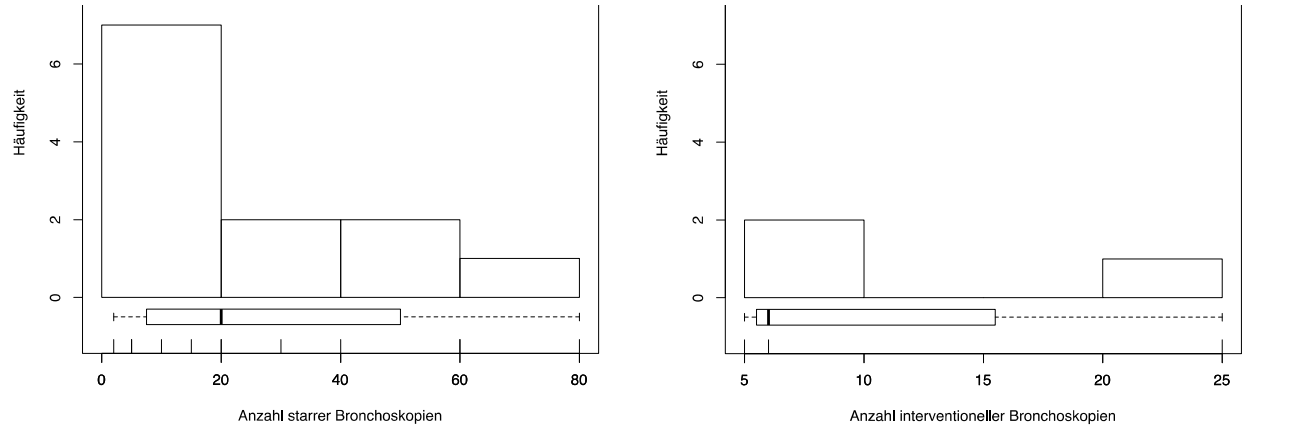
Abbildung 4.6: Ausbildung in flexibler Bronchoskopie



Links: Anzahl der flexiblen Bronchoskopien in der Ausbildung. Rechts: Klinikart der Ausbildung.

Alle Befunder wurden in flexibler Bronchoskopie ausgebildet und haben im Median während der Ausbildung 40 flexible Bronchoskopien durchgeführt, im Mittel 76. Nur 3 Befunder absolvierten über 100 Bronchoskopien in der Ausbildung. Das Minimum waren 4 flexible Bronchoskopien.

Abbildung 4.7: Ausbildung in starrer und interventioneller Bronchoskopie



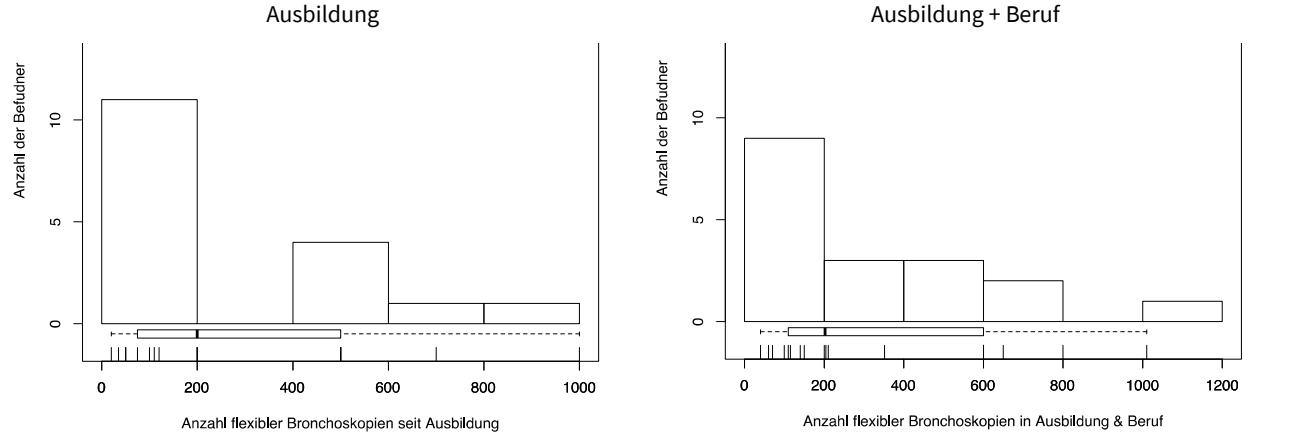
Anzahl starrer und interventioneller Bronchoskopien in der Ausbildung.

Zwei Drittel der Befunder wurden in starrer Bronchoskopie ausgebildet. Im Median wurden dabei 20, im Mittel 42 starre Bronchoskopien absolviert. Eine praktische Anleitung in interventioneller Bronchoskopie erhielten 3 von 20 Befundern. Der Median der starren Bronchoskopien in der Ausbildung liegt bei 6.

4.1.4 Erfahrung

Drei Viertel der Befunder dieser Studie erwarben Ihre praktische Berufserfahrung vorwiegend an einer Universitätsklinik. Von 18 Befundern, die Angaben zu Ihrer Erfahrung machten, gaben alle an, Erfahrung in flexibler Bronchoskopie gesammelt zu haben. Im Median wurden dabei 200 flexible Bronchoskopien seit der Ausbildung absolviert. Berechnet man die Zahl der Bronchoskopien während der Ausbildung in die Erfahrung mit ein, verschiebt sich die Verteilung nur unwesentlich. Der Median liegt weiterhin bei 200, das Mittel steigt von 285 auf 334.

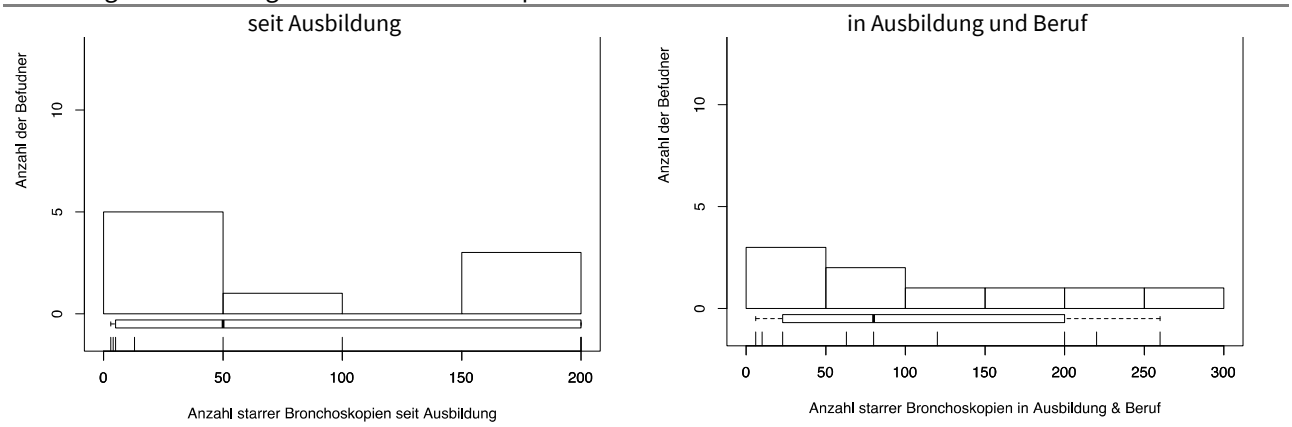
Abbildung 4.8: Erfahrung in flexibler Bronchoskopie



Links: Erfahrung in flexibler Bronchoskopie seit der Ausbildung. Rechts: Ausbildung und berufliche Erfahrung zusammen.

Nur knapp die Hälfte der Befunder verfügt über Erfahrung in starrer Bronchoskopie. Der Median für die Erfahrung in starren Bronchoskopien liegt bei 50.

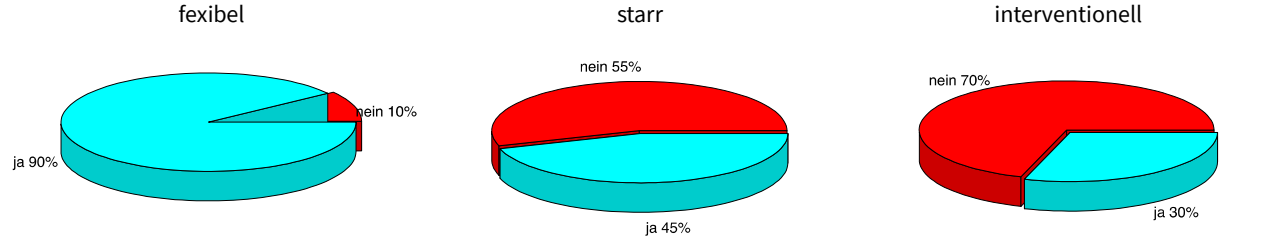
Abbildung 4.9: Erfahrung in starrer Bronchoskopie



Links: Erfahrung in starrer Bronchoskopie seit der Ausbildung. Rechts: Ausbildung und berufliche Erfahrung zusammen.

Anders als bei der flexiblen Bronchoskopie verschiebt sich die Verteilung bei der starren Bronchoskopie durch Einbeziehung der Bronchoskopien in der Ausbildung: Der Median steigt auf 80 Untersuchungen, die Verteilung ist erkennbar homogener.

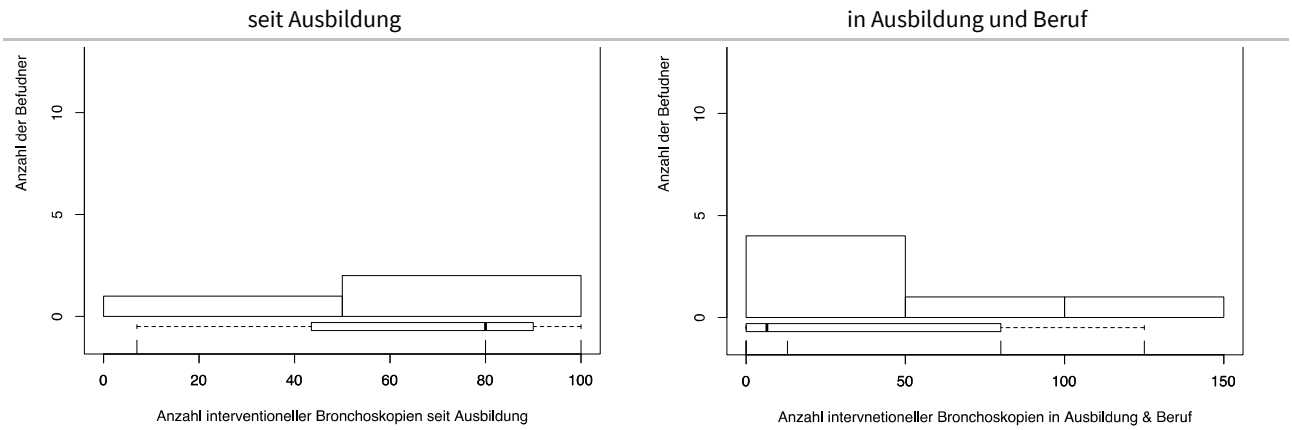
Abbildung 4.10: Erfahrung in Disziplinen der Bronchoskopie



Erfahrung in den verschiedenen Disziplinen der Bronchoskopie.

Erfahrung in interventioneller Bronchoskopie haben weniger als ein Drittel der Befunder. Von 6 Befundern, die Erfahrung in interventioneller Bronchoskopie angegeben haben, machten 3 Angaben zur Anzahl seit der Ausbildung, sodass kein seriöser Mittelwert angegeben werden kann.

Abbildung 4.11: Erfahrung in interventioneller Bronchoskopie



Links: Erfahrung in interventioneller Bronchoskopie seit der Ausbildung. Rechts: Ausbildung und berufliche Erfahrung zusammen.

Zusammenfassung 4.1: Arztfragebogen

Die 20 Teilnehmer dieser Studie rekrutierten sich aus der Arbeitsgruppe Bronchoskopie der Gesellschaft für pädiatrische Pneumologie (GPP). Sie waren zum Zeitpunkt der Befundung bei einer Spannbreite von 34 bis 59 Jahren im Median 42 Jahre alt und hatten fast ausschließlich bereits die Facharztqualifikation erreicht. Knapp zwei Drittel (62 %) absolvierten in ihrer Ausbildung im Median 2,5 Hospitationen. Sie dauerten im Median 28 Tage und fanden zu 75 % an einer Universi-

tätsklinik statt. Ebenfalls zwei Drittel absolvierten einen Bronchoskopiekurs von im Mittel 6 Tagen Dauer. Wiederum zwei Drittel der Studienteilnehmer wurden an einer Universitätsklinik ausgebildet. In einer medianen Ausbildungszeit von 2 Jahren wurden im Median 40 flexible und 20 starre Bronchoskopien durchgeführt wurden. Drei Viertel der Befunder erwarben ihre praktische Berufserfahrung an einer Universitätsklinik und haben nach der Ausbildung im Median 200 flexible Bronchoskopien durchgeführt. Etwa die Hälfte der Befunder sammelte auch Erfahrung in starrer Bronchoskopie mit im Median 50 Eingriffen.

4.2 Befundfragebogen

Zusätzlich zum medizinischen Befund wurde am Anfang des Befundfragebogens eine subjektive Einschätzung der Qualität des vorgelegten Videomaterials erhoben. Der daran anschließende bronchoskopische Befund, ist thematisch in die Abschnitte

- Stenosen
- Schleimhaut
- Entzündung
- Malazie
- Pulsationen und
- Kompressionen

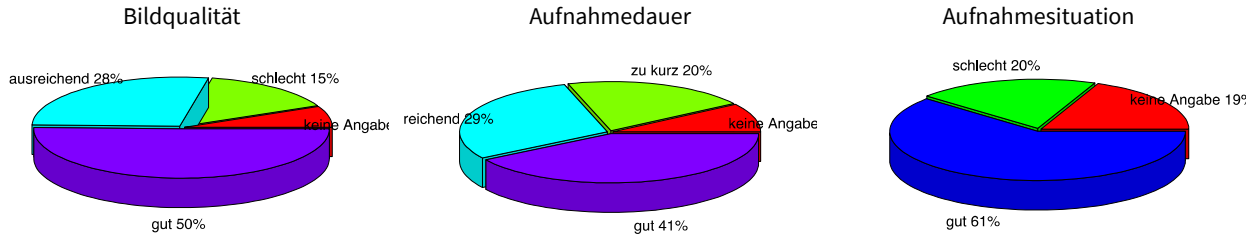
gegliedert.

4.2.1 Videoqualität

Neben der rein technischen Bildqualität sind im Kontext medizinischer Befundung auch Rahmenbedingungen wie Aufnahmedauer und Aufnahmesituation von Bedeutung. Diese Faktoren wurden mit der Bildqualität unter dem Begriff der „Videoqualität“ zusammengeführt. Der Befundfragebogen erhob die Videoqualität als subjektiven Eindruck der Befunder in den Kategorien

- Bildqualität,
- Aufnahmedauer und
- Aufnahmesituation.

Abbildung 4.12: Videoqualität



Subjektive Bewertung von Bildqualität, Aufnahmedauer und Aufnahmesituation.

Bildqualität, Aufnahmedauer und Aufnahmesituation wurden insgesamt mehrheitlich als „gut“ beurteilt. Eine genauere Untersuchung der einzelnen Kategorien zeigt, dass auf Ebene der einzelnen Videos deutliche Differenzen bestehen.

4.2.1.1 Bildqualität

Die technische Bildqualität wurde mit der subjektiven Skala „schlecht – ausreichend – gut“ geschätzt. Die Bildqualität der Hälfte aller Videos wurde als „gut“ eingestuft, etwas mehr als ein Viertel als „ausreichend“ und ein gutes Achtel als „schlecht“. Bei 61 Befunden wurden keine An-

gaben zur Bildqualität gemacht. Die Bewertungen „gut“ und „ausreichend“ verteilen sich auf fast alle Videos (39 bzw. 40 von 42) „schlecht“ auf noch fast 80 % der Videos.

Tabelle 4.1: Subjektiver Eindruck der Bildqualität

Befund	**Anzahl unterschiedlicher ...**				**Konkordanz** der Untersucher untereinander		
	Befunde (max. 20)		**Befunder** (max. 20)	**Videos** (max. 20)	Ø positive Übereinstimmung	Kappa **Fleiss**	modifizierte Klassifikation **nach Landis**
gesamt	840	[100 %]	20	42	36,8	0,136	gering
NA	61	[7,3 %]	14	32	13,1	-0,002	keine
schlecht	126	[15 %]	20	33	23,8	0,154	gering
ausreichend	233	[27,7 %]	20	39	32,9	0,069	kaum
gut	420	[50 %]	20	40	53,8	0,217	schwach

Die Tabelle betrachtet die Übereinstimmung in der Beurteilung der Bildqualität. Der rein deskriptiven Befundverteilung folgt die Übereinstimmung innerhalb der Befunder mit dem Kappa nach Fleiss als Kennzahl.

4.2.1.2 Aufnahmedauer

Auch für die Aufnahmedauer wurde eine einfache dreistufige Skala („zu kurz“ – „ausreichend“ - gut“) abgefragt. Bei knapp 40 % der Videos wurde die Aufnahmedauer als „gut“ eingestuft, bei 30 % als ausreichend und bei 20 % als zu kurz. Etwa 10 % der Videos wurden nicht beurteilt. Während die Aufnahmedauer so gut wie aller Videos teils als gut (41/42), teils als ausreichend (42/42) empfunden wurde, verteilt sich der Befund „zu kurz“ auf 80 % der Videos. Hinsichtlich der Kategorie „zu kurz“ besteht bei einem Kappa Fleiss von 0,206 die einheitlichste Bewertung der Aufnahmedauer.

Tabelle 4.2: Subjektiver Eindruck Aufnahmedauer

Befund	**Anzahl unterschiedlicher...**				**Konkordanz** der Untersucher untereinander		
	Befunde (max. 20)		**Befunder** (max. 20)	**Videos** (max. 20)	Ø positive Übereinstimmung	Kappa **Fleiss**	modifizierte Klassifikation **nach Landis**
gesamt	840	[100 %]	20	42	32,5	0,117	gering
NA	83	[10 %]	17	37	14,6 %	0,007	keine
zu kurz	175	[21 %]	19	35	33,2 %	0,206	schwach
ausreichend	256	[30 %]	20	42	20,6 %	0,043	keine
gut	326	[39 %]	20	41	43,6 %	0,163	gering

Die Tabelle betrachtet die Übereinstimmung in der Beurteilung der Aufnahmedauer. Der rein deskriptiven Befundverteilung folgt die Übereinstimmung innerhalb der Befunder (Konkordanz).

4.2.1.3 Aufnahmesituation

Die Aufnahmesituation konnte mit den Kategorien „gut“ und „schlecht“ beurteilt werden. Mit 60 % wurde die Aufnahmesituation der überwiegenden Anzahl von Videos als „gut“ eingestuft, in 20 % der Videos war die Aufnahmesituation in den Augen der Befunder „schlecht“. In ebenfalls etwa 20 % der Videos wurden keine Angaben zu Aufnahmesituation gemacht.

Tabelle 4.3: Subjektiver Eindruck Aufnahmesituation

Befund	**Anzahl unterschiedlicher ...**				**Konkordanz** der Untersucher untereinander		
	Befunde (max. 20)		**Befunder** (max. 20)	**Videos** (max. 20)	Ø positive Übereinstimmung [%]	Kappa **Fleiss**	modifizierte Klassifikation **nach Landis**
gesamt	840	[100 %]	20	42	52,2	0,096	keine
NA	165	[19,6 %]	19	42	21,6	0,003	keine
schlecht	173	[20,6 %]	20	36	25,9	0,154	gering
gut	502	[59,8 %]	20	42	61,2	0,118	gering

Die Tabelle betrachtet die Übereinstimmung in der Beurteilung der Aufnahmesituation. Der rein deskriptiven Befundverteilung folgt die Übereinstimmung innerhalb der Befunder (Konkordanz). Außer Fehlwerten wurden bei der durchschnittlichen positiven Übereinstimmung alle Angaben als „positiv" gewertet.

Zusammenfassung 4.2: Videoqualität

Die Videoqualität wurde von den Untersuchern sehr variabel empfunden. Erkennbare, aber schwache Übereinstimmungen bei der Beurteilung gibt es nur hinsichtlich der Bildqualität in der Kategorie „gut" und hinsichtlich der Aufnahmedauer in der Kategorie „zu kurz". Bildqualität und Aufnahmedauer wurden in etwa 70 % der Videomitschnitte als gut oder ausreichend bewertet, die Aufnahmesituation in knapp 60 % als gut.

4.2.2 Hauptdiagnose

Die Hauptdiagnose wurde im Fragebogen als Freitext erfasst und im Rahmen der Auswertung in die vier Kategorien

- „keine Angabe"
- „andere Diagnose"
- „ähnliche Diagnose" und
- „gleiche bzw. synonyme Diagnose"

klassifiziert.

Tabelle 4.4: Richtigkeit klassifizierter Hauptdiagnosen

	keine Angabe	falsch	richtig	ähnlich	Summe
Anzahl	92	238	380	130	840
Prozent	11 %	28,3 %	45,2 %	15,5 %	100 %

Die von den Befundern im Freitext angegebenen Hauptdiagnosen wurden für die Auswertung klassifiziert.

Befundverteilung: Richtige, ähnliche und falsche Diagnosen kommen bei fast allen Befundern vor. Nur 3 Befunder haben bei allen Videos eine Diagnose genannt, alle übrigen Befunder stellten bei einzelnen Videos keine Diagnose. Richtige und falsche Diagnosen verteilen sich auf alle Videos. Zum Goldstandard ähnliche Diagnosen kommen nur in der Hälfte aller Videos vor. Auch fehlende Diagnosen betreffen fast alle Videos.

Präzision: Die höchste positive Übereinstimmung findet sich bei richtigen Befunden. Mit Werten um 0,3 liegt das Kappa nach Fleiss für richtige und ähnliche Befunde in etwa gleich auf. Bei falschen Befunden liegt das Kappa nach Fleiss mit 0,248 etwas darunter. Bildet man aus richtigen und ähnlichen Befunden eine gemeinsame Klasse, steigt die durchschnittliche positive Übereinstimmung zwar auf 60 %, das Kappa nach Fleiss sinkt jedoch auf 0,243 und damit auf das Niveau der falschen Befunde.

Tabelle 4.5: Inter-Beobachter-Varaibilität der Hauptdiagnose

Befund	Befundverteilung				Konkordanz Übereinstimmung der Befunder untereinander			
	Referenz	Befunder				Kappa Fleiss		
	Anzahl verschiedener				Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**		modifizierte Klassifikation **nach Landis**
	Befunde (= Videos)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)				
gesamt	42 [840]	748	20	42	NA	0,252	0,214	schwach
richtig	42 [840]	380	20	38	50,0	0,297	0,243	schwach
ähnlich	NA	130	19	22	32,0	0,308		mäßig
					60,7			schwach
falsch	NA	238	20	35	37,7	0,248		schwach
keine Angabe	NA	92	17	40	16,8	0,072		kaum

Die Richtigkeit der Befunde im Sinne der Übereinstimmung mit dem Goldstandard ist bereits Bestandteil der für die Auswertung vorgenommenen Klassifikation der Freitextangaben aus den Befundfragebögen. Die Tabelle stellt neben der Befundverteilung die Konkordanz (Präzision) der Untersucher untereinander dar.

Richtigkeit: Knapp die Hälfte aller Diagnosen stimmen mit der des Goldstandards überein, sind also korrekt. Weitere 15 % sind ähnlich zur Diagnose der Referenz. Weniger als ein Drittel der Diagnosen differiert deutlich vom Goldstandard, sodass sie im Kontext dieser Studie als falsch eingestuft wurden. In gut 10 % wurde vom Befunder keine Diagnose genannt. Nimmt man die ähnlichen und die richtigen Diagnosen zusammen, wurden etwa 60 % der Videos im Sinne des Goldstandards richtig befundet und 30 % divergierend.

Zusammenfassung 4.3: Hauptdiagnose

Die Hauptdiagnose wurde in knapp der Hälfte der Fälle korrekt erkannt (45 %). Bezieht man ähnliche Diagnosen mit ein, ergibt sich eine Trefferquote von 60 %. In gut 10 % der Fälle wurde keine Angabe zur Diagnose gemacht.

4.2.3 Stenosen

Stenosen wurde mit dem Befundfragebogen anhand der klinisch wichtigen Beurteilungskriterien

- Stenosegrad
- Lokalisation und
- Form

verglichen. In fast 90 % aller Videomitschnitte waren Stenosen zu sehen (Prävalenz gemäß Goldstandard 37/42 Videos). Die Befunder sahen hingegen nur in etwas mehr als der Hälfte der Videos Stenosen (441/840 Befunde).

4.2.3.1 Stenosegrad

Der Stenosegrad ist im Befundfragebogen als Prozentangabe im Freitext erfasst und wurde im Rahmen der Auswertung nach einer modifizierten Myer-Cotton-Klassifikation kategorisiert, die unauffällige Befunde in einer eigenen Kategorie berücksichtigt und die Klasse höchstgradiger Stenosen der Klasse IV verbreitert (90 – 100 %, statt Totalverlegung von 100 %)

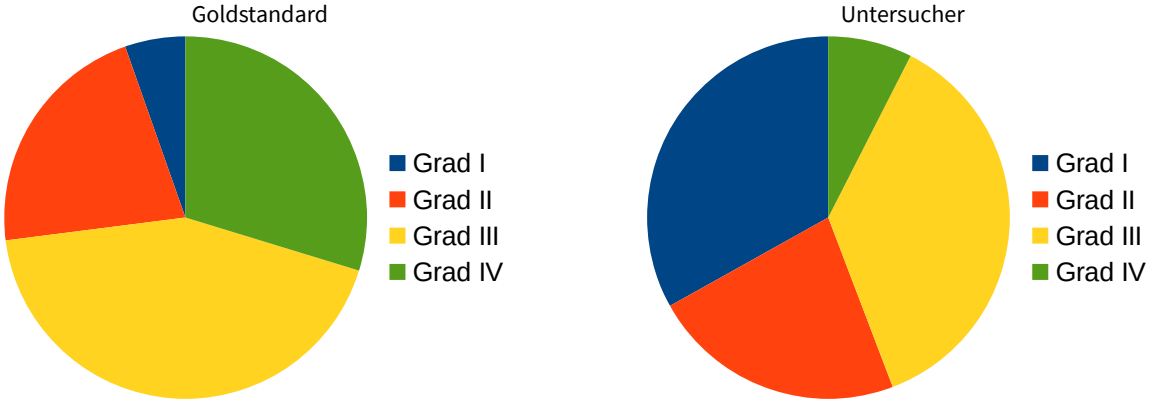
Befundverteilung: Der Goldstandard ordnet die Stenosen überwiegend (16/37 ≙ 43 %) der Klasse III der Myer-Cotton-Klassifikation zu, gefolgt von höchstgradigen Stenosen der Klasse IV (11/37 ≙ 30 %) und Stenosen der Klasse II (8/37 ≙ 22 %). Geringgradige Stenosen der Klasse I spielen mit 2/37 (≙ 5 %) eine untergeordnete Rolle. Die Beurteilung der Befunder zeichnet ein etwas anderes Bild: Auch hier führen zwar Stenosen der Klasse III mit 163/441 Befunden (≙ 37 %), ihnen folgen jedoch Stenosen < 50 % der Klasse I mit 146/441 Befunden (≙ 33 %). Stenosen der Klasse II zwi-

schen 50 und 70 % machen bei den Befundern 100/441 Befunde (≙ 23 %) aus. Subtotale Stenosen > 90 % sind mit nur 33/441 Befunden (≙ 7 %) selten.

Die Befunde der Klassen I bis III werden von so gut wie allen Befundern getragen, während an der Klasse IV nur etwa ⅔ (14/20) beteiligt sind. Subtotale Stenosen > 90 % werden von Goldstandard und Befundern übereinstimmend in 11 verschiedenen Videos, also etwa einem Viertel des Testfeldes, gesehen. Alle übrigen Klassen verteilen sich auf erheblich mehr Videos. Der Goldstandard sieht über die Klassen I bis III einen deutlichen Anstieg der Häufigkeit (2, 8, 16). Dagegen sehe die Befunder jeweils konstant um die 30 Videos betroffen (36, 32, 28). Die Streubreite der Befunder beträgt ein Vielfaches der Referenz des Goldstandards: Waren laut Goldstandard Stenosen< 50 % in nur 2 Videos zu sehen, befundeten die Untersucher solche Stenosen in 36 verschiedenen Videos. Der Goldstandard beschrieb Stenosen der Klasse II in 8 verschiedene Videos, die Untersucher in 32. In Klasse III stehen 16 Videos beim Goldstandard 28 Videos bei den Befunder gegenüber.

Der relative Anteil der Klasse II, also Stenosen zwischen 50 % und 70 %, wird ähnlich eingeschätzt, während die Randklassen I und IV genau entgegengesetzt befundet werden: Die Befunder sehen zahlreiche Stenosen < 50 % aber nur wenige > 90 %. Der Goldstandard sieht viele Stenosen > 90 %, aber nur wenige unter 50 %. Im direkten Vergleich eines gepaarten Histogramms ist der Unterschied klarer zu erkennen (Abbildung 4.14a). Bei diesem Vergleich muss jedoch beachtet werden, dass in Abbildung 4.13 die Kategorie der Fehlwerte bzw. „keine Stenose“ außer Acht gelassen werden, die bei den Befundern fast die Hälfte aller Videos ausmachen. Die absoluten – das heißt unter Einbeziehung der Fehlwerte auf die Gesamtzahl von 840 Befunden bezogenen – Zahlenverhältnisse zeigt das Histogramm in Abbildung 4.14b. Angesichts der eben beschriebenen relativen Anteile der Befundklassen bei den Befundern und dem Goldstandard ist die beste Vergleichbarkeit für Stenosen zwischen 50 und 70 % (Klasse II nach Myer-Cotton) gegeben. Die Vergleichbarkeit in den übrigen Klassen (I, III und IV) ist durch die divergierenden Randsummen (bzw. Prävalenzen) eingeschränkt.

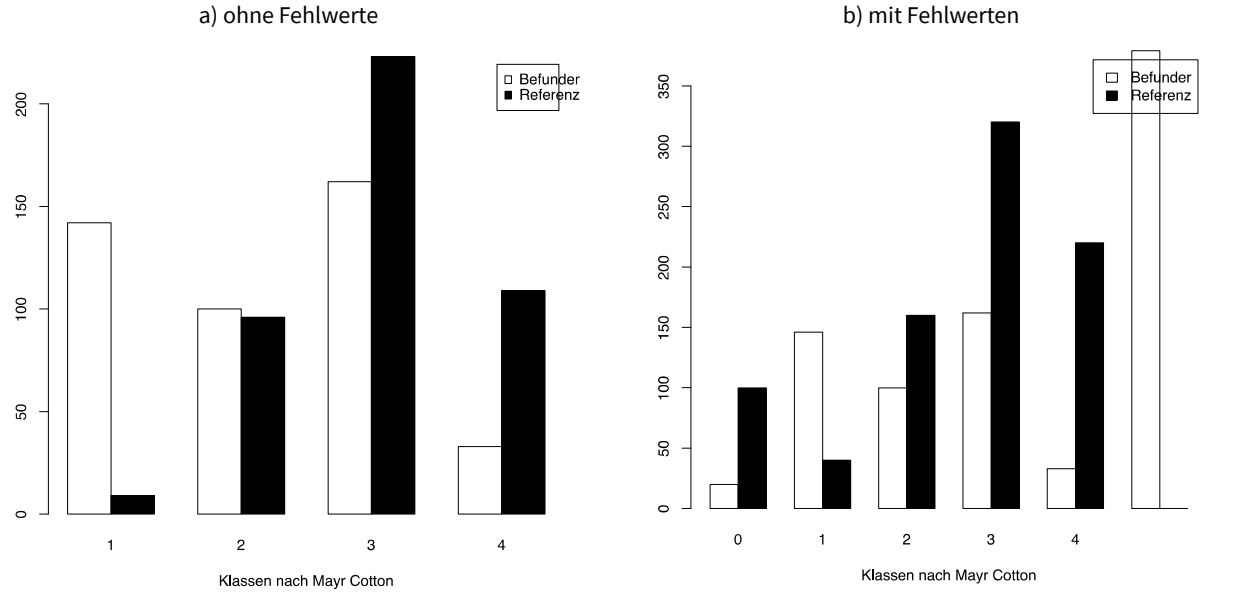
Abbildung 4.13: Vergleich Stenosegrade bei modifizierter Myer-Cotton-Klassifikation



Relative Anteile der Stenosegrade gemäß der Myer-Cotton Klassifikation an sämtlichen Stenosen. Unauffällige Befunde („keine Stenose“) und Fehlwerte werden in dieser Darstellung nicht berücksichtigt.

Dieser Eindruck wird durch den McNemar Test bestätigt: lässt man Fehlwerte unberücksichtigt, ergeben sich mit Ausnahme Stenosen II für alle Klassen signifikante Unterschiede der Befundhäufigkeiten zwischen Befundern und dem Goldstandard (Tabelle 4.10). Alleine aufgrund der unterschiedlichen Prävalenzen in den Klassen I, III und IV ist also eine verhältnismäßig geringe Übereinstimmung zu erwarten. Das lässt sich auch anhand der Kontingenztafel (Tabelle 4.9) nachvollziehen: Bei einer Gesamtzahl von 461 gemeinsamen Befunden finden sich bei „keinen Stenosen“ (13 gegenüber 20) und bei Stenosen Grad II (98 gegenüber 100) ähnliche Anteile, während der Anteil bei Stenosen Grad I fast um den Faktor 15 divergiert (10 gegenüber 146 Befunden).

Abbildung 4.14: Histogramm Stenosegrade nach modifizierter Myer-Cotton-Klassifikation



Die Balkendiagramme stellen die Befundhäufigkeiten der Stenosegrade gemäß der Myer-Cotton-Klassifikation bei den Befundern denen des Goldstandards gegenüber.

Präzision: Innerhalb der Befunder bestehen hinsichtlich der Einschätzung des Stenosegrades – trotz der vergleichsweise einfachen Myer-Cotton Klassifikation mit nur 4 Kategorien – erhebliche Differenzen. Ein Kappa Fleiss von insgesamt 0,201 entspricht lediglich einer schwachen Übereinstimmung. Stenosen > 70 % werden deutlich einheitlicher befundet als Stenosen < 70 %.

Richtigkeit: Sämtliche Befunde zur Richtigkeit müssen aufgrund der signifikant unterschiedlichen Befundhäufigkeiten in den einzelnen Klassen zwischen Befundern und dem Goldstandard mit Vorsicht interpretiert werden: der Bhapkar-Test, der für die Kontingenztafel insgesamt gilt, ist signifikant. Mit Ausnahme der Klassen 0 und II fällt auch der McNemar-Test, der für einzelne Klassen gilt, signifikant aus. Im Vergleich zum Goldstandard besteht eine systematische Überschätzung in der die Klasse I und eine systematische Unterschätzung in der Klasse IV. Hieraus ergibt sich von vornherein eine Verzerrung der Befundübereinstimmung, die bei der Interpretation der Maßzahlen zur Richtigkeit berücksichtigt werden sollte. Da Fehlwerte nicht berücksichtigt werden können, decken die Maßzahlen der Richtigkeit insgesamt nur 54,9 % der Daten ab. Fehlwerte treten gehäuft bei unauffälligen Befunden auf, sodass die Datenabdeckung bei den Stenosegraden I bis III bei über 70 % liegt. Insgesamt wurden Stenosen in etwa einem Drittel der Fälle korrekt mit ihrer Graduierung erkannt (Sensitivität Tabelle 4.6). Wurde von den Befundern eine Stenose gesehen, handelte es sich in jedem 3. Fall um eine korrekte Diagnose im Sinne des Goldstandards. Korrigiert man für zufällige Übereinstimmung, ist bei einem Kappa Cohen von 0,139 allerdings nur eine geringe Übereinstimmung erkennbar. Die höchste Sensitivität besteht mit 60 % bei Stenosen der Klasse I. Im Übrigen nimmt die Sensitivität mit zunehmendem Stenosegrad über die Klassen II bis III zu, erreicht aber bei höchstgradigen Stenosen > 90 % in Klasse IV ihr Minimum. Stenosen der Klasse III werden in fast der Hälfte der Fälle als solche erkannt, Stenosen der Klassen II in etwas weniger als einem Drittel, Stenosen > 70 % in nur etwa 10 % der Fälle. Verlässlich waren positive Befunde ebenfalls in einem Drittel der Fälle (ppv 34,5 %). Den besten positiven prädiktiven Wert haben Stenosen III mit 65,4 %. Sie setzen sich damit deutlich von Stenosen Grad IV und II (33,3 % und 29 %) ab. Stenosen Grad I werden mit einem ppv von 4,1 % am unzuverlässigsten diagnostiziert. Grob betrachtet steigt die Spezifität mit dem Stenosegrad an, wobei die Klassen II und III mit um die 80 % in etwa gleich auf liegen (Tabelle 4.10).

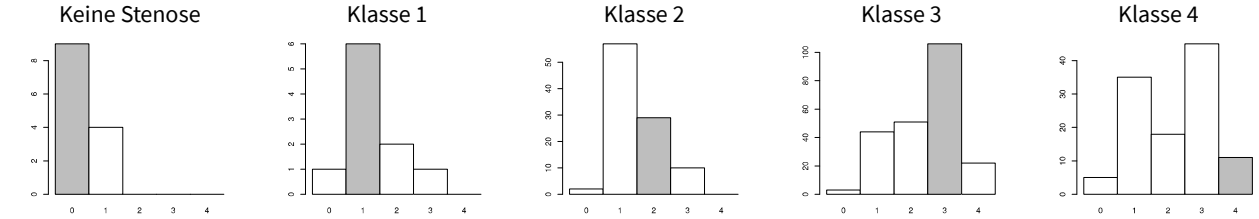
Tabelle 4.6: Inter-Beobachter-Variabilität maximaler Stenosegrad

Befund	**Befundverteilung**				**Präzision** Übereinstimmung der Befunder			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz				
	Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
	Anzahl verschiedener											
Klassifikation nach Myer-Cotton	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**	Datenabdeckung
gesamt	42 [840]	441	20	40	25,0	0,201	schwach	33,9	34,5	0,139	gering	54,9
Keine Angabe	NA	379	20	42	46,1	0,272	schwach	NA	NA	NA	NA	NA
keine Stenose	5 [100]	20	4	14	11,0	0,013	keine	69,2	45,0	0,529	gut	21,6
Grad I (<50%)	2 [40]	146	20	36	23,3	0,111	gering	60,0	4,1	0,038	keine	83,3
Grad II (51-70%)	8 [160]	100	19	32	20,5	0,078	kaum	29,6	29,0	0,1	kaum	73,2
Grad III (71-90%)	16 [320]	162	20	28	32,9	0,264	schwach	46,9	65,4	0,232	schwach	75,0
Grad IV (> 90%)	11 [220]	33	14	11	23,3	0,275	schwach	9,7	33,3	0,043	kaum	56,2

Der deskriptiven Befundverteilung folgt eine Analyse der Übereinstimmung innerhalb der Befunder (Präzision) sowie im Vergleich zum Goldstandard (Richtigkeit). Zur besseren Vergleichbarkeit zwischen Referenz und Befundern ist in eckigen Klammern die gewichtete Anzahl der Befunde des Goldstandards angegeben (Faktor 20 gegenüber Befundern). Fehlende Angaben des Stenosegrades wurden bei den Maßzahlen der Richtigkeit nicht berücksichtigt.

Aufgrund des verzerrenden Effektes der differierenden Randsummen scheint die direkte Interpretation der Kontingenztafel aufschlussreicher als die Maßzahlen der Richtigkeit. Die Kontingenztafel (Tabelle 4.9) zeigt, dass Stenosen von den Befundern tendenziell eher niedriger eingestuft werden als vom Goldstandard. Dabei liegt, insbesondere bei den Stenosegraden II bis IV, eine erhebliche Streuung der Graduierung vor. Beispielsweise wurden Stenosen zwischen 70 und 90 % (Klasse III) fast ebenso häufig in die Klassen I (< 50 %) und II (50 – 70 %) eingeordnet. Selbst subtotale Stenosen > 90 % werden in etwa einem Drittel (35/114) Fällen von den Befundern als < 50 % geschätzt (Abbildung 4.15). Die systematische Unterschätzung ist auch im Bangdiwala Diagramm (Abbildung 4.16 Seite 104) erkennbar: Die Schnittpunkte der Rechtecke liegen unterhalb der Diagonalen.

Abbildung 4.15: Befundverteilungen der Untersucher in der Myer-Cotton-Klassfikation



Die Säulendiagramme der Befundverteilungen der Untersucher in den einzelnen Klassen der Myer-Cotton Klassifikation.

Im Befundfragebogen wurde der „maximale Stenosegrad“ erfragt. Gleichzeitig konnten im multiple choice Verfahren mehrere Stenoselokalisationen angegeben werden. Bei mehrfachen Stenosen kann der maximale Stenosegrad daher nicht mehr eindeutig einer Stenose zugeordnet werden. Somit ist es möglich, dass sich die Angaben zum maximalen Stenosegrad auf unterschiedliche Stenosen beziehen. Da explizit nach dem maximalen Stenosegrad gefragt wurde, kann diese Situtation jedoch nur eintreten, wenn Uneinigkeit über die als maximal zu beurteilende Stenose besteht oder Stenosen übersehen wurden. Tabelle 4.7 zeigt, dass vom Goldstandard wie auch den Untersuchern zahlreiche Mehrfachstenosen befundet wurden.

Tabelle 4.7: Mehrfachstenosen

Anzahl der Stenosen	Goldstandard	Untersucher
0	6 [120]	294
1	13 [260]	342
2	17 [340]	148
3	5 [100]	44
4	keine	8
5	1 [20]	3
6	keine	1
Summe	42 [840]	840

Die Anzahl der Stenosen wurde über die Befunde zur Stenoselokalisation in alle anatomischen Abschnitte bestimmt.

Tabelle 4.8: maximaler Stenosegrad Subgruppe singuläre Stenosen

Mc Nemar Test	Befund	Befundverteilung: Referenz Befunde (& Videos) (max. 42)	Befundverteilung: Befunder Anzahl verschiedener Befunde (max. 840)	Befundverteilung: Befunder Anzahl verschiedener Befunder (max. 20)	Befundverteilung: Befunder Anzahl verschiedener Videos (max. 42)	Präzision: Übereinstimmung der Befunder: Ø positive Übereinstimmung [%]	Präzision: Kappa Fleiss: Kappa **nach Fleiss**	Präzision: Kappa Fleiss: modifizierte Klassifikation **nach Landis**	Richtigkeit: Übereinstimmung mit Goldstandard als Referenz: Sensitivität [%]	Richtigkeit: Übereinstimmung mit Goldstandard als Referenz: positiver prädiktiver Wert [%]	Richtigkeit: Kappa Cohen: Kappa **nach Cohen**	Richtigkeit: Kappa Cohen: modifizierte Klassifikation **nach Landis**
	gesamt	42 [840]				25,6	0,108	gering			0,0785	kaum
	keine Angabe	0	498	20	42	57,2	0,167	gering	NA	NA	NA	NA
M	keine Stenose	5 [100]	92	18	34	18,2	0,038	keine	99	16,8	0,105	gering
M	Grad I	2 [40]	87	18	30	20,3	0,072	kaum	15	6,9	0,031	keine
M	Grad II	8 [160]	56	17	26	14,7	0,032	keine	10	28,6	0,055	kaum
M	Grad III	16 [320]	90	19	25	21,5	0,097	kaum	17,5	62,2	0,127	gering
M	Grad IV	11 [220]	17	10	8	21,7	0,212	schwach	1,4	17,6	-0,013	keine

Übereinstimmung in der Subgruppe der singulären Stenosen

Um zu prüfen, ob Mehrfachstenosen die Beurteilung des maximalen Stenosegrades beeinträchtigen könnten, wurde eine Subgruppenanalyse mit singulären Stenosen durchgeführt. Tabelle 4.8 kann den Verdacht, dass die Beurteilung des maximalen Stenosegrades durch Mehrfachstenosen beeinträchtigt werden könnte, nicht stützen. Die Beschränkung auf singuläre Stenosen führt sowohl hinsichtlich der Präzision als auch hinsichtlich der Richtigkeit zu deutlich schlechteren Werten als im Gesamtdatensatz mit mehrfachen Stenosen.

Tabelle 4.9: Kontingenztafel maximaler Stenosegrad

Mc Nemar <0,01		Goldstandard keine Stenose	Grad I <50%	Grad II 51-70%	Grad III 71-90%	Grad IV >90%	Randsummen
Befunder	keine Angabe	87	30	62	94	106	379
	keine Stenose	9	1	2	3	5	20
	Grad I < 50%	4	6	57	44	35	146
	Grad II 51-70%	0	2	29	51	18	100
	Grad III 71-90%	0	1	10	106	45	162
	Grad IV >90%	0	0	0	22	11	33
Randsummen		13	10	98	226	114	461
		100	40	160	320	220	840

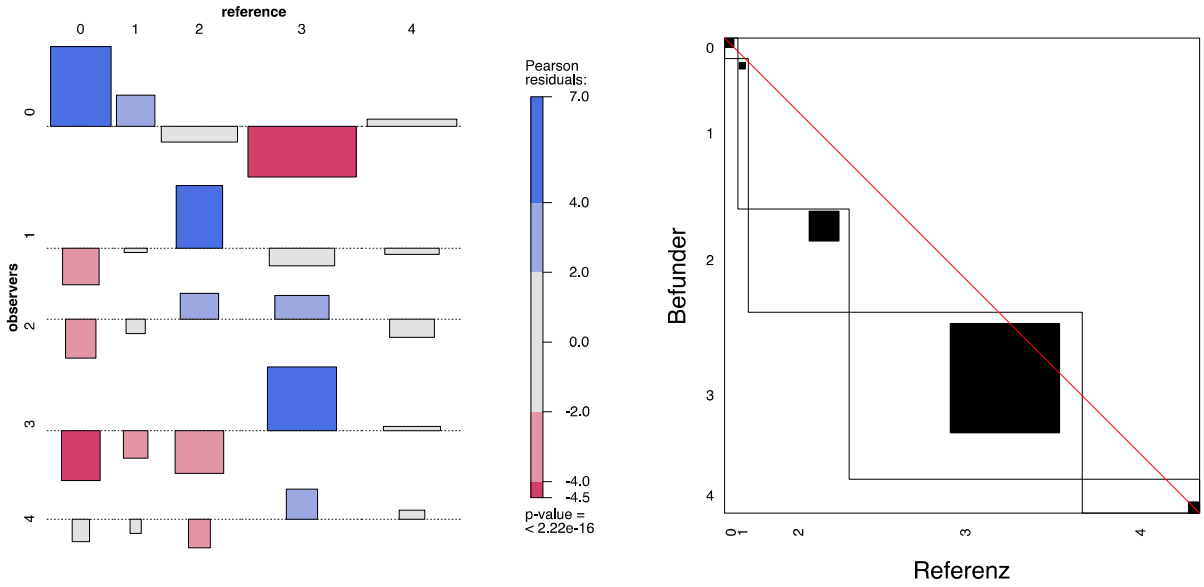
Der McNemar Test prüft die Homogenität der Randsummen. Das signifikante Testergebnis belegt erhebliche divergierende Befundverteilungen zwischen den Befundern und dem Goldstandard.

Tabelle 4.10: Kennwerte maximaler Stenosegrad

Grad	pre	McN	nag	pag	spe	sen	acc	npv	ppv	B	LR-	LR+	AUC	OR	Q	Y
0	2,8	0,121	98,3	54,5	97,5	69,2	96,7	99,1	45,0	0,966	0,315	28,2	0,834	89,4	0,978	0,809
I	2,2	<0,05	81,2	7,7	69,0	60,0	68,8	98,7	4,1	0,674	0,580	1,93	0,645	3,3	0,538	0,292
II	21,3	0,933	80,7	29,3	80.4	29,6	69,6	80,9	29,0	0,611	0,875	1,51	0,550	1,7	0,267	0,136
III	49	<0,05	67,0	54,6	76,2	46,9	61,8	59,9	65,4	0,405	0,697	1,97	0,615	2,8	0,477	0,254
IV	24,7	<0,05	83,9	15,0	93,7	9,7	72,9	75,9	33,3	0,694	0,965	1,52	0,517	1,6	0,224	0,113

Prävalenz (pre), McNemar (McN) Test, negative (nag) und positive (pag) Übereinstimmung, Spezifität (spe), Sensitivität (sen), Genauigkeit (acc), negativer (npv) und positiver prädiktiver Wert (ppv), Bangdiwala (B), negative (LR-) und positive Likelihood ratio (LR+), area under the curve (AUC), odds ratio (OR), Youden Q und Y.

Abbildung 4.16: Assoziations- & Bangdiwaladiagramm maximaler Stenosegrad



Das Assoziationsdiagramm zeigt das Verhältnis der beobachteten Werte zu den Erwartungswerten. Im Bangdiwaladiagramm werden horizontal Spezifität und Sensitivität, vertikal negativer und positiver prädiktiver Wert illustriert.

Zusammenfassung 4.4: Stenosegrad

Selbst beim maximalen Stenosegrad ist die Übereinstimmung innerhalb der Untersucher, wie auch im Vergleich zum Goldstandard gering. Die Konkordanz der Befunder ist bei Stenosen III. und IV. Grades erheblich höher als bei niedriggradigeren Stenosen, absolut gesehen aber nur schwach ausgeprägt. Mit dem Goldstandard gibt es nur bei Stenosen Grad III eine erkennbare Übereinstimmung. Insgesamt wurden Stenosen in einem Drittel der Fälle mit Ihrer Graduierung erkannt. Die Diagnose von Stenosen war in einem Drittel der Fälle korrekt. Ein Teil der Uneinigkeit ist durch erheblich divergierenden Randverteilungen erklärbar: Während der Goldstandard zahlreiche höchstgradig Stenosen vorgibt, wurden diese von den Untersuchern nicht gesehen. Die Befunder sahen viele niedriggradige Stenosen, die gemäß der Referenz nicht existieren. Nur im Mittelfeld II. und III.gradiger Stenosen gibt es annähernd vergleichbare Prävalenzen. Anders als erwartet wurden Stenosen nicht bevorzugt in die jeweils benachbarte Klassen eingruppiert: Höchstgradige Stenosen Grad IV wurden z. B. in etwa einem Drittel der Fälle als Grad I beurteilt.

4.2.3.2 Stenoselokalisation

Die Stenoselokalisation wird zunächst getrennt nach den anatomischen Abschnitten Larynx, Trachea, Hauptbronchien und Lappenbronchien betrachtet. Zuerst werden jeweils die einzelnen Befunde („Symptome“) unabhängig voneinander betrachtet, anschließend wird auf die Kombinationen der Einzelbefunde („Syndrome“) eingegangen. Abschließend folgt eine Betrachtung der Stenoselokalisation, quer über alle Abschnitte des Bronchialbaumes.

4.2.3.2.1 Stenoselokalisation Larynx

Tabelle 4.11: Einzelbefunde Stenoselokalisation Larynx

Befund	**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz			
	Referenz	Befunder Anzahl verschiedener			Ø positive Übereinstimmung [%]	**Kappa Fleiss**		Sensitivität [%]	positiver prädiktiver Wert [%	**Kappa Cohen**	
	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)		Kappa **nach Fleiss**	Klassifikation **modifiziert nach Landis**			Kappa **nach Cohen**	Klassifikation **modifiziert nach Landis**
supraglottisch	8 [160]	62	17	17	33,1	0,261	schwach	23,8	61,3	0,264	schwach
glottisch	11 [220]	80	19	18	33,2	0,250	schwach	27,7	76,3	0,310	mäßig
subglottisch	11 [220]	170	20	20	46,1	0,569	gut	58,6	75,8	0,561	gut

Die Tabelle betrachtet die Übereinstimmung der Einzelbefunde. Befundkombinationen werden gesondert behandelt. Der rein deskriptiven Befundverteilung folgen mittig die Übereinstimmung innerhalb der Befunder (Präzision) und rechts die Übereinstimmung mit dem Goldstandard (Richtigkeit). Die Befunde des Goldstandards müssen beim Vergleich mit den Befundern mit 20 multipliziert werden (eckige Klammern).

Befundverteilung Einzelbefunde: In etwas mehr, als der Hälfte aller Videos (23 von 42) sind gemäß dem Goldstandard Larynxstenosen zu sehen. Die Befunder dagegen beobachteten insbesondere supraglottische und glottische Stenosen deutlich seltener (62 versus 160 bzw. 80 versus 220), aber in ungefähr doppelt so vielen Videos (17 versus 8 bzw. 18 versus 11) wie der Goldstandard. Fast alle Befunder waren an der Diagnose von Larynxstenosen beteiligt (zwischen 17 und 20/20). Diese inhomogene Randsummenverteilung in allen drei Larynxabschnitten werden im McNemar-Test bestätigt.

Präzision Einzelbefunde: Bei supraglottischen und glottischen Stenosen besteht bei einem Kappa Fleiss um 0,250 eine schwache Übereinstimmung innerhalb der Befunder. Subglottische Stenosen werden dagegen mit einem Kappa Fleiss von 0,569 recht einheitlich beurteilt.

Richtigkeit Einzelbefunde: Subglottische Stenosen wurden in fast 60 % der Fälle erkannt, glottische in nur 28 % und supraglottische in 24 %. Die Spezifität, also die Richtigkeit unauffälliger Befunde, liegt in allen 3 Lokalisationen deutlich über 90 %. Korrigiert man für zufällige Übereinstimmung, werden bei subglottischen und glottischen Stenosen 2 von 3 unauffälligen Befunden (67,3 und 67,8 %) richtig vergeben, bei supraglottischen Befunden die Hälfte (52,2 %). Pathologische Befunde werden bei supraglottischen und glottischen Stenosen in einem Fünftel der Fälle richtig erkannt (17,7 und 20,1 %), bei subglottischen Stenosen in der Hälfte der Fälle.

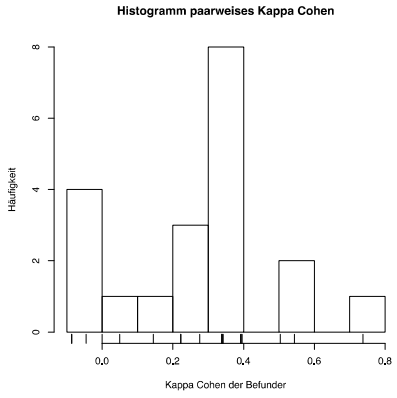
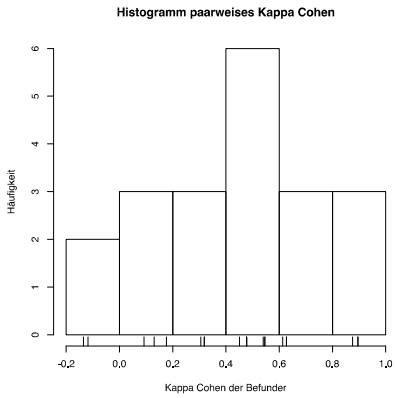
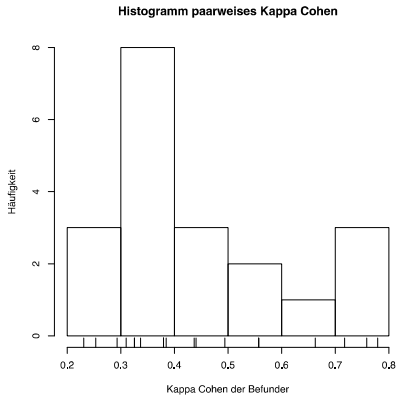
Mit einem positiven prädiktiven Wert von ca. 76 % werden subglottische Stenosen etwa genauso verlässlich diagnostiziert wie glottische Stenosen, die jedoch nur halb so oft erkannt wurden. Supraglottische Stenosen schneiden mit einem positiven prädiktiven Wert von 61,3 % deutlich schlechter ab und wurden gleichzeitig am seltensten erkannt. Bei den negativen prädiktiven

Werten ist die subglottische Stenose mit 86 % Spitzenreiter und wird von supraglottischen Stenosen mit 84 % gefolgt. In beiden Fällen liegen die negativen prädiktiven Werte erkennbar über den positiven prädiktiven Werten. Unauffällige Befunde waren also verlässlicher als pathologische. Dabei darf natürlich die stark divergierende Prävalenz negativer und positiver Befunde nicht außer Acht gelassen werden: Die Erwartungswerte der jeweiligen prädiktiven Werte zeigen, dass die positiven prädiktiven Werte deutlich weiter über dem Zufallsniveau liegen, als die negativ prädiktiven Werte. Bei glottischen Stenosen sind positive und negative Befunde in etwa ähnlich aussagekräftig.

Für subglottische Stenosen findet sich bei einem Kappa Cohen von 0,561 eine gute Übereinstimmung mit dem Goldstandard, die sich von der mäßigen Übereinstimmung bei glottischen (Kappa Cohen 0,310) und schwachen Übereinstimmung bei supraglottischen Stenosen abhebt (Kappa Cohen 0,264). In der Stichprobe nehmen also sowohl Präzision, als auch Richtigkeit bei Larynxstenosen von proximal nach distal zu.

Auch bei einer prävalenzunabhängigen Betrachtung führen subglottische Stenosen mit einer odds ratio von 20 gegenüber glottischen (OR 12) und supraglottischen Stenosen (OR 8,5). In den ROCs setzen sich subglottische Stenosen mit einer AUC von 0,760 gegenüber glottischen (AUC 0,623) und supraglottischen Stenosen (AUC 0,601) ab.

Tabelle 4.12: Paarweises Kappa Cohen Stenoselokalisation Larynx

supraglottisch				**glottisch**				**subglottisch**			
vereintes Kappa Cohen				vereintes Kappa Cohen				vereintes Kappa Cohen			
0,264				**0,310**				**0,561**			
paarweises Kappa Cohen				paarweises Kappa Cohen				paarweises Kappa Cohen			
min.	Ø	median	max.	min.	Ø	median	max.	min.	Ø	median	max.
0	**0,250**	0,236	0,549	-0,056	**0,292**	0,253	0,701	0,206	**0,556**	0,594	0,754
											

Die Tabelle vergleicht die Ergebnisse der beiden Berechnungsvarianten „vereintes“ und „paarweises“ Kappa Cohen.

Tabelle 4.13: Vierfeldertafeln Einzelbefunde Stenoselokalisation Larynx

supraglottisch					glottisch					subglottisch				
McN <0,01		Referenz 0	1	Summen	McN <0,01		Referenz 0	1	Summen	McN <0,01		Referenz 0	1	Summen
Befunder	0	656	122	778	Befunder	0	601	159	760	Befunder	0	579	91	670
	1	24	38	62		1	19	61	80		1	41	129	170
Summen		680	160	840	Summen		620	220	840	Summen		620	220	840

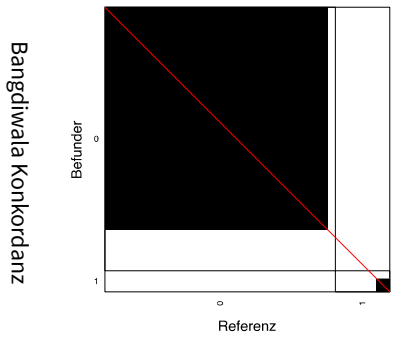
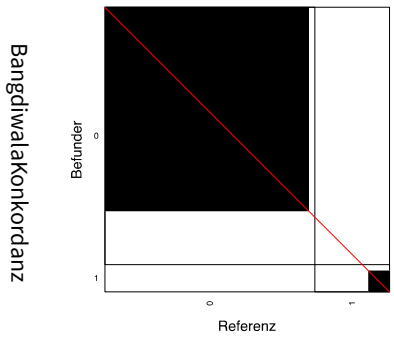
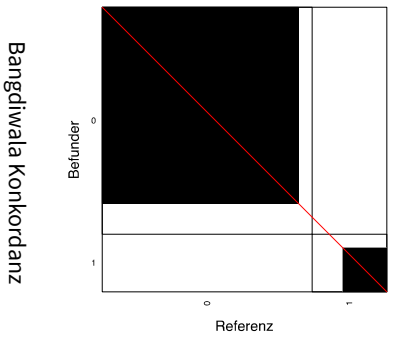
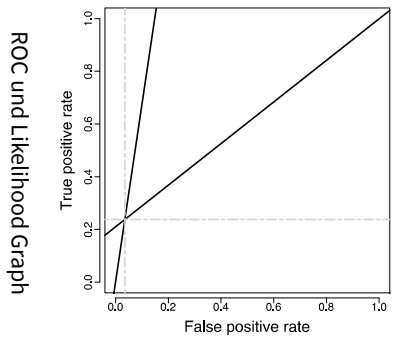
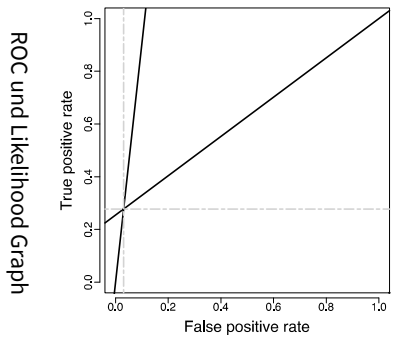
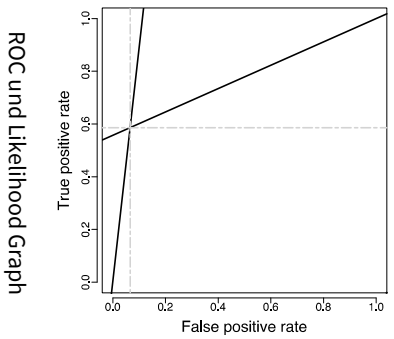
Ein signifikanter Mc Nemar Test (McN) zeigt eine ungleiche Verteilung der Randsummen an.

Tabelle 4.14: Kennwerte Einzelbefunde Stenoselokalisation Larynx

supraglottisch				glottisch				subglottisch			
Prävalenz	19,0			Prävalenz	26,2			Prävalenz	26,2		
	beo.	erw.	kor.		beo.	erw.	kor.		beo.	erw.	kor.
neg. Übereinst.	90,0	86,4	=κ	neg. Übereinst.	87,1	81,2	=κ	neg. Übereinst.	89,8	76,7	=κ
pos. Übereinst.	34,2	10,6	=κ	pos. Übereinst.	40,7	14,0	=κ	pos. Übereinst.	66,2	22,8	=κ
Spezifität	96,5	92,6	52,2	Spezifität	96,9	90,5	67,8	Spezifität	93,4	79,8	67,3
Sensitivität	23,8	7,4	17,7	Sensitivität	27,7	9,5	20,1	Sensitivität	58,6	20,2	48,1
Genauigkeit	82,6	76,4	26,4=κ	Genauigkeit	78,8	69,3	31,0=κ	Genauigkeit	84,3	64,2	56,1=κ
odds ratio (Yule Q/Y)	8,5 (0,790/0,490)			odds ratio (Yule Q/Y)	12,1 (0,848/0,554)			odds ratio (Yule Q/Y)	20,0 (0,905/0,635)		

Beobachtete Kennwerte (beo.), Erwartungswerte gemäß Randsummen (erw.) & für erwartete Übereinstimmung korrigierte Werte (kor.) in Prozent. Kappa Cohen bei Untersuchern als ein fusionierter Befunder & bei paarweisem Vergleich mit dem Goldstandard (min., Ø, max.). Odds ratio und ihre Transformationen (Yule Q & Y).

Abbildung 4.17: Diagramme Einzelbefunde Stenoselokalisation Larynx

supraglottisch	glottisch	subglottisch
		
Bangdiwala 0,801 (0,736) npv 84,3 (81,0) ppv 61,3 (19,0)	Bangdiwala 0,747 (0,645) npv 79,1 (73,8) ppv 76,3 (26,2)	Bangdiwala 0,777 (0,544) npv 86,4 (73,8) ppv 75,9 (26,2)
		
area under curve (AUC) 0,601 LR+ 6,73 LR- 0,79	area under curve (AUC) 0,623 LR+ 9,05 LR- 0,75	area under curve (AUC) 0,760 LR+ 8,87 LR- 0,44

Die Bangdiwala Diagramme illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Die Receiver Operator Characteristic (ROC) trägt die Rate falsch Positiver (=1-Spezifität, fpr) gegen die Rate richtig Positiver (=Sensitivität, tpr) auf. Die Steigung der Tangenten durch den Ursprung bzw. (1,1) definieren die Likelihood ratios (LR+/LR-), die Fläche darunter die Area Under the Curve (AUC).

Tabelle 4.15: Befundkombinationen Stenoselokalisation Larynx

Befund				Befundverteilung				Präzision Übereinstimmung der Befunder untereinander			Richtigkeit Übereinstimmung mit dem Goldstandard in den überschneidenden Befunden				
				Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
				Anzahl verschiedener											
Anzahl der Stenosen	supraglottisch	glottisch	subglottisch	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	**Kappa nach Cohen**	modifizierte Klassifikation **nach Landis**	Datenabdeckung [%]
x	**gesamt**			**23 [460]**	**278**	**20**	**32**	**26,0**	**0,418**	**moderat**	**29,8**	**49,3**	**0,357**	**mäßig**	**100**
0	0	0	0	19 [380]	562	20	42	71,7	0,463	moderat	91,8	62,1	0,437	moderat	100
			1	6 [120]	151	20	21	45,6	0,563	gut	66,7	53,0	0,513	gut	100
1		1	0	5 [100]	47	19	14	25,6	0,225	schwach	24,2	51,1	0,271	schwach	100
	1	0	0	5 [100]	46	17	13	28,6	0,264	schwach	24,2	52,2	0,274	schwach	100
		1	1	4 [80]	18	11	8	15,0	0,086	kaum	37,5	16,7	0,027	keine	100
2	1	0	1	1 [20]	1	1	1	NA	NA	NA	5,0	100	0,093	kaum	100
	1	1	0	2 [40]	15	10	7	15,0	0,089	kaum	12,5	33,3	0,160	gering	100

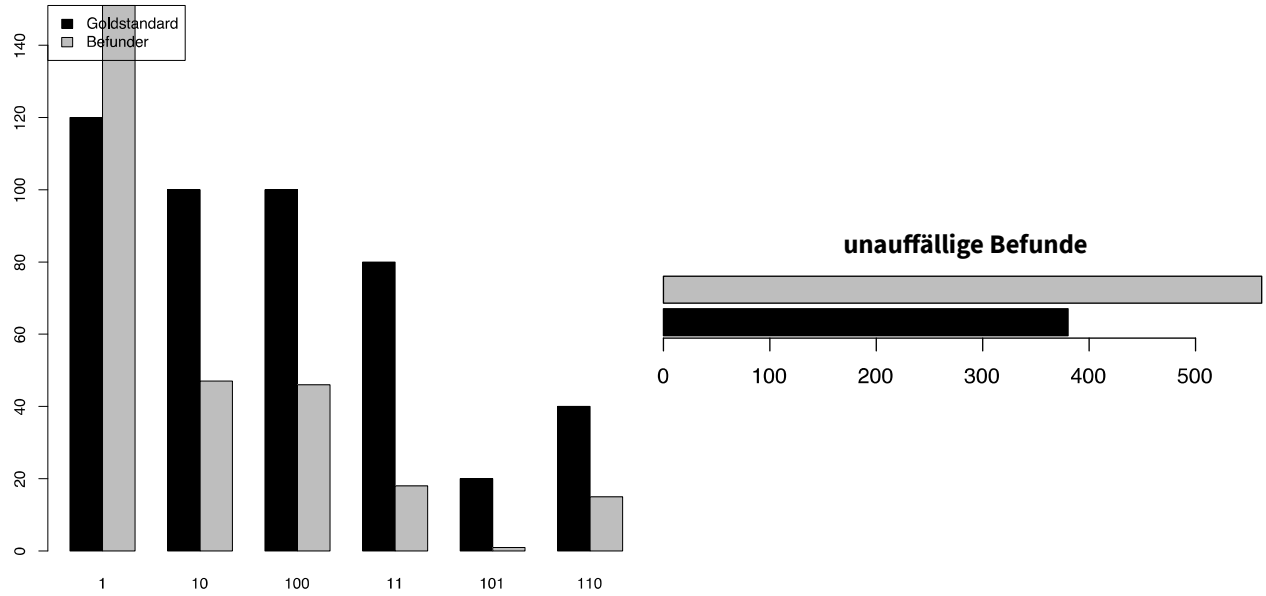
Die Tabelle analysiert die Befundkombinationen in den einzelnen Larynxabschnitten im Sinne eines Gesamtbefundes des Larynx. Sie gliedert sich in 3 Hauptbereiche: Befundverteilung, Präzision und Richtigkeit. Der Abschnitt Befundverteilung beschreibt Häufigkeiten. Der Abschnitt Präzision zeigt die Übereinstimmung der Befunder untereinander (also unabhängig vom Goldstandard!). Dem gegenüber beschreibt der Abschnitt Präzision die Übereinstimmung der Befunder mit dem Goldstandard als Referenz.
In der Spalte „Befund" sind sämtliche Befundkombinationen aufgelistet. Ein im multiple-choice-Fragebogen gesetztes Kreuz wird durch eine 1 repräsentiert, ein fehlendes Kreuz durch eine 0. Die Befundkombinationen wurden nach der Anzahl der Befunde gruppiert. Innerhalb der sich hieraus ergebenden Befundgruppen wurden die Codes der Befundkombinationen als Zahlen behandelt und aufsteigend sortiert.
Die Spalte „Befundverteilung" gibt – getrennt nach Befundern und Goldstandard – die Anzahl der Befunde und Videos bei der jeweiligen Befundkombination an. Die Angaben der Referenz sind auch als Prävalenz der Befunde zu interpretieren. Da jeder der 20 Befunder mit dem Goldstandard verglichen wird, wiegt seine Bewertung 20fach. Für den direkten Vergleich der Befundverteilung der Befunder mit der des Goldstandards müssen die Befunde des Goldstandards deshalb mit 20 multipliziert werden (Angaben in eckigen Klammern).
Die Spalte Präzision gibt die Konkordanz der Befunder untereinander, also unabhängig vom Goldstandard an. Als Maßzahlen hierfür wurden die durchschnittliche positive Übereinstimmung und das Kappa nach Fleiss gewählt. Letzteres wird gemäß einer modifizierten Klassifikation nach Landis bewertet.
Die Spalte Richtigkeit stellt die Übereinstimmung der Befunder mit dem Goldstandard als Referenz dar. Die Sensitivität gibt die prozentuale Wahrscheinlichkeit an, dass ein tatsächlich vorliegender Befund (= Befund des Goldstandards) von den Befundern erkannt wurde. Der positive prädiktive Wert gibt die prozentuale Wahrscheinlichkeit an, dass der Befund eines Befunders tatsächlich vorliegt (= dem des Goldstandards entspricht). Das Kappa nach Cohen ist ein Maß der Übereinstimmung zwischen Befundern und Goldstandard, das für zufällig zu erwartende Übereinstimmungen korrigiert. Kappa Cohen wurde ebenfalls nach der modifizierten Landis Klassifikation eingestuft.

Befundverteilung Befundkombinationen: Singuläre Stenosen sind etwas mehr als doppelt so häufig wie Mehrfachstenosen (16 versus 7). Die Befunder sahen ca. 1,5 mal so viele unauffällige Befunde wie der Goldstandard (562 versus 380). Singuläre subglottische Stenosen sahen die Befunder etwas häufiger als der Goldstandard (151 versus 120), singuläre glottische und supraglottische Stenosen hingegen nur halb so häufig wie der Goldstandard (47 versus 100 bzw. 46 versus 100). Kombinierte Stenosen wurden von den Befundern deutlich seltener erhoben als vom Goldstandard. Der größte Unterschied in der Befundprävalenz findet sich für kombinierte glottische

und subglottische Stenosen (18 versus 80). Entsprechend der ungleichen Randsummen fällt der Bhapkar-Test für die Kontingenztafel der Befunde (Tabelle 4.16) signifikant aus. Der McNemar-Test für die einzelnen Befundkombinationen ist ebenfalls signifikant und zeigt damit an, dass sich die Befundprävalenzen in allen Befundklassen erheblich unterscheiden (Abbildung 4.18). Die Befunde singulärer Larynxstenosen verteilen sich auf jeweils etwa dreimal so viele verschiedene Videos wie vom Goldstandard vorgegeben[26] und werden von fast allen Befundern getragen. An multiplen Stenosen war nur die Hälfte der Befunder beteiligt.

Präzision Befundkombinationen: Die beste Übereinstimmung innerhalb der Befunder wird bei subglottischen Stenosen mit einem Kappa Fleiss von 0,563 erreicht. Supraglottische und glottische Stenosen liegen mit einem Kappa Fleiss von 0,264, respektive 0,225 eine Größenordnung darunter. Auch hinsichtlich unauffälliger Befunde wird mit einem Kappa Fleiss von 0,463 eine beachtliche Konkordanz erzielt.

Abbildung 4.18: Randverteilungen Befundkombinationen Stenoselokalisation Larynx



Gegenüberstellung der Befundhäufigkeiten der Befunder (grau) mit denen des Goldstandards (schwarz).

Richtigkeit Befundkombinationen: Da Befunder und Goldstandard die gleichen Befundkombinationen wählten, konnten alle Daten in die Analyse der Befundrichtigkeit mit einbezogen werden. Die Datenabdeckung liegt in allen Befundkategorien bei 100 %. Die Sensitivität der singulären subglottischen Stenosen liegt mit 66,7 % deutlich über der von subglottischen Stenosen mit insgesamt 58,6 %. Rein supraglottische und glottische Stenosen erreichen eine Sensitivität um 24 %, was sich nicht wesentlich von den Werten auf Ebene der Einzelbefunde unterscheidet. Die Spezifität singulärer subglottischer Stenosen liegt mit 90,1 % – wie auf der Symptomebene – etwas niedriger als die glottischer und supraglottischer Stenosen.

Bei den positiven prädiktiven Werten sind keine signifikanten Unterschiede zwischen singulären supraglottischen, glottischen und subglottischen Stenosen erkennbar: Alle Befunde waren in gut der Hälfte der Fälle richtig. Spiegelbildlich zur etwas niedrigeren Spezifität liegt der negative prädiktive Wert subglottischer Stenosen mit 94,2 % etwas höher als der glottischer und supraglottische mit 90,4 %. Wie bei den Einzelbefunden hebt sich das Kappa Cohen für subglottische Stenosen mit 0,523 deutlich von dem glottischer und supraglottischer Stenosen ab und erreicht das Niveau einer guten Übereinstimmung. Bei glottischen und supraglottischen Stenosen besteht mit einem Kappa um 0,27 eine schwache Übereinstimmung. Das geringere Kappa Cohen

[26] *supraglottisch 5 gegenüber 13, glottisch 5 gegenüber 14 und subglottisch 6 gegenüber 21*

bei glottischen und supraglottischen Stenosen ist auch, auf die im Vergleich zu subglottischen Stenosen stärker divergierenden Befundprävalenzen zwischen Befundern und Goldstandard, zurückzuführen. Bei mehrfachen Stenosen besteht im Gegensatz zu singulären Stenosen keine erkennbare Übereinstimmung: Weder bei den Befundern untereinander, noch im Vergleich zum Goldstandard wurden nennenswerte Kappawerte erzielt.

Die prävalenzunabhängigen Maßzahlen odds ratio und AUC belegen ebenfalls die überlegene Richtigkeit bei der Befundung subglottischer Stenosen. Die odds ratio kombinierter supraglottischer und glottischer Stenosen, erreicht das Niveau singulärer glottischer und supraglottischer Stenosen, obwohl positive Übereinstimmung, Sensitivität, positiver prädiktiver Wert, Kappa Fleiss und Cohen deutlich darunter liegen.

Bei Betrachtung der Kontingenztafel fällt auf, dass singuläre Stenosen von den Befundern zwar zu einem erheblichen Anteil nicht erkannt, aber verhältnismäßig selten einer anderen Lokalisation zugeordnet werden. Uneinigkeit zwischen Befundern und Goldstandard besteht also in erster Linie hinsichtlich des Vorhandenseins der Stenose, weniger aber hinsichtlich ihrer Lage.

Zusammenfassung 4.5: Stenoselokalisation Larynx

Die Übereinstimmung der Befunde nimmt, sowohl innerhalb der Befunder, als auch im Vergleich zum Goldstandard, von proximal nach distal zu. Subglottische Stenosen werden mit Abstand am sichersten erkannt. Lässt man den bei glottischen und supraglottischen Stenosen den erheblichen Anteil falsch-negativer Befunde außer Acht, nivellieren sich jedoch die Unterschiede. Insgesamt werden Larynxstenosen recht homogen beurteilt. Der größte Anteil der Differenzen zwischen Untersuchern und dem Goldstandard entsteht bei der Frage, ob es sich um eine singuläre oder mehrere benachbarte Stenosen handelt. Hinsichtlich der Lage der Stenose gibt es hingegen wenig abweichende Beurteilungen.

Tabelle 4.16: Kontingenztafel Befundkombinationen Stenoselokalisation Larynx

Bhapkar <0,01		Referenz 0	1			2			Summen
		0	1	10	100	11	101	110	
Befunder 0	0	349	31	60	66	40	1	15	562
	1	25	80	1	3	22	17	3	151
1	10	0	1	24	3	8	1	10	47
	100	3	1	6	24	7	0	5	46
	11	1	6	5	1	3	0	2	18
2	101	0	0	0	0	0	1	0	1
	110	2	1	4	3	0	0	5	15
Summen		380	120	100	100	80	20	40	840

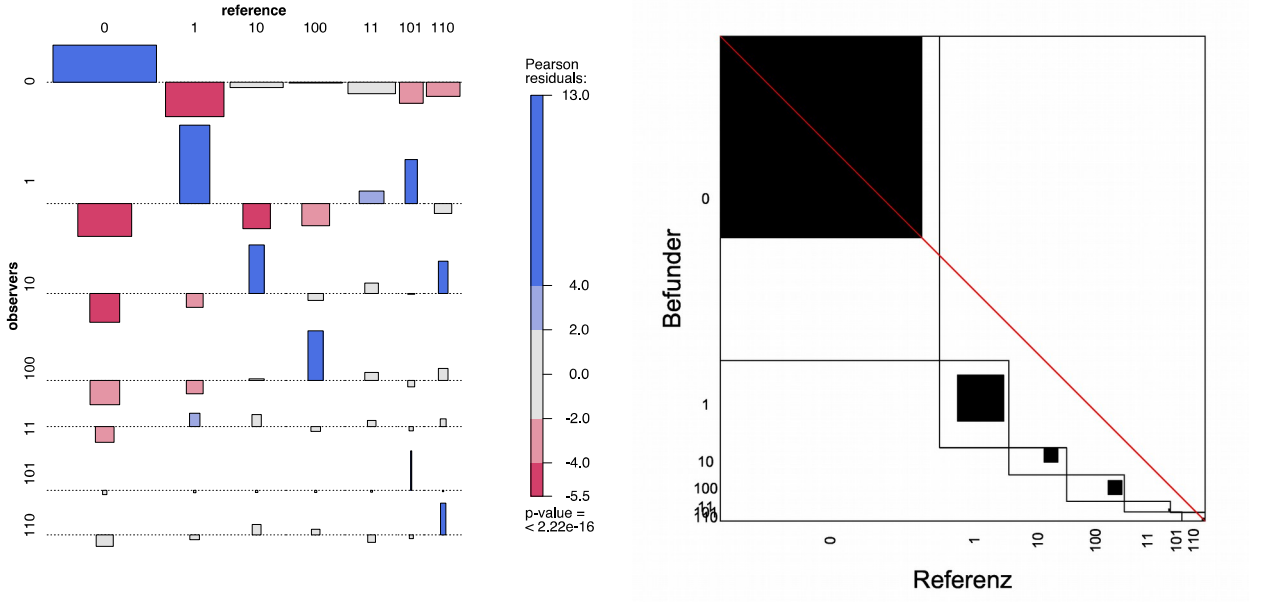
Der signifikante Bhapkar Test zeigt eine ungleiche Verteilung der Randsummen an.

Tabelle 4.17: Kennwerte Befundkombinationen Stenoselokalisation Larynx

	Befund	pre	McN	nag	pag	tnr	tpr	acc	npv	ppv	B	LR-	LR+	AUC	OR	Q	Y
0	0	45,2	<0,01	66,9	74,1	53,7	91,8	71,0	88,8	62,1	0,535	0,152	1,98	0,728	13,1	0,858	0,566
	1	14,3	0,004	92,1	59,0	90,1	66,7	86,8	94,2	53,0	0,832	0,370	6,76	0,784	18,3	0,896	0,621
1	10	11,9	<0,01	93,5	32,7	96,9	24,0	88,2	90,4	51,1	0,870	0,784	7,72	0,604	9,8	0,816	0,517
	100	11,9	<0,01	93,6	32,9	97,0	24,0	88,3	90,4	52,2	0,872	0,783	8,07	0,605	10,3	0,823	0,525
	11	9,5	<0,01	94,2	6,1	98,0	37,5	89,0	90,6	16,7	0,886	0,982	1,9	0,509	1,9	0,319	0,164
2	101	2,4	<0,01	98,9	9,5	100	5,0	97,7	97,7	100	0,977	0,950	Inf	0,525	Inf	NA	NA
	110	4,8	<0,01	97,2	18,2	98,8	12,5	94,6	95,8	33,3	0,945	0,886	10	0,556	11,3	0,837	0,541

Prävalenz (pre), McNemar Test (McN), negative (nag) und positive (pag) Übereinstimmung, Spezifität (tnr), Sensitivität (tpr), Genauigkeit (pac), negativer (npv) und positiver prädiktiver Wert (ppv), Bangdiwala (B), negative (LR-) und positive Likelihood ratio (LR+), area under the curve (AUC), odds ratio (OR), Youden Q und Y.

Abbildung 4.19: Diagramme Befundkombinationen Stenoselokalisation Larynx



Das Assoziationsdiagramm zeigt das Verhältnis der beobachteten Werte zu den Erwartungswerten. Das Bangdiwala-Diagramm stellt für jede Befundklasse die beobachtete Übereinstimmung (schwarz) der maximal möglichen (umgebendes Rechteck) gegenüber. Die Seitenverhältnisse illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Je mehr die Rechtecke von einem Quadrat abweichen um so ungleicher sind die Prävalenzen in der betreffenden Befundkategorie.

4.2.3.2.2 Stenoselokalisation Trachea

Die Lokalisation innerhalb der Trachea wurde im Befundfragebogen auf ein proximales, mittleres und distales Drittel abgebildet.

Tabelle 4.18: Inter-Beobachter-Variabilität Einzelbefunde Stenoselokalisation Trachea

Befund	**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz			
	Referenz	Befunder Anzahl verschiedener			Ø positive Übereinstimmung [%]	**Kappa Fleiss**		Sensitivität [%]	positiver prädiktiver Wert [%]	**Kappa Cohen**	
Stenoselokalisation in der Trachea	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)		Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**			Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**
proximales Drittel	10 [200]	75	19	20	37,8	0,150	gering	26,0	69,3	0,285	schwach
mittleres Drittel	5 [100]	90	20	17	35,4	0,299	schwach	48,0	53,3	0,442	moderat
distales Drittel	9 [180]	108	20	16	54,2	0,393	mäßig	43,3	72,2	0,454	moderat

Die Tabelle betrachtet die Übereinstimmung der Einzelbefunde und ist in die Kriterien Befundverteilung, Präzision und Richtigkeit gegliedert. Details siehe Kommentar Tabelle 4.11 Seite 106.

Befundverteilung Einzelbefunde: Die Prävalenz von Trachealstenosen variiert von 23,8 % im proximalen Drittel, über 11,9 % im mittleren Drittel, bis 21,4 % im distalen Drittel. Stenosen des mittleren Drittels werden von Goldstandard und Befundern vergleichbar häufig beobachtet. Stenosen im proximalen Drittel sehen die Befunder fast 3 mal weniger und im distalen Drittel etwas mehr als halb so häufig wie der Goldstandard. In allen drei Abschnitten wurden Trachealstenosen von nahezu allen Befundern dokumentiert. Proximale Trachealstenosen wurden in knapp der Hälfte aller Videos gesehen, Stenosen im mittleren und distalen Drittel in jeweils um 40 % der Videos. Trachealstenosen im proximalen und distalen Drittel wurden damit in doppelt so vielen verschiedenen Videos gesehen wie vom Goldstandard. Stenosen im mittleren Drittel in etwa 3 mal so vielen Videos. Der McNemar-Test zeigt, dass die Übereinstimmung bei Stenosen im proximalen und distalen Tracheadrittel – anders als im mittleren Drittel – durch signifikant inhomogene Randsummen beeinträchtigt wird.

Präzision Einzelbefunde: Bei proximalen Trachealstenosen ist innerhalb der Befunder nur eine schwache Übereinstimmung zu erkennen. Deutlich mehr Einigkeit wird hinsichtlich Trachealstenosen im mittleren und insbesondere distalen Drittel erzielt. Die Übereinstimmung nimmt von proximal nach distal zu.

Richtigkeit Einzelbefunde: Knapp die Hälfte der Trachealstenosen im mittleren Drittel wird erkannt. Ebenfalls in etwa der Hälfte der Fälle ist der Befund einer Trachealstenose im mittleren Drittel verlässlich. Die Quote der erkannten Stenosen im distalen Drittel liegt etwas darunter, die diagnostische Aussagekraft mit einem prädiktiven Wert von ca. 70 % aber deutlich darüber. Nur jede vierte Trachealstenose im proximalen Drittel wird erkannt. Ein positiver Befund ist jedoch ähnlich zuverlässig wie der einer Stenose im distalen Drittel.

Hinsichtlich der Übereinstimmung mit dem Goldstandard liegen mittlere und distale Trachealstenosen mit einem Cohen Kappa von 0,442 bzw. 0,454 gleich auf. Proximale Trachealstenosen fallen dagegen mit einem Cohen Kappa von nur 0,285 deutlich ab. Die odds ratios als prävalenzunabhängige Maßzahlen vermitteln mit Werten von 15,3 bzw. 16,1 gegenüber 9,4 das gleiche Bild.

Tabelle 4.19: Vierfeldertafeln Einzelbefunde Stenoselokalisation Trachea

proximales Drittel					mittleres Drittel					distales Drittel				
McN <0,01		Referenz 0	Referenz 1	Summen	McN 0,353		Referenz 0	Referenz 1	Summen	McN <0,01		Referenz 0	Referenz 1	Summen
Befunder	0	617	148	765	Befunder	0	698	52	750	Befunder	0	630	102	732
	1	23	52	75		1	42	48	90		1	30	78	108
Summen		640	200	840	Summen		740	100	840	Summen		660	180	840

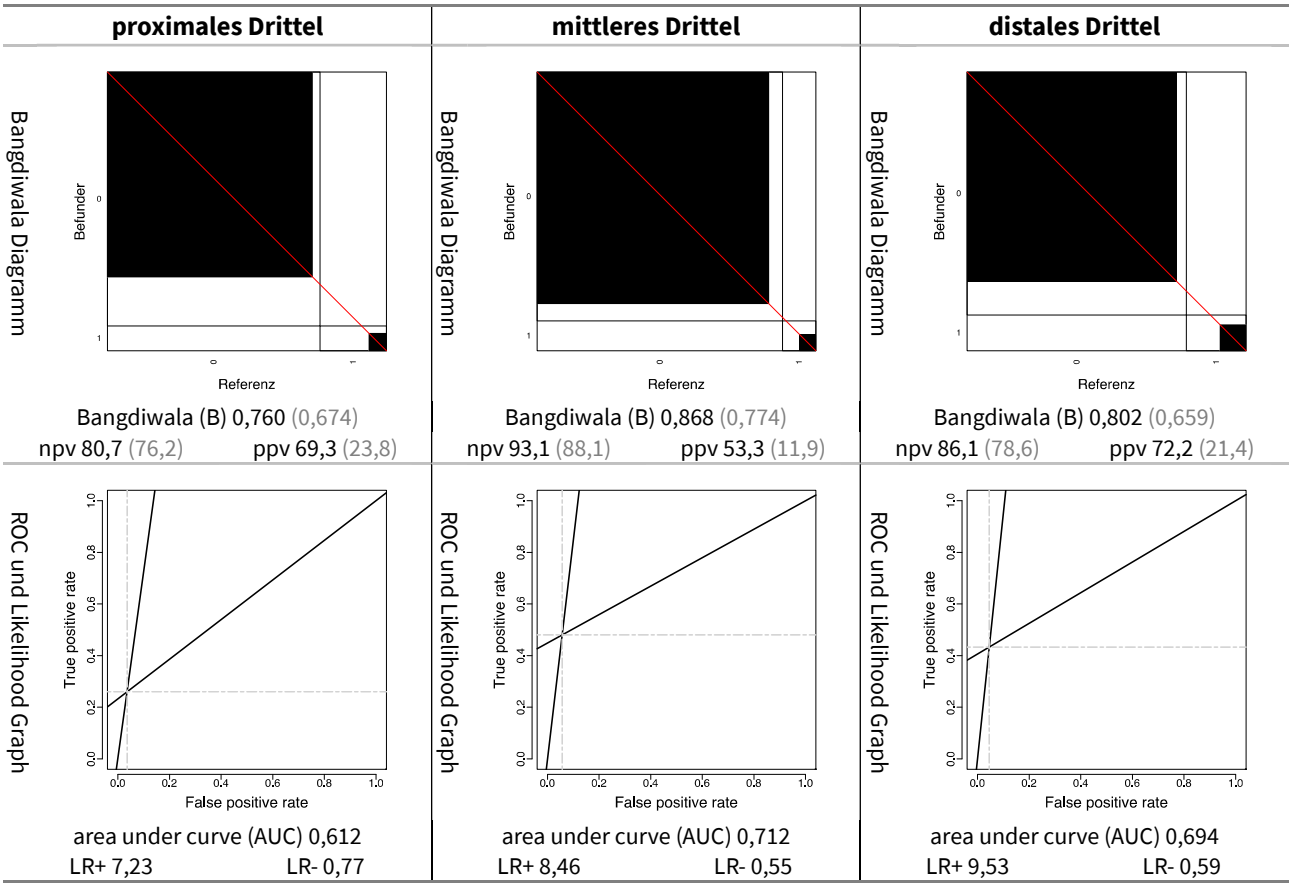
Die Vier-Felder-Tafeln und ihre Randsummen sind Berechnungsgrundlage sämtlicher Kennwerte und Diagramme.

Tabelle 4.20: Kennwerte Einzelbefunde Stenoselokalisation Trachea

proximales Drittel				mittleres Drittel				distales Drittel			
Prävalenz	23,8			Prävalenz	11,9			Prävalenz	21,4		
	beo.	erw.	kor.		beo.	erw.	kor.		beo.	erw.	kor.
neg. Übereinst.	87,8	83,0	=κ	neg. Übereinst.	93,7	88,7	=κ	neg. Übereinst.	90,5	82,6	=κ
pos. Übereinst.	37,8	13,0	=κ	pos. Übereinst.	50,5	11,3	=κ	pos. Übereinst.	54,2	16,1	=κ
Spezifität	96,4	91,1	59,8	Spezifität	94,3	89,3	47,0	Spezifität	95,5	87,1	64,6
Sensitivität	26,0	8,9	18,7	Sensitivität	48,0	10,7	41,8	Sensitivität	43,3	12,9	35,0
Genauigkeit	79,6	71,5	28,5=κ	Genauigkeit	88,8	79,9	44,2=κ	Genauigkeit	84,3	71,2	45,4=κ
odds ratio (Yule Q/Y)	9,4 (0,808/0,509)			odds ratio (Yule Q/Y)	15,3 0,878 / 0,593			odds ratio (Yule Q/Y)	16,1 0,883 / 0,601		

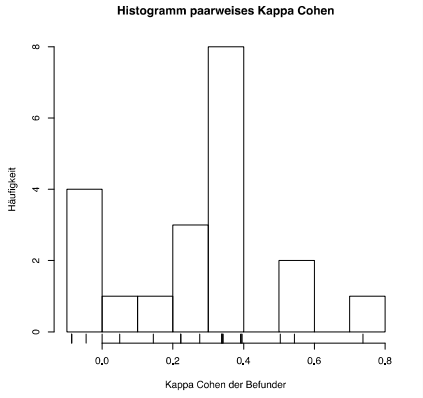
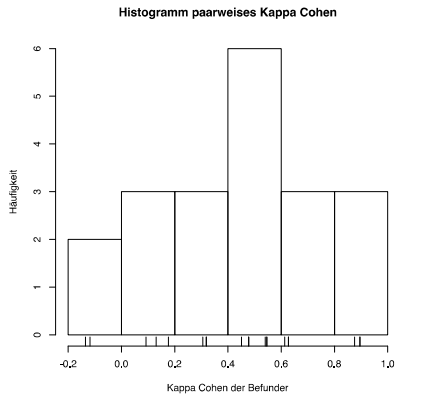
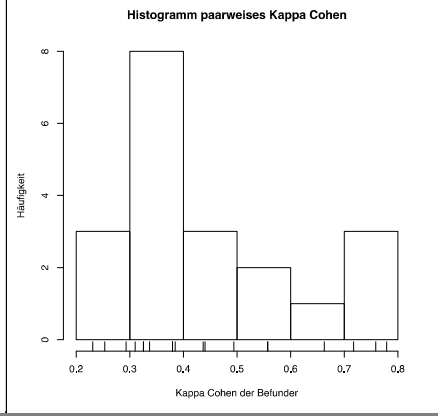
Beobachtete Kennwerte (beo.), Erwartungswerte gemäß Randsummen (erw.) & für erwartete Übereinstimmung korrigierte Werte (kor.) in Prozent. Die odds ratio und ihre Transformationen (Yule Q & Y) sind prävalenzunabhängig.

Abbildung 4.20: Diagramme Einzelbefunde Stenoselokalisation Trachea



Die Bangdiwala-Diagramme stellen für jede Befundklasse die beobachtete Übereinstimmung (schwarz) der maximal möglichen (umgebendes Rechteck) gegenüber.Die Seitenverhältnisse illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Die Receiver Operator Characteristic (ROC) trägt die Rate falsch Positiver (=1-Spezifität, fpr) gegen die Rate richtig Positiver (=Sensitivität, tpr) auf. Die Steigung der Tangenten durch den Ursprung bzw. (1,1) definieren die Likelihood ratios (LR+/LR-), die Fläche darunter die Area Under the Curve (AUC).

Tabelle 4.21: Paarweises Kappa Cohen Stenoselokalisation Trachea

proximales Drittel	mittleres Drittel	distales Drittel
vereintes Kappa Cohen	vereintes Kappa Cohen	vereintes Kappa Cohen
0,285	**0,442**	**0,454**
paarweises Kappa Cohen	paarweises Kappa Cohen	paarweises Kappa Cohen
min. Ø median max.	min. Ø median max.	min. Ø median max.
-0,086 **0,268** 0,339 0,738	-0,135 **0,433** 0,478 0,895	0,230 **0,450** 0,382 0,779
		

Die Tabelle vergleicht die Ergebnisse der beiden Berechnungsvarianten „vereintes“ und „paarweises“ Kappa Cohen.

Tabelle 4.22: Inter-Beobachter-Variabilität Befundkombinationen Stenoselokalisation Trachea

:Befund				Befundverteilung				Präzision Übereinstimmung der Befunder untereinander			Richtigkeit Übereinstimmung mit dem Goldstandard in den überschneidenden Befunden				
				Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
Anzahl der Stenosen	proximales Drittel	mittleres Drittel	distales Drittel	Anzahl verschiedener Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	**Kappa nach Cohen**	modifizierte Klassifikation **nach Landis**	Datenabdeckung [%]
x	**gesamt**			**17 [340]**	**227**	**20**	**28**	**21,0**	**0,324**	**mittelmäßig**	**30,3**	**49,4**	**0,385**	**mittelmäßig**	**91,8**
0			0	25 [500]	613	20	40	76,6	0,474	beachtlich	93,4	76,9	0,467	beachtlich	98,0
			1	4 [80]	80	19	14	35,5	0,325	mittelmäßig	50,7	43,8	0,413	beachtlich	91,2
1		1	0	NA	52	19	15	24,0	0,232	mittelmäßig	NA	NA	NA	NA	NA
	1	0	0	7 [140]	57	17	18	21,2	0,115	schwach	25,9	64,8	0,300	mittelmäßig	95,1
		1	1	3 [60]	20	13	9	15,0	0,078	schwach	18,6	40,0	0,227	mittelmäßig	76,4
2	1	0	1	1 [20]	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	1	1	0	1 [20]	10	9	7	12,5	0,031	schwach	14,3	11,1	0,116	schwach	51,7
3	1	1	1	1 [20]	8	5	3	17,5	0,110	schwach	20,0	80,0	0,313	mittelmäßig	87,5

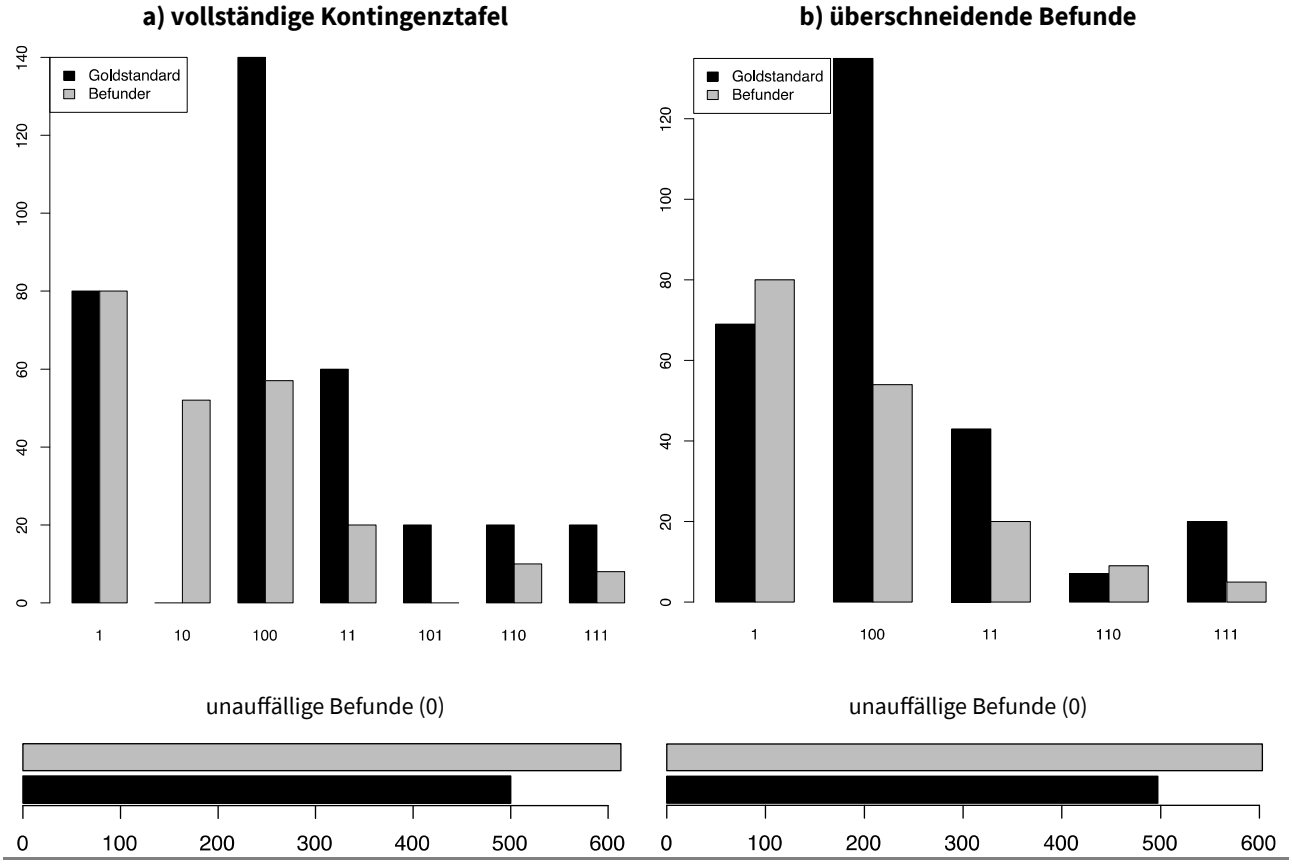
Die Tabelle analysiert die Befundkombinationen. Sie gliedert sich in die 3 Hauptbereiche Befundverteilung Präzision und Richtigkeit. Details siehe Kommentar Tabelle 4.15 auf Seite 109.

Befundverteilung Befundkombinationen: Für Trachealstenosen insgesamt ergibt sich gemäß dem Goldstandard eine Prävalenz von 40,5 % (17/42), wobei singuläre Stenosen des distalen Drittels einen Anteil von 9,5 % (4/42) und singuläre Stenosen im proximalen Drittel einen Anteil von 16,6 % haben. Stenosen im mittleren Drittel waren, gemäß der Referenz, nur in Kombination mit Stenosen im Distalen Drittel zu sehen. Der übrige Anteil von 14,3 % entfällt auf multiple Stenosen. Kombinierte Stenosen im mittleren und distalen Drittel stellen dabei mit 7,1 % den größten Anteil. Singuläre distale Trachealstenosen werden von den Befundern genauso häufig, wie

vom Goldstandard gesehen, verteilen sich aber dabei auf mehr als 3 mal so viele Videos. Trachealstenosen im proximalen Drittel machen bei den Befundern nur 41 % (57/140) der des Goldstandards aus und betreffen 2,5 mal so viele Videos. Die Befunde singulärer Trachealstenosen werden von dem Großteil der Befunder getragen (mind. 17/20). Doppelte Stenosen wurden von 45 bis 65 % der Befunder erhoben und machen insgesamt nur einen Anteil von 3,6 % aus. Beim Goldstandard entspricht der Anteil doppelter Trachealstenosen 11,9 %. Globale Trachealstenosen sind mit unter 1 % bei den Befundern die absolute Ausnahme, die von nur einem Viertel der Befunder gesehen wurde.

Präzision Befundkombinationen: Singuläre proximale Stenosen weisen mit einem Kappa Fleiss von 0,115 die geringste Präzision auf. Bei Stenosen im mittleren und distalen Drittel besteht eine mittelmäßige Übereinstimmung innerhalb der Befunder, wobei Stenosen im distalen Drittel mit einem Kappa Fleiss von 0,325 am einheitlichsten befundet werden. Bei mehrfachen Stenosen wird nur hinsichtlich globaler Stenosen über alle 3 Abschnitte eine gerade noch messbare Übereinstimmung innerhalb der Befunder erzielt (Kappa Fleiss 0,110).

Abbildung 4.21: Randverteilungen Befundkombinationen Stenoselokalisation Trachea



Die Balkendiagramme stellen die Befundhäufigkeiten der Befunder denen des Goldstandards gegenüber.

Richtigkeit Befundkombinationen: Isolierte Trachealstenosen im mittleren Drittel können nicht auf ihre Richtigkeit untersucht werden, da sie gemäß dem Goldstandard in den vorgelegten Videomitschnitten nicht zu sehen waren. Kombinierte Stenosen im proximalen und distalen Drittel waren zwar nach Einschätzung des Goldstandards in einem Video zu sehen, wurden jedoch von keinem Befunder beobachtet.

In den bisherigen Analysen waren die von Befundern und Goldstandard gewählten Befundkombinationen deckungsgleich. Bei Trachealstenosen finden sich erstmals Differenzen, sodass ein Teil der Daten in die Berechnung von Maßzahlen der Richtigkeit nicht mit einbezogen werden kann. Mit einer Datenabdeckung von insgesamt knapp 92 % und einer ähnlichen Verteilung der Fehlwerte auf die einzelnen Befundkombinationen, kann die Analyse jedoch trotzdem als valide

angesehen werden. Lediglich beim Befund der kombinierten proximalen und mittleren Trachealstenose, dessen Prävalenz ohnehin bei unter 1 % liegt, ist ein Verlust von knapp der Hälfte der Daten zu verzeichnen. In die Berechnung singulärer Stenosen des proximalen und distalen Drittels flossen jeweils über 90 % der originären Daten ein.

Trachealstenosen im distalen Drittel wurden zur Hälfte erkannt, Stenosen im proximalen Drittel nur zu einem Viertel. Die Spezifität liegt bei proximalen Stenosen mit 97 % etwas höher als bei distalen. Proximale Trachealstenosen sind den distalen Stenosen hinsichtlich ihres positiven prädiktiven Wertes mit 64,8 % gegenüber 43,8 % überlegen. Auch in Bezug auf den negativen prädiktiven Wert sind sie mit 86,1 % gegenüber 95,1 % unterlegen. Unter Berücksichtigung der zufällig zu erwartenden Übereinstimmung, werden distale Trachealstenosen bei einem Kappa Cohen von 0,413 sicherer befundet, als Stenosen im proximalen Drittel (Kappa Cohen 0,300).

Es wird zwar nur jede 5. globale Trachealstenose als solche erkannt, positive Befund erreichen aber einen positiven prädiktiven Wert von 80 % und das Kappa Cohen steht mit 0,313 für eine mittelmäßige Übereinstimmung. Abseits globaler Stenosen findet sich bei mehrfachen Stenosen kein nennenswerter Konsens zwischen den Befundern und dem Goldstandard. Die Übereinstimmung hinsichtlich proximaler Trachealstenosen wird fast ausschließlich durch die Frage ihrer Existenz beeinträchtigt wird. Betrachtet man die Gruppe der positiven Befunde, kommen Fehlklassifikationen kaum vor: Nur in etwa 10 % der Fälle (5/43) wird die proximale Trachealstenose im mittleren Drittel gesehen, nie im distalen Drittel. In 7 % (3/43) wird statt einer singulären proximalen Stenose eine kombinierte im proximalen und mittleren Drittel angegeben. Der mit 92,4 % (97/105) weitaus überwiegende Teil der divergenten Befunde entsteht bei der Frage, ob überhaupt eine Stenose vorliegt. Innerhalb der Gruppe der vom Goldstandard vorgegebenen singulären proximalen Stenosen wurden von den Befundern mehr als doppelt so viele Befunde als unauffällig angesehen (97) als mit einer Stenose bewertet (43).

Auch bei distalen Stenosen macht die Kategorie „fehlende Stenose“ mit 18/80 den größten Anteil aus, es besteht eine deutlich breitere Streuung über andere Klassen. In 11/62 Fällen wurde die Stenose statt im distalen im mittleren Drittel gesehen. In 4/62 Fällen sogar im proximalen Drittel. In 10/62 Fällen wurde eine kombinierte Stenose im distalen und proximalen Drittel angegeben. Mehr als die Hälfte der erkannten Stenosen (35/62) wurden korrekt im proximalen Drittel befundet.

Subgruppenanalyse positive Befunde: Beschränkt man die Betrachtung auf die Befunde, bei denen von Befundern und dem Goldstandard Trachealstenosen erhoben wurden[27], steigt das Kappa Cohen von 0,385 auf 0,456 an. Während sich das Kappa Cohen bei singulären Stenosen im distalen Drittel nicht verändert[28], ist bei Stenosen im proximalen Drittel ein sprunghafter Anstieg von 0,300 auf 0,775 zu beobachten. Bei globalen Stenosen besteht ein ähnlicher Effekt: Kappa Cohen erhöht sich von 0,313 auf 0,652. Dieser Anstieg ist durch den verhältnismäßig geringen Anteil fehlklassifizierter Stenosen im Vergleich zu falsch negativen, respektive falsch positiven Befunden bei Trachealstenosen im proximalen Drittel, sowie globalen Trachealstenosen zu erklären. Der Großteil fehlklassifizierter Befunde fällt somit bei der Subgruppenanalyse positiver Befunde weg (siehe Kontingenztafel 4.23 Seite 118). Die Randsummen sind innerhalb der positiven Befunde deutlich homogener verteilt: Der McNemar Test fällt nur bei kombinierten Stenosen im mittleren und distalen Tracheadrittel positiv aus.

[27] *Die Anzahl der in die Analyse einbezogenen Befunde sinkt hierdurch von 771 auf 135.*

[28] *Kappa Cohen 0,413 versus 0,415*

Tabelle 4.23: Kontingenztafel Befundkombinationen Stenoselokalisation Trachea

Bhapkar <0,01			Referenz							Summen	
				1			2		3		
			0	1	100	11	101	110	111		
Befunder		0	464	18	97	11	10	0	13	603	613
		1	23	35	0	16	0	6	0	80	80
	1	10	3	11	5	17	3	13	0		52
		100	9	4	35	5	3	0	1	54	57
	2	11	1	10	0	8	0	0	1	20	20
		110	0	1	3	3	1	1	1	9	10
	3	111	0	1	0	0	3	0	4	5	8
Summen			497	69	135	43		7	20	771	
			500	80	140	60	20	20	20		840

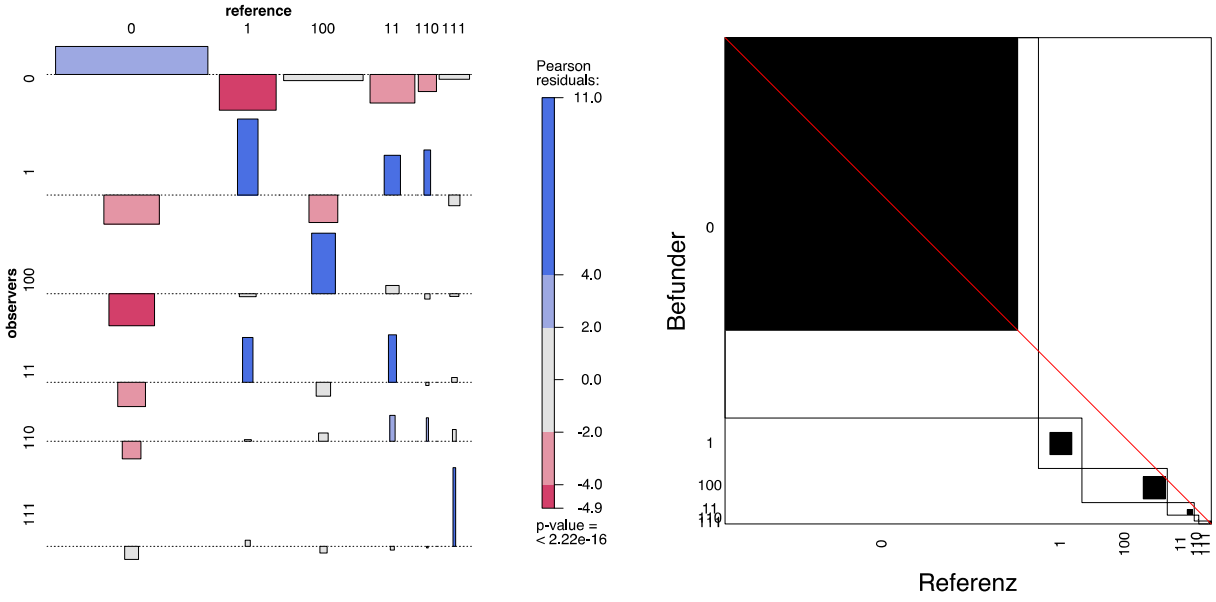
Die Kontingenztafel mit ihren Randsummen ist Berechnungsgrundlage sämtlicher Kennwerte und Diagramme.

Tabelle 4.24: Kennwerte überschneidende Befundkombinationen Stenoselokalisation Trachea

Befund		pre	McN	nag	pag	spe	sen	acc	npv	ppv	B	LR-	LR+	AUC	OR	Q	Y
0	0	64,5	<0,01	61,1	84,4	49,3	93,4	77,7	80,4	76,9	0,675	0,135	1,84	0,713	13,7	0,864	0,574
1	1	8,9	0,261	94,3	47	93,6	50,7	89,8	95,1	43,8	0,882	0,527	7,91	0,722	15,0	0,875	0,59
	100	17,5	<0,01	91,2	37	97,0	25,9	84,6	86,1	64,8	0,824	0,764	8,68	0,615	11,4	0,838	0,542
2	11	5,6	<0,01	96,8	25,4	98,4	18,6	93,9	95,3	40,0	0,936	0,828	11,4	0,585	13,6	0,863	0,574
	110	0,9	0,789	99,1	12,5	99,0	14,3	98,2	99,2	11,1	0,982	0,866	13,6	0,566	15,8	0,881	0,597
3	111	2,6	<0,01	98,9	32	99,9	20,0	97,8	97,9	80,0	0,978	0,801	150	0,599	188	0,989	0,864

Prävalenz (pre), McNemar Test (McN), negative (nag) und positive (pag) Übereinstimmung, Spezifität (spe), Sensitivität (sen), Genauigkeit (acc), negativer (npv) und positiver prädiktiver Wert (ppv), Bangdiwala (B), negative (LR-) und positive Likelihood ratio (LR+), area under the curve (AUC), odds ratio (OR), Youden Q und Y.

Abbildung 4.22: Assoziationsdiagramme Stenoselokalisation Trachea.



Links: Beobachtungen im Vergleich zu den Erwartungswerten in der Kontingenztafel der Befundkombinationen. Rechts: Das Bangdiwala-Diagramm stellt für jede Befundklasse die beobachtete Übereinstimmung (schwarz) der maximal möglichen (umgebendes Rechteck) gegenüber. Die Seitenverhältnisse illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Je mehr die Rechtecke von einem Quadrat abweichen um so ungleicher sind die Prävalenzen in der betreffenden Befundkategorie.

Zusammenfassung 4.6: Stenoselokalisation Trachea

Bei Betrachtung der Einzelbefunde nehmen Präzision und Richtigkeit von proximal nach distal zu, ein Trend, der sich so auch für singuläre Stenoselokalisation bei syndromaler Betrachtung bestätigt. Diese Übereinstimmung wird erheblich von falsch-negativen Befunden, also der Frage, ob überhaupt eine Stenose vorliegt, beeinträchtigt. Der starke Einfluss falsch-negativer Befunde ist auch bei globalen Trachealstenosen klar zu erkennen. Bei kombinierten Stenosen findet sich keine wesentliche Übereinstimmung.

4.2.3.2.3 Stenoselokalisation Bronchien

Bei der Lokalisation von Stenosen im Hauptbronchus wurde zwischen „rechts“ und „links“ unterschieden, weiter distal konnte der Lappen angegeben werden. Rechts standen Ober-, Mittel- und Unterlappen, links Oberlappen, Unterlappen und Lingula zur Auswahl (siehe Abbildung 4.47 auf Seite 179).

4.2.3.2.3.1 Hauptbronchus

Tabelle 4.25: Einzelbefunde Stenoselokalisation Hauptbronchus

Befund	**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz			
	Referenz	Befunder Anzahl verschiedener			Ø positive Übereinstimmung [%]	**Kappa Fleiss**		Sensitivität [%]	positiver prädiktiver Wert [%]	**Kappa Cohen**	
	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)		Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**			Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**
rechts	4 [80]	60	20	10	36,9	0,443	moderat	50,0	66,7	0,533	gut
links	6 [120]	98	20	14	40,4	0,509	gut	50,0	61,2	0,484	moderat

Die Tabelle betrachtet die Stenoselokalisationen des Hauptbronchus unabhängig von möglichen Kombinationen. Der deskriptiven Befundverteilung folgen Betrachtungen der Übereinstimmung innerhalb der Befunder (Präzision) sowie im Vergleich zum Goldstandard (Richtigkeit). Zur besseren Vergleichbarkeit zwischen Referenz und Befundern ist in eckigen Klammern die gewichtete Anzahl der Befunde des Goldstandards angegeben (Faktor 20 im Vgl. zu den Befundern).

Befundverteilung Einzelbefunde: Stenosen im rechten Hauptbronchus waren in knapp 10 % der Videos, Stenosen im linken Hauptbronchus in knapp 15 % der Videos zu sehen. Bronchiale Stenosen wurden von den Befundern vergleichbar häufig wie vom Goldstandard befundet. Alle Befunder haben Stenosen im Hauptbronchus diagnostiziert. Die Diagnosen der Befunder verteilen sich auf mehr als doppelt so viele Videos wie die des Goldstandards.

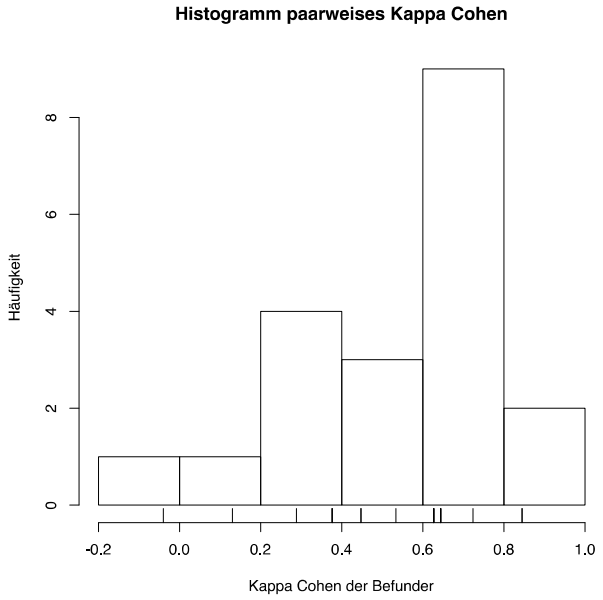
Präzision Einzelbefunde: Bei linksseitigen Trachealstenosen besteht bei einem Kappa Fleiss von 0,509 eine etwas höhere Einigkeit innerhalb der Befunder, als bei rechtsseitigen Trachealstenosen.

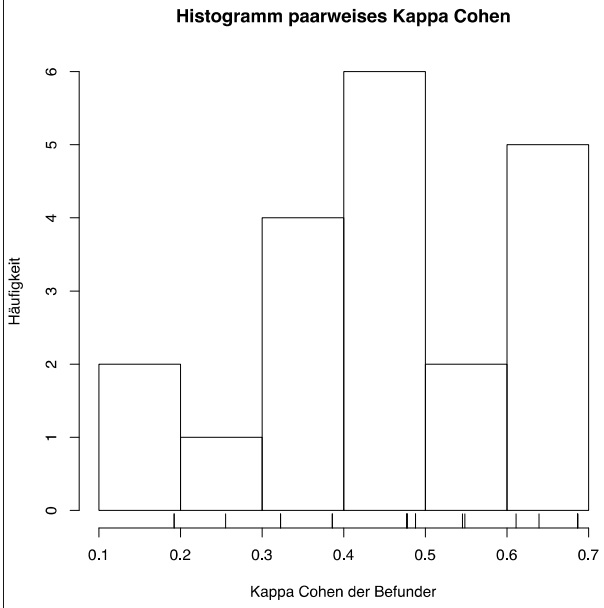
Richtigkeit Einzelbefunde: In beiden Hauptbronchien wurden Stenosen nur zur Hälfte erkannt. Der positive prädiktive Wert liegt bei knapp 67 % für den rechten und 61 % für den linken Hauptbronchus und ist damit nicht höher als bei laryngealen oder trachealen Stenosen. Spezifität und negativer prädiktiver Wert liegen rechtsseitig ebenfalls etwas höher, als links. Sowohl innerhalb der Befunder, als auch im Vergleich zum Goldstandard stehen die Kappa Werte für eine gute Übereinstimmung. Die Übereinstimmung mit dem Goldstandard ist rechtsseitig etwas stärker

ausgeprägt, als links. Diese Überlegenheit drückt sich auch im Bangdiwala B und ganz knapp auch in der AUC aus. Die Unterschiede in der AUC gehen auf die deutlich höheren LR+ bei rechtsseitigen Hauptbronchusstenosen zurück, die LR- ist auf beiden Seiten gleich. Noch deutlicher kommen die Unterschiede in der OR und Yule Q / Y zum Ausdruck.

Tabelle 4.26: Paarweises Kappa Stenoselokalisation Hauptbronchus

rechts				**links**			
vereintes Kappa Cohen				vereintes Kappa Cohen			
0,533				**0,484**			
paarweises Kappa Cohen				paarweises Kappa Cohen			
min.min.	Ø	median	max.	min.	Ø	median	max.
-0,040	**0,522**	0,627	0,844	0,192	**0,467**	0,478	0,687





Die Tabelle vergleicht die Ergebnisse der beiden Berechnungsvarianten „vereintes“ und „paarweises“ Kappa Cohen.

Tabelle 4.27: Vier-Felder-Tafeln Einzelbefunde Stenoselokalisation Hauptbronchus

rechts					links				
McN 0,014		Referenz 0	Referenz 1	Summen	McN 0,034		Referenz 0	Referenz 1	Summen
Befunder	0	740	40	780	Befunder	0	682	60	742
	1	20	40	60		1	38	60	98
Summen		760	80	840	Summe		720	120	840

Die Vier-Felder-Tafeln und ihre Randsummen sind Berechnungsgrundlage sämtlicher Kennwerte und Diagramme.

Tabelle 4.28: Kennwerte Einzelbefunde Stenoselokalisation Hauptbronchus

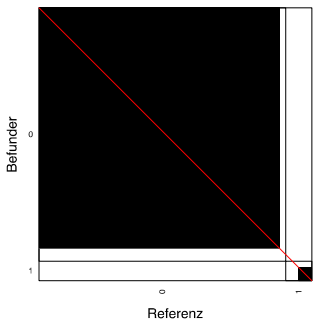
rechts				links			
Prävalenz		9,5		Prävalenz		14,3	
	beo.	erw.	kor.		beo.	erw.	kor.
neg. Übereinstim.	96,1	91,7	=κ	neg. Übereinstim.	93,3	87,0	=κ
pos. Übereinstim.	57,1	8,2	=κ	pos. Übereinstim.	55,0	12,8	=κ
Spezifität	97,4	92,9	63,2	Spezifität	94,7	88,3	54,8
Sensitivität	50,0	7,1	46,2	Sensitivität	50,0	11,7	43,4
Genauigkeit	92,9	84,7	53,3=κ	Genauigkeit	88,3	77,4	48,4=κ
Odds ratio		36,41		Odds ratio		17,9	
Yule Q / Y		0,947 / 0,718		Yule Q / Y		0,894 / 0,618	

Beobachtete Kennwerte (beo.), Erwartungswerte gemäß Randsummen (erw.) & für erwartete Übereinstimmung korrigierte Werte (kor.) in Prozent. Die odds ratio und ihre Transformationen (Yule Q & Y) sind prävalenzunabhängig.

Abbildung 4.23: Diagramme Einzelbefunde Stenoselokalisation Hauptbronchus

rechts

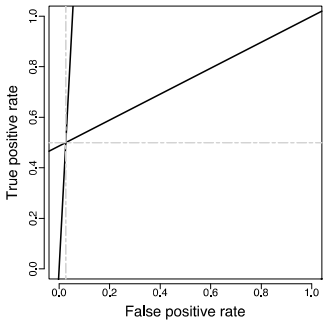
Bangdiwala Diagramm



Bangdiwala (B) 0,919 (0,833)

npv 94,9 (90,5) ppv 66,7 (9,5)

ROC und Likelihood Graph

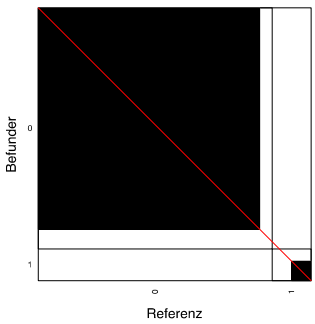


area under curve (AUC) 0,737

LR+ 19,0 LR- 0,51

links

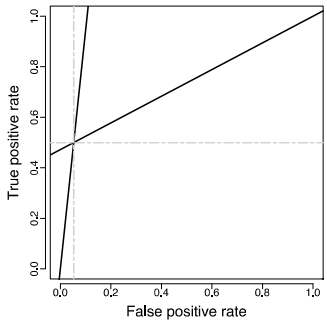
Bangdiwala Diagramm



Bangdiwala (B) 0,858 (0,741)

npv 91,9 (85,7) ppv 61,2 (14,3)

ROC und Likelihood Graph



area under curve (AUC) 0,724

LR+ 9,47 LR- 0,53

Die Bangdiwala-Diagramme stellen für jede Befundklasse die beobachtete Übereinstimmung (schwarz) der maximal möglichen (umgebendes Rechteck) gegenüber.Die Seitenverhältnisse illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Die Receiver Operator Characteristic (ROC) trägt die Rate falsch Positiver (=1-Spezifität, fpr) gegen die Rate richtig Positiver (=Sensitivität, tpr) auf. Die Steigung der Tangenten durch den Ursprung bzw. (1,1) definieren die Likelihood ratios (LR+/LR-), die Fläche darunter die Area Under the Curve (AUC).

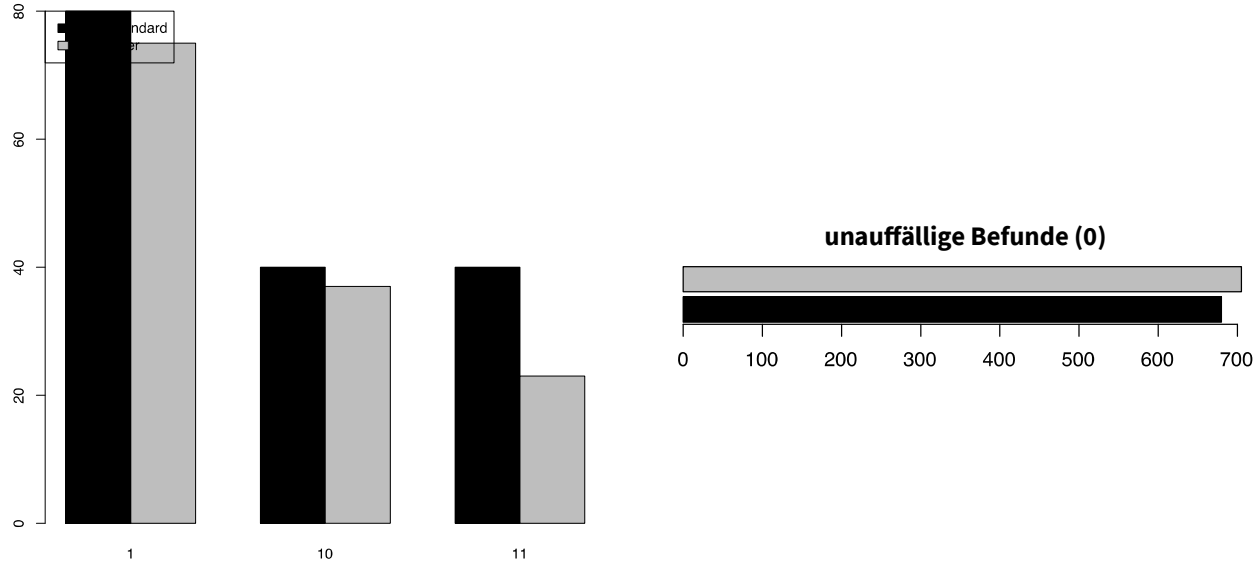
Tabelle 4.29: Befundkombinationen Stenoselokalisation Hauptbronchus

Befund			**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz				
			Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
			Anzahl verschiedener											
Anzahl der Stenosen	rechts	links	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**	Datenabdeckung [%]
x	**gesamt**		**8 [160]**	**135**	**20**	**16**	**30,6**	**0,487**	**moderat**	**29,8**	**49,3**	**0,452**	**moderat**	**100**
0	0	0	34 [680]	705	20	42	83,9	0,533	gut	93,5	90,2	0,536	gut	100
1	0	1	4 [80]	75	20	12	36,5	0,542	gut	33,8	36	0,282	schwach	100
	1	0	2 [40]	37	18	8	29,2	0,356	mäßig	55	59,5	0,551	gut	100
2	1	1	2 [40]	23	15	6	26,2	0,297	schwach	32,5	56,5	0,392	mäßig	100

Die Tabelle analysiert die Befundkombinationen im Sinne eines Gesamtbefundes. Sie ist nach den 3 Kriterien Befundverteilung, Präzision und Richtigkeit gegliedert. Details siehe Kommentar Tabelle 4.15 auf Seite 109.

Befundverteilung Befundkombinationen: Stenosen in den Hauptbronchien haben insgesamt eine Prävalenz von 19 % (160/840). Drei Viertel (6/8) hiervon sind einseitige Stenosen, ein Viertel beidseitige. Linksseitige und rechtsseitige Stenosen, sowie unauffällige Befunde, wurden von den Befundern und dem Goldstandard in etwa gleich häufig erhoben. Beidseitige Stenosen wurden von den Befundern etwas mehr als halb so häufig wie vom Goldstandard gesehen. An der Befundung singulärer Stenosen beteiligten sich mit mind. 18/20 fast alle Befunder. Globale Stenosen wurden von 75 % der Befunder gesehen. Die Befunde linksseitiger und globaler Stenosen verteilen sich bei den Befundern auf dreimal so viele Videos, wie beim Goldstandard, die Befunde rechtsseitiger Stenosen auf viermal so viele Videos.

Abbildung 4.24: Randverteilungen Befundkombinationen Stenoselokalisation Hauptbronchus



Das Säulendiagramm vergleicht die Befundverteilung der Befunder (grau) mit der des Goldstandards (schwarz).

Präzision Befundkombinationen: Wie schon auf Ebene der Einzelbefunde, errechnet sich für linksseitige Stenosen mit einem Kappa Fleiss von 0,542 eine einheitlichere Befundung innerhalb der Befunder, als für die rechte Seite.

Richtigkeit Befundkombinationen: Ein Drittel der linksseitigen und etwas mehr als die Hälfte der rechtsseitigen Stenosen im Hauptbronchus wird als solche erkannt. Stenosen im linken Hauptbronchus erreichen einen positiven prädiktiven Wert von 36 %, Stenosen, im rechten Hauptbronchus einen positiven prädiktiven Wert von knapp 60 %. Hinsichtlich der Spezifität (98,1 % gegenüber 93,7 %) und des negativen prädiktiven Wertes (97,8 % versus 93,1 %) sind rechtsseitige Stenosen im Hauptbronchus linksseitigen überlegen. Insgesamt ergibt sich — spiegelbildlich zur Präzision — bei rechtsseitigen Stenosen im Hauptbronchus eine wesentlich höhere Übereinstimmung mit dem Goldstandard als bei linksseitigen (Kappa Cohen 0,551 versus 0,282).

Die Divergenz bei der Befundung singulärer Stenosen geht — wie bei trachealen und laryngealen Stenosen — im Wesentlichen auf die Frage zurück, ob überhaupt eine Stenose vorliegt. Fehlklassifikationen treten dahinter deutlich zurück. Beidseitige Stenosen wurden von denen Befundern überwiegend als isolierte rechtsseitige Stenosen eingestuft. Deutlich weniger Befunder sahen eine isolierte linksseitige Stenose. Zu einem erheblichen Anteil wurde jedoch auch hier überhaupt keine Stenose befundet.

Die Randsummen singulärer Stenosen sind homogen. Dem entsprechend liegt die Teststatistik nach McNemar abseits des Signifikanzniveaus. Nur bei globalen Stenosen zeigt sich mit 40 Befunden beim Goldstandard gegenüber 23 Befunden bei den Untersuchern eine erhebliche Differenz (McNemar p = 0,009). Das Gleiche gilt für unauffällige Befunde (McNemar p=0,024).

Subgruppenanalyse positive Befunde: Lässt man unauffällige Befunde außer Acht, steigt Kappa Cohen insgesamt von 0,452 auf 0,518. Linksseitige Stenosen verbessern sich hinsichtlich Kappa Cohen von 0,282 auf 0,468, rechtsseitige Stenosen von 0,551 auf 0,761. Bei globalen Stenosen fällt Kappa Cohen von 0,392 auf 0,336.

Zusammenfassung 4.7: Stenoselokalisation Hauptbronchus

Sowohl auf Ebene der Einzelbefunde als auch auf Ebene der Kombinationsbefunde zeigte sich im linken Hauptbronchus, eine im Vergleich zu rechts, einheitlichere Befundung der Untersucher untereinander. Die Übereinstimmung mit dem Goldstandard ist hingegen bei Einzelbefunden wie auch Kombinationsbefunden rechtsseitig besser. Falsch negative Befunde machen bei linksseitigen wie auch bei rechtsseitigen Stenosen des Hauptbronchus einen erheblichen Anteil der Fehlklassifikationen aus, sodass die Kappawerte in der Subgruppenanalyse positiver Befunde deutlich ansteigen.

Tabelle 4.30: Kontingenztafel Befundkombinationen Stenoselokalisation Hauptbronchus

Bhapkar <0,01			Referenz 0	Referenz 1		Referenz 2	Summen
			0 (keine)	1 (links)	10 (rechts)	11 (beidseits)	
Befunder	0	0 (keine)	**636**	**15**	**46**	**8**	**705**
	1	1 (links)	**9**	**22**	**2**	**4**	**37**
		10 (rechts)	**31**	**2**	**27**	**15**	**75**
	2	11 (beidseits)	**4**	**1**	**5**	**13**	**23**
Summen			680	40	80	40	840

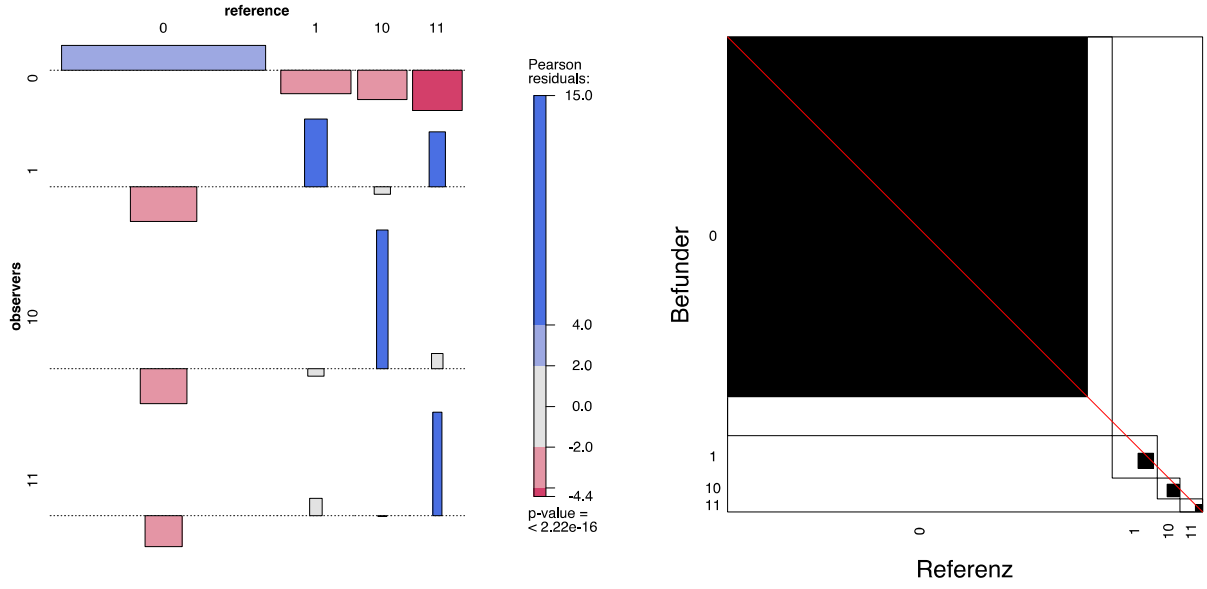
Die Kontingenztafel mit ihren Randsummen ist Berechnungsgrundlage sämtlicher Kennwerte und Diagramme.

Tabelle 4.31: Kennwerte Befundkombinationen Stenoselokalisation Hauptbronchus

	Befund	pre	McN	nag	pag	spe	sen	acc	npv	ppv	B	LR-	LR+	AUC	OR	Q	Y
0	0	81,0	0,024	61,7	91,8	56,9	93,5	86,5	67,4	90,2	0,824	0,114	2,17	0,752	19,1	0,9	0,627
1	1	4,8	0,691	93,4	34,8	93,7	33,8	88	93,1	36,0	0,864	0,707	5,34	0,637	7,6	0,766	0,467
	10	9,5	0,728	97,9	57,1	98,1	55,0	96,1	97,8	59,5	0,958	0,459	29,3	0,766	64	0,969	0,778
2	11	4,8	0,009	97,7	41,3	98,8	32,5	95,6	96,7	56,5	0,954	0,684	26	0,656	38	0,949	0,721

Prävalenz (pre), McNemar (McN) Test, negative (nag) und positive (pag) Übereinstimmung, Spezifität (spe), Sensitivität (sen), Genauigkeit (acc), negativer (npv) und positiver prädiktiver Wert (ppv), Bangdiwala (B), negative (LR-) und positive Likelihood ratio (LR+), area under the curve (AUC), odds ratio (OR), Youden Q und Y.

Abbildung 4.25: Diagramme Befunde Stenoselokalisation Hauptbronchus



Das Assoziationsdiagramm zeigt das Verhältnis der beobachteten Werte zu den Erwartungswerten. Das Bangdiwala-Diagramm stellt für jede Befundklasse die beobachtete Übereinstimmung (schwarz) der maximal möglichen (umgebendes Rechteck) gegenüber.

4.2.3.2.3.2 Lappenbronchien

Die Betrachtung der Stenoselokalisation in den Lappenbronchien erfolgt separat nach der Körperhälfte.

4.2.3.2.3.2.1 Lappenbronchus rechts

Tabelle 4.32: Einzelbefunde Stenoselokalisation Lappenbronchus rechts

Befund	**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz			
	Referenz	Befunder				**Kappa Fleiss**				**Kappa Cohen**	
	Anzahl verschiedener										
Stenoselokalisation Lappenbronchus rechts	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**
Oberlappen rechts	NA	13	8	6	16,7	0,099	kaum	NA	NA	NA	NA
Mittellappen	NA	6	6	5	10,0	0,011	keine	NA	NA	NA	NA
Unterlappen rechts	1 [20]	7	6	5	10,0	0,022	keine	10,0	29,0	0,137	gering

Die gewichtete Anzahl der Befunde des Goldstandards ist in eckigen Klammern angegeben.

Befundverteilung Einzelbefunde: Stenosen in den rechten Lappenbronchien können aufgrund der Prävalenz nur eingeschränkt untersucht werden. Gemäß dem Goldstandard beinhalteten die Videos nur eine einzige Stenose im rechten Unterlappen, sodass für den rechten Oberlappen und Mittellappen keine Referenz zur Verfügung steht. Stenosen im rechten Unterlappen wurden in einem Sechstel der Fälle (7/42) von einem knappen Drittel der Befunder (6/20) in insgesamt 5 verschiedenen Videos erkannt. Schon diese Zahlen lassen darauf schließen, dass sich hier so gut wie keine Übereinstimmung findet. Bei Stenosen im Mittellappen findet sich eine zum Unterlappen fast identische Verteilung. Stenosen im rechten Oberlappen sind gemäß den Befundern mit knapp einem Drittel der Videos (13/42) fast doppelt so häufig.

Präzision Einzelbefunde: Die durchschnittliche positive Übereinstimmung ist bei Stenosen des Oberlappen etwas höher, als im Mittellappen und Unterlappen. Dieser Trend ist auch beim Kappa nach Fleiss erkennbar, absolut gesehen findet sich jedoch in keinem Abschnitt der rechten Lappenbronchien eine nennenswerte Übereinstimmung innerhalb der Befunder.

Richtigkeit Einzelbefunde: Nur ein Zehntel der Stenosen im rechten Unterlappen wurde erkannt. Die Diagnose einer Stenose im rechten Unterlappen war in weniger als einem Drittel der Fälle korrekt. Die Spezifität lag mit 99,4 nur 0,2 % über dem Erwartungswert, bei Zufälligkeit von 99,2 %. „Zufallsbereinigt“ wurde nur etwa jeder 4. unauffällige Befund als solcher erkannt. Dem entsprechend liegt auch der negative prädiktive Wert mit 97,8 % nur geringfügig höher als erwartet (97,6 %). Der positiv prädiktive Wert besagt mit 28,6 %, dass weniger als jeder 3. positive Befund den tatsächlichen Gegebenheiten entsprach. Das liegt erheblich über dem zufällig zu erwartenden Niveau von 2,4 %. Entsprechend steht eine mit 16,4 starke positive Likelihood ratio, einer mit 0,91 ausgesprochen schwachen negativen Likelihood ratio gegenüber. In der ROC folgt die LR- fast der Diagonalen – dem bei Zufälligkeit zu erwartenden Bild. Die AUC liegt daher mit 0,547 nur wenig über dem Zufallsniveau von 0,5.

Tabelle 4.33: Kontingenztafeln Einzelbefunde Stenoselokalisation Lappenbronchus rechts

Oberlappen rechts					Mittellappen					Unterlappen rechts				
NA		Referenz 0	1	Summen	NA		Referenz 0	1	Summen	McN 0,012		Referenz 0	1	Summen
Befunder	0	827	0	827	Befunder	0	834	0	834	Befunder	0	815	18	833
	1	13	0	13		1	6	0	6		1	5	2	7
Summen		840	0	840	Summen		840	0	840	Summen		820	20	840

Die Vier-Felder-Tafeln und ihre Randsummen sind Berechnungsgrundlage sämtlicher Kennwerte und Diagramme.

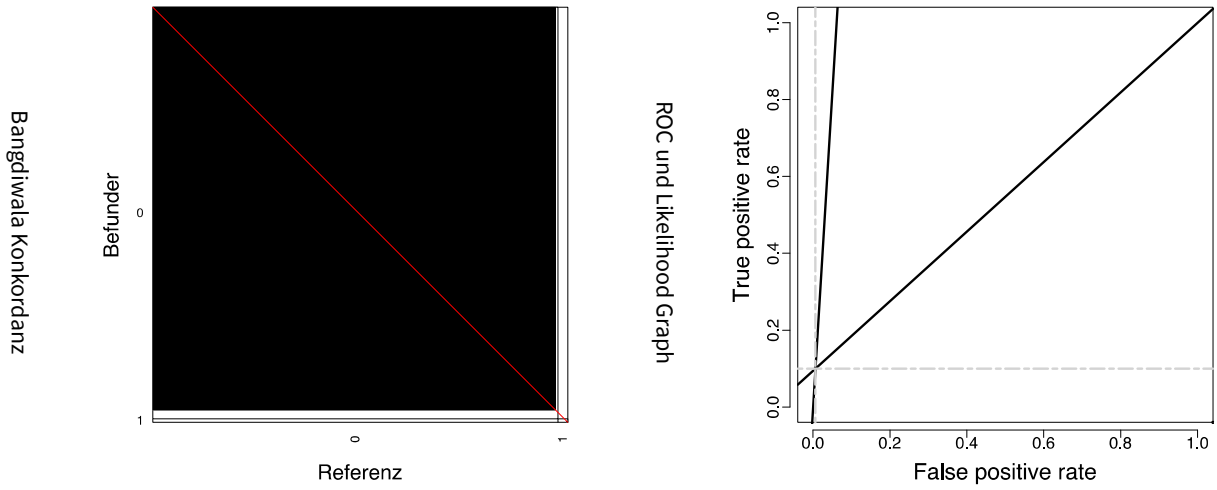
Tabelle 4.34: Kennwerte Einzelbefunde Stenoselokalisation Lappenbronchus rechts

Oberlappen rechts		Mittellappen		Unterlappen rechts			
Prävalenz	NA	Prävalenz	NA	Prävalenz	2,8		
					beo.	erw.	
neg. Übereinst.	NA	neg. Übereinst.	NA	neg. Übereinst.	98,6	98,4	
pos. Übereinst.	NA	pos. Übereinst.	NA	pos. Übereinst.	14,8	1,2	
Spezifität	NA	Spezifität	NA	Spezifität	99,4	99,2	26,8
Sensitivität	NA	Sensitivität	NA	Sensitivität	10,0	0,8	9,2
Genauigkeit	NA	Genauigkeit	**NA**	Genauigkeit	97,3	96,8	13,8
odds ratio	NA	odds ratio	NA	odds ratio		18,6	
Yule Q / Y	NA	Yule Q / Y	NA	Yule Q / Y		0,895 / 0,619	

Beobachtete Kennwerte (beo.), Erwartungswerte gemäß Randsummen (erw.) & für erwartete Übereinstimmung korrigierte Werte (kor.) in Prozent. Die odds ratio und ihre Transformationen (Yule Q & Y) sind prävalenzunabhängig.

Abbildung 4.26: Diagramme Einzelbefunde Stenoselokalisation Lappenbronchus rechts

Unterlappen rechts



Bangdiwala (B) 0,972 (0,968) area under curve (AUC) 0,547

npv 97,8 (97,6) ppv 28,6 (2,4) LR+ 16,4 LR- 0,91

Die Bangdiwala Diagramme illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Die Receiver Operator Characteristic (ROC) trägt die Rate falsch Positiver (=1-Spezifität, fpr) gegen die Rate richtig Positiver (=Sensitivität, tpr) auf. Die Steigung der Tangenten durch den Ursprung bzw. (1,1) definieren die Likelihood ratios (LR+/LR-), die Fläche darunter die Area Under the Curve (AUC).

Tabelle 4.35: Befundkombinationen Stenoselokalisation Lappenbronchus rechts

Befund				**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit dem Goldstandard in den überschneidenden Befunden				
				Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
				Anzahl verschiedener											
Anzahl der Stenosen	Oberlappen rechts	Mittellappen	Unterlappen rechts	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	**Kappa nach Cohen**	modifizierte Klassifikation **nach Landis**	Datenabdeckung [%]
x	**gesamt**			**1 [20]**	**21**	**11**	**10**	**13,8**	**0,068**	**schwach**	**29,8**	**49,3**	**0,159**	**schwach**	**98,1**
0			0	41 [820]	819	20	42	97,5	0,093	schwach	99,6	97,9	0,159	schwach	98,2
			1	1 [20]	5	4	4	10,0	0,015	schwach	10,5	40,0	0,159	schwach	95,7
1		1	0	NA	2	2	2	NA	-0,002	kaum	NA	NA	NA	NA	NA
	1	0	0	NA	10	7	5	17,5	0,084	schwach	NA	NA	NA	NA	NA
2		1	1	NA	1	1	1	NA	NA	keine	NA	NA	NA	NA	NA
	1	1	1	NA	2	2	2	NA	-0,002	keine	NA	NA	NA	NA	NA
3	1	1	1	NA	1	1	1	NA	NA	keine	NA	NA	NA	NA	NA

Die Tabelle analysiert die Befundkombinationen im Sinne eines Gesamtbefundes. Sie gliedert sich in die 3 Hauptbereiche Befundverteilung Präzision und Richtigkeit. Details siehe Kommentar Tabelle 4.15 auf Seite 109.

Befundverteilung Befundkombinationen: Während unauffällige Befunde von Befundern und dem Goldstandard gleich häufig vergeben wurden (820 versus 819), bestehen bei den positiven Befunden im Bereich der rechten Lappenbronchien erkennbare Unterschiede: Der Goldstandard gibt nur eine einzige, singuläre Stenose im rechten Unterlappen vor. Ein Befund, der nur von einer Minderheit der Befunder (4/20) und deutlich seltener als vom Goldstandard (5 versus 20) gewählt wurde. Singuläre Stenosen im rechten Oberlappen sind gemäß den Befundern mit immerhin 10 Befunden doppelt so häufig. Singuläre Stenosen im Mittellappen und mehrfache Stenosen sind mit Prävalenzen von maximal 2/840 Befunden die absolute Ausnahme.

Präzision Befundkombinationen: Eine Übereinstimmung innerhalb der Befunder ist bei keiner der Befundkombinationen der rechtsseitigen Lappenbronchien erkennbar.

Richtigkeit Befundkombinationen: Sensitivität und Spezifität sowie positiver und negativer prädiktiver Wert der Befundkombinationen stimmen weitgehend mit denen der Einzelbefunde überein. Die Kontingenztafel der Befundkombinationen (Tabelle 4.36 Seite 128) belegt, dass die uneinheitliche Befundung bei singulären rechtsseitigen Unterlappenstenosen fast ausschließlich durch falsch negative Befunde (17/20) verursacht wird. Nur ein Befund entfällt auf eine andere Lokalisation, zwei Befunde waren richtig. Der Anteil falsch-positiver an den negativen Befunden fällt mit 18/820 Befunden verhältnismäßig gering aus.

Tabelle 4.36: Kontingenztafel Befundkombinationen Stenoselokalisation Lappenbronchus rechts

	Bhapkar		Referenz 0	Referenz 1	Summen	
Befunder	0	0	802	17	819	819
	1	1	3	2	5	5
		10	1	1		2
		100	10	0		10
	2	11	1	0		1
		110	2	0		2
	3	111	1	0		1
Summen			805	19	824	
			820	20		840

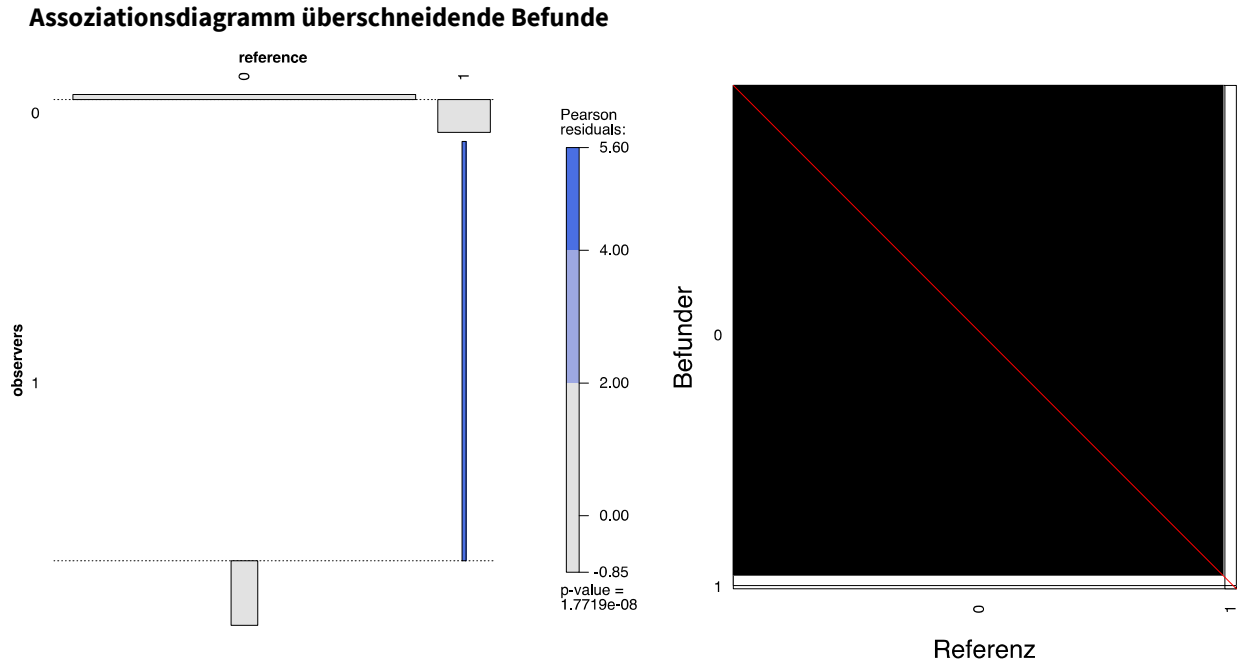
Die Kontingenztafel mit ihren Randsummen ist Berechnungsgrundlage sämtlicher Kennwerte und Diagramme.

Tabelle 4.37: Kennwerte Befundkombinationen Stenoselokalisation Lappenbronchus rechts

Befund	pre	McN	nag	pag	spe	sen	acc	npv	ppv	B	LR-	LR+	AUC	OR	Q	Y
0	97,7	0,004	16,7	98,8	10,5	99,6	97,6	40	97,9	0,975	0,035	1,11	0,551	31,5	0,938	0,69
1	2,31	0,004	98,8	16,7	99,6	10,5	97,6	97,9	40	0,975	0,898	28,2	0,551	31,5	0,938	0,697

Prävalenz (pre), McNemar Test (McN), negative (nag) und positive (pag) Übereinstimmung, Spezifität (spe), Sensitivität (sen), Genauigkeit (acc), negativer (npv) und positiver prädiktiver Wert (ppv), Bangdiwala (B), negative (LR-) und positive Likelihood ratio (LR+), area under the curve (AUC), odds ratio (OR), Youden Q und Y.

Abbildung 4.27: Assoziationsdiagramme Stenoselokalisation Lappenbronchus rechts



Links: Illustration der Beobachtungen im Vergleich zu den Erwartungswerten in der Kontingenztafel der Befundkombinationen. Rechts: Kontingenztafel der überschneidenden Befundkombinationen.

4.2.3.2.3.2.2 Lappenbronchus links

Tabelle 4.38: Einzelbefunde Stenoselokalisation Lappenbronchus links

Befund	**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz			
	Referenz	Befunder Anzahl verschiedener			Ø positive Übereinstimmung [%]	**Kappa Fleiss**		Sensitivität [%]	positiver prädiktiver Wert [%]	**Kappa Cohen**	
Stenoselokalisation Lappenbronchus links	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)		Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**			Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**
Oberlappen links	NA	14	9	7	13,8	0,067	keine	NA	NA	NA	NA
Unterlappen links	2 [40]	27	16	6	40,0	0,418	beachtlich	50,0	74,1	0,581	beachtlich
Lingula	NA	13	9	7	20,0	0,091	keine	NA	NA	NA	NA

Die Tabelle betrachtet die Stenoselokalisationen innerhalb der linken Lappenbronchien unabhängig von möglichen Kombinationen. Der deskriptiven Befundverteilung folgen Betrachtungen der Übereinstimmung innerhalb der Befunder (Präzision) sowie im Vergleich zum Goldstandard (Richtigkeit). Zur besseren Vergleichbarkeit zwischen Referenz und Befundern ist in eckigen Klammern die gewichtete Anzahl der Befunde des Goldstandards angegeben. Da jeder Befund mit dem Goldstandard verglichen wird, ergibt sich bei 20 Befundern für die Gewichtung der Befunde des Goldstandards der Faktor 20.

Befundverteilung Einzelbefunde: Analog zu den Lappenbronchien auf der rechten Seite ist die Auswertung aufgrund der Prävalenz eingeschränkt: Gemäß den Befunden des Goldstandards liegen nur 2 (singuläre) Stenosen im linken Unterlappen vor. Die Befunder sahen knapp drei Viertel (27/40) der vom Goldstandard befundeten Stenosen im linken Unterlappen. Eine Diagnose, die von etwas mehr als drei Viertel der Befunder in dreimal so vielen Videos wie vom Goldstandard erhoben wurde. Stenosen im linken Oberlappen bzw. der Lingula sahen die Befunder nur halb so häufig wie Stenosen im Unterlappen. An diesen Befunden waren nur knapp die Hälfte der Befunder beteiligt.

Präzision Einzelbefunde: Die Stenosen des linken Mittellappens werden von den Befundern bei einem Kappa Fleiss von 0,418 „beachtlich“ einheitlich beurteilt. In den anderen linksseitigen Lappenbronchien liegt die Konkordanz auf Zufallsniveau.

Richtigkeit Einzelbefunde: Die Hälfte der Stenosen im Unterlappen links wurde erkannt. Positive Befunde waren in drei Viertel der Fälle korrekt. Selbst bei Korrektur für zufällige Übereinstimmung liegt die Spezifität noch bei über 70 %. Insgesamt ergibt sich ein Kappa Cohen von 0,581. Ähnlich wie bei Stenosen in den Lappenbronchien rechts besteht insbesondere bei positiven Befunden eine starke Übereinstimmung. Die LR+ erreicht einen Wert von 57,1, die true positive rate zeigt in der ROC dem entsprechend einen fast senkrechten Anstieg. Die OR liegt bei 113,3.

Tabelle 4.39: Kontingenztafeln Einzelbefunde Stenoselokalisation Lappenbronchus links

Oberlappen links					Unterlappen links					Lingula				
McN NA		Referenz 0	1	Summen	McN 16,056		Referenz 0	1	Summen	McN NA		Referenz 0	1	Summen
Befunder	0	**826**	0	826	Befunder	0	**793**	20	813	Befunder	0	**827**	0	827
	1	14	**0**	14		1	7	**20**	27		1	13	**0**	13
Summen		840	0	840	Summen		800	40	840	Summe		840	0	840

Die Vier-Felder-Tafeln und ihre Randsummen sind Berechnungsgrundlage sämtlicher Kennwerte und Diagramme.

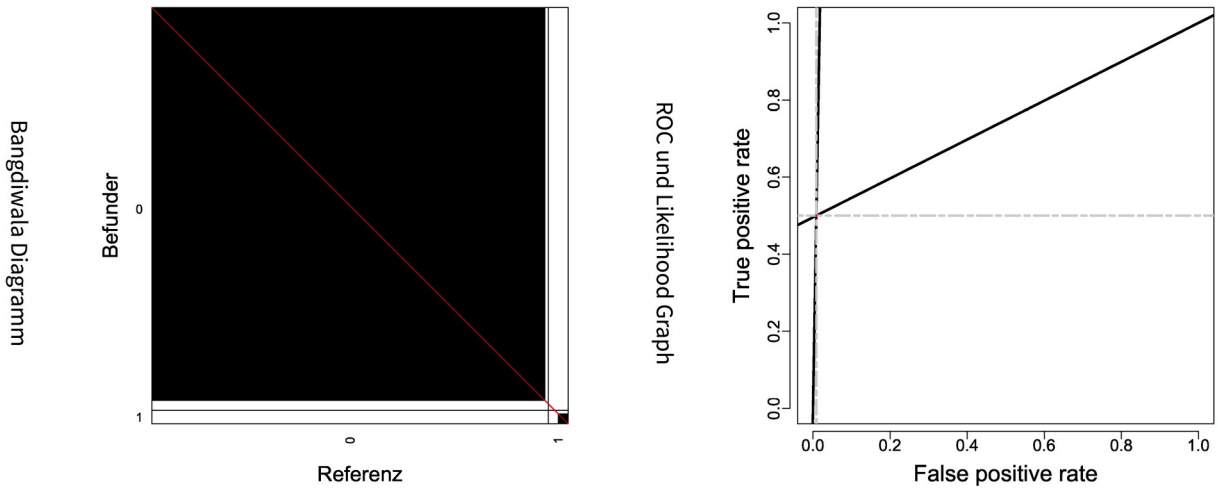
Tabelle 4.40: Kennwerte Einzelbefunde Stenoselokalisation Lappenbronchus links

Oberlappen links		Unterlappen links				Lingula	
Prävalenz	0	Prävalenz	4,8			Prävalenz	0
			beo.	erw.	kor.		
neg. Übereinst.		neg. Übereinst.	98,3	96,0	=κ	neg. Übereinst.	
pos. Übereinst.	NA	pos. Übereinst.	59,7	3,8	=κ	pos. Übereinst.	NA
Spezifität	98,3	Spezifität	99,1	96,8	72,8	Spezifität	98,5
Sensitivität	NA	Sensitivität	50,0	3,2	48,3	Sensitivität	NA
Genauigkeit	NA	Genauigkeit	96,8	92,3	58,1	Genauigkeit	NA
Odds ratio	NA	Odds ratio		113,3		Odds ratio	NA
Yule Q / Y	NA	Yule Q / Y		0,982 / 0,828		Yule Q / Y	NA

Beobachtete Kennwerte (beo.), Erwartungswerte gemäß Randsummen (erw.) & für erwartete Übereinstimmung korrigierte Werte () in Prozent. Die odds ratio und ihre Transformationen (Yule Q & Y) sind prävalenzunabhängig.

Abbildung 4.28: Diagramme Einzelbefunde Stenoselokalisation Lappenbronchus links

Unterlappen links



Bangdiwala (B) 0,966 (0,920) area under curve (AUC) 0,746

npv 97,5 (95,2) ppv 74,1 (4,8) LR+ 57,1 LR- 0,504

Die Bangdiwala Diagramme illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Die Receiver Operator Characteristic (ROC) trägt die Rate falsch Positiver (=1-Spezifität, fpr) gegen die Rate richtig Positiver (=Sensitivität, tpr) auf. Die Steigung der Tangenten durch den Ursprung bzw. (1,1) definieren die Likelihood ratios (LR+/LR-), die Fläche darunter die Area Under the Curve (AUC).

Tabelle 4.41: Befundkombinationen Stenoselokalisation Lappenbronchus links

Befund				**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit dem Goldstandard in den überschneidenden Befunden				
				Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
				Anzahl verschiedener											
Anzahl der Stenosen	Oberlappen links	Unterlappen links	Lingula	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert	**Kappa nach Cohen**	modifizierte Klassifikation **nach Landis**	Datenabdeckung [%]
x	**gesamt**			2 [40]	35	17	8	16,5 %	0.297	schwach	45.5	100	0.615	stark	97.6
0			0	40 [800]	805	20	42	95.8 %	0.424	moderat	100	97.8	0.615	stark	98.4
			1	NA	1	1	1	NA	-0.001	keine	NA	NA	NA	NA	NA
1		1	0	2 [40]	15	12	2	37.5 %	0.375	mäßig	45.5	100	0.615	stark	82.5
	1	0	0	NA	5	2	4	10.0 %	0.015	keine	NA	NA	NA	NA	NA
		1	1	NA	5	5	3	15.0 %	0.058	kaum	NA	NA	NA	NA	NA
2	1	0	1	NA	2	2	2	NA	-0.002	keine	NA	NA	NA	NA	NA
	1	1	0	NA	2	2	1	10.0 %	0.050	keine	NA	NA	NA	NA	NA
3	1	1	1	NA	5	3	4	10.0 %	0.015	keine	NA	NA	NA	NA	NA

Die Tabelle analysiert die Befundkombinationen im Sinne eines Gesamtbefundes. Sie gliedert sich in die 3 Hauptbereiche Befundverteilung Präzision und Richtigkeit. Details siehe Kommentar Tabelle 4.15 auf Seite 109.

Befundverteilung Befundkombinationen: Die Prävalenz von Stenosen in den linken Lappenbronchien beträgt bei den Befundern 4,1 % (35/840), was gut mit der Prävalenz des Goldstandards von 5 % (40/800) vergleichbar ist. Allerdings konzentrieren sich die Befunde des Goldstandards auf Stenosen im linken Unterlappen. Die Befunde der Untersucher verteilen sich dagegen zusätzlich auf Befunde im linken Oberlappen, der Lingula sowie Mehrfachbefunde. Mit 17/20 Befundern hat eine überwiegende Mehrheit Stenosen in den linken Lappenbronchien befundet. Insgesamt haben die Befunder Stenosen im linken Lappenbronchus in 8 verschiedenen Videomitschnitten gesehen.

Befundverteilung Befundkombinationen: Bei Stenosen im linken Unterlappen findet sich mit einem Kappa Fleiss von 0,375 eine der besten Übereinstimmungen in dieser Studie. Die übrige Konkordanz bewegt sich – abseits unauffälliger Befunde – auf Zufallsniveau.

Richtigkeit Befundkombinationen: Stenosen im linken Unterlappen wurden in knapp der Hälfte der Fälle (45,5 %) als solche erkannt. Von den Untersuchern diagnostizierte Stenosen im linken Unterlappen waren zu 100 % korrekt, was sie zur validesten Diagnose der Studie macht. Das Kappa Cohen von Stenosen im linken Unterlappen erhöht sich bei Betrachtung auf Ebene der Befundkombinationen auf 0,615, was einer starken Übereinstimmung entspricht. Das drückt sich auch darin aus, dass die 15 Befunde der Untersucher in dieser Klasse von 12 unterschiedlichen Befundern in nur 2 verschiedenen Videos erhoben wurden. Die nummerische Erhöhung des Kappa Cohen ist auf die höhere Anzahl möglicher Befundklassen auf Ebene der Befundkombinationen zurückzuführen.

Zusammenfassung 4.8: Stenoselokalisation Lappenbronchien

Insgesamt gesehen sind die für die rechten Lappenbronchien erhobenen Zahlen aufgrund der geringen Prävalenz als nicht repräsentativ einzustufen. Interessant ist jedoch, dass sich der we-

sentliche Anteil der Differenzen auch hier aus falsch-negativen Befunden speist. Die Aussagekraft der Studie zu Stenosen im linken Lappenbronchus ist eingeschränkt, da sie gemäß dem Referenzbefund des Goldstandards nur im linken Unterlappen zu sehen waren. Dieser Befund erreicht allerdings sowohl innerhalb der Untersucher mit einem Kappa Fleiss von 0,375, als auch im Vergleich zum Goldstandard mit einem Kappa Cohen von 0,615 Spitzenwerte. Mit einem prädiktiven Wert von 100 % sind Stenosen im linken Unterlappen der valideste Befund der Studie. Abseits von Stenosen im linken Unterlappen bewegt sich die Konkordanz der Untersucher auf Zufallsniveau.

Tabelle 4.42: Kontingenztafel Befundkombinationen Stenoselokalisation Lappenbronchus links

Mc Nemar 16,056		Referenz 0	Referenz 1	Summen	
		0	**10**		
0		**787**	**18**	805	805
1	1	1	0		1
	10	**0**	**15**	15	15
	100	4	1		5
2	11	1	4		5
	101	1	1		2
	110	2	0		2
3	111	4	1		5
Summen		787	33	820	
		800	40		840

Befunder

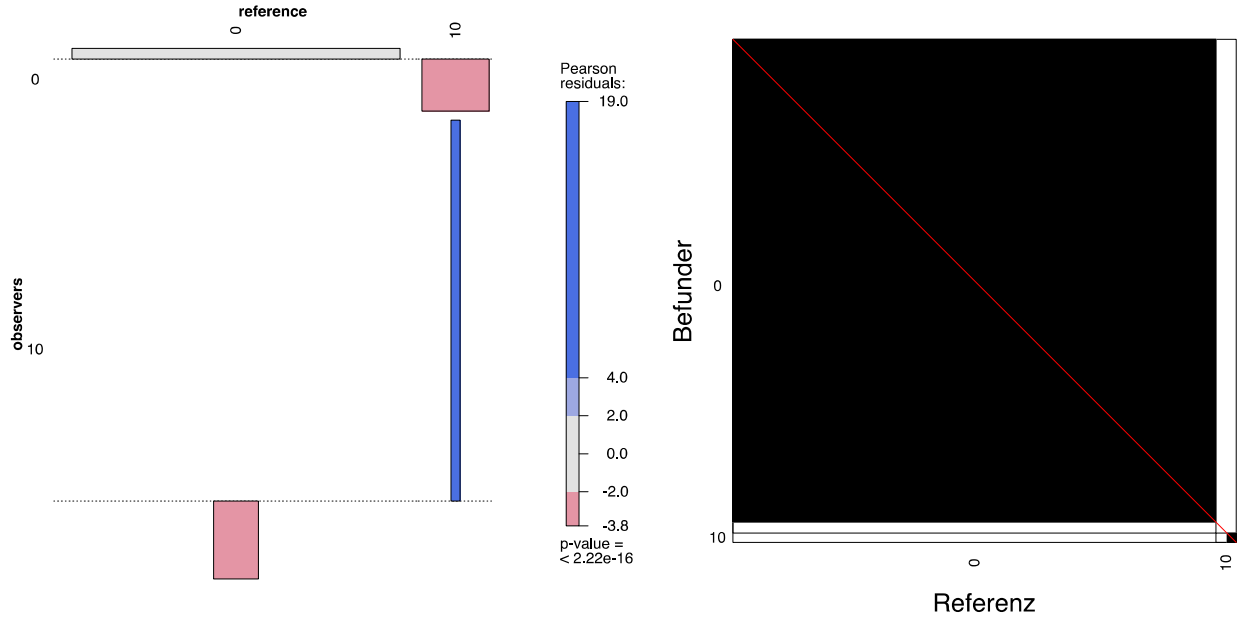
Die Kontingenztafel mit ihren Randsummen ist Berechnungsgrundlage sämtlicher Kennwerte und Diagramme. Zwischen Goldstandard und den Untersuchern überlappende Befundkategorien sind fett hervorgehoben, nicht überlappende Kategorien grau dargestellt. Analog hierzu zeigen die fetten Randsummen die Summen der überlappenden Kategorien, die grauen Randsummen die Summe aller Kategorien.

Tabelle 4.43: Kennwerte Befundkombinationen Stenoselokalisation Lappenbronchus links

	Befund	**pre**	**McN**	**nag**	**pag**	**spe**	**sen**	**acc**	**npv**	**ppv**	**B**	**LR-**	**LR+**	**AUC**	**OR**	**Q**	**Y**
0	0	96	16,056	62,5	98,9	45,5	100	97,8	100	97,8	0,977	0	1,83	0,727	NA	NA	NA
1	10	4	16,056	98,9	62,5	100	45,5	97,8	97,8	100	0,977	0,545	Inf	0,727	NA	NA	NA

Prävalenz (pre), McNemar Test (McN), negative (nag) und positive (pag) Übereinstimmung, Spezifität (spe), Sensitivität (sen), Genauigkeit (acc), negativer (npv) und positiver prädiktiver Wert (ppv), Bangdiwala (B), negative (LR-) und positive Likelihood ratio (LR+), area under the curve (AUC), odds ratio (OR), Youden Q und Y.

Abbildung 4.29: Assoziationsdiagramme Stenoselokalisation Lappenbronchus links.



Links: Illustration der Beobachtungen im Vergleich zu den Erwartungswerten in der Kontingenztafel der Befundkombinationen. Rechts: Kontingenztafel der überschneidenden Befundkombinationen.

4.2.3.2.4 Vergleich der anatomischen Abschnitte der Stenoselokalisation

Beim Vergleich der anatomischen Abschnitte werden Stenoselokalisationen im Hauptbronchus am präzisesten und abgesehen von Stenosen im linken Lappenbronchus auch am genauesten erkannt. Ähnlich einheitlich werden Larynxstenosen von den Untersuchern beurteilt, Trachealstenosen fallen dem gegenüber leicht ab. Deutlich darunter liegen in der Präzision Stenosen in den Lappenbronchien, wobei bei linksseitigen Stenosen erheblich mehr Befundeinigkeit besteht als bei rechtsseitigen. Dabei muss berücksichtigt werden, dass Stenosen in Larynx und Trachea am häufigsten befundet wurden. Stenosen in den Lappenbronchien waren dem gegenüber etwa um den Faktor 5 bzw. 10 seltener. Die Übereinstimmung mit dem Goldstandard ist in Larynx und Trachea mäßig, im Hauptbronchus moderat. Überraschend ist die starke Übereinstimmung bei Stenosen im linksseitigen Lappenbronchus, die auf Stenosen im linken Unterlappen basieren. Dieser Befund wurde trotz niedriger Prävalenz am genauesten erkannt. Die Überlegenheit bei der Beurteilung des Hauptbronchus und des Larynx hängt vermutlich mit den visuell eindeutig bestimmbaren anatomischen Lokalisationen zusammen. In diesem Kontext ist es überraschend, dass trotz der vagen Untergliederung der Trachea in proximales, mittleres und distales Drittel eine verhältnismäßig einheitliche Beurteilung stattfindet.

Tabelle 4.44: Stenoselokalisation

Befund		**Befundverteilung**				**Präzision** Übereinstimmung der Befunder			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz			
		Referenz	Befunder				Kappa Fleiss				Kappa Cohen	
		Anzahl verschiedener										
Typ der Untersuchung		Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**
syn.	Larynx	23 [460]	278	20	32	26,0	0,418	moderat	29,8	49,3	0,357	mäßig
einzeln	supraglottisch	8 [160]	62	17	17	33,1	0,261	schwach	23,8	61,3	0,264	schwach
einzeln	glottisch	11 [220]	80	19	18	33,2	0,250	schwach	27,7	76,3	0,310	mäßig
einzeln	subglottisch	11 [220]	170	20	20	46,1	0,569	gut	58,6	75,8	0,561	gut
syn.	Trachea	17 [340]	227	20	28	21,0	0,324	mäßig	30,3	49,4	0,385	mäßig
einzeln	proximales Drittel	10 [200]	75	19	20	37,8	0,150	gering	26,0	69,3	0,285	schwach
einzeln	mittleres Drittel	5 [100]	90	20	17	35,4	0,299	schwach	48,0	53,3	0,442	moderat
einzeln	distales Drittel	9 [180]	108	20	16	54,2	0,393	mäßig	43,3	72,2	0,454	moderat
syn.	Hauptbronchus	8 [160]	135	20	16	30,6	0,487	moderat	29,8	49,3	0,452	moderat
einzeln	rechts	4 [80]	60	20	10	36,9	0,443	moderat	50,0	66,7	0,533	gut
einzeln	links	6 [120]	98	20	14	40,4	0,509	gut	50,0	61,2	0,484	moderat
syn.	Lappenb. rechts	1 [20]	21	11	10	13,8	0,068	schwach	29,8	49,3	0,159	schwach
einzeln	Oberlappen re.	na	13	8	6	16,7	0,099	kaum	na	na	na	na
einzeln	Mittellappen	na	6	6	5	10,0	0,011	keine	na	na	na	na
einzeln	Unterlappen re.	1 [20]	7	6	5	10,0	0,022	keine	10,0	29,0	0,137	gering
syn.	Lappenb. links	2 [40]	35	17	8	16,5	0,297	schwach	45,5	100	0,615	stark
einzeln	Oberlappen links	na	14	9	7	13,8	0,067	schwach	na	na	na	na
einzeln	Unterlappen links	2 [40]	27	16	6	40,0	0,418	beachtlich	50,0	74,1	0,581	beachtlich
einzeln	Lingula	na	13	9	7	20,0	0,091	schwach	na	na	na	na

Übersicht über die für jeden Einzelbefund und für die Befundkombinationen der anatomischen Abschnitte berechneten Kappa-Werte. Die Kappa Werte der einzelnen Kombinationsbefunde sind im Ergebnisteil genau aufgeschlüsselt.

4.2.3.2.5 Kombinationsbefund Stenoselokalisation

In den vorausgehenden Abschnitten wurden die Stenoselokalisationen, getrennt nach dem jeweiligen anatomischen Abschnitt, betrachtet. Durch Kombination der Einzelbefunde („Symptome“) zu Befundkombinationen („Syndrome“) konnten Kappawerte für den jeweiligen gesamten anatomischen Abschnitt angegeben werden und zwischen Einfach- und Mehrfachbefunden differenziert werden. In diesem Abschnitt werden die Stenoselokalisationen übergreifend — also über alle anatomischen Abschnitte von Larynx bis hin zu den Lappenbronchien — betrachtet. Hierzu wurden nicht die Befundkombinationen der einzelnen anatomischen Abschnitte aggregiert, denn die Analyse innerhalb der anatomischen Abschnitte zeigt, dass sich bei Mehrfachbefunden innerhalb eines anatomischen Abschnittes bereits eine erhebliche Variabilität an Befundkombinationen ergibt. Stattdessen wurde der Befund eines übergeordneten anatomischen Abschnitts positiv gesetzt, wenn in einem der untergeordneten Abschnitte mindestens ein positiver Befund erhoben wurde. Tabelle 4.45 vergleicht die syndromale Berechnung der Befundkombinationen mit der Vereinfachung auf binäre Befunde. Durch diese nachträgliche Vereinfachung der Befundung erhöhen sich in den meisten Fällen die Übereinstimmungsmaße.

Tabelle 4.45: Vergleich Berechnungsmodi Stenoselokalisation

Befund			**Befundverteilung**				**Präzision** Übereinstimmung der Befunder			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz				
			Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
			Anzahl verschiedener											
Vergleich der Randverteilungen	Lokalisation	Berechnungsmodus	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**	Datenabdeckung
B	Larynx	syndromal	23 [460]	278	20	32	26,0	0,418	moderat	29,8	49,3	0,357	mäßig	100
M		binär					52,3	0,463	moderat	53,7	88,8	0,437	moderat	100
B	Trachea	syndromal	17 [340]	227	20	28	21,0	0,324	mittelmäßig	30,3	49,4	0,385	mittelmäßig	91,8
M		binär					46,5	0,474	moderat	56,2	84,1	0,517	gut	100
B	Hauptbr.	syndromal	42 [840]	135	20	16	30,6	0,487	moderat	29,8	49,3	0,452	moderat	100
M		binär					47,5	0,533	gut	56,9	67,4	0,536	gut	100
B	Lappenbr. rechts	syndromal	1 [20]	21	11	10	13,8	0,068	schwach	29,8	49,3	0,159	schwach	98,1
M		binär					16,0	0,039	kaum	15,0	14,3	0,125	gering	100
B	Lappenbr. links	syndromal	2 [40]	35	17	8	16,5	0,297	schwach	45,5	100	0,615	stark	97,6
M		binär					32,0	0,424	moderat	55	62,9	0,567	gut	100

Vergleich von Maßzahlen der Präzision und Übereinstimmung bei Berechnung abschnittsweiser Kombinationsbefunde (syndromal) und einer Vereinfachung (binär), bei der die Kombinationsbefunde in eine binäre Klassifikation überführt wurden. Ein übergeordneter anatomische Abschnitt wurde dabei als positiv gewertet, wenn in mindestens einem Unterabschnitt ein Befund erhoben wurde.

Tabelle 4.46: Kombinationsbefund Stenoselokalisation

Befund							Befundverteilung				Präzision Übereinstimmung der Befunder			Richtigkeit Übereinstimmung mit Goldstandard als Referenz				
Anzahl der Befunde	Mc Nemar Test	Larynx	Trachea	Hauptbronchus	Lappenbronchien rechts	Lappenbronchien links	Ref. Befunde (& Videos)	Befunder Anzahl verschiedener Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa Fleiss		Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa Cohen		Datenabdeckung
												Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**			Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**	
x	B	gesamt					36	546	20	42	NA	0.328	mäßig	38.5	51.7	0.346	mäßig	95.0
0						0	6	294	20	37	46.5	0.382	mäßig	93.2	37.8	0.415	moderat	98.4
						1	1	15	11	4	32.5	0.382	mäßig	5	6.67	0.036	keine	100.0
					1	0	na	6	4	4	10.0	0.028	keine	na	na	na	na	na
1				1	0	0	2	55	19	11	32.5	0.378	mäßig	30.3	24.4	0.235	schwach	75.3
			1	0	0	0	6	112	20	19	40.8	0.373	mäßig	52.6	54.5	0.458	moderat	97.7
		1	0	0	0	0	13	225	20	27	50.0	0.378	mäßig	53.5	61.8	0.388	mäßig	100.0
					1	1	na	2	2	2	na	-0.002	keine	na	na	na	na	na
				1	0	1	1	12	7	3	20.0	0.172	gering	27.8	41.7	0.321	mäßig	92.6
				1	1	0	1	3	3	1	15.0	0.102	gering	15.8	100	0.268	schwach	95.0
2			1	1	0	0	2	49	17	11	24.0	0.216	schwach	45.7	34	0.358	mäßig	90.4
		1	0	1	0	0	1	na	na	na	na	na	na	na	na	na	na	na
		1	1	0	0	0	8	50	15	20	20.0	0.098	kaum	16.4	52	0.17	gering	99.5
				1	1	1	na	1	1	1	na	na	keine	na	na	na	na	na
			1	0	1	1	na	1	1	1	na	na	keine	na	na	na	na	na
3			1	1	0	1	na	4	3	3	10.0	0.022	keine	na	na	na	na	na
			1	1	1	0	na	8	4	5	12.5	0.044	keine	na	na	na	na	na
		1	1	1	0	0	1	3	3	3	NA	-0,004	keine	5	50	0,087	kaum	95,5

Die Tabelle zeigt die Kennzahlen abschnittsübergreifender Befundkombinationen der Stenoselokalisation. Dafür wurden die Befunde der einzelnen Abschnitte auf ein binäres Format vereinfacht.

Befundverteilung: Betrachtet man sämtliche anatomischen Abschnitte gemeinsam[29], erhebt der Goldstandard Stenoselokalisationen in 36/42 Videos (85,7 %), die Untersucher in 546/840 (65 %). Die Befunder sehen also nur 75 % der Stenoselokalisationen. Stenoselokalisationen in nur einem anatomischen Abschnitt haben beim Goldstandard einen relativen Anteil von 22/36 (61,1 %), bei den Befundern mit 413/546 (75,6 %) einen höheren Anteil. Der relative Anteil an Mehrfachbefunden fällt beim Goldstandard mit 14/42 (33,3 %) höher aus, als bei den Untersuchern mit 133/546 (24,4 %).

Präzision: Stenoselokalisationen in nur einem anatomischen Abschnitt werden bei einem Kappa Fleiss um zumeist 0,380 mit „mäßiger“ Übereinstimmung der Befunder untereinander beurteilt. Die Ausnahme bilden singuläre Stenoselokalisationen im rechten Lappenbronchus bei denen keine Übereinstimmung zu erkennen ist — ein Befund der gemäß dem Goldstandard in der Videobibliothek nicht zu sehen war.

Richtigkeit: Bei singulären Stenosen wird die Lokalisation in der Trachea mit der besten Übereinstimmung zum Goldstandard angegeben, gefolgt von unauffälligen Befunden. Die Richtigkeit

[29] *Indem man die Stenoselokalisationen jedes anatomischen Abschnittes (Larynx, Trachea, Bronchien und Lappenbronchien) auf eine binäre Klassifikation („vorhanden“/„nicht vorhanden“) vereinfacht.*

von Larynxstenosen setzt sich deutlich gegen die des Hauptbronchus ab. Keine Übereinstimmung mit dem Goldstandard findet sich bei Stenoselokalisationen im Lappenbronchus. Damit zeigt sich bei syndromaler Betrachtung, quer über alle anatomischen Abschnitte, ein etwas anderes Bild als beim Vergleich der abschnittsweisen Kappa-Werte. Die gute Richtigkeit bei der Stenoselokalisation im Lappenbronchus links relativiert sich bei der syndromalen Betrachtung: Weder bei singulären Stenosen im linken Lappenbronchus, noch bei Kombinationsbefunden, die den linken Lappenbronchus mit einbeziehen, findet sich eine Übereinstimmung, die an die Spitzenwerte bei der Betrachtung innerhalb des Lappenbronchus heranreicht.

Zusammenfassung 4.9: Stenoselokalisation

Im Larynx werden distale subglottische Stenosen gegenüber glottischen und supraglottischen mit Abstand am präzisesten und auch genausten lokalisiert. Glottische und supraglottische Stenosen liegen hinsichtlich Präzision und Richtigkeit in etwa gleichauf. Auch in der Trachea nehmen Präzision und Richtigkeit vom proximalen bis zum distalen Drittel zu - bei Betrachtung der Einzelbefunde wie auch der Kombinationsbefunde. Im Hauptbronchus ist die Befundung innerhalb der Untersucher links, die Übereinstimmung mit dem Goldstandard rechts am größten.
Die Übereinstimmung bei den Angaben zur Stenoselokalisation ist bei singulären Befunden höher als bei Kombinationsbefunden, bei denen kaum eine einheitliche Befundung zu erkennen ist. Den wesentlichen Anteil an Fehlbefunden machen in allen Abschnitten falsch-negative Befunde aus, also die Frage, ob überhaupt eine Stenose vorhanden ist. Beim Vergleich der anatomischen Abschnitte wird die Stenoselokalisation Hauptbronchus insgesamt am präzisesten und genausten erkannt[30].

4.2.3.3 Stenoseform

Stenosen konnten im multiple-choice-Verfahren von den Befundern in Bezug auf ihre Länge als kurz- oder langstreckig und in Bezug auf die Morphologie als membranös und ringförmig klassifiziert werden.

Befundverteilung Einzelbefunde: Kurzstreckige Stenosen lagen in 26,2 % der Videos vor, wurden von den Befundern aber nur in 18,7 % als solche wahrgenommen. Einigkeit zwischen Goldstandard und Befundern herrscht hingegen bei der Prävalenz langstreckiger Stenosen (100 versus 109), die in knapp 12 % der Videos gesehen wurden. Membranöse Stenosen lagen in 7,1 % der Videos vor, ringförmige waren mit 14,3 % doppelt so häufig. Als membranös wurden Stenosen vom Goldstandard doppelt so häufig bezeichnet, wie von den Befundern, als ringförmig etwa 2,5 mal so oft. Angaben zur Länge der Stenosen wurden von fast allen Befundern gemacht (min. 18). Eine darüber hinaus gehende Charakterisierung als membranös oder ringförmig wurde von nur ca. ¾ der Befunder vorgenommen. Kurzstreckige Stenosen wurden in gegenüber dem Goldstandard knapp dreimal so vielen Videos gesehen. Langstreckige Stenosen verteilen sich auf mehr als viermal so viele Videos. Auch bei membranösen Stenosen findet sich der Faktor 4. Die geringste Variation zeigen ringförmige Stenosen mit dem Faktor 2.

Die Verteilung der Randsummen kann nur bei „langstreckig“ als homogen gelten. In den übrigen Kategorien fällt der McNemar-Test signifikant aus. Das lässt sich anhand der Konkordanz-Diagramme auch graphisch nachvollziehen: Der Schnittpunkt der durch die Randsummen definierten horizontalen und vertikalen Gerade liegt bei „langstreckig“ nahe der Diagonalen, während er in den übrigen Diagrammen deutlich davon abweicht.

[30] Ausnahme ist die Richtigkeit der Stenoselokalisation „Lappenbronchus links“ mit einem Kappa Cohen von 0,615.

Tabelle 4.47: Inter-Beobachter-Variabilität Einzelbefunde Stenoseform

Befund	**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz			
	Referenz	Befunder Anzahl verschiedener				**Kappa Fleiss**				**Kappa Cohen**	
Stenoseform	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landiss**	Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**
kurzstreckig	11 [220]	157	18	32	26,7	0,119	gering	31,4	43,9	0,189	gering
langstreckig	5 [100]	109	19	23	30,6	0,230	schwach	33,0	30,3	0,219	schwach
membranös	3 [60]	32	14	12	38,3	0,278	schwach	31,7	59,4	0,382	mäßig
ringförmig	6 [120]	49	16	12	35,8	0,269	schwach	26,7	65,3	0,323	mäßig

Die Tabelle betrachtet die Übereinstimmung der Einzelbefunde. Befundkombinationen wurden hier nicht berücksichtigt und werden gesondert behandelt. Der rein deskriptiven Befundverteilung folgen mittig die Übereinstimmung innerhalb der Befunder (Präzision) und rechts die Übereinstimmung mit dem Goldstandard (Richtigkeit). Die Befunde des Goldstandards müssen beim Vergleich mit den Befundern mit 20 multipliziert werden (eckige Klammern).

Präzision Einzelbefunde: Die Übereinstimmung innerhalb der Befunde ist hinsichtlich der Stenoseform generell schwach ausgeprägt. Bei morphologischen Charakteristika wie membranös und ringförmig besteht erkennbar mehr Einigkeit als bei der Einschätzung der Stenoselänge.

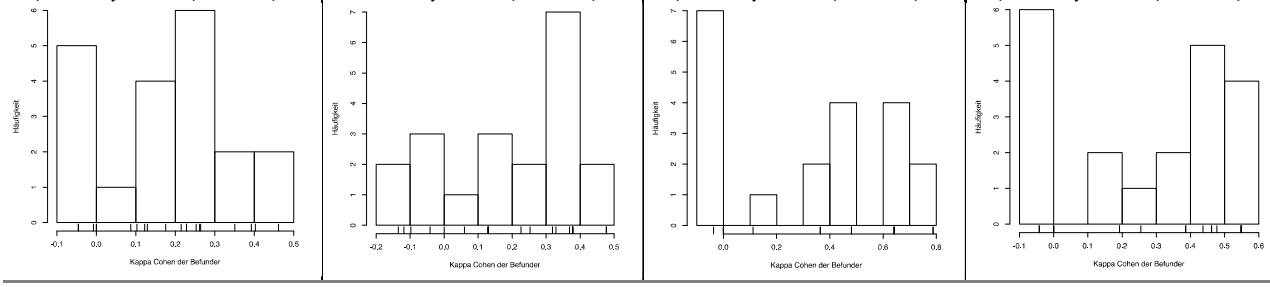
Richtigkeit Einzelbefunde: Kurzstreckige und langstreckige Stenosen wurden in einem knappen Drittel der Fälle erkannt. In ebenfalls einem knappen Drittel der Fälle war die Diagnose einer langstreckigen Stenose zutreffend. Die Diagnose einer kurzstreckigen Stenose war mit gut 2/5 richtigen Befunden etwas zuverlässiger. Hinsichtlich der Spezifität erreichen kurzstreckige und langstreckige Stenosen mit 85,8 % bzw. 89,7 % eine ähnliche Größenordnung. Der negative prädiktive Wert langstreckiger Stenosen liegt mit 90,8 % deutlich über dem kurzstreckiger Stenosen von 77,9 %. Somit wird insgesamt bei beiden Längenangaben ein Kappa Cohen von etwa 0,2 erreicht. Unter Berücksichtigung zufälliger Übereinstimmung war also nur jeder fünfte Befund der Stenoselänge korrekt.

Die Sensitivität des Merkmals „membranös“ liegt ebenfalls bei knapp einem Drittel, die von ringförmig bei nur gut einem Viertel. Die positiven prädiktiven Werte liegen mit knapp 60 % bei „membranös“ und 65 % bei „ringförmig“ wesentlich höher, als bei der Stenoselänge. Auch die Spezifitäten erreichen deutlich höhere Werte. Der negative prädiktive Wert von „membranös“ ist mit knappen 95 % der Spitzenreiter der Stenoseform, „ringförmig“ schneidet mit 88,9 % nur knapp schlechter ab, als „langstreckig“.

Insbesondere bei der positiven Befundung ist die Prognosekraft bei „membranös“ und „ringförmig“ höher als bei „kurzstreckig“ und „langstreckig“: Die positiven prädiktiven Werte von „membranös“ und „ringförmig“ liegen erheblich über ihren Erwartungswerten. Während die der LR- entsprechenden rechten Geraden der ROC bei allen Charakteristika ähnlich steil sind, verlaufen die linken, der LR+ entsprechenden Geraden bei „membranös“ und „ringförmig“ deutlich steiler, als bei „kurzstreckig“ und „langstreckig“. Der hierdurch bedingte Zugewinn bei der AUC fällt gering aus, die odds ratios von „membranös“ und „ringförmig“ liegen jedoch bei einem Vielfachen der von „kurzstreckig“ und „langstreckig“.

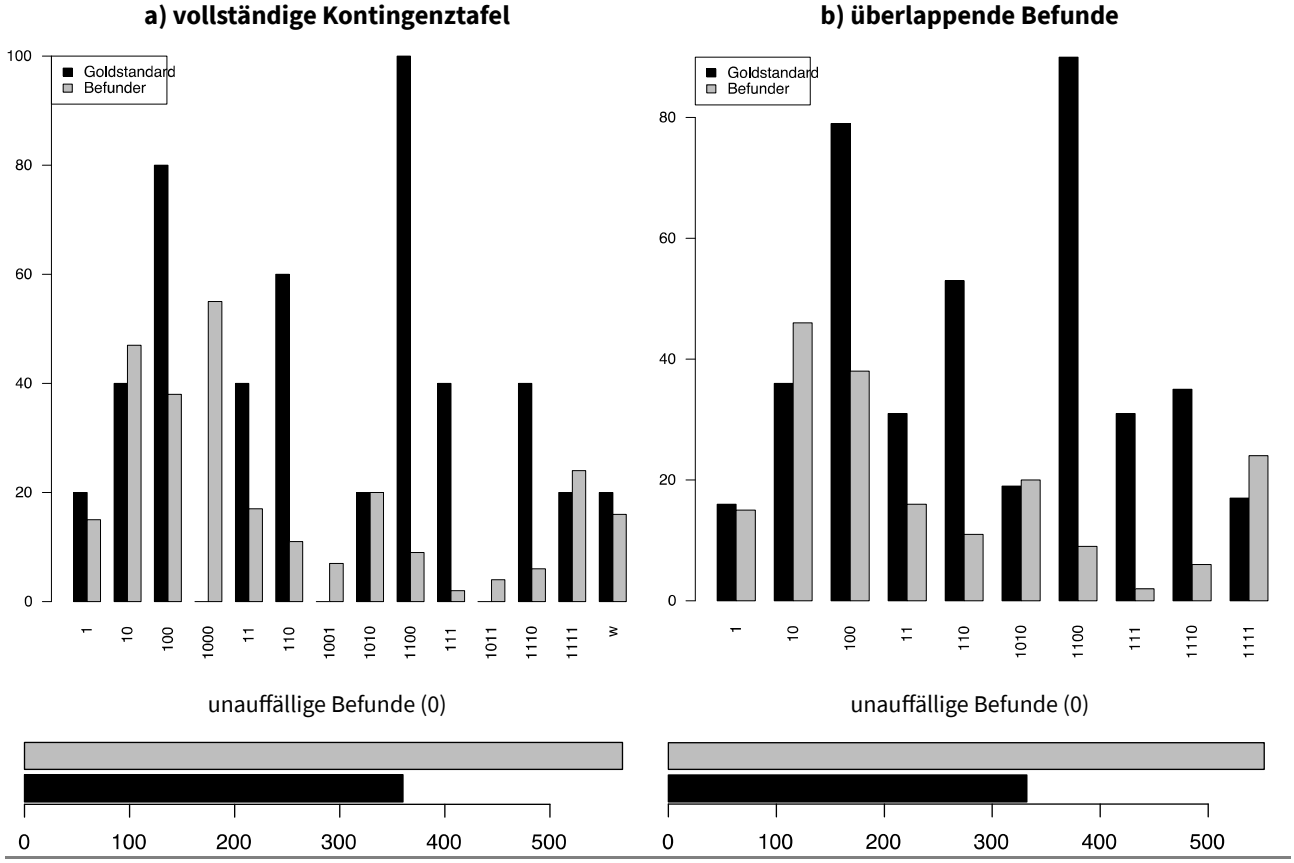
Tabelle 4.48: Paarweises Kappa Cohen Stenoseform

kurzstreckig				**langstreckig**				**membranös**				**ringförmig**			
vereintes Kappa				vereintes Kappa				vereintes Kappa				vereintes Kappa			
0,189				**0,219**				**0,382**				**0,323**			
paarweises Kappa				paarweises Kappa				paarweises Kappa				paarweises Kappa			
min.	Ø	med.	max.	min.	Ø	med.	max.	min.	Ø	med.	max.	min.	Ø	med.	max.
-0,046	**0,179**	0,195	0,462	-0.135	**0,201**	0,239	0,478	-0,037	**0,343**	0,423	0,788	-0,423	**0,291**	0,386	0,548



Die Tabelle vergleicht die Ergebnisse der beiden Berechnungsvarianten „vereintes“ und „paarweises“ Kappa Cohen.

Abbildung 4.30: Randverteilungen Befundkombinationen Stenoseform



Die Balkendiagramme stellen die Befundhäufigkeiten der Befunder denen des Goldstandards gegenüber.

Tabelle 4.49: Inter-Beobachter-Variabilität Befundkombinationen Stenoseform

Befund				Befundverteilung				Präzision Übereinstimmung der Befunder untereinander			Richtigkeit Übereinstimmung mit dem Goldstandard in den überschneidenden Befunden				
				Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
				Anzahl verschiedener											
Anzahl der Qualitäten	kurzstreckig	langstreckig	membranös	ringförmig	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**	Datenabdeckung [%]
x	**gesamt**				**42 [840]**	**840**	**20**	**42**	**24,7**	**0,202**	**schwach**	**21,3**	**24,4**	**0,245**	**schwach**	**96,5**
0				0	25 [500]	536	20	42	63,8	0,303	schwach	81,5	75,4	0,405	mittel	99,4
1				1	1 [20]	20	11	7	18,0	0,110	kaum	26,3	25	0,238	schwach	97,1
			1	0	NA	11	8	6	15,0	0,093	kaum	NA	NA	NA	NA	NA
		1	0	0	5 [100]	103	19	23	28,5	0,212	schwach	32	30,1	0,213	schwach	98,3
	1	0	0	0	6 [120]	134	17	32	23,9	0,092	kaum	20,5	17,9	0,044	kaum	98,7
2			1	1	NA	8	8	4	15,0	0,083	kaum	NA	NA	NA	NA	NA
		1	0	1	NA	3	3	3	NA	-0,004	keine	NA	NA	NA	NA	NA
		1	1	0	NA	2	1	2	NA	-0,002	keine	NA	NA	NA	NA	NA
	1	0	0	1	2 [40]	11	7	6	13,3	0,064	kaum	5,6	18,2	0,066	kaum	91,8
	1	0	1	0	NA	4	3	4	NA	-0,005	keine	NA	NA	NA	NA	NA
	1	1	0	0	NA	1	1	1	NA	-0,001	keine	NA	NA	NA	NA	NA
3	1	0	1	1	3 [60]	7	6	4	20,0	0,083	kaum	10,9	71,4	0,176	kaum	77,4

Die Tabelle analysiert die Befundkombinationen im Sinne eines Gesamtbefundes. Sie gliedert sich in die 3 Hauptbereiche Befundverteilung Präzision und Richtigkeit. Details siehe Kommentar Tabelle 4.15 auf Seite 109.

Befundverteilung Befundkombinationen: Isolierte kurzstreckige Stenosen haben mit 14,4 % die höchste Prävalenz, dicht gefolgt von ausschließlich langstreckigen Stenosen mit 12 %. Singuläre ringförmige Stenosen fallen dem gegenüber mit nur 2,3 % stark ab. Isoliert membranöse Stenosen lagen gemäß der Vorgabe des Goldstandards überhaupt nicht vor. Membranöse Stenosen kamen nur in Kombination mit kurzstreckigen ringförmigen Stenosen vor (5,7 %) - ein Befund, der bei den Befundern wesentlich seltener erhoben wurde (0,8 %). Membranöse Stenosen spielen auch bei den Befundern mit nur 1,3 % ebenfalls so gut wie keine Rolle. Bemerkenswert ist die insgesamt gute Übereinstimmung hinsichtlich der Prävalenzen bei singulären Befunden.

Im Gegensatz zur Symptomebene treten auch bei der Betrachtung der Randsummen bei singulären Befunden keine signifikanten Verzerrungen auf. Die Teststatistik nach McNemar liegt weitab des Signifikanzniveaus. Während sich an der Einschätzung der Länge fast alle Befunder beteiligten (mind. 17/20), wurden die Stenoseform nur von etwa der Hälfte der Befunder als „membranös“ oder „ringförmig“ bezeichnet. Die Streuung der Befunde der Befunder gegenüber dem Goldstandard erreicht bei singulär ringförmigen Stenosen den Faktor 7, bei singulär langstreckigen Stenosen nicht ganz den Faktor 5 und bei singulär kurzstreckigen Stenosen etwas mehr als den Faktor 5. Dem gegenüber ist bei den wesentlich selteneren Mehrfachbefunden die Streuung erkennbar geringer.

Präzision Syndromebene: Eine erkennbare Übereinstimmung innerhalb der Befunder besteht nur bei singulär langstreckigen Stenosen mit einem Kappa Fleiss von 0,212. Mit Ausnahme un-

auffälliger Befunde ist bei sämtlichen anderen Kategorien von einer allenfalls zufälligen Befundübereinstimmung auszugehen.

Richtigkeit Syndromebene: Gut ein Viertel der singulär ringförmigen Stenosen wurde erkannt. Bei singulär langstreckigen Stenosen war es nur ein knappes Drittel, bei singulär kurzstreckigen nur jede Fünfte. Den besten positiven prädiktiven Wert abseits unauffälliger Befunde (75,4 %) erreichen kombiniert kurzstreckige, membranös, ringförmige Stenosen mit 71,4 %. Bei Einzelbefunden sind langstreckige Stenosen mit einem positiven prädiktiven Wert von 30 % der Spitzenreiter, gefolgt von ringförmigen Stenosen mit 25 % und kurzstreckigen Stenosen mit 17,9 %. Die höchste Spezifität bei Einfachbefunden findet sich bei alleinig ringförmigen Stenosen 98,1 % die sich gegen einzig langstreckige Stenosen (89,9 %) und singulär kurzstreckige Stenosen (84,1 %) deutlich absetzen. Bei ringförmigen und langstreckigen Stenosen zeigt sich eine schwache Übereinstimmung mit dem Goldstandard (Kappa Cohen 0,238 & 0,213).

Bei Mehrfachbefunden wurden von den zahlreichen denkbaren Kombinationen vom Goldstandard lediglich zwei gewählt: die Kombination aus kurzstreckig und ringförmig sowie die Kombination aus kurzstreckig, ringförmig und membranös. Das Befundspektrum der Befunder fällt deutlich heterogener aus: Alle sechs möglichen Befundkombinationen bei dualen Befunden werden ausgeschöpft. Bei dreifachen Befunden besteht zwischen Goldstandard und den Befundern Einigkeit darin, dass nur die Kombination kurzstreckig, ringförmig und membranös vorkommt. Damit wurde nur eine von vier theoretisch denkbaren Kombinationen realisiert. Insgesamt haben multiple Befunde bei den Befundern eine Prävalenz von nur 4,3 % (36/840), singuläre hingegen 31,9 % (268/840). Singuläre Befunde sind also mehr als siebenmal so häufig.

Die Kontingenztafel zeigt, dass bei Einzelbefunden – wie bei der Befundung der Stenoselokalisation – der wesentliche Anteil divergierender Befunde auf falsch negative Befunde zurück geht. Abgesehen von falsch negativen Befunden wurden kurzstreckige Stenosen fast ausschließlich als langstreckig fehlklassifiziert und andersherum.

Die Datenabdeckung ist mit insgesamt 96,5 % gut. Die Abdeckung der einzelnen Klassen ist bei singulären Stenosen mit Werten über 97 % gleichmäßig. Lediglich bei Dreifachbefunden fällt die Abdeckung mit 77,4 % unter die 90 %-Marke ab.

Zusammenfassung 4.10: Stenoseform

Auch wenn die Aussagekraft durch eine geringe Prävalenz eingeschränkt wird, scheinen die morphologischen Befunde „membranös“ und „ringförmig“ zuverlässiger, als die Einschätzung der Länge der Stenose, selbst wenn letztere nur durch die einfachst mögliche Einteilung in „langstreckig“ und „kurzstreckig“ erfolgt. Befundkombinationen aus mehreren Merkmalen spielen mit einem Anteil von nur 4,3 % an allen Befunden so gut wie keine Rolle. Wie schon bei der Lage der Stenosen haben falsch negative Befunde einen erheblichen Anteil an der Fehlklassifikation.

Tabelle 4.50: Vier-Felder-Tafeln Einzelbefunde Stenoseform

kurzstreckig				langstreckig				membranös				ringförmig			
McN <0,01	Referenz 0	1	∑	McN 0,503	Referenz 0	1	∑	McN <0,01	Referenz 0	1	∑	McN <0,01	Referenz 0	1	∑
Befunder 0	532	151	683	Befunder 0	664	67	731	Befunder 0	767	41	808	Befunder 0	703	88	791
Befunder 1	88	69	157	Befunder 1	76	33	109	Befunder 1	13	19	32	Befunder 1	17	32	49
∑	620	220	840	∑	740	100	840	∑	780	60	840	∑	720	120	840

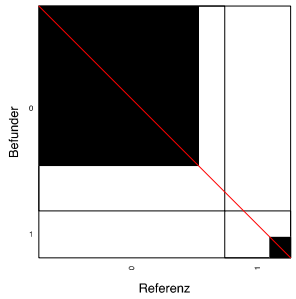
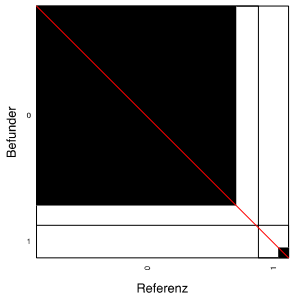
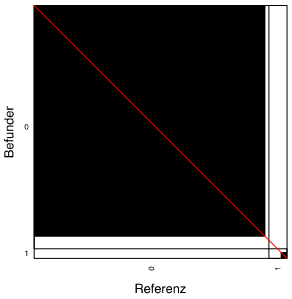
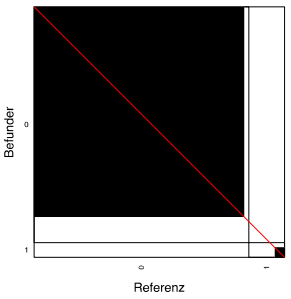
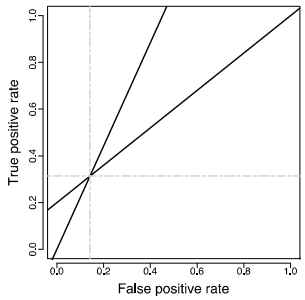
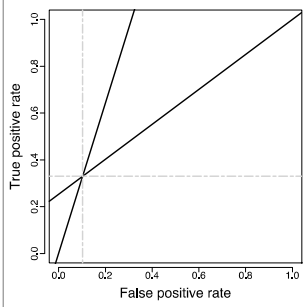
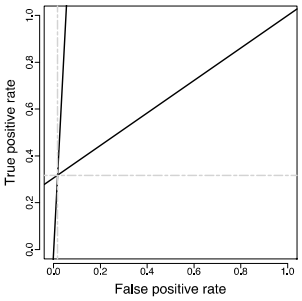
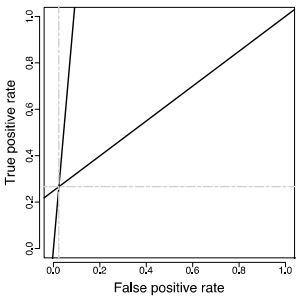
Die Vier-Felder-Tafeln und ihre Randsummen sind Berechnungsgrundlage sämtlicher Kennwerte und Diagramme.

Tabelle 4.51: Kennwerte Einzelbefunde Stenoseform

kurzstreckig				langstreckig				membranös				ringförmig			
Prävalenz	26,2			Prävalenz	11,9			Prävalenz	7,1			Prävalenz	14,3		
	beo.	erw.	kor.		beo.	erw.	kor.		beo.	erw.	kor.		beo.	erw.	kor.
neg. Übst.	81,7	77,4	=κ	neg. Übst.	90,3	87,6	=κ	neg. Übst.	96,6	94,5	=κ	neg. Übst.	93,1	89,7	=κ
pos. Übst.	36,6	21,8	=κ	pos. Übst.	31,6	12,4	=κ	pos. Übst.	41,3	5,0	=κ	pos. Übst.	37,9	8,3	=κ
Spezifität	85,8	81,3	24,1	Spezifität	89,7	87,0	20,9	Spezifität	98,3	96,2	56,2	Spezifität	97,6	94,2	59,5
Sensitivität	31,4	18,7	15,6	Sensitivität	33,0	13,0	23,0	Sensitivität	31,7	3,8	29,0	Sensitivität	26,7	5,8	22,1
Genauigkeit	71,6	64,9	18,9*	Genauigkeit	83,0	78,2	21,9*	Genauigkeit	93,6	89,6	38,2*	Genauigkeit	87,5	81,5	32,3*
Odds ratio	2,8			Odds ratio	4,3			Odds ratio	27,3			Odds ratio	15,0		
Yule Q / Y	0,468 / 0,249			Yule Q / Y	0,623 / 0,349			Yule Q / Y	0,929 / 0,679			Yule Q / Y	0,875 / 0,590		

Beobachtete Kennwerte (beo.), Erwartungswerte gemäß Randsummen (erw.) & für erwartete Übereinstimmung korrigierte Werte (kor.) in Prozent. Die odds ratio und ihre Transformationen (Yule Q & Y) sind prävalenzunabhängig.

Abbildung 4.31: Diagramme Einzelbefunde Stenoseform

kurzstreckig	langstreckig	membranös	ringförmig
Bangdiwala Diagramm	Bangdiwala Diagramm	Bangdiwala Diagramm	Bangdiwala Diagramm
			
Bangdiwala 0,628	Bangdiwala 0,801	Bangdiwala 0,931	Bangdiwala 0,861
npv 77,9(73,8) ppv 43,9(26,2)	npv 90,8(88,1) ppv 30,3(11,9)	npv 94,9(92,9) ppv 59,4(7,1)	npv 88,9(85,7) ppv 65,3(14,3)
ROC & Likelihood Graph	ROC & Likelihood Graph	ROC & Likelihood Graph	ROC & Likelihood Graph
			
AUC 0,586	AUC 0,614	AUC 0,650	AUC 0,622
LR+ 2,21 LR- 0,800	LR+ 3,21 LR- 0,747	LR+ 19,0 LR- 0,695	LR+ 11,3 LR- 0,751

Die Bangdiwala Diagramme illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Die Receiver Operator Characteristic (ROC) trägt die Rate falsch Positiver (=1-Spezifität, fpr) gegen die Rate richtig Positiver (=Sensitivität, tpr) auf. Die Steigung der Tangenten durch den Ursprung bzw. (1,1) definieren die Likelihood ratios (LR+/LR-), die Fläche darunter die Area Under the Curve (AUC).

Tabelle 4.52: Kontingenztafeln der Befundkombinationen Stenoseform

			Referenz						
			0	1			2	3	Summen
			0	1	100	1000	1001	1011	
Befudner	0	0	404	11	43	48	14	16	536 536
	1	1	4	5	0	4	3	4	20 20
		10	2	0	0	1	1	7	11
		100	27	1	31	36	2	6	103 103
		1000	60	1	23	24	14	12	134 134
	2	11	0	1	0	0	1	6	8
		101	1	0	1	1	0	0	3
		110	1	0	0	1	0	0	2
		1001	0	1	0	5	2	3	11 11
		1010	0	0	1	0	2	1	4
		1100	0	0	1	0	0	0	1
	3	1011	1	0	0	0	1	5	7 7
	Summen		496 500	19 20	97 100	117 120	36 40	46 60	811 840

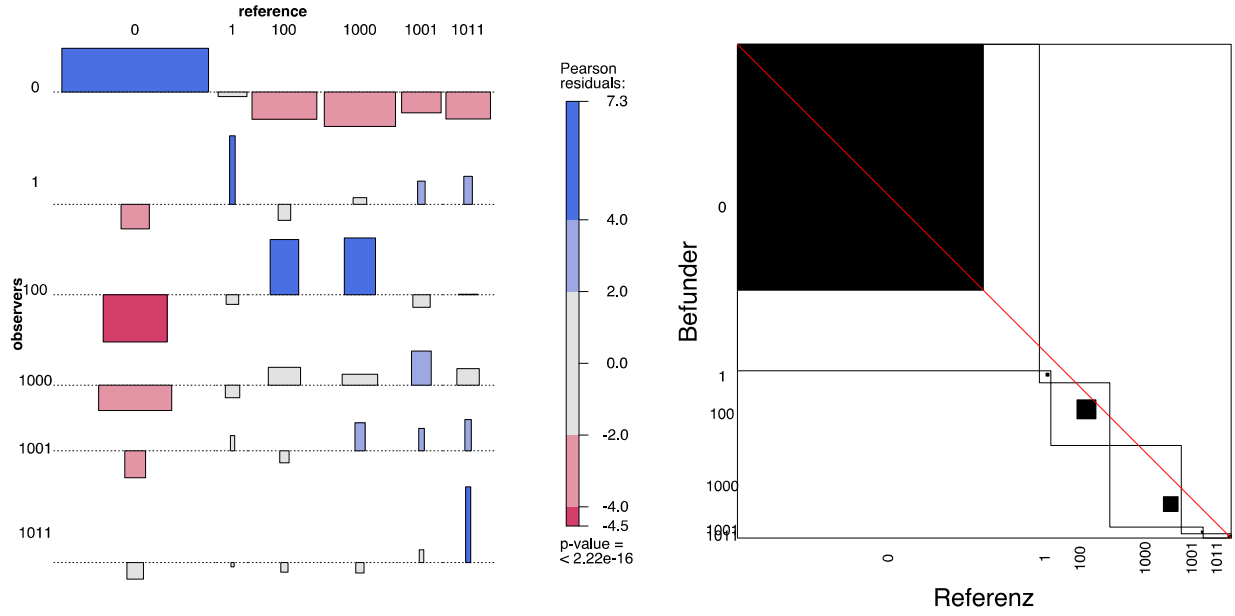
Nicht überlappende Kategorien sind grau, überlappende schwarz dargestellt.

Tabelle 4.53: Kennwerte überlappende Befundkombinationen Stenoseform

Befund		pre	McN	nag	pag	spe	sen	acc	npv	ppv	B	LR-	LR+	AUC	OR	Q	Y
0	0	61,2	0,009	62	78,3	58,1	81,5	72,4	66,5	75,4	0,558	0,319	1,94	0,698	6,1	0,718	0,423
1	1	2,3	1	98,2	25,6	98,1	26,3	96,4	98,2	25	0,963	0,751	13,9	0,622	18,5	0,897	0,623
	100	12	0,67	90,3	31	89,9	32	83	90,7	30,1	0,801	0,757	3,17	0,609	4,2	0,615	0,344
	1000	14,4	0,26	85,2	19,1	84,1	20,5	75	86,3	17,9	0,704	0,945	1,29	0,523	1,4	0,156	0,079
2	1001	4,4	<0,01	97,3	8,5	98,8	5,56	94,7	95,8	18,2	0,946	0,956	4,78	0,522	5	0,667	0,382
3	1011	5,7	<0,01	97,3	18,9	99,7	10,9	94,7	94,9	71,4	0,946	0,894	41,6	0,553	46,5	0,958	0,744

Prävalenz (pre), McNemar Test (McN), negative (nag) und positive (pag) Übereinstimmung, Spezifität (spe), Sensitivität (sen), Genauigkeit (acc), negativer (npv) und positiver prädiktiver Wert (ppv), Bangdiwala (B), negative (LR-) und positive Likelihood ratio (LR+), area under the curve (AUC), odds ratio (OR), Youden Q und Y.

Abbildung 4.32: Diagramme Befundkombinationen Stenoseform



Links: Illustration der Beobachtungen im Vergleich zu den Erwartungswerten in der Kontingenztafel der Befundkombinationen. Rechts: Darstellung für Kontingenztafel der überschneidenden Befundkombinationen.

4.2.4 Spezielle Stenosen

Spezielle Varianten von Stenosen wurden separat erfragt. Dazu zählen Malazien, Pulsationen und Kompressionen.

4.2.4.1 Malazie

Tabelle 4.54: Inter-Beobachter-Variabilität Einzelbefunde Malazie

Befund	**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz				
	Referenz	Befunder Anzahl verschiedener			Ø positive Übereinstimmung [%]	**Kappa Fleiss**		Sensitivität [%]	positiver prädiktiver Wert [%]	**Kappa Cohen**		Datenabdeckung [%]
	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)		Kappa **nach Fleiss**	**modifizierte** Klassifikation **nach Landis**			Kappa **nach Cohen**	**modifizierte** Klassifikation **nach Landis**	
Stenosebereich	4 [80]	57	16	18	18,9	0,106	gering	12,5	17,5	0,072	kaum	99,8
Larynx	5 [100]	32	13	13	18,6	0,131	gering	17,2	53,1	0,214	schwach	
Trachea	8 [160]	73	17	15	29,2	0,241	schwach	28,1	61,6	0,303	mäßig	
Bronchus	2 [40]	46	14	15	17,9	0,110	gering	17,5	15,2	0,118	gering	

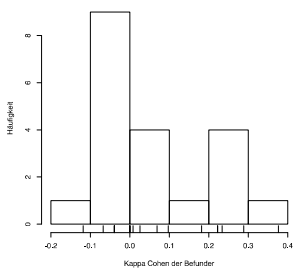
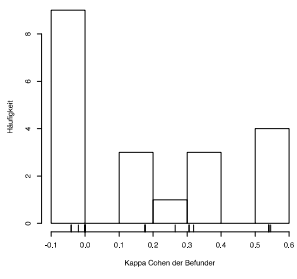
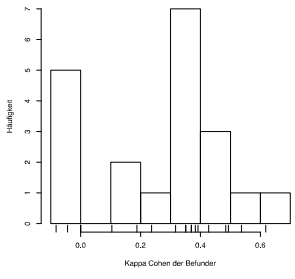
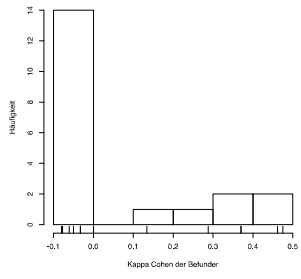
Die Tabelle betrachtet die Übereinstimmung der Einzelbefunde und ist in die Kriterien Befundverteilung, Präzision und Richtigkeit gegliedert. Details siehe Kommentar Tabelle 4.11 Seite 106.

Befundverteilung Einzelbefunde: Die Prävalenz von Malazien in den Videomitschnitten liegt gemäß dem Goldstandard im Stenosebereich und im Larynx bei ca. 10 %. Tracheale Malazien sind mit etwa 20 % knapp doppelt so häufig, bronchiale Malazien mit etwa 5 % halb so häufig. Im Stenosebereich, Larynx und Trachea werden Malazien von den Befundern deutlich seltener befundet, im Bronchus hingegen etwas häufiger. Die Befunde verteilen sich dabei auf deutlich mehr verschiedene Videos, als beim Goldstandard (Faktor 2 bis 7).

Präzision Einzelbefunde: Die beste Übereinstimmung bei der Beurteilung der Malazie innerhalb der Befunder findet sich bei trachealen Malazien. Die Übereinstimmung in den übrigen Abschnitten liegt auf etwa dem gleichen niedrigen Niveau.

Richtigkeit Einzelbefunde: Die Befundübereinstimmung mit dem Goldstandard ist bei Malazien im Bereich der Trachea sowohl hinsichtlich der Sensitivität als auch des positiven prädiktiven Wertes der Befundübereinstimmung in den anderen Abschnitten überlegen. Kongruent hierzu erreicht auch das Kappa nach Cohen mit 0,303 – entsprechend einer mäßigen Übereinstimmung - den höchsten Wert innerhalb der Malazien.

Tabelle 4.55: Paarweises Kappa Cohen Malazien

Stenosebereich	**Larynx**	**Trachea**	**Bronchus**
vereintes Kappa Cohen	vereintes Kappa Cohen	vereintes Kappa Cohen	vereintes Kappa Cohen
0,072	**0,214**	**0,303**	**0,118**
paarweises Kappa Cohen	paarweises Kappa Cohen	paarweises Kappa Cohen	paarweises Kappa Cohen
min. Ø med. max.	min. Ø med. max.	min. Ø med. max.	min. Ø med. max.
-0,118 **0,069** 0,004 0,376	-0.041 **0,189** 0,176 0,546	-0,082 **0,275** 0,351 0,618	-0.08 **0,085** 0 0,475
			

Die Tabelle vergleicht die Ergebnisse der beiden Berechnungsvarianten „vereintes“ und „paarweises“ Kappa Cohen.

Tabelle 4.56: Inter-Beobachter-Variabilität Befundkombinationen Malazie

Befund					**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit Goldstandard in den überschneidenden Befunden				
					Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
					Anzahl verschiedener											
Anzahl der Malazien	Stenosebereich	Larynx	Trachea	Bronchus	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**	Datenabdeckung [%]
x	**gesamt**				**13 [260]**	**161**	**20**	**28**	**16,1**	**0,171**	**gering**	**13,6**	**27,8**	**0,221**	**schwach**	**92,0**
0				0	29 [580]	677	20	42	80,6	**0,256**	**schwach**	93,3	77,6	**0,302**	**mäßig**	94,7
1				1	1 [20]	17	7	7	15,0	**0,093**	**kaum**			**-0,022**	**keine**	91,9
			1	0	3 [60]	34	16	11	24,2	**0,171**	**gering**	32,7	50,0	**0,361**	**mäßig**	89,6
		1	0	0	3 [60]	29	12	12	17,1	**0,145**	**gering**	12,3	33,3	**0,146**	**gering**	86,6
	1	0	0	0	NA	38	11	16	14,2	**0,060**	**kaum**	NA	NA	**NA**	**NA**	NA
2			1	1	1 [20]	22	10	11	18,8	**0,081**	**kaum**	5,0	55,0	**0,021**	**keine**	100
		1	1	0	1 [20]	2	2	1	10,0	**0,050**	**keine**			**-0,005**	**keine**	95,5
	1	0	0	1	NA	4	4	4	NA	**-0,005**	**keine**	NA	NA	**NA**	**NA**	NA
	1	0	1	0	3 [60]	12	7	7	13,3	**0,057**	**kaum**	9,1	41,7	**0,127**	**gering**	92,5
	1	1	0	0	1 [20]	NA	NA	NA	NA	**NA**	**NA**	NA	NA	**NA**	**NA**	NA
3	1	0	1	1	NA	2	2	2	NA	**-0,002**	**keine**	NA	NA	**NA**	**NA**	NA
4	1	1	1	1	NA	1	1	1	NA	**-0,001**	**keine**	NA	NA	**NA**	**NA**	NA
n	ω				NA	2	2	2	NA	**-0,002**	**keine**	NA	NA	**NA**	**NA**	NA

Die Tabelle analysiert die Befundkombinationen im Sinne eines Gesamtbefundes und ist in die Kriterien Befundverteilung, Präzision und Richtigkeit gegliedert. Details siehe Kommentar Tabelle 4.15 auf Seite 109.

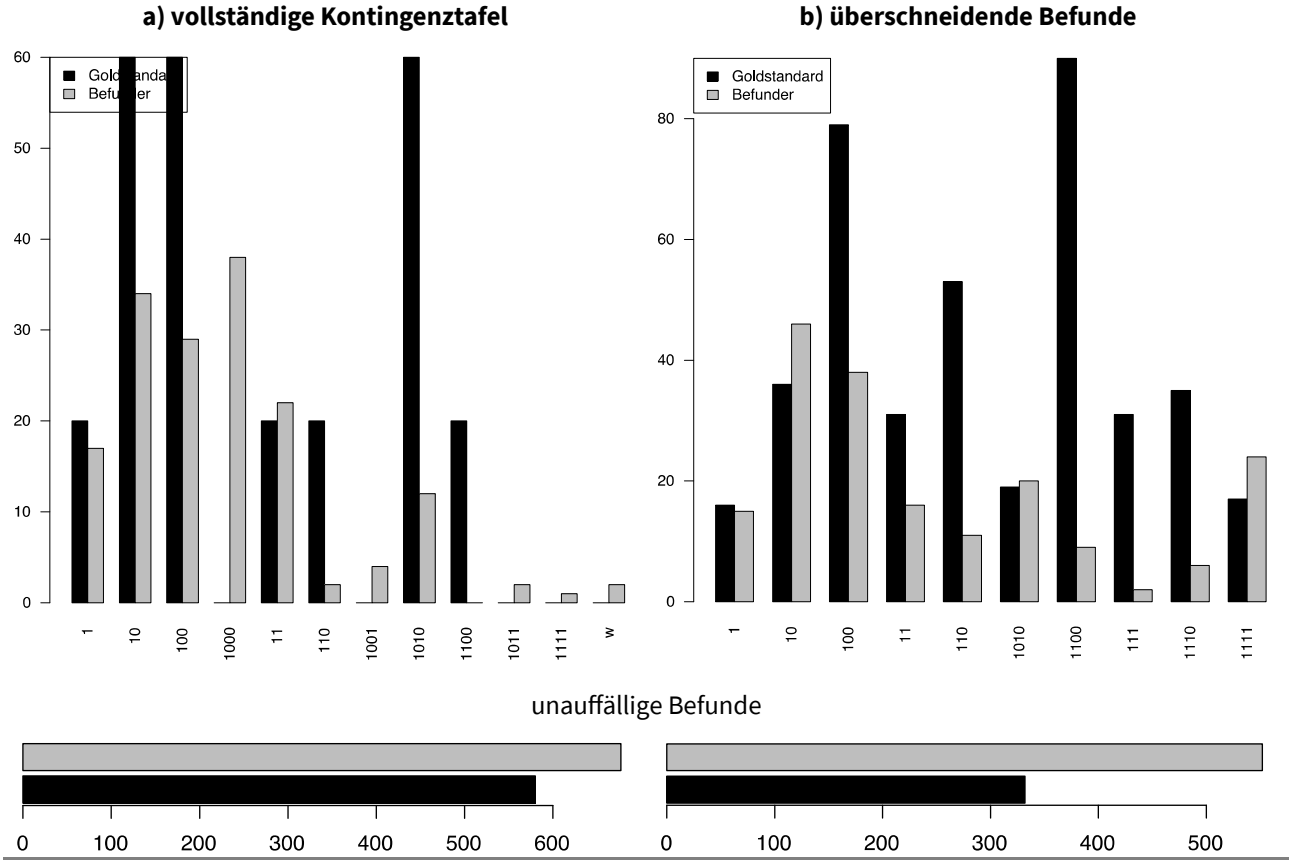
Befundverteilung Befundkombinationen: Gemäß dem Goldstandard sind singuläre Malazien in Larynx und Trachea zusammen mit kombinierten Malazien in Stenosebereich und Trachea am häufigsten. Die Befunder sahen am häufigsten singuläre Malazien im Stenosebereich, die vom

Goldstandard in keinem Fall gesehen wurden. Am zweithäufigsten diagnostizierten die Befunder singuläre Stenosen in Larynx und Trachea. Kombinierte Malazien sind demgegenüber erkennbar seltener und werden insbesondere in Bronchus und Trachea sowie in Stenosebereich und Trachea gesehen.

Präzision Befundkombinationen: Eine erkennbare Übereinstimmung innerhalb der Befunder besteht – abseits unauffälliger Befunde – nur bei singulären Malazien in Trachea und Larynx. Hier erreicht das Kappa nach Fleiss eine „geringe“ Übereinstimmung in der Klassifikation nach Landis.

Richtigkeit Befundkombinationen: Eine nennenswerte Sensitivität und ein nennenswerter positiver prädiktiver Wert bestehen nur bei singulären trachealen Malazien. Das Kappa nach Cohen erreicht hier mit dem Spitzenwert von 0,361 eine „mäßige“ Übereinstimmung. Die Datenabdeckung liegt bei insgesamt 92 % und fällt nie wesentlich unter 90 % ab (min. 86,6 % bei singulären Larynxstenosen).

Abbildung 4.33: Vergleich Randverteilungen Befundkombinationen Entzündungsbereich



Die Assoziationsdiagramme zeigen das Verhältnis der beobachteten Werte zu den Erwartungswerten.

Tabelle 4.57: Vierfeldertafeln Einzelbefunde Malazie

Stenosebereich		Referenz			∑
		0	1	NA	
Befunder	0	711	70	0	781
	1	47	10	0	57
	NA	2	0	0	2
∑		760	80	0	840

Larynx		Referenz			∑
		0	1	NA	
Befunder	0	724	82	0	806
	1	15	17	0	32
	NA	1	1	0	2
∑		740	100	0	840

Trachea		Referenz			∑
		0	1	NA	
Befunder	0	650	115	0	765
	1	28	45	0	73
	NA	2	0	0	2
∑		680	160	0	840

Bronchus		Referenz			∑
		0	1	NA	
Befunder	0	759	33	0	792
	1	39	7	0	46
	NA	2	0	0	2
∑		800	40	0	840

Der Mc Nemar Test (McN) prüft auf signifikant differierende Randsummen. P-Werte < 0,05 sind schwarz hervorgehoben.

Tabelle 4.58: Kennwerte Einzelbefunde Malazie

Stenosebereich			
Prävalenz	9,5		
	beo.	erw.	kor.
neg. Übst.	92,4	91,8	=κ
pos. Übst.	14,6	7,9	=κ
Spezifität	93,8	93,2	8,8
Sensitivität	12,5	6,8	6,1
Genauigkeit	86,0	85,0	7,2
Odds ratio	2,2		
Yule Q / Y	0,367 / 0,190		

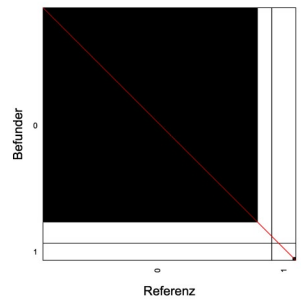
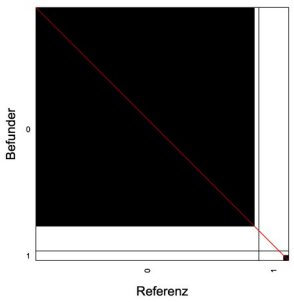
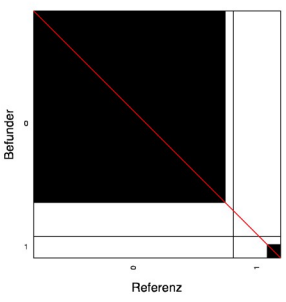
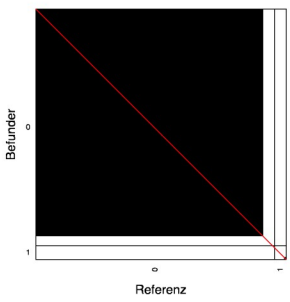
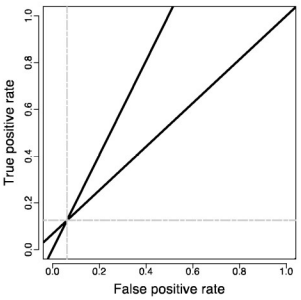
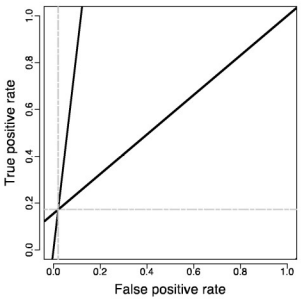
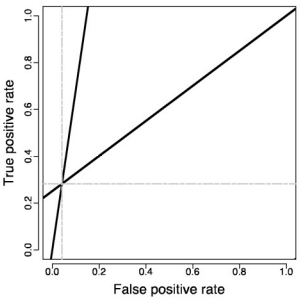
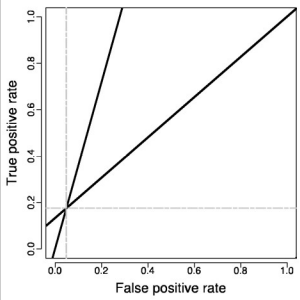
Larynx			
Prävalenz	11,8		
	beo.	erw.	kor.
neg. Übst.	93,7	92,0	=κ
pos. Übst.	26,0	5,8	=κ
Spezifität	98,0	96,2	46,8
Sensitivität	17,2	3,8	13,9
Genauigkeit	88,4	85,3	21,3
Odds ratio	10		
Yule Q / Y	0,818 / 0,520		

Trachea			
Prävalenz	19,1		
	beo.	erw.	kor.
neg. Übst.	90,1	85,8	=κ
pos. Übst.	38,6	12,0	=κ
Spezifität	95,9	91,3	52,6
Sensitivität	28,1	8,7	21,3
Genauigkeit	82,9	75,5	30,3
Odds ratio	9,1		
Yule Q / Y	0,802 / 0,502		

Bronchus			
Prävalenz	4,8		
	beo.	erw.	kor.
neg. Übst.	95,5	94,9	=κ
pos. Übst.	16,3	5,1	=κ
Spezifität	95,1	94,5	11,0
Sensitivität	17,5	5,5	12,7
Genauigkeit	91,4	90,3	11,8
Odds ratio	4,1		
Yule Q / Y	0,610 / 0,340		

Beobachtete Kennwerte (beo.), Erwartungswerte gemäß Randsummen (erw.) & für erwartete Übereinstimmung korrigierte Werte (kor.) in Prozent. Die odds ratio und ihre Transformationen (Yule Q & Y) sind prävalenzunabhängig.

Abbildung 4.34: Diagramme Einzelbefunde Malazie

Stenosebereich	Larynx	Trachea	Bronchus
Bangdiwala Konkordanz	Bangdiwala Konkordanz	Bangdiwala Konkordanz	Bangdiwala Konkordanz
			
Bangdiwala 0,848 (0,837)	Bangdiwala 0,876 (0,844)	Bangdiwala 0,800 (0,723)	Bangdiwala 0,909 (0,897)
npv 91,0(90,5) ppv 17,5(9,6)	npv 89,8(88,2) ppv 53,1(11,8)	npv 85,0(80,9) ppv 61,6(19,1)	npv 95,8(95,2) ppv 15,2(4,8)
ROC & Likelihood Graph	ROC & Likelihood Graph	ROC & Likelihood Graph	ROC & Likelihood Graph
			
AUC 0,531	**AUC 0,576**	**AUC 0,620**	**AUC 0,563**
LR+ 2,02 LR- 0,933	**LR+ 8,46 LR- 0,845**	**LR+ 6,81 LR- 0,750**	**LR+ 3,58 LR- 0,867**

Die Bangdiwala Diagramme illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Die Receiver Operator Characteristic (ROC) trägt die Rate falsch Positiver (=1-Spezifität, fpr) gegen die Rate richtig Positiver (=Sensitivität, tpr) auf. Die Steigung der Tangenten durch den Ursprung bzw. (1,1) definieren die Likelihood ratios (LR+/LR-), die Fläche darunter die Area Under the Curve (AUC).

Tabelle 4.59: Kontingenztafel Befundkombinationen Malazie

		Bhapkar	Referenz 0	Referenz 1			Referenz 2				Summen	
			0	1	10	100	11	110	1010	1100		
Befunder	0	0	516	10	28	44	15	7	45	12	665	677
	1	1	10	0	2	5	0	0	0	0	17	17
		10	7	2	17	0	0	7	1	0	34	34
		100	9	0	0	7	2	2	1	8	21	29
		1000	22	2	7	2	0	1	4	0		38
	2	11	9	5	2	1	1	1	3	0	22	22
		110	0	0	0	0	2	0	0	0	2	2
		1001	2	1	0	0	0	0	1	0		4
		1010	2	0	3	0	0	2	5	0	12	12
	3	1011	2	0	0	0	0	0	0	0		2
	4	1111	0	0	1	0	0	0	0	0		1
	x	ω	1	0	0	1	0	0	0	0		2
Summe überl.			553	17	52	57	20	19	55		773	
Summe			580	20	60	60	20	20	60	20		840

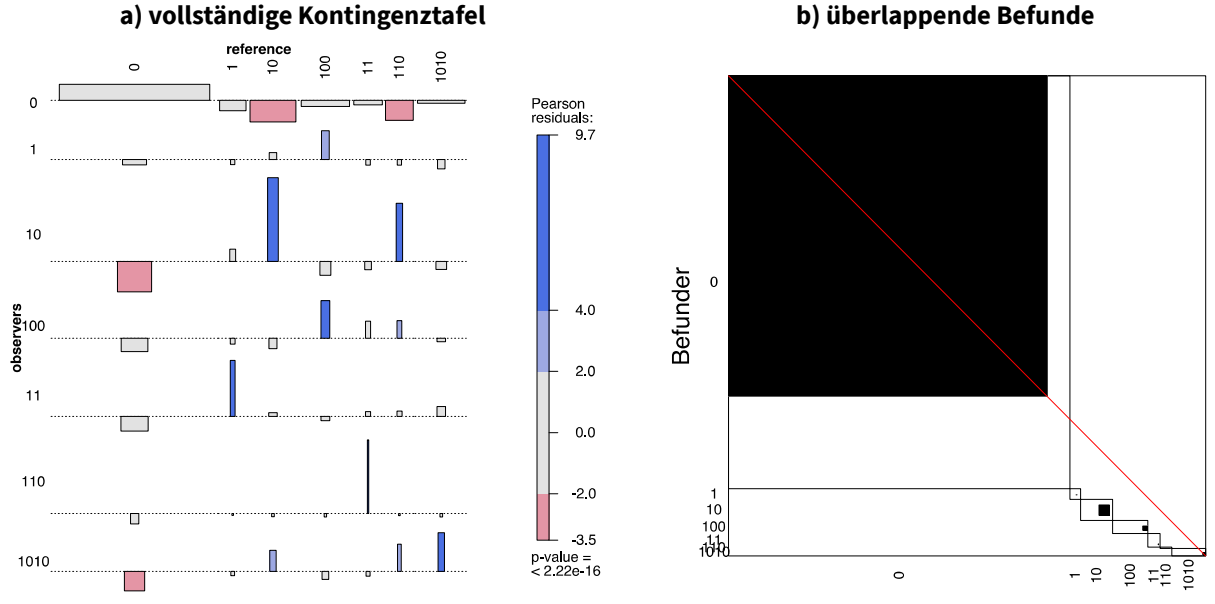
Nicht überlappende Kategorien sind grau, überlappende schwarz dargestellt.

Tabelle 4.60: Kennwerte Befundkombinationen Malazie

Befund		pre	McN	nag	pag	spe	sen	acc	npv	ppv	B	LR-	LR+	AUC	OR	Q	Y
0	0	71,5	0	43,3	84,7	32,3	93,3	75,9	65,7	77,6	0,693	0,207	1,38	0,628	6,6	0,738	0,441
	1	2,2	1	97,8	0	97,8	0	95,6	97,8	0	0,955	1,02	0	0,489	n.a.	n.a	n.a
1	10	6,7	0,018	96,4	39,5	97,6	32,7	93,3	95,3	50	0,928	0,689	13,9	0,652	20,1	0,905	0,635
	100	7,4	0	95,6	17,9	98	12,3	91,7	93,4	33,3	0,913	0,895	6,28	0,552	7	0,751	0,452
	11	2,6	0,874	97,3	4,8	97,2	5	94,8	97,5	4,55	0,947	0,977	1,79	0,511	1,8	0,294	0,151
2	110	2,5	0	98,6	0	99,7	0	97,3	97,5	0	0,973	1	0	0,499	n.a.	n.a	n.a
	1010	7,1	0	96,1	14,9	99	9,09	92,6	93,4	41,7	0,924	0,918	9,32	0,541	10,2	0,821	0,522

Prävalenz (pre), McNemar Test (McN), negative (nag) und positive (pag) Übereinstimmung, Spezifität (spe), Sensitivität (sen), Genauigkeit (acc), negativer (npv) und positiver prädiktiver Wert (ppv), Bangdiwala (B), negative (LR-) und positive Likelihood ratio (LR+), area under the curve (AUC), odds ratio (OR), Youden Q und Y.

Abbildung 4.35: Diagramme Befundkombinationen Malazie



Die Assoziationsdiagramme zeigen das Verhältnis der beobachteten Werte zu den Erwartungswerten.

Zusammenfassung 4.11: Malazie

Auf Ebene der Einzelbefunde wurden tracheale Malazien mit einem Kappa nach Fleiss von 0,241 am einheitlichsten befundet und erreichten bei einem Kappa Cohen von 0,303 die beste Übereinstimmung mit dem Goldstandard. In der Rangliste folgen sowohl hinsichtlich Präzision wie auch der Richtigkeit Larynxmalazien. Bronchusmalazien bilden das Schlusslicht. Auch bei syndromaler Betrachtung zeigten singuläre Malazien im Bereich der Trachea in Präzision und Richtigkeit die höchste Übereinstimmung und werden auch hier von laryngealen Malazien gefolgt. Bei kombinierte Malazien wurden weder innerhalb der Untersucher noch im Vergleich zum Goldstandard eine erkennbare Übereinstimmung erzielt. Dabei wurde nur jede dritte singuläre Tracheomalazie als solche erkannt und nur jede 10. Laryngomalazie. Der Befund einer singulären Tracheomalazie war in 50 % der Fälle, der Befund der Laryngomalazie in ⅓ der Fälle korrekt.

4.2.4.2 Pulsationen

Tabelle 4.61: Inter-Beobachter-Variabilität Einzelbefunde Pulsationen

Befund	**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz				
	Referenz	Befunder Anzahl verschiedener			Ø positive Übereinstimmung [%]	**Kappa Fleiss**		Sensitivität [%]	positiver prädiktiver Wert [%]	**Kappa Cohen**		Datenabdeckung [%]
	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)		Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**			Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**	
Stenosebereich	7 [140]	74	18	12	33,2	0,324	**mäßig**	33,1	62,2	0,358	**mäßig**	99,9
Larynx	NA	9	6	6	12,5	0,037	**kaum**	NA	NA	NA	**NA**	
Trachea	7 [140]	47	16	11	30,7	0,273	**schwach**	27,3	80,9	0,355	**mäßig**	
Bronchus	5 [100]	39	16	10	23,1	0,172	**gering**	26,0	66,7	0,329	**mäßig**	

Die Tabelle betrachtet die Übereinstimmung der Einzelbefunde und ist in die Kriterien Befundverteilung, Präzision und Richtigkeit gegliedert. Details siehe Kommentar Tabelle 4.11 Seite 106.

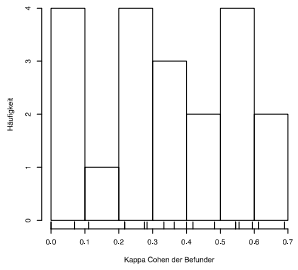
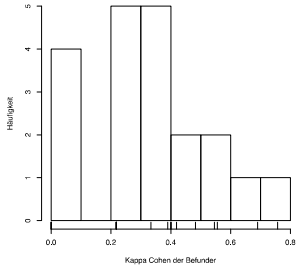
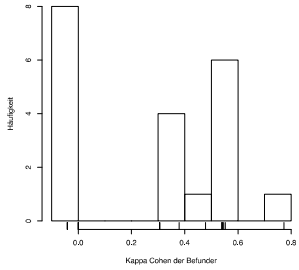
Befundverteilung Einzelbefunde: Pulsationen verteilen sich gemäß dem Goldstandard in etwa zu je einem Drittel auf Stenosebereich, Trachea und Bronchus. Pulsationen im Larynx waren laut dem Goldstandard in den Videos nicht zu sehen. Auch bei den Befundern spielen Pulsationen im Larynx eine untergeordnete Rolle. Die Untersucher befundeten am häufigsten Pulsationen im Stenosebereich, gefolgt von trachealen Pulsationen. An dem häufigsten Befund, der Pulsation im Stenosebereich, waren bis auf 2 alle 20 Befunder beteiligt. Der seltenste Befund – die Pulsation im Larynx – wurde hingegen nur von 6 Untersuchern gesehen. Die Diagnose von Pulsationen in Trachea und Bronchus wurde von etwas mehr als drei Viertel der Befunder gestellt. Dabei verteilen sich die Befunde der Untersucher bei Pulsationen in Stenosebereich, Trachea und Bronchus auf ca. 10 Videos. Pulsationen im Larynx wurden bei 6 Videos gesehen.

Präzision Einzelbefunde: Pulsationen im Stenosebereich als häufigster Befund erzielen mit einem Kappa Fleiss von 0,324 gleichzeitig auch die beste Übereinstimmung innerhalb der Befunder. Es folgen Pulsationen der Trachea mit einem Kappa Fleiss von 0,273 und Pulsationen im Bronchus mit einem Kappa Fleiss von 0,172. Schlusslicht sind Pulsationen im Larynx bei denen mit einem Kappa Fleiss von 0,037 keine nennenswerte Übereinstimmung mehr zu erkennen ist. Somit folgt die Rangliste der Übereinstimmung innerhalb der Befunder den Häufigkeiten.

Richtigkeit Einzelbefunde: Knapp ein Drittel der Pulsationen im Stenosebereich wurde erkannt, Pulsationen in Trachea und Bronchus nur zu einem guten Viertel. Am verlässlichsten wurden mit einem positiven prädiktiven Wert von 80 % Pulsationen der Trachea beschrieben; Pulsationen in Bronchus und dem Stenosebereich sind immerhin noch in knapp 2/3 der Fälle korrekt befundet. Unter Berücksichtigung der unterschiedlichen Erwartungswerte für die einzelnen Befunde, liegen alle Pulsationen mit einem Kappa Cohen von um 0,35 in etwa gleich auf, was einer mäßigen Übereinstimmung mit dem Goldstandard entspricht.

Bei prävalenzunabhängiger Betrachtung führen Pulsationen in der Trachea mit einer odds ratio von 28,9 vor Pulsationen im Bronchus mit einer odds ratio 19,6 und Pulsationen im Stenosebereich mit einer odds ratio von 11,9.

Tabelle 4.62: Paarweises Kappa Cohen Pulsationen

Stenosebereich				**Larynx**				**Trachea**				**Bronchus**			
vereintes Kappa Cohen				vereintes Kappa Cohen				vereintes Kappa Cohen				vereintes Kappa Cohen			
0,358				**na**				**0,355**				**0,329**			
paarweises Kappa Cohen				paarweises Kappa Cohen				paarweises Kappa Cohen				paarweises Kappa Cohen			
min.	Ø	med.	max.	min.	Ø	med.	max.	min.	Ø	med.	max.	min.	Ø	med.	max.
0	**0,339**	0,348	0,690	na	**na**	na	na	0	**0,323**	0,362	0,757	-0.041	**0,282**	0.306	0,773
				na											

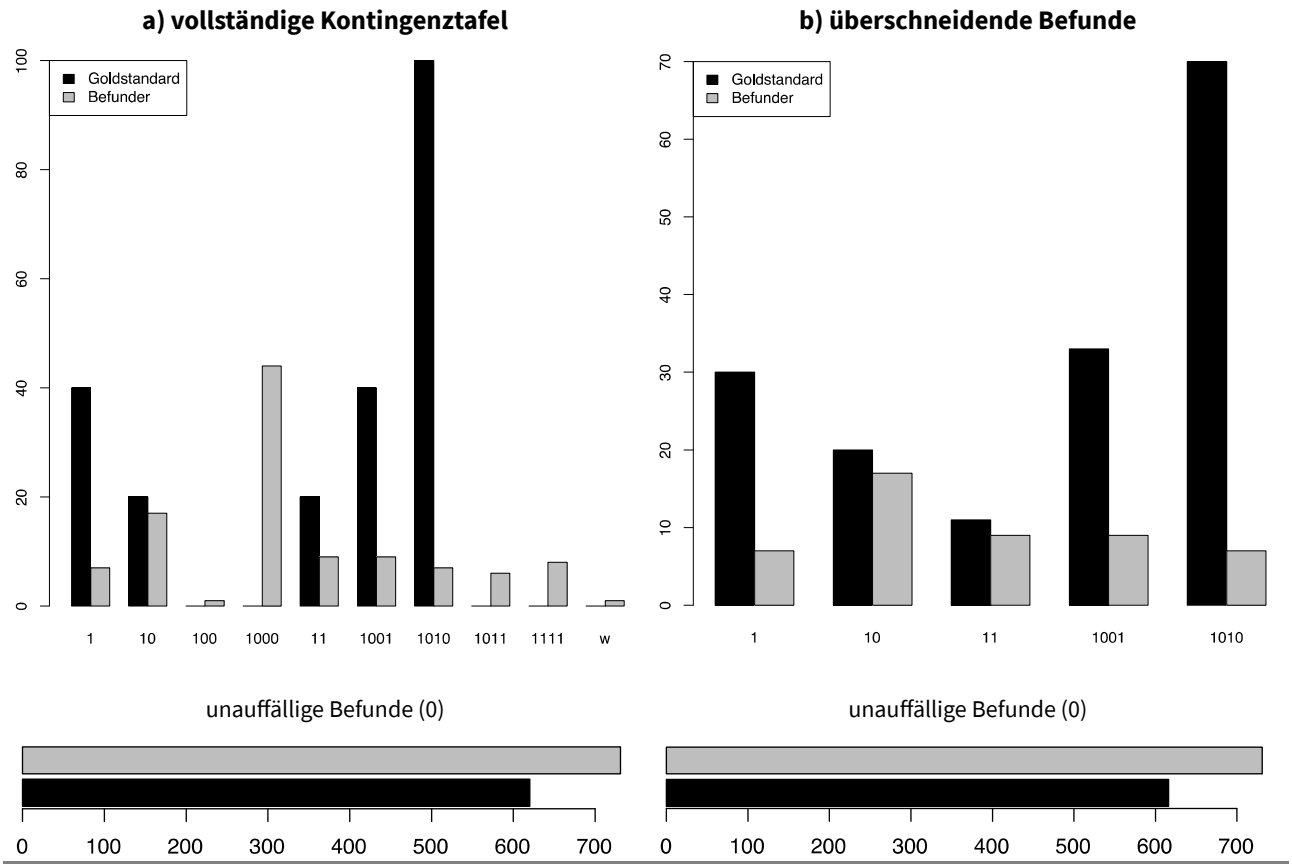
Die Tabelle vergleicht die Ergebnisse der beiden Berechnungsvarianten „vereintes“ und „paarweises“ Kappa Cohen.

Befundverteilung Befundkombinationen: Gemäß dem Goldstandard waren in 11 von 42 Videos (26 %) Pulsationen zu sehen, wobei Kombinationsbefunde (8/42; 19 %) die Zahl der Einzelbefunde (3/42; 7 %) deutlich überwogen. Unter den pathologischen Befunden sind gemäß dem Goldstandard kombinierte Pulsationen in Trachea und Stenosebereich (5/42; 12 %) am häufigsten. Die Befunder sehen insgesamt nur halb so viele Pulsationen (108/840; 13 %) wie der Goldstandard, wobei singuläre Pulsationen im Stenosebereich führen (44/108; 41 %) – ein Befund der vom Goldstandard nicht erhoben wurde. Im Vergleich zum Goldstandard ist der Anteil an Kombinationsbefunden (25/108; 23 %) bei den Befundern etwas geringer. Dreifache und vierfache Pulsationen wurden ausschließlich von den Befundern dokumentiert. Wie die Säulendiagramme in Abbildung 4.36 zeigen, differieren die Häufigkeiten fast aller Befundkombinationen zwischen Goldstandard und Befundern erheblich. Nur Pulsationen in der Trachea wurden ähnlich häufig beobachtet. Dem entsprechend ist der Bhapkar-Test für die Kontingenztafel 4.66 signifikant.

Betrachtet man zur Bestimmung der Richtigkeit nur die überschneidenden Befundkombinationen, bleiben insgesamt zwar nur 7,1 % der Befunde unberücksichtigt. Dabei sind jedoch vorwiegend positive Befunde von der Beschneidung der Kontingenztafel betroffen, sodass nur zwischen 67,9 % und 84,4 % in die Berechnung der Richtigkeit einfließen.

Wie die Kontingenztafel 4.66 zeigt, liegt keineswegs nur eine Fehlklassifikation der Lokalisation von Pulsationen vor, sondern mit 77/114 wurden 67 % der vom Goldstandard vorgegebenen Kombinationsbefunde durch die Untersucher als unauffällig gewertet. Bei den Einzelbefunden machen falsch negative Befunde 40/70 entsprechend 57 % aus. Der Anteil hinsichtlich Anzahl und Ort fehlklassifizierter Pulsationen tritt also hinter falsch negativen Befunden deutlich zurück.

Abbildung 4.36: Randverteilungen Befundkombinationen Pulsationen



Die Balkendiagramme stellen die Befundhäufigkeiten der Befunder denen des Goldstandards gegenüber.

Präzision Befundkombinationen: Eine erkennbare Übereinstimmung innerhalb der Befunder liegt nur für singuläre Pulsationen im Stenosebereich mit einem Kappa Fleiss von 0,212 und eventuell noch für singuläre tracheale Pulsationen mit einem Kappa Fleiss von 0,137 vor. Beide Diagnosen erreichen damit jedoch lediglich das Niveau einer „schwachen“ bzw. „geringen“ Übereinstimmung. Insgesamt besteht – aufgrund der „guten“ Übereinstimmung bei unauffälligen Befunden mit einem Kappa Fleiss von 0,317 – eine mäßige Übereinstimmung.

Richtigkeit Befundkombinationen: Die Anzahl der erkannten Pulsationen schwankt um etwa 10 %. Der höchste prädiktive Wert abseits unauffälliger Befunde tritt mit 85,7 % bei in Trachea und Stenosebereich kombinierten Stenosen auf. Nur bei singulären bronchialen Pulsationen ist mit einem Kappa Cohen von 0,205 eine „schwache“ Übereinstimmung mit dem Goldstandard feststellbar. Angesichts einer Datenabdeckung von nur 76,7 % und stark differierender Befundprävalenzen zwischen Goldstandard und Befundern (40 : 7) ist dieses Ergebnis mit Vorsicht zu interpretieren. Insgesamt erreicht die Richtigkeit von Pulsationen mit einem Kappa Cohen von 0,248 das Niveau einer schwachen Übereinstimmung.

Tabelle 4.63: Inter-Beobachter-Variabilität Befundkombinationen Pulsationen

Befunde					**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz bei überschneidenden Befunden				
					Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
					Anzahl verschiedener											
Anzahl der Pulsationen	Stenosebereich	Larynx	Trachea	Bronchus	Befunde (max. 20) (& Videos)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	Klassifikation **modifiziert nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa **nach Cohen**	Klassifikation **modifiziert nach Landis**	Datenabdeckung [%]
x	gesamt				11 [220]	108	19	15	24.2	0,317	mäßig	9,2	30,6	0,248	schwach	92,9
0				0	31 [620]	731	20	42	87,0	0,522	gut	99,7	84,0	0,381	mäßig	99,5
1				1	2 [40]	7	4	3	15,0	0,098	kaum	13,3	57,1	0,205	schwach	76,7
			1	0	1 [20]	17	10	8	20,0	0,137	gering	0	0	-0,024	keine	100
		1	0	0	NA	1	1	1	NA	NA	NA	NA	NA	NA	NA	NA
	1	0	0	0	NA	44	15	10	26,2	0,212	schwach	NA	NA	NA	NA	NA
2			1	1	1 [20]	9	5	5	15,0	0,072	kaum	9,1	11,1	0,088	kaum	67,9
	1	0	0	1	2 [40]	9	7	4	17,5	0,096	kaum	12,1	44,4	0,176	gering	84,4
	1	0	1	0	5 [100]	7	6	5	10,0	0,022	keine	8,6	85,7	0,142	gering	70,3
3	1	0	1	1	NA	6	6	4	15,0	0,046	keine	NA	NA	NA	NA	NA
4	1	1	1	1	NA	8	6	5	12,5	0,044	keine	NA	NA	NA	NA	NA
n	ω				NA	1	1	1	NA	NA	NA	NA	NA	NA	NA	NA

Die Tabelle analysiert die Befundkombinationen im Sinne eines Gesamtbefundes. Sie gliedert sich in die 3 Hauptbereiche Befundverteilung Präzision und Richtigkeit. Details siehe Kommentar Tabelle 4.15 auf Seite 109.

Zusammenfassung 4.12: Pulsationen

Auf Ebene der Einzelbefunde wurde die häufigste Diagnose, die Pulsation im Stenosebereich, sowohl in der Übereinstimmung untereinander (Kappa Fleiss 0,324), als auch zum Goldstandard (Kappa Cohen 0,358) am besten bewertet. Die weitere Rangliste folgt sowohl hinsichtlich Präzision wie auch hinsichtlich der Richtigkeit den Häufigkeiten der Befunde bei den Untersuchern: Pulsationen in der Trachea (Kappa Fleiss 0,273, Kappa Cohen 0,355), Pulsationen im Bronchus (Kappa Fleiss 0,172; Kappa Cohen 0,329) und zuletzt Pulsationen im Larynx (Kappa Fleiss 0,370), die vom Goldstandard nicht gesehen wurden (Kappa Cohen daher nicht erhebbar).
Auf Ebene der Befundkombinationen waren Übereinstimmungen nur bei singulären Pulsationen erkennbar. Die beste Konkordanz der Befunder findet sich bei Pulsationen im Stenosebereich mit einem schwachen Kappa Fleiss von 0,212 – einem Befund der vom Goldstandard nicht erhoben wurde. Eine schwache Übereinstimmung mit dem Goldstandard wurde nur bei Pulsationen im Bronchus beobachtet. Allerdings gingen bei der Berechnung der Richtigkeit wegen der geringen Überschneidung der Befundkategorien etwa ⅓ der Daten verloren.

Tabelle 4.64: Vier-Felder-Tafeln Einzelbefunde Pulsationen

Stenosebereich		Referenz			∑
		0	1	ω	
Befunder	0	672	93	**0**	765
	1	28	46	**0**	74
	ω	**0**	**1**	**0**	1
∑		700	140	0	840

Larynx		Referenz			∑
		0	1	ω	
Befunder	0	830	0	**0**	830
	1	9	0	**0**	9
	ω	**1**	**0**	**0**	1
∑		840	0	0	840

Trachea		Referenz			∑
		0	1	ω	
Befunder	0	691	101	**0**	792
	1	9	38	**0**	47
	ω	**0**	**1**	**0**	1
∑		700	140	0	840

Bronchus		Referenz			∑
		0	1	ω	
Befunder	0	726	74	**0**	800
	1	13	26	**0**	39
	ω	**1**	**0**	**0**	1
∑		740	100	0	840

Die Vier-Felder-Tafeln und ihre Randsummen sind Berechnungsgrundlage sämtlicher Kennwerte und Diagramme.

Tabelle 4.65: Kennwerte Einzelbefunde Pulsationen

Stenosebereich			
Prävalenz	16,6		
	beo.	erw.	kor.
neg. Übst.	91,7	87,1	=κ
pos. Übst.	43,2	11,5	=κ
Spezifität	96,0	91,2	54,6
Sensitivität	33,1	8,8	26,6
Genauigkeit	85,6	77,5	35,8
Odds ratio	11,9		
Yule Q/ Y	0,845 / 0,550		

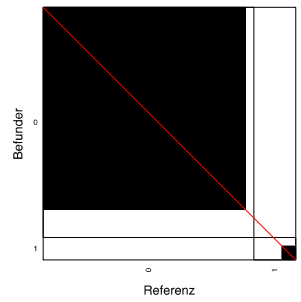
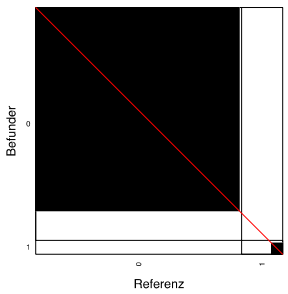
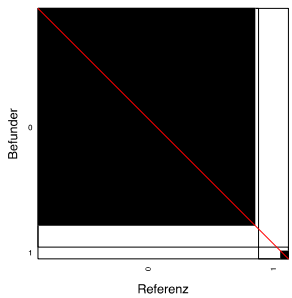
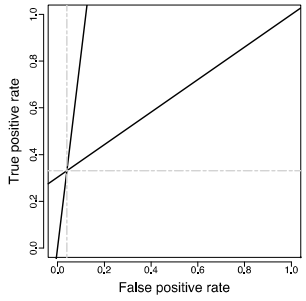
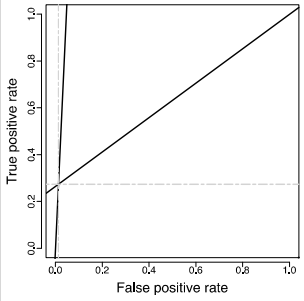
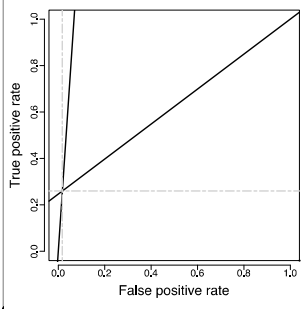
Larynx			
Prävalenz			
	beo.	erw.	kor.
neg. Übst.		NA	
pos. Übst.		NA	
Spezifität		NA	
Sensitivität		NA	
Genauigkeit		NA	
Odds ratio		NA	
Yule Q / Y		NA	

Trachea			
Prävalenz	16,6		
	beo.	erw.	kor.
neg. Übst.	92,6	88,6	=κ
pos. Übst.	40,9	8,4	=κ
Spezifität	98,7	94,4	77,0
Sensitivität	27,3	5,6	23,0
Genauigkeit	86,9	79,7	35,5
Odds ratio	28,9		
Yule Q / Y	0,933 / 0,686		

Bronchus			
Prävalenz	11,9		
	beo.	erw.	kor.
neg. Übst.	94,3	91,6	=κ
pos. Übst.	37,4	6,7	=κ
Spezifität	98,2	95,4	62,2
Sensitivität	26,0	4,7	22,4
Genauigkeit	89,6	84,5	32,9
Odds ratio	19,6		
Yule Q / Y	0,903 / 0,632		

Beobachtete Kennwerte (beo.), Erwartungswerte gemäß Randsummen (erw.) & für erwartete Übereinstimmung korrigierte Werte (kor.) in Prozent. Die odds ratio und ihre Transformationen (Yule Q & Y) sind prävalenzunabhängig.

Abbildung 4.37: Diagramme Einzelbefunde Pulsationen

Stenosebereich	**Larynx**	**Trachea**	Bronchus
Bangdiwala Konkordanz	Bangdiwala Konkordanz	Bangdiwala Konkordanz	Bangdiwala Konkordanz
	NA		
Bangdiwala 0,831 (0,747)	Bangdiwala NA	Bangdiwala 0,854 (0,779)	Bangdiwala 0,887 (0,834)
npv 87,8(16,6) ppv 62,2(16,6)	npv NA ppv NA	npv 87,2(83,4) ppv 80,9(16,6)	npv 90,8(88,1) ppv 66,7(11,9)
ROC & Likelihood Graph	ROC & Likelihood Graph	ROC & Likelihood Graph	ROC & Likelihood Graph
	NA		
AUC 0,645	**AUC NA**	**AUC 0,630**	**AUC 0,621**
LR+ 8,27 LR- 0,645	**LR+ NA LR- NA**	**LR+ 21,3 LR- 0,736**	**LR+ 14,8 LR- 0,753**

Die Bangdiwala Diagramme illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Die Receiver Operator Characteristic (ROC) trägt die Rate falsch Positiver (=1-Spezifität, fpr) gegen die Rate richtig Positiver (=Sensitivität, tpr) auf. Die Steigung der Tangenten durch den Ursprung bzw. (1,1) definieren die Likelihood ratios (LR+/LR-), die Fläche darunter die Area Under the Curve (AUC).

Tabelle 4.66: Kontingenztafel Befundkombinationen Pulsationen

Bhapkar <0,01			Referenz 1			Referenz 2			Summen	
			0	**1**	**10**	**11**	**1001**	**1010**		
0	**0**		614	20	20	8	26	43	**731**	731
1	**1**		0	4	0	0	3	0	**7**	7
	10		2	0	0	1	0	14	**17**	17
	100		1	0	0	0	0	0		1
	1000		2	6	0	6	6	24		44
2	**11**		0	1	0	1	0	7	**9**	9
	1001		0	5	0	0	4	0	**9**	9
	1010		0	0	0	1	0	6	**7**	7
3	1011		0	1	0	3	0	2		6
4	1111		1	3	0	0	1	3		8
x	ω		0	0	0	0	0	1		1
Summen			**616**	**30**	**20**	**11**	**33**	**70**	**780**	
			620	40	20	20	40	100		840

Befunder

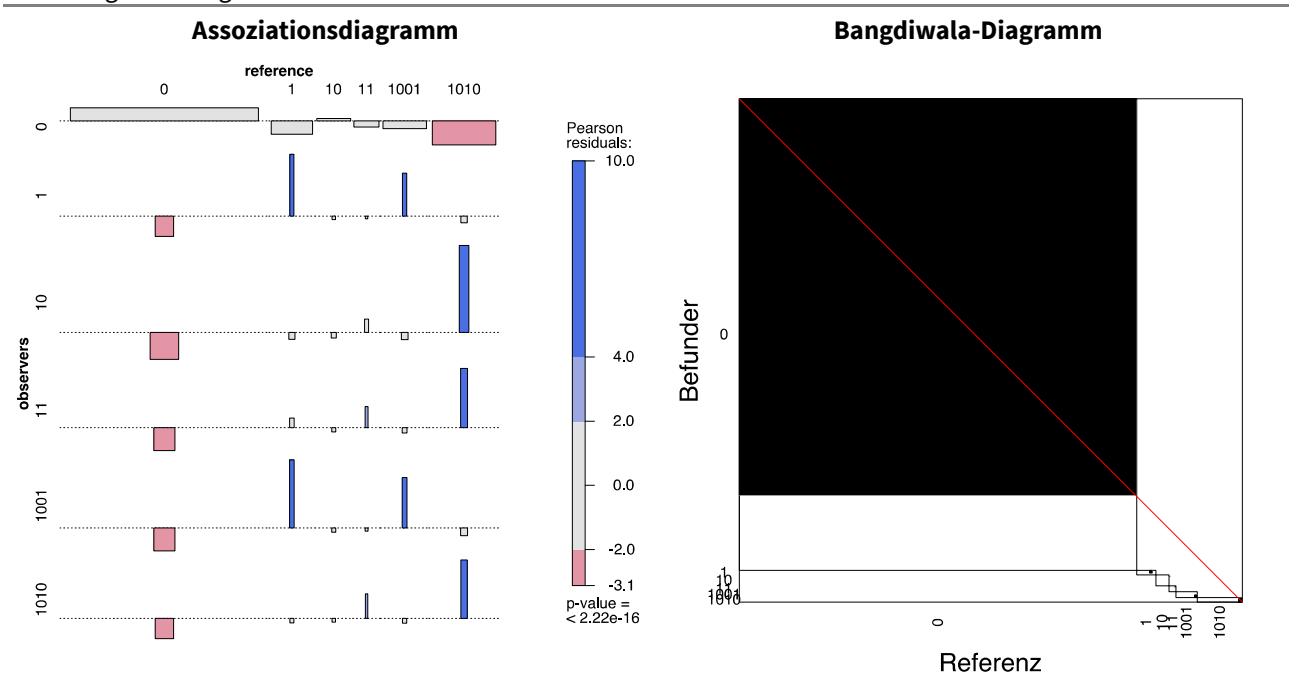
Nicht überlappende Kategorien sind grau, überlappende schwarz dargestellt.

Tabelle 4.67: Kennwerte Befundkombinationen Pulsationen

	Befund	pre	McN	nag	pag	spe	sen	pac	npv	ppv	B	LR-	LR+	AUC	OR	Q	Y
0	0	79	<0,01	44,1	91,2	28,7	99,7	84,7	95,9	84	0,827	0,011	1,4	0,642	123,3	0,984	0,835
1	1	3,8	<0,01	98,1	21,6	99,6	13,3	96,3	96,6	57,1	0,962	0,87	33,3	0,565	38,3	0,949	0,722
	10	2,6	0,742	97,6	0	97,8	0	95,3	97,4	0	0,951	1,02	0	0,489	n.a.	n.a	n.a
	11	1,4	0,814	98,8	10	99	9,09	97,7	98,7	11,1	0,977	0,919	8,74	0,54	9,5	0,81	0,51
2	1001	4,2	<0,01	97,8	19	99,3	12,1	95,6	96,2	44,4	0,955	0,885	18,1	0,557	20,5	0,907	0,638
	1010	9	<0,01	95,6	15,6	99,9	8,57	91,7	91,7	85,7	0,915	0,916	60,9	0,542	66,5	0,97	0,781

Prävalenz (pre), McNemar Test (McN), negative (nag) und positive (pag) Übereinstimmung, Spezifität (spe), Sensitivität (sen), Genauigkeit (acc), negativer (npv) und positiver prädiktiver Wert (ppv), Bangdiwala (B), negative (LR-) und positive Likelihood ratio (LR+), area under the curve (AUC), odds ratio (OR), Youden Q und Y.

Abbildung 4.38: Diagramme Befundkombinationen Pulsationen



Die Diagramme zeigen das Verhältnis der beobachteten Werte zu den Erwartungswerten.

4.2.4.3 Kompressionen

Tabelle 4.68: Inter-Beobachter-Variabilität Einzelbefunde Kompressionen

Befund	Befundverteilung				Präzision Übereinstimmung der Befunder untereinander			Richtigkeit Übereinstimmung mit Goldstandard als Referenz				
	Referenz	Befunder Anzahl verschiedener			Ø positive Übereinstimmung [%]	Kappa Fleiss		Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa Cohen		Datenabdeckung [%]
	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)		Kappa **nach Fleiss**	Klassifikation **modifiziert nach Landis**			Kappa **nach Cohen**	Klassifikation **modifiziert nach Landis**	
Stenosebereich	7 [140]	75	18	17	24,3	0,165	gering	21,4	40,0	0,184	gering	100
Larynx	NA	8	6	5	12,5	0,044	kaum	NA	NA	NA	NA	
Trachea	4 [80]	46	16	12	22,0	0,174	gering	26,2	45,7	0,284	schwach	
Bronchus	6 [120]	26	14	10	16,4	0,102	gering	17,5	80,8	0,249	schwach	

Die Tabelle betrachtet die Übereinstimmung der Einzelbefunde und ist in die Kriterien Befundverteilung, Präzision und Richtigkeit gegliedert. Details siehe Kommentar Tabelle 4.11 Seite 106.

Befundverteilung Einzelbefunde: Die meisten Kompressionen waren laut dem Goldstandard im Stenosebereich zu sehen, gefolgt von Stenosen im Bronchus und der Trachea. Im Larynx waren, gemäß Goldstandard, keine Kompressionen zu sehen. Die Befunder stimmen mit dem Goldstandard darin überein, dass die meisten Kompressionen im Stenosebereich zu sehen sind, befundeten dort aber nur etwas mehr als halb so viele Kompressionen wie der Goldstandard. Am zweithäufigsten sahen die Befunder Kompressionen in der Trachea, gefolgt von Bronchus und Larynx. Dabei beobachteten die Befunder in der Trachea knapp 60 % (46/80) der Kompressionen des Goldstandards, im Bronchus nur 22 % (26/120). Einige wenige Kompressionen wurden von den Befundern auch im Bereich des Larynx erhoben, ein Befund, der vom Goldstandard nicht gesehen wurde. Die Kompressions-Befunde im Stenosebereich wurden von fast allen Befundern erhoben (18/20), die übrigen Befunde werden von weniger Befundern getragen. Die Anzahl der Befunder nimmt mit der Befundhäufigkeit ab. Auch die Anzahl der Videos fällt mit der Befundhäufigkeit, wobei sich die Befunde bei den Befundern auf deutlich mehr Videos verteilen, als beim Goldstandard.

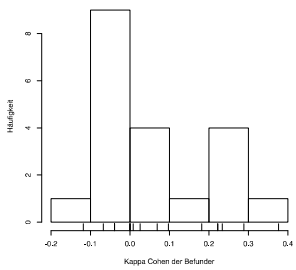
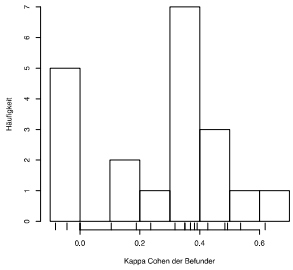
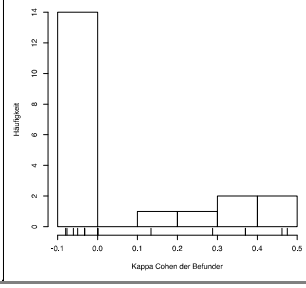
Präzision Einzelbefunde: Die beste Übereinstimmung für Kompressionen innerhalb der Befunder liegt in der Trachea mit einem Kappa Fleiss von 0,174 vor, dicht gefolgt von Kompressionen im Stenosebereich mit einem Kappa Fleiss von 0,165. Die Übereinstimmung im Bronchus fällt mit einem Kappa Fleiss von 0,102 erkennbar dagegen ab. Im Larynx liegt mit einem Kappa Fleiss von 0,044 kaum eine Übereinstimmung vor.

Richtigkeit Einzelbefunde: Bei einer Sensitivität von 26,2 % wird in der Trachea jede Vierte Kompression erkannt. Mit 45,7 % stimmten knapp die Hälfte der Kompressions-Befunde in der Trachea mit dem Goldstandard überein. Unauffällige Befunde waren in 92,6 % der Fälle korrekt, was nur knapp über dem Erwartungswert von 90,5 % liegt.

Im Stenosebereich wurden bei einer Sensitivität von 21,4 % etwas mehr als jede 5. Kompression erkannt. Unauffällige Befunde wurden in 93,6 % erkannt. Kompressionen im Stenosebereich stimmten in 40 % der Fälle mit der Diagnose des Goldstandards überein. Die Sensitivität für Kompressionen im Bronchus lag bei nur 17,5 %. Die Spezifität lag mit 99,3 % bei Kompressionen im Bronchus am höchsten. Wie bei der Präzision findet sich auch bei der Richtigkeit der beste

Kappa-Wert für die Trachea, die hier aber nicht vom Stenosebereich, sondern dem Bronchus gefolgt wird. Bei der odds ratio führt hingegen der Bronchus vor Trachea und Stenosebereich.

Tabelle 4.69: Paarweises Kappa Cohen Kompressionen

Stenosebereich				**Larynx**				**Trachea**				**Bronchus**			
vereintes Kappa Cohen				vereintes Kappa Cohen				vereintes Kappa Cohen				vereintes Kappa Cohen			
0,184				**na**				**0,284**				**0,249**			
paarweises Kappa Cohen				paarweises Kappa Cohen				paarweises Kappa Cohen				paarweises Kappa Cohen			
min.	Ø	med.	max.	min.	Ø	med.	max.	min.	Ø	med.	max.	min.	Ø	med.	max.
-0.111	**0,161**	0,432	0,189	na	**na**	na	na	-0,068	**0,246**	0,222	0,844	0	**0,225**	0,255	0,611
				na											

Die Tabelle vergleicht die Ergebnisse der beiden Berechnungsvarianten „vereintes“ und „paarweises“ Kappa Cohen.

Tabelle 4.70: Inter-Beobachter-Variabilität Befundkombinationen Kompressionen

Befund					**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz				
					Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
					Anzahl verschiedener											
Anzahl der Kompressionen	Stenosebereich	Larynx	Trachea	Bronchus	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	**Kappa** nach Fleiss	**Klassifikation** modifiziert nach Landis	Sensitivität [%]	positiver prädiktiver Wert [%]	**Kappa** nach Cohen	**Klassifikation** modifiziert nach Landis	Datenabdeckung [%]
x	**gesamt**				**9 [180]**	**124**	**20**	**19**	**13,4**	**0,193**	**gering**	**7,6**	**34,5**	**0,181**	**gering**	**88,7**
0				0	33 [660]	716	20	42	85,2	**0,307**	**mäßig**	99,2	85,1	**0,252**	**schwach**	94,0
1				1	1 [20]	11	7	6	11,2	**0,045**	**keine**	15,4	18,2	**0,153**	**gering**	75,9
			1	0	NA	26	11	11	17,5	**0,098**	**kaum**	NA	NA	**NA**	**NA**	NA
		1	0	0	NA	4	4	2	15,0	**0,075**	**kaum**	NA	NA	**NA**	**NA**	NA
	1	0	0	0	NA	57	17	16	20,8	**0,131**	**gering**	NA	NA	**NA**	**NA**	NA
2			1	1	1 [20]	7	4	4	12,5	**0,052**	**kaum**	11,1	14,3	**0,116**	**gering**	57,7
		1	1	0	NA	1	1	1	NA	**-0,001**	**keine**	NA	NA	**NA**	**NA**	NA
	1	0	0	1	4 [80]	4	4	3	10,0	**0,022**	**keine**	6,0	100	**0,104**	**gering**	83,8
	1	0	1	0	3 [60]	7	4	5	10,0	**0,022**	**keine**	7,1	42,9	**0,108**	**gering**	71,9
	1	1	0	0	NA	2	2	2	NA	**-0,002**	**keine**	NA	NA	**NA**	**NA**	NA
3	1	0	1	1	NA	4	4	3	10,0	**0,022**	**keine**	NA	NA	**NA**	**NA**	NA
	1	1	1	0	NA	1	1	1	NA	**-0,001**	**keine**	NA	NA	**NA**	**NA**	NA

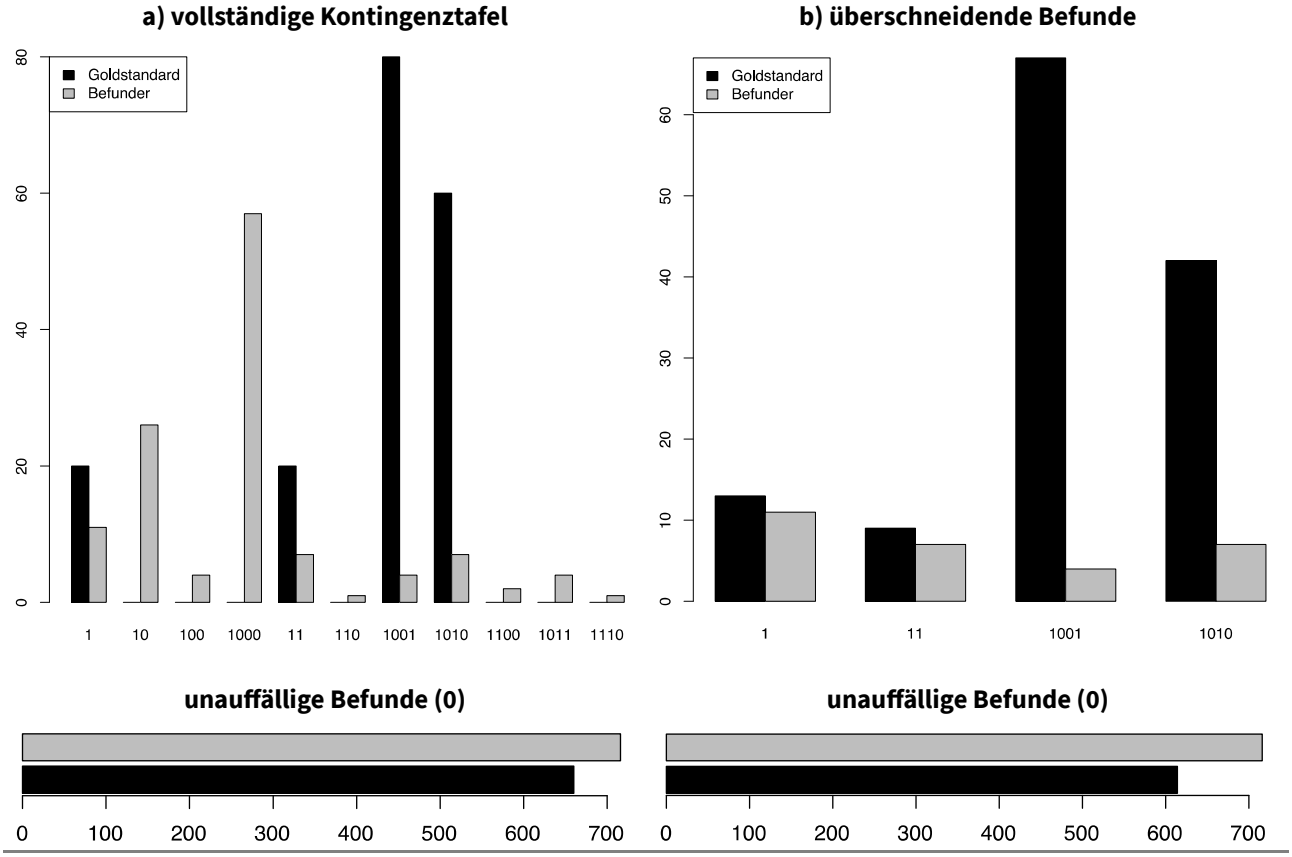
Die Tabelle analysiert die Befundkombinationen im Sinne eines Gesamtbefundes und ist in die Kriterien Befundverteilung, Präzision und Richtigkeit gegliedert. Details siehe Kommentar Tabelle 4.15 auf Seite 109.

Befundverteilung Befundkombinationen: Gemäß dem Goldstandard waren in insgesamt 9 Videos Kompressionen zu sehen: Eine isolierte bronchiale Kompression, eine kombinierte tracheo-

bronchiale Kompression, vier Kompressionen im Bronchus bei gleichzeitiger Kompression im Stenosebereich und drei Kompressionen in der Trachea, bei gleichzeitiger Kompression im Stenosebereich. Der überwiegende Teil der Kompressionen (7/9; 77,8 %) war also mit Kompressionen im Stenosebereich assoziiert. Bei den Befundern ist der Anteil von Kompressionen im Stenosebereich deutlich geringer (70/124; 56,5 %). Insgesamt werden mit 124 gegenüber 180 Befunden bei den Befundern weniger Kompressionen diagnostiziert als beim Goldstandard, wobei die Befundkombinationen weiter streuen und auch Kompressionen im Larynx mit einbeziehen, die vom Goldstandard nicht beobachtet wurden. Alle Befunder haben Kompressionen diagnostiziert und sahen in insgesamt knapp der Hälfte aller Videos (19/42) Kompressionen. Am häufigsten waren Kompressionen im Stenosebereich, gefolgt von isolierten trachealen (26) und isolierten bronchialen (11) Kompressionen. Anders als beim Goldstandard, der fast ausschließlich Kombinationsbefunde vorgibt (8/9; 88,9 %), überwiegen bei den Befundern Einzelbefunde (98/124; 79 %). Wie Abbildung 4.39 zeigt, differieren die Befunde zwischen Goldstandard und Befundern erheblich. Der Bhapkar-Test fällt signifikant aus. In keiner Befundklasse bestehen ähnliche Prävalenzen.

Bei der Beschneidung der Kontingenztafel auf die überschneidenden Befunde zur Bestimmung der Richtigkeit gehen 11,3 % der Befunde verloren. Kompressionen im Bronchus (1) und kombinierte bronchotracheale Kompressionen (11) werden durch die Beschneidung verzerrt: die Prävalenzen in diesen Befundklassen gleichen sich an.

Abbildung 4.39: Randverteilungen Befundkombinationen Kompressionen



Die Balkendiagramme stellen die Befundhäufigkeiten der Befunder denen des Goldstandards gegenüber.

Präzision Befundkombinationen: Eine erkennbare Übereinstimmung innerhalb der Befunder besteht nur für Kompressionen im Stenosebereich mit einem Kappa Fleiss von 0,131, fällt aber gering aus.

Richtigkeit Befundkombinationen: Die höchste Sensitivität wird für isolierte bronchiale Kompressionen erreicht. Der prädiktive Wert liegt bei nur 18,2 % – nicht einmal jede 5. Diagnose der

Befunder stimmte also mit dem Goldstandard überein. Dennoch wird hier mit einem Kappa von 0,153 der höchste Wert für Kompressionen erreicht. Ein Viertel der Befunde konnte in dieser Befundklasse bei der Berechnung der Richtigkeit wegen fehlender Überschneidungen von Befunden zwischen den Untersuchern und dem Goldstandard nicht mit einbezogen werden.

Kombinierte tracheobronchiale Kompressionen wurden in etwas mehr als jedem 10. Fall (11,1 %) erkannt, die Diagnose einer tracheobronchialen Kompression stimmte in 14,3 % der Fälle mit der des Goldstandards überein. Das Kappa Cohen ist mit 0,116 gering. Diese Daten sollten jedoch nur eingeschränkt beurteilt werden, da die Datenabdeckung für tracheobronchiale Kompressionen bei nur 57,7 % liegt. Stenosen im Bronchus bei gleichzeitiger Kompression im Stenosebereich wurden in nur 7,1 % der Fälle erkannt. Wenn sie erkannt wurden, war die Diagnose jedoch zu 100 % korrekt. Kappa Cohen entspricht mit 0,104 einer geringen Übereinstimmung. Die Datenabdeckung ist mit 83,3 % die höchste der Kompressionsbefunde. Falsch Negative Befunde machen mit 107/180 (59,4 %) den überwiegenden Anteil der fehlklassifizierten Befunde aus. Die zweitgrößte Quelle von Fehlklassifizierungen ist die fehlende Angabe der Lokalisation zusätzlich zur Kompression im Stenosebereich. Von 20 kombinierten tracheobronchialen Kompressionen (11) wurden 5 als Kompression im Stenosebereich (1000) eingestuft. Bei Kompressionen im Bronchus mit gleichzeitiger Kompression im Stenosebereich (1001) sind es 12 von 80 Befunden (15 %) und bei Larynxkompressionen mit gleichzeitiger Kompression im Stenosebereich 10 von 60 Befunden (16,7 %) die auf Kompressionen im Stenosebereich (1000) entfallen.

Insgesamt wurden 7,6 % der von Goldstandard vorgegebenen Befunde erkannt und ein Drittel der Diagnosen der Befunder (34,5 %) waren korrekt. Die Übereinstimmung der Befunder untereinander (Kappa Fleiss 0,193) wie auch die Übereinstimmung mit dem Goldstandard (Kappa Cohen 0,181) ist gering.

Zusammenfassung 4.13: Kompressionen

Bei Betrachtung der Einzelbefunde wird die Präzision von Kompressionen in der Trachea (Kappa Fleiss 0,174) dicht gefolgt von Kompressionen im Stenosebereich (Kappa Fleiss 0,165). Auch hinsichtlich der Richtigkeit führen Kompressionen in der Trachea (Kappa Cohen 0,284), hier allerdings gefolgt von Kompressionen im Bronchus (Kappa Cohen 0,249). Laryngeale Kompressionen wurden im Gegensatz zu den Befundern vom Goldstandard nicht gesehen. Auf der Ebene der Befundkombinationen sieht der Goldstandard überwiegen Kombinationsbefunde der Kompressionen, während bei den Untersuchern Einzelbefunde überwiegen. Die bei den Befundern häufigsten Befundklasse – Kompressionen im Stenosebereich – bei der zugleich mit 0,131 die höchste Präzision besteht, wurde vom Goldstandard nicht gesehen. Gemäß dem Goldstandard treten Kompressionen hauptsächlich in Assoziationen mit Kompressionen im Stenosebereich auf (7/9 Befunden). Eine Übereinstimmung mit dem Goldstandard ist nur bei singulären Kompressionen im Bronchus erkennbar (Kappa Cohen 0,153), wobei die Datenabdeckung nur bei 75,9 % liegt.

Tabelle 4.71: Vierfeldertafeln Einzelbefunde Kompressionen

Stenosebereich		Referenz			∑
		0	1	ω	
Befunder	0	655	110	0	765
	1	45	30	0	75
	ω	0	0	0	0
∑		700	140	0	840

Larynx		Referenz			∑
NA		0	1	ω	
Befunder	0	832	0	0	832
	1	8	0	0	8
	ω	0	0	0	0
∑		840	0	0	840

Trachea		Referenz			∑
		0	1	ω	
Befunder	0	735	59	0	794
	1	25	21	0	46
	ω	0	0	0	0
∑		760	80	0	840

Bronchus		Referenz			∑
		0	1	ω	
Befunder	0	715	99	0	814
	1	5	21	0	26
	ω	0	0	0	0
∑		720	120	0	840

Ein signifikanter Mc Nemar Test (McN) zeigt eine ungleiche Verteilung der Randsummen an.

Tabelle 4.72: Kennwerte Einzelbefunde Kompressionen

Stenosebereich			
Prävalenz	16,7		
	beo.	erw.	kor.
neg. Übst.	89,4	87,0	
pos. Übst.	27,9	11,6	
Spezifität	93,6	91,1	28,0
Sensitivität	21,4	8,9	13,7
Genauigkeit	81,5	77,4	18,4
Odds ratio	4,0		
Yule Q / Y	0,598 / 0,332		

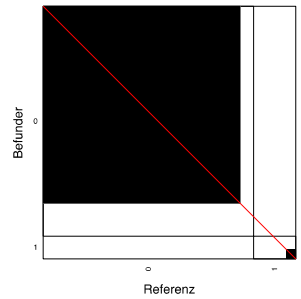
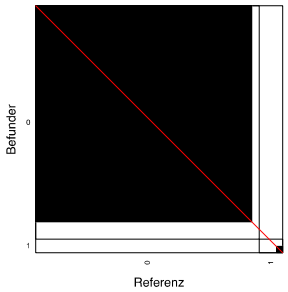
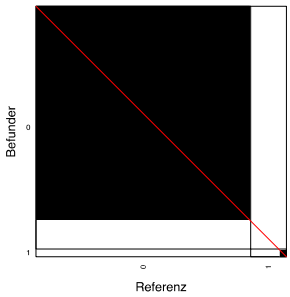
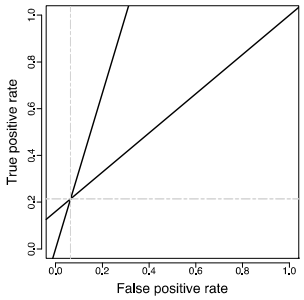
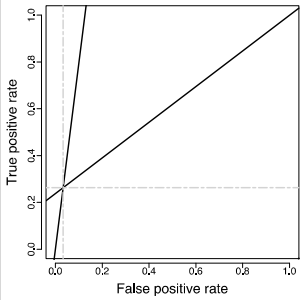
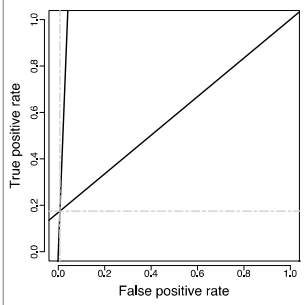
Larynx			
Prävalenz	0		
	beo.	erw.	kor.
neg. Übst.		NA	
pos. Übst.		NA	
Spezifität		NA	
Sensitivität		NA	
Genauigkeit		NA	
Odds ratio	NA		
Yule Q / Y	NA		

Trachea			
Prävalenz	9,5		
	beo.	erw.	kor.
neg. Übst.	94,6	92,5	
pos. Übst.	33,3	7,0	
Spezifität	96,7	94,5	39,9
Sensitivität	26,2	5,5	22,0
Genauigkeit	90,0	86,0	28,4
Odds ratio	10,5		
Yule Q / Y	0,826 / 0,528		

Bronchus			
Prävalenz	14,3		
	beo.	erw.	kor.
neg. Übst.	93,2		
pos. Übst.	28,8		
Spezifität	99,3	96,9	77,6
Sensitivität	17,5	3,1	14,9
Genauigkeit	87,6	83,5	24,9
Odds ratio	30,3		
Yule Q / Y	0,936 / 0,693		

Beobachtete Kennwerte (beo.), Erwartungswerte gemäß Randsummen (erw.) & für erwartete Übereinstimmung korrigierte Werte (kor.) in Prozent. Die odds ratio und ihre Transformationen (Yule Q & Y) sind prävalenzunabhängig.

Abbildung 4.40: Diagramme Einzelbefunde Kompressionen

Stenosebereich	Larynx	Trachea	Bronchus
Bangdiwala Konkordanz	Bangdiwala Konkordanz	Bangdiwala Konkordanz	Bangdiwala Konkordanz
	**NA**		
Bangdiwala 0,787 (0,745)	Bangdiwala NA	Bangdiwala 0,891 (0,850)	Bangdiwala 0,868 (0,826)
npv 85,6(83,3) ppv 40,0(16,7)	npv NA ppv NA	npv 92,6(90,5) ppv 45,7(9,5)	npv 87,8(85,7) ppv 80,8(14,3)
ROC & Likelihood Graph	ROC & Likelihood Graph	ROC & Likelihood Graph	ROC & Likelihood Graph
	**NA**		
AUC 0,575	AUC NA	AUC 0,615	AUC 0,584
LR+ 3,33 LR- 0,84	LR+ NA LR- NA	LR+ 7,98 LR- 0,763	LR+ 25,2 LR- 0,831

Die Bangdiwala Diagramme illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Die Receiver Operator Characteristic (ROC) trägt die Rate falsch Positiver (=1-Spezifität, fpr) gegen die Rate richtig Positiver (=Sensitivität, tpr) auf. Die Steigung der Tangenten durch den Ursprung bzw. (1,1) definieren die Likelihood ratios (LR+/LR-), die Fläche darunter die Area Under the Curve (AUC).

Tabelle 4.73: Kontingenztafel Befundkombinationen Kompressionen

Bhapkar <0,01		Referenz 0	1	2			Summen	
		0	**1**	**11**	**1001**	**1010**		
Befunder 0	**0**	609	11	6	54	36	716	716
Befunder 1	**1**	1	2	2	6	0	11	11
	10	14	0	4	0	8		26
	100	4	0	0	0	0		4
	1000	23	7	5	12	10		57
Befunder 2	**11**	0	0	1	3	3	7	7
	110	1	0	0	0	0		1
	1001	0	0	0	4	0	4	4
	1010	4	0	0	0	3	7	7
	1100	2	0	0	0	0		2
Befunder 3	1011	1	0	2	1	0		4
	1110	1	0	0	0	0		1
Summen		614	13	9	67	42	745	
		660	20	20	80	60		840

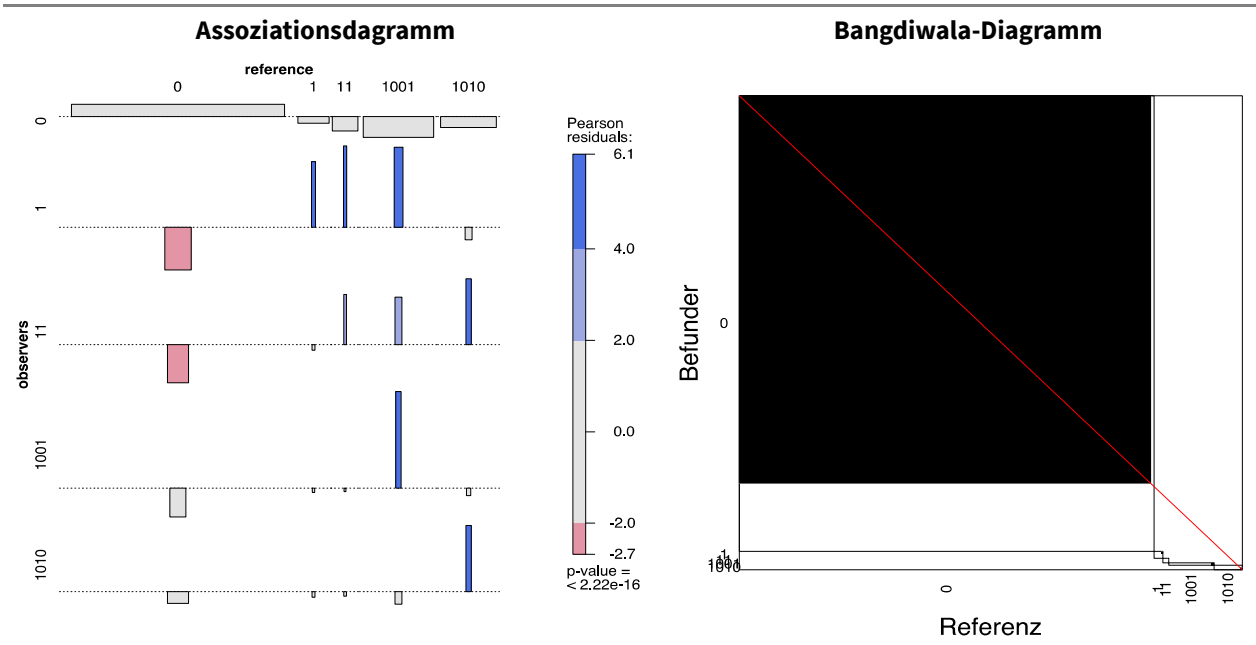
Nicht überlappende Kategorien sind grau, überlappende schwarz dargestellt.

Tabelle 4.74: Kennwerte Befundkombinationen Kompressionen

Befund		pre	McN	nag	pag	spe	sen	acc	npv	ppv	B	LR-	LR+	AUC	OR	Q	Y
0	0	82,4	<0,01	30	91,6	18,3	99,2	85	82,8	85,1	0,838	0,044	1,21	0,588	27,3	0,929	0,679
1	1	1,7	0,823	98,6	16,7	98,8	15,4	97,3	98,5	18,2	0,973	0,857	12,5	0,571	14,6	0,872	0,585
	11	1,2	0,789	99,1	12,5	99,2	11,1	98,1	98,9	14,3	0,981	0,896	13,6	0,551	15,2	0,877	0,592
2	1001	9,0	<0,01	95,6	11,3	100	5,97	91,5	91,5	100	0,915	0,94	Inf	0,53	Inf	NaN	NaN
	1010	5,6	<0,01	97	12,2	99,4	7,14	94,2	94,7	42,9	0,941	0,934	12,6	0,533	13,4	0,862	0,571

Prävalenz (pre), McNemar Test (McN), negative (nag) und positive (pag) Übereinstimmung, Spezifität (spe), Sensitivität (sen), Genauigkeit (acc), negativer (npv) und positiver prädiktiver Wert (ppv), Bangdiwala (B), negative (LR-) und positive Likelihood ratio (LR+), area under the curve (AUC), odds ratio (OR), Youden Q und Y.

Abbildung 4.41: Diagramme Befundkombinationen Kompressionen



Das Assoziationsdiagramm zeigt das Verhältnis der beobachteten Werte zu den Erwartungswerten

4.2.5 Schleimhaut

Die Schleimhaut wurde nach den drei Kriterien Schwellung, Hyperämie und Hypersekretion beurteilt. Entzündung wird in einem eigenen Abschnitt untersucht.

4.2.5.1 Schwellung

Tabelle 4.75: Einzelbefunde Schleimhautschwellung

Befund		**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz				
		Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
		Anzahl verschiedener											
Kodierung		Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung	Kappa **nach Fleiss**	Klassifikation **modifiziert nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	**Kappa nach Cohen**	Klassifikation **modifiziert nach Landis**	Datenabdeckung [%]
x	**gesamt**	**42 [840]**	**840**	**20**	**42**	**NA**	**0,153**	**kaum**	**63,9**	**69,5**	**0,283**	**schwach**	**72,3**
						NA	**0,231**	**schwach**	**49,1**	**69,5**	**0,250**	**schwach**	**100**
0	nein	20 [400]	296	20	41	37,7	0,107	kaum	64,7	58,8	0,283	schwach	74,9
	NA als nein		529	20	42	65,9	0,231	schwach	76,2	57,7	0,250	schwach	100
1	ja	22 [440]	311	20	39	41,8	0,231	schwach	63,9	69,5	0,283	schwach	80,9
	ja bei NA als nein								49,1	69,5	0,250	schwach	100
NA	keine Antwort	NA	233	20	41	31,7	0,114	kaum	NA	NA	NA	NA	NA

Zur besseren Vergleichbarkeit zwischen Referenz und Befundern ist in eckigen Klammern die gewichtete Anzahl der Befunde des Goldstandards angegeben (Faktor 20 im Vgl. zu den Befundern).

Befundverteilung Schleimhautschwellung: In 72,3 % (607/840) der Videos wurde ein Befund zur Schleimhautschwellung abgegeben. Eine Schleimhautschwellung lag gemäß Goldstandard in 52,4 % der Videos vor, was sich mit der Einschätzung der Untersucher von 51,2 % (311/607) deckt. An der Beurteilung der Schleimhaut beteiligten sich alle Befunder. Die Befunde der Befunder verteilen sich auf fast doppelt so viele Videos (39) wie beim Goldstandard (22). Fehlwerte kamen bei allen Befundern und in allen Videos vor.

Präzision Schleimhautschwellung: Innerhalb der Befunder besteht nur eine geringe Übereinstimmung hinsichtlich Schleimhautschwellung. Die durchschnittliche positive Übereinstimmung beträgt 41,8 %. Das Kappa Fleiss positiver Befunde steht mit 0,231 für eine schwache Übereinstimmung, unter Einbeziehung der negativen Befunde sowie der Fehlwerte ergibt sich ein Kappa Fleiss von 0,153.

Richtigkeit Schleimhautschwellung: Sensitivität und Spezifität liegen mit jeweils ca. 64 % gleich auf. Es wurden also jeweils ⅔ der negativen und positiven Befunde erkannt. Der Erwartungswert für Spezifität und Sensitivität beträgt hier ca. 50 %. Der positive prädiktive Wert liegt für positive Befunde mit 69,5 % höher als für unauffällige Befunde (58,8 %). Kappa Cohen erreicht einen Wert von 0,283. Unter Berücksichtigung der zufällig zu erwartenden Übereinstimmung wurde etwas weniger als jeder Dritte Befund richtig erhoben (28,3 %).

Die Receiver Operator Charakteristik (ROC) illustriert die in etwa gleiche Vorhersagekraft von negativen und positiven Befunden: Der Betrag der Steigungen der Geraden relativ zur Diagonalen ist in etwa gleich. Die Likelihood ratio Graphen spannen unter sich eine Area Under the Curve (AUC) von 0,643 auf. Die odds ratio beträgt 3,2.

Betrachtet man Fehlwerte als unauffällige Befunde, steigt die Spezifität auf 76,3 % während die Sensitivität auf 49,1 % abfällt. Negative und positiver prädiktiver Wert bleiben hingegen so gut wie unverändert. Die Vorhersagekraft des Testes verschiebt sich leicht zugunsten positiver Befunde (LR+ steigt von 1,81 auf 2,07), wobei die Prognose des Testes insgesamt - .gemessen an der AUC – etwas abnimmt. Dem entsprechend sinkt auch das Kappa nach Cohen auf 0,250, womit - nach Korrektur für zufällige Übereinstimmung - also nur noch jeder Vierte Befund richtig erhoben wird.

Tabelle 4.76: Vier-Felder-Tafeln Schleimhautschwellung

mit Fehlwerten (NA)						Fehlwerte (NA) als „nein“ (0)					
McN p=0,078		Referenz 0	Referenz 1	Referenz NA	Summe	McN p<0,01		Referenz 0	Referenz 1	Referenz NA	Summe
Befunder	0	174	122	0	296	Befunder	0	305	224	0	529
	1	95	216	0	311		1	95	216	0	311
	NA	131	102	0	233		NA	als 0	als 0	0	NA
Summe		400	440	0	840	Summe		400	440	0	840

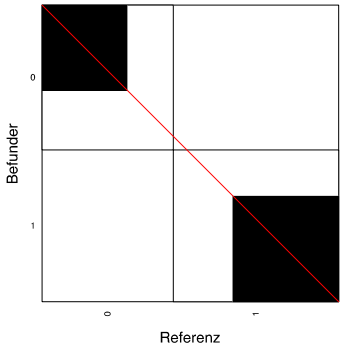
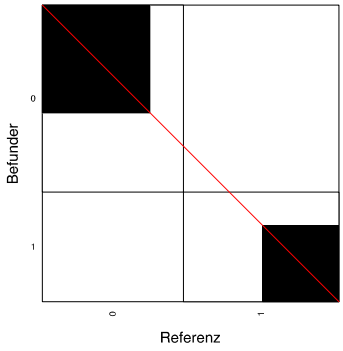
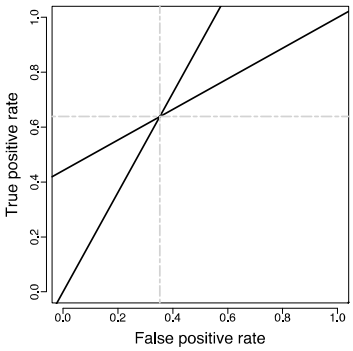
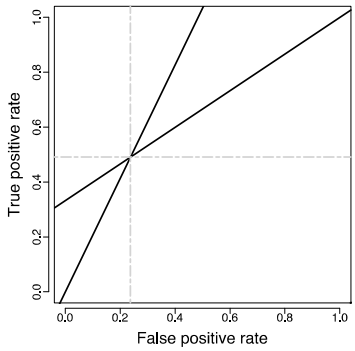
Die Vier-Felder-Tafeln und ihre Randsummen sind Berechnungsgrundlage sämtlicher Kennwerte und Diagramme.

Tabelle 4.77: Kennwerte Schleimhautschwellung

mit Fehlwerten (NA)				Fehlwerte als „nein“ (0)			
Prävalenz		55,7		Prävalenz		52,4	
	beo.	erw.	kor.		beo.	erw.	kor.
negative Übereinstimmung	61,6	46,4	=κ	negative Übereinstimmung	65,7	54,2	=κ
positive Übereinstimmung	66,6	53,3	=κ	positive Übereinstimmung	57,5	43,4	=κ
Spezifität	64,7	48,8	31,1	Spezifität	76,3	63,0	35,9
Sensitivität	63,9	51,2	26,0	Sensitivität	49,1	37,0	19,2
Genauigkeitauigkeit	64,3	50,1	28,3=κ	Genauigkeitauigkeit	62,0	49,4	25,0=κ
Odds ratio		3,2		Odds ratio		3,1	
Yule Q / Y		0,529 / 0,286		Yule Q / Y		0,512 / 0,275	

Beobachtete Kennwerte (beo.), Erwartungswerte gemäß Randsummen (erw.) & für erwartete Übereinstimmung korrigierte Werte (kor.) in Prozent. Die odds ratio und ihre Transformationen (Yule Q & Y) sind prävalenzunabhängig.

Abbildung 4.42: Diagramme Schleimhautschwellung

mit Fehlwerten (NA)	Fehlwerte als „nein“ (0)
Konkordanz-Diagramm	Konkordanz-Diagramm
	
Bangdiwala 0,416 (0,255)	Bangdiwala 0,401 (0,258)
npv 58,8 (44,3) ppv 69,5 (55,7)	npv 57,7 (47,6) ppv 69,5 (52,4)
ROC und Likelihood Graph	ROC und Likelihood Graph
	
area under curve (AUC) 0,643	area under curve (AUC) 0,627
LR+ 1,81 LR- 0,558	LR+ 2,07 LR- 0,668

Die Bangdiwala Diagramme illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Die Receiver Operator Characteristic (ROC) trägt die Rate falsch Positiver (=1-Spezifität, fpr) gegen die Rate richtig Positiver (=Sensitivität, tpr) auf. Die Steigung der Tangenten durch den Ursprung bzw. (1,1) definieren die Likelihood ratios (LR+/LR-), die Fläche darunter die Area Under the Curve (AUC).

4.2.5.2 Hyperämie

Tabelle 4.78: Einzelbefunde Hyperämie

Befund		Befundverteilung				Präzision Übereinstimmung der Befunder untereinander			Richtigkeit Übereinstimmung mit Goldstandard als Referenz				
		Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
		Anzahl verschiedener											
Kodierung		Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	Klassifikation **modifiziert nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	**Kappa nach Cohen**	Klassifikation **modifiziert nach Landis**	Datenabdeckung [%]
x	**gesamt**	**42 [840]**	**840**	**20**	**42**	**NA**	**0,175**	**schwach**	**62,8**	**75,1**	**0,269**	**mittelmäßig**	**72,4**
						NA	**0,271**	**mittelmäßig**	**49,6**	**75,1**	**0,262**	**mittelmäßig**	**100**
0	nein	18	291	20	40	39,9	0,120	schwach	65,5	51,5	0,269	mittelmäßig	73,9
	NA als nein		523	20	40	65,4	0,271	mittelmäßig	78,1	53,7	0,262	mittelmäßig	100
1	ja	24 [480]	317	20	39	44,7	0,271	mittelmäßig	62,8	75,1	0,269	mittlelmäßig	81,9
	Ja bei NA als nein								49,6	75,1	0,262	mittelmäßig	100
NA	keine Antwort	NA	232	20	39	31,1	0,124	schwach	NA	NA	NA	NA	NA

Zur besseren Vergleichbarkeit zwischen Referenz und Befundern ist in eckigen Klammern die gewichtete Anzahl der Befunde des Goldstandards angegeben (Faktor 20 im Vgl. zu den Befundern).

Befundverteilung Hyperämie: Mit 72,4 % (608/840) entspricht der Anteil der bei Hyperämie abgegebenen Befunde dem der Schleimhautschwellung. Die durch den Goldstandard vorgegebene Prävalenz ist mit 57,1 % (480/840) etwas höher. Die Befunder sehen die Prävalenz mit 52,1 % (317/608) etwas geringer. Alle Befunder haben Hyperämien gesehen. Dabei verteilen sich die Befunde auf fast alle Videos (39/42). Fehlwerte kommen ebenfalls bei allen Befundern und in nahezu allen Videos (39/42) vor.

Präzision Hyperämie: Mit einer durchschnittlichen positiven Übereinstimmung von 44,7 % und einem Kappa Fleiss von 0,175 ist der Befund Hyperämie kaum präziser als der Schleimhautschwellung. Auch hier wird das etwas höhere Kappa Fleiss für positive Befunde von 0,271 durch das deutlich geringere Kappa Fleiss für negative Befunde (0,120) und Fehlwerte (0,124) abgeschwächt.

Richtigkeit Hyperämie: Spezifität und Sensitivität halten sich mit 65,5 % bzw. 62,8 % die Waage. Der positive prädiktive Wert liegt mit 75,1 % deutlich über dem negativen prädiktiven Wert von 51,5 %. Kappa erreicht mit 0,269 eine mittelmäßige Übereinstimmung. Die odds ratio ist 3,2. Bei Einbeziehung von Fehlwerten als negative Befunde verschiebt sich die Prognosekraft zu Gunsten positiver Befunde.

Tabelle 4.79: Vier-Felder-Tafel Hyperämie

mit Fehlwerten (NULL)						Fehlwerte (NULL) als „nein“ (0)					
McN <0,01		Referenz 0	Referenz 1	Referenz NULL	Summe	McN <0,01		Referenz 0	Referenz 1	Referenz NULL	Summe
Befunder	0	150	141	0	291	Befunder	0	281	242	0	523
	1	79	238	0	317		1	79	238	0	317
	NULL	131	101	0	232		NULL	als 0	als 0	0	NA
Summe		360	480	0	840	Summe		360	480	0	840

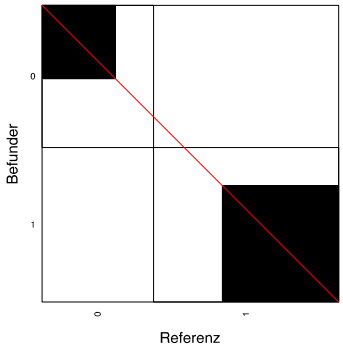
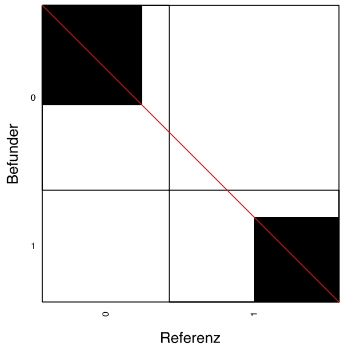
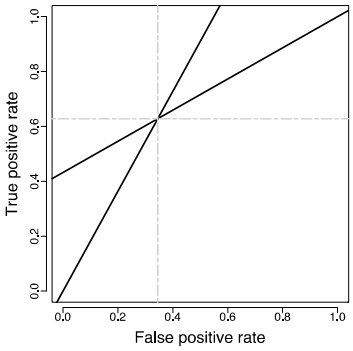
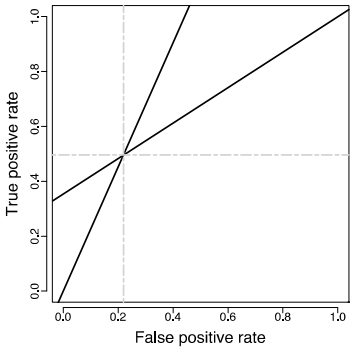
Die Vier-Felder-Tafeln und ihre Randsummen sind Berechnungsgrundlage sämtlicher Kennwerte und Diagramme.

Tabelle 4.80: Kennwerte Hyperämie

mit Fehlwerten (NA)				Fehlwerte als „nein“ (0)			
Prävalenz		62,3		Prävalenz		57,1	
	beo.	erw.	kor.		beo.	erw.	kor.
negative Übereinstimmung	57,7	42,2	=κ	negative Übereinstimmung	63,6	50,8	=κ
positive Übereinstimmung	68,4	56,8	=κ	positive Übereinstimmung	59,7	45,5	=κ
Spezifität	65,5	47,9	33,8	Spezifität	78,1	62,3	41,9
Sensitivität	62,8	52,1	22,3	Sensitivität	49,6	37,7	19,0
Genauigkeitauigkeit (≙	63,8	50,5	26,9=κ	Genauigkeitauigkeit	61,8	48,2	26,2=κ
Odds ratio		3,2		Odds ratio		3,5	
Yule Q / Y		0,524 / 0,283		Yule Q / Y		0,555 / 0,303	

Beobachtete Kennwerte (beo.), Erwartungswerte gemäß Randsummen (erw.) & für erwartete Übereinstimmung korrigierte Werte (kor.) in Prozent. Die odds ratio und ihre Transformationen (Yule Q & Y) sind prävalenzunabhängig.

Abbildung 4.43: Diagramme Hyperämie

mit Fehlwerten (NULL)	Fehlwerte als „nein“ (0)
Bangdiwala Konkordanz	Bangdiwala Konkordanz
	
Bangdiwala 0,424 (0,273)	Bangdiwala 0,398 (0,244)
npv 51,5 (37,7) ppv 75,1 (62,3)	npv 53,7 (42,9) ppv 75,1 (57,1)
ROC und Likelihood Graph	ROC und Likelihood Graph
	
area under curve (AUC) 0,641	area under curve (AUC) 0,638
LR+ 1,82 LR- 0,568	LR+ 2,26 LR- 0,646

Die Bangdiwala Diagramme illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Die Receiver Operator Characteristic (ROC) trägt die Rate falsch Positiver (=1-Spezifität, fpr) gegen die Rate richtig Positiver (=Sensitivität, tpr) auf. Die Steigung der Tangenten durch den Ursprung bzw. (1,1) definieren die Likelihood ratios (LR+/LR-), die Fläche darunter der Area Under the Curve (AUC).

4.2.5.3 Hypersekretion

Tabelle 4.81: Einzelbefunde Hypersekretion

Befund		Befundverteilung				Präzision Übereinstimmung der Befunder untereinander			Richtigkeit Übereinstimmung mit Goldstandard als Referenz				
		Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
		Anzahl verschiedener											
Kodierung		Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	Klassifikation **modifiziert nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	**Kappa nach Cohen**	Klassifikation **modifiziert nach Landis**	Datenabdeckung [%]
x	**gesamt**	**42 [840]**	**840**	**20**	**42**	**NA**	**0,211**	**mittelmäßig**	**67,8**	**81,4**	**0,384**	**mittelmäßig**	**75,0**
						NA	**0,333**	**mittelmäßig**	**56,5**	**81,4**	**0,373**	**mittelmäßig**	**100**
0	nein	18 [360]	297	20	40	38,8	0,114	schwach	73,0	56,6	0,384	mittelmäßig	73,4
	NA als nein		507	20	40	64,9	0,333	mittelmäßig	82,8	58,8	0,373	mittelmäßig	100
1	ja	24 [280]	333	20	38	48,4	0,333	mittelmäßig	67,8	81,4	0,384	mittelmäßig	85,2
	Ja bei NA als nein								56,5	81,4	0,373	mittelmäßig	100
NA	keine Antwort	NA	210	20	36	31,4	0,175	schwach	NA	NA	NA	NA	NA

Zur besseren Vergleichbarkeit zwischen Referenz und Befundern ist in eckigen Klammern die gewichtete Anzahl der Befunde des Goldstandards angegeben (Faktor 20 im Vgl. zu den Befundern).

Befundverteilung: Auch bei der Hypersekretion wurden, wie bei Schleimhautschwellung und Hyperämie, etwa drei Viertel der Befunde aktiv erhoben (75 %, 630/840). Die Prävalenz der Hypersekretion liegt bei einem Drittel (280/840), die Befunder sehen in gut der Hälfte der beurteilten Befunde 52,9 % (333/630) eine Hypersekretion. Die Hypersekretion gehört zum Befundspektrum aller Befunder und wurde in 90 % der Videomitschnitte gesehen. Fehlwerte kommen bei allen Befundern in insgesamt 85,7 % der Videomitschnitte vor.

Präzision Hypersekretion: Die Präzision der Hypersekretion liegt mit einem Kappa Fleiss von 0,211 über der von Schleimhautschwellung und Hyperämie. Bei einem Kappa Fleiss von 0,333 für den positiven Befund wird innerhalb der Schleimhautbeurteilung die beste Übereinstimmung innerhalb der Befunde erreicht. Kappa Fleiss insgesamt wird durch die niedrigeren Werte bei negativen Befunden und Fehlwerten auf 0,211 reduziert.

Richtigkeit Hypersekretion: Gut zwei Drittel der Befunde wurden richtig erkannt, wobei sich Spezifität und Sensitivität mit 67,8 % bzw. 73,0 % in der gleichen Größenordnung bewegen. Der positive prädiktive Wert liegt mit 81,4 % erheblich über dem negativen prädiktiven Wert von 56,6 %. Mit einem Kappa Cohen von 0,384 und einer Odds ratio von 5,7 ist die Befundverlässlichkeit von Hypersekretion der Hyperämie und Schwellung klar überlegen.

Tabelle 4.82: Kontingenztafeln Hypersekretion

mit Fehlwerten (NULL)						Fehlwerte (NULL) als „nein“ (0)					
McNemar p<0,01		Referenz 0	Referenz 1	Referenz NULL	Summe	McNemar p<0,01		Referenz 0	Referenz 1	Referenz NULL	Summe
Befunder	0	168	129	0	297	Befunder	0	298	209	0	507
	1	62	271	0	333		1	62	271	0	333
	NULL	130	80	0	210		NULL	als 0	als 0	0	NA
Summe		360	480	0	840	Summe		360	480	0	840

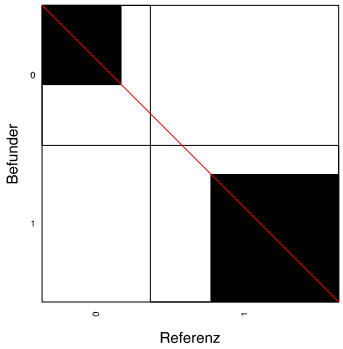
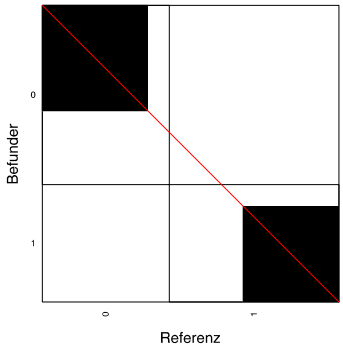
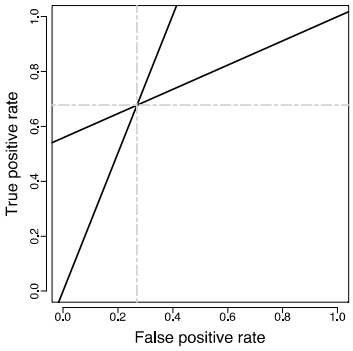
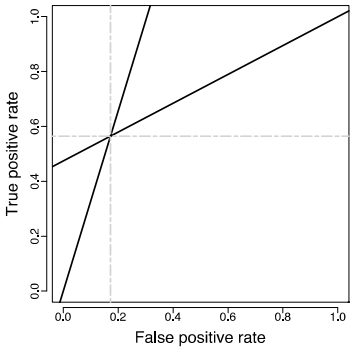
Die Vier-Felder-Tafeln und ihre Randsummen sind Berechnungsgrundlage sämtlicher Kennwerte und Diagramme.

Tabelle 4.83: Kennwerte Hypersekretion

mit Fehlwerten (NA)				Fehlwerte als „nein“ (0)			
Prävalenz		63,5		Prävalenz		57,1	
	beo.	erw.	kor.		beo.	erw.	kor.
negative Übereinstimmung	63,8	41,1	=κ	negative Übereinstimmung	68,7	50,1	=κ
positive Übereinstimmung	73,9	57,7	=κ	positive Übereinstimmung	66,7	46,8	=κ
Spezifität	73,0	47,1	49,0	Spezifität	82,8	60,4	56,6
Sensitivität	67,8	52,9	31,6	Sensitivität	56,5	39,6	27,9
Genauigkeit	69,7	50,8	38,4=κ	Genauigkeit	67,7	48,5	37,3=κ
Odds ratio		5,7		Odds ratio		6,2	
Yule Q / Y		0,701 / 0,409		Yule Q / Y		0,723 / 0,428	

Beobachtete Kennwerte (beo.), Erwartungswerte gemäß Randsummen (erw.) & für erwartete Übereinstimmung korrigierte Werte (kor.) in Prozent. Die odds ratio und ihre Transformationen (Yule Q & Y) sind prävalenzunabhängig.

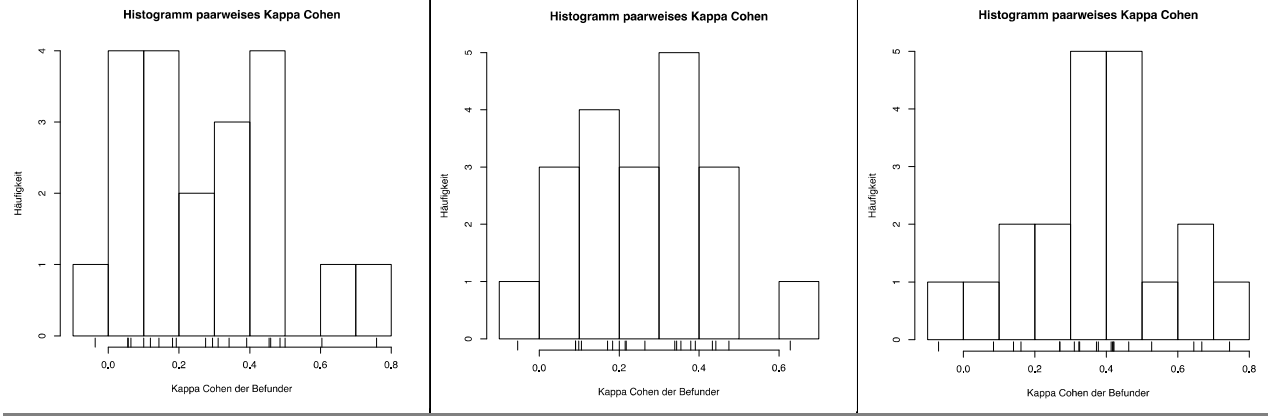
Abbildung 4.44: Diagramme Hypersekretion

mit Fehlwerten (NA)	Fehlwerte als „nein“ (0)
Bangdiwala	Bangdiwala
	
Bangdiwala 0,505 (0,280)	Bangdiwala (B) 0,474
npv 56,6 (36,5) ppv 81,4 (63,5)	npv 58,8 (42,9) ppv 81,4 (57,1)
ROC und Likelihood Graph	ROC und Likelihood Graph
	
area under curve (AUC) 0,704	area under curve (AUC) 0,696
LR+ 2,51 LR- 0,442	LR+ 3,28 LR- 0,526

Die Bangdiwala Diagramme illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Die Receiver Operator Characteristic (ROC) trägt die Rate falsch Positiver (=1-Spezifität, fpr) gegen die Rate richtig Positiver (=Sensitivität, tpr) auf. Die Steigung der Tangenten durch den Ursprung bzw. (1,1) definieren die Likelihood ratios (LR+/LR-), die Fläche darunter die Area Under the Curve (AUC).

Tabelle 4.84: Paarweises Kappa Cohen Schleimhaut

Schwellung				Hyperämie				Hypersekretion			
vereintes Kappa Cohen				vereintes Kappa Cohen				vereintes Kappa Cohen			
0,283				**0,269**				**0,384**			
paarweises Kappa Cohen				paarweises Kappa Cohen				paarweises Kappa Cohen			
min.	Ø	median	max.	min.	Ø	median	max.	min.	Ø	median	max.
-0,037	**0,288**	0,285	0,759	-0,054	**0,268**	0,241	0,629	-0.069	**0,364**	0,374	0,744



Die Tabelle vergleicht die Ergebnisse der beiden Berechnungsvarianten „vereintes“ und „paarweises“ Kappa Cohen.

4.2.5.4 Gesamtbefund Schleimhaut

Die Befunde bzw. Symptome Schwellung, Hyperämie und Hypersekretion können zu einem gemeinsamen Schleimhautbefund („Syndrom“) zusammengefasst werden. Die Zusammenstellung dieser Befundkombination ist insbesondere hinsichtlich des Befundes „Entzündung“ interessant. Denn Schwellung, Hyperämie und Hypersekretion sind die entscheidenden morphologischen Merkmale einer Entzündung und können mit dem sogenannten Bronchitis Index (BI) semiquantitativ erfasst werden (Thompson u. a., 1993).

Tabelle 4.85: Inter-Beobachter-Variabilität Befundkombinationen Schleimhaut

Befund				**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit dem Goldstandard in den überschneidenden Befunden				
				Referenz	Befunder				**Kappa Fleiss**				**Kappa Cohen**		
				Anzahl verschiedener											
Anzahl der Befunde	Schleimhautschwellung	Hyperämie	Hypersekretion	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	**Kappa nach Cohen**	modifizierte Klassifikation **nach Landis**	Datenabdeckung [%]
x	**gesamt**			**29**				**20,9**	**0,196**	**gering**	**25,9**	**34,8**	**0,217**	**schwach**	**96,8**
0			0	13	394	20	40	50,4	0,326	mäßig	78	49,5	0,368	mäßig	97,8
			1	2	70	19	30	19,3	0,080	kaum	23,1	12,9	0,110	gering	99,0
1		1	0	2	31	15	21	13,3	0,022	keine	0	0	-0,044	keine	97,2
	1	0	0	na	27	14	19	10,7	0,003	keine	na	na	na	na	na
		1	1	3	34	15	20	13,8	0,032	keine	0	0	-0,055	keine	96,8
2	1	0	1	3	32	16	16	15,0	0,063	kaum	6,9	12,5	0,04	keine	97,7
	1	1	0	3	55	16	27	12,4	0,010	keine	13,8	14,5	0,078	kaum	98,1
3	1	1	1	16	197	20	36	32,8	0,259	schwach	39.9	63,5	0,274	schwach	98,2

Die Tabelle analysiert die Befundkombinationen. Sie gliedert sich in die 3 Hauptbereiche Befundverteilung Präzision und Richtigkeit. Details siehe Kommentar Tabelle 4.15 auf Seite 109.

Nennenswerte Übereinstimmung zwischen den Befundern untereinander gibt es nur bei unauffälligen Schleimhautbefunden mit einem „mäßigen“ Kappa Fleiss von 0,326 und beim Vollbild aus Schwellung, Hyperämie und Hypersekretion mit einem „schwachen“ Kappa Fleiss von 0,259. Bei Einzelbefunden liegt keine größere Übereinstimmung als bei Kombinationsbefunden vor – mit Ausnahme des bereits erwähnten Vollbildes aus allen 3 Befunden. Schwellungen ohne Hyperämie oder Hypersekretion wurden vom Goldstandard nicht beobachtet. Signifikante Übereinstimmungen zwischen Befundern und dem Goldstandard finden sich ebenfalls nur bei unauffälligen Schleimhautbefunden (Kappa Cohen 0,368, „mäßig“) und beim Vollbild (Kappa Cohen 0,274, „schwach“). Anders als bei den Befundern untereinander heben sich singuläre Hypersekretionen mit einem Kappa Cohen von 0,110 erkennbar von den übrigen 1fach- und 2fach-Befunden ab.

Zusammenfassung 4.14: Schleimhaut

Bei Betrachtung der Einzelbefunde führt die Hypersekretion sowohl hinsichtlich der Präzision mit einem Kappa Fleiss von 0,211, als auch der Richtigkeit mit einem Kappa Cohen von 0,384. Der positive prädiktive Wert liegt bei 81,4 %. Die Konkordanz innerhalb der Befunder hinsichtlich Schwellung und Hyperämie ist mit einem Kappa Fleiss von 0,153 respektive 0,175 ähnlich. Der positive prädiktive Wert der Hyperämie liegt mir 75,1 % über dem der Schwellung mit 69,5 %. Das Kappa nach Cohen verhält sich mit Kappa Cohen von 0,283 für die Schwellung und 0,269 für Hyperämie gegenläufig. Die Einbeziehung von Fehlwerten als unauffällige Befunde ergibt kein wesentlich anderes Bild der Befundrichtigkeit. Auf Ebene der Kombinationsbefunde erkennt man weder bei singulärer Hypersekretion, Hyperämie oder Schleimhautschwellung noch bei Zweierkombinationen eine Übereinstimmung — weder bei den Untersuchern untereinander, noch im Vergleich zum Goldstandard. Nur beim Vollbefund aus Schwellung, Hyperämie und Hypersekretion besteht mit einem Kappa Fleiss von 0,259 eine erkennbare Präzision und mit einem Kappa Cohen von 0,274 auch eine schwache Übereinstimmung zum Goldstandard. Der prädiktive Wert des Vollbefundes liegt bei 63,5 %.

4.2.6 Entzündung

Neben dem Vorhandensein von Entzündung wurde die Lage der Entzündung beurteilt. Im Rahmen der Auswertung wurde untersucht, in wieweit der pauschale Befund Entzündung mit dem Syndrom der Schleimhautbefunde Schwellung, Hyperämie und Sekretion korreliert.

4.2.6.1 Entzündung als pauschaler Befund

Tabelle 4.86: Einzelbefunde Entzündung

Befund		**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz				
		Referenz	Befunder			Ø positive Übereinstimmung [%]	Kappa Fleiss		Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa Cohen		Datenabdeckung [%]
		Anzahl verschiedener											
Kodierung		Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)		Kappa **nach Fleiss**	Klassifikation **modifiziert nach Landis**			**Kappa nach Cohen**	Klassifikation **modifiziert nach Landis**	
x	**gesamt**	**24 [480]**	**289**	**20**	**39**	**39,7**	**0,155**	**gering**	**60,7**	**74,4**	**0,268**	**schwach**	**69,2**
						NA	**0,262**	**schwach**	**44,8**	**74,4**	**0,277**	**schwach**	**100**
0	nein	18 [360]	292	20	40	38,2	0,093	kaum	67,4	52,4	0,268	schwach	73,3
	NA als nein		551	20	41	68,8	0,262	schwach	79,4	51,9	0,227	schwach	100
1	ja	24 [480]	289	20	39	39,7	0,262	schwach	60,7	74,4	0,268	schwach	77,3
	ja bei NA als nein								44,8	74,4	0,277	schwach	100
NA	keine Antwort	NA	259	20	41	32,2	0,108	gering	NA	NA	NA	NA	NA

Zur besseren Vergleichbarkeit zwischen Referenz und Befundern ist in eckigen Klammern die gewichtete Anzahl der Befunde des Goldstandards angegeben (Faktor 20 im Vgl. zu den Befundern).

Befundverteilung: Gemäß dem Goldstandard ist die Prävalenz von Entzündung 57,1 % (24/42). Die Befunder sahen in der Hälfte der aktiv beurteilten Videos eine Entzündung (49,7 %, 289/581), was absolut gesehen 34,4 % (289/840) entspricht. Entzündungen wurden von allen Befundern in 39 von 42 Videos befundet.

Präzision: Das Kappa Fleiss steht mit 0,155 für eine insgesamt geringe Übereinstimmung. Bei positiven Befunden ist die Übereinstimmung mit 0,262 immerhin schwach.

Richtigkeit: Die Befundrichtigkeit ist mit einem Kappa Cohen von 0,268 schwach. Bei der Analyse der Richtigkeit gehen gut 30 % der Daten verloren. Auch wenn man fehlende Angaben zur Entzündung als Abwesenheit von Entzündung interpretiert und so alle Daten mit einbezieht, steigt Kappa Cohen nur geringfügig auf 0,277 an.

Tabelle 4.87: Kontingenztafeln Entzündung

mit Fehlwerten (NA)						Fehlwerte (NA) als „nein“ (0)					
Mc Nemar < 0,01		Referenz 0	Referenz 1	Referenz NA	Summe	Mc Nemar < 0,01		Referenz 0	Referenz 1	Referenz NA	Summe
Befunder	0	153	139	0	292	Befunder	0	286	265	0	551
	1	74	215	0	289		1	74	215	0	289
	NA	133	126	0	259		NA	als 0	als 0	0	NA
Summe		360	480	0	840	Summe		360	480	0	840

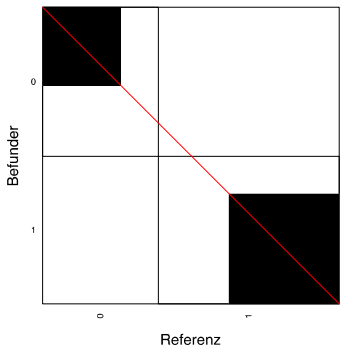
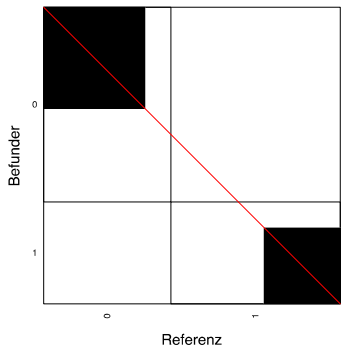
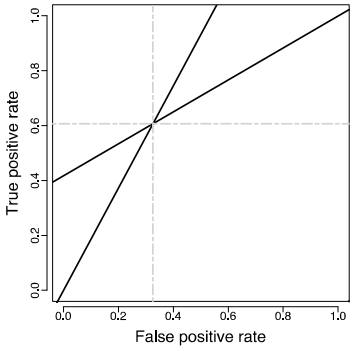
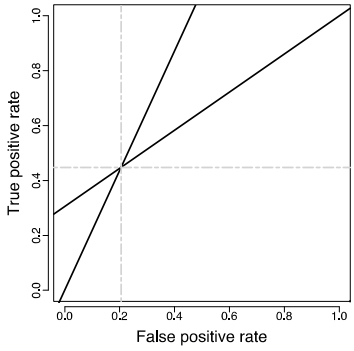
Die Vier-Felder-Tafeln und ihre Randsummen sind Berechnungsgrundlage sämtlicher Kennwerte und Diagramme.

Tabelle 4.88: Kennwerte Entzündung

mit Fehlwerten (NA)				Fehlwerte als „nein“ (0)			
Prävalenz		60,9		Prävalenz		57,1	
	beo.	erw.	kor.		beo.	erw.	kor.
negative Übereinstimmung	59,0	44,0	=κ	negative Übereinstimmung	62,8	51,8	=κ
positive Übereinstimmung	66,9	54,8	=κ	positive Übereinstimmung	55,9	43,0	=κ
Spezifität	67,4	50,3	34,5	Spezifität	79,4	65,6	40,3
Sensitivität	60,7	49,7	21,9	Sensitivität	44,8	34,4	15,8
Genauigkeit	63,3	49,9	26,8=κ	Genauigkeit	59,6	47,8	22,7=κ
Odds ratio		3,2		Odds ratio		3,1	
Yule Q / Y		0,524 / 0,283		Yule Q / Y		0,516 / 0,278	

Beobachtete Kennwerte (beo.), Erwartungswerte gemäß Randsummen (erw.) & für erwartete Übereinstimmung korrigierte Werte (kor.) in Prozent. Die odds ratio und ihre Transformationen (Yule Q & Y) sind prävalenzunabhängig.

Abbildung 4.45: Diagramme Entzündung

mit Fehlwerten (NA)	Fehlwerte als „nein“ (0)
Bangdiwala Konkordanz	Bangdiwala Konkordanz
	
Bangdiwala 0,413 (0,261)	Bangdiwala 0,380 (0,246)
npv 52,4 (39,1) ppv 74,4 (60,9)	npv 51,9 (42,9) ppv 74,4 (57,1)
ROC und Likelihood Graph	ROC und Likelihood Graph
	
area under curve (AUC) 0,641	area under curve (AUC) 0,621
LR+ 1,86 LR- 0,583	LR+ 2,18 LR- 0,695

Die Bangdiwala Diagramme illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Die Receiver Operator Characteristic (ROC) trägt die Rate falsch Positiver (=1-Spezifität, fpr) gegen die Rate richtig Positiver (=Sensitivität, tpr) auf. Die Steigung der Tangenten durch den Ursprung bzw. (1,1) definieren die Likelihood ratios (LR+/LR-), die Fläche darunter die Area Under the Curve (AUC).

4.2.6.2 Entzündung als Syndrom der Schleimhautbefunde

Im vorausgehenden Abschnitt wurde Entzündung als pauschale Diagnose gesehen. In diesem Abschnitt wird Entzündung gemäß dem Konzept des Bronchitis Index (BI) diskutiert (Thompson u. a., 1993), als ein aus den Einzelbefunden Schwellung, Hyperämie und Sekretion zusammengesetztes Syndrom. Dabei wird untersucht in wie weit die Symptome Schwellung, Hyperämie und Sekretion bzw. Kombinationen aus diesen Befunden mit dem Befund Entzündung korrelieren. Verletzlichkeit (engl. friability) der Schleimhaut als viertes Kriterium des Bronchitis Index wurde nicht einbezogen, da es gemäß Studienergebnissen von Thompson vergleichsweise wenig zur Diagnose Entzündung beiträgt. Die eigentlich dreistufige Skala für Schwellung, Hyperämie und Sekretion wurde auf das Vorhandensein bzw. Fehlen der einzelnen Befunde vereinfacht. Neben der Kongruenz des globalen Schleimhautbefundes zum Entzündungsbefund innerhalb der Untersucher, wurde der Bezug zum Referenzbefund des Goldstandards analysiert.

4.2.6.2.1 Konsistenz von Schleimhaut- und Entzündungsbefund der Befunder

Die Hyperämie wird in 90,6 % der mit Entzündung befundeten Videos beschrieben, Schwellung in 85,1 % und Hypersekretion in 81,5 % der Videos (Sensitivität). Wurde keine Entzündung diagnostiziert, fehlte in 87,2 % auch der Befund einer Hyperämie, in 82,3 % der einer Schwellung und in 79,9 % der einer Hypersekretion (Spezifität). In 88,8 % korrelierte der Befund einer Hyperämie mit dem einer Entzündung, bei Schwellung betrug die Korrelation 83,7 %, bei Hypersekretion 80,7 % (Genauigkeit). Unter den Einzelbefunden hat die Hyperämie mit einem positiven prädiktiven Wert von 41,9 % die höchste Vorhersagekraft hinsichtlich eines gleichzeitig erhobenen Entzündungsbefundes. Hypersekretion hat einen positiven prädiktiven Wert von 28,6 %, die Schleimhautschwellung von nur 11,1 %. Auch für das Fehlen eines Entzündungsbefundes war die Abwesenheit einer Hyperämie mit einem negativen prädiktiven Wert von 90,9 % der beste Prädiktor, gefolgt von fehlender Schwellung mit 85,6 % und fehlender Hypersekretion mit 81,6 %. Bei sämtlichen Maßzahlen der Richtigkeit ergibt sich also die Rangliste

- Hyperämie
- Schwellung
- Hypersekretion

und zwar sowohl im Hinblick auf das Vorhandensein einer Entzündung, wie auch im Hinblick auf deren Fehlen. Dieser Eindruck bleibt auch bei Betrachtung mit den prävalenzunabhängigen Maßzahlen bestehen. Hyperämie hat die größte positive Likelihood ratio und die kleinste negative Likelihood ratio, Schleimhautschwellung und Hypersekretion folgen ihr nach. Dem entsprechend folgt auch die AUC dieser Reihenfolge, wie auch visuell an den ROC erkennbar ist. Die odds ratio zeigt, dass die Hyperämie sich mit 65,7 klar von den Befunden Schwellung mit 26,6 und Hypersekretion mit 17,5 absetzt.

Tabelle 4.89: Vier-Felder-Tafeln Schleimhautphänomene versus Entzündung

Schwellung					Hyperämie					Hypersekretion				
McN 0,295	0	Entzündung	NA	∑	McN 0,162	0	Entzündung	NA	∑	McN 0,634	0	Entzündung	NA	∑
0	237	40	19	296	0	251	25	15	291	0	230	52	15	297
Schwellung	51	229	31	311	Hyperämie	37	242	38	317	Hypersekr.	58	229	46	333
NA	4	20	209	233	NA	4	22	206	232	NA	4	8	198	210
∑	292	289	259	840	∑	292	289	259	840	∑	292	289	259	840

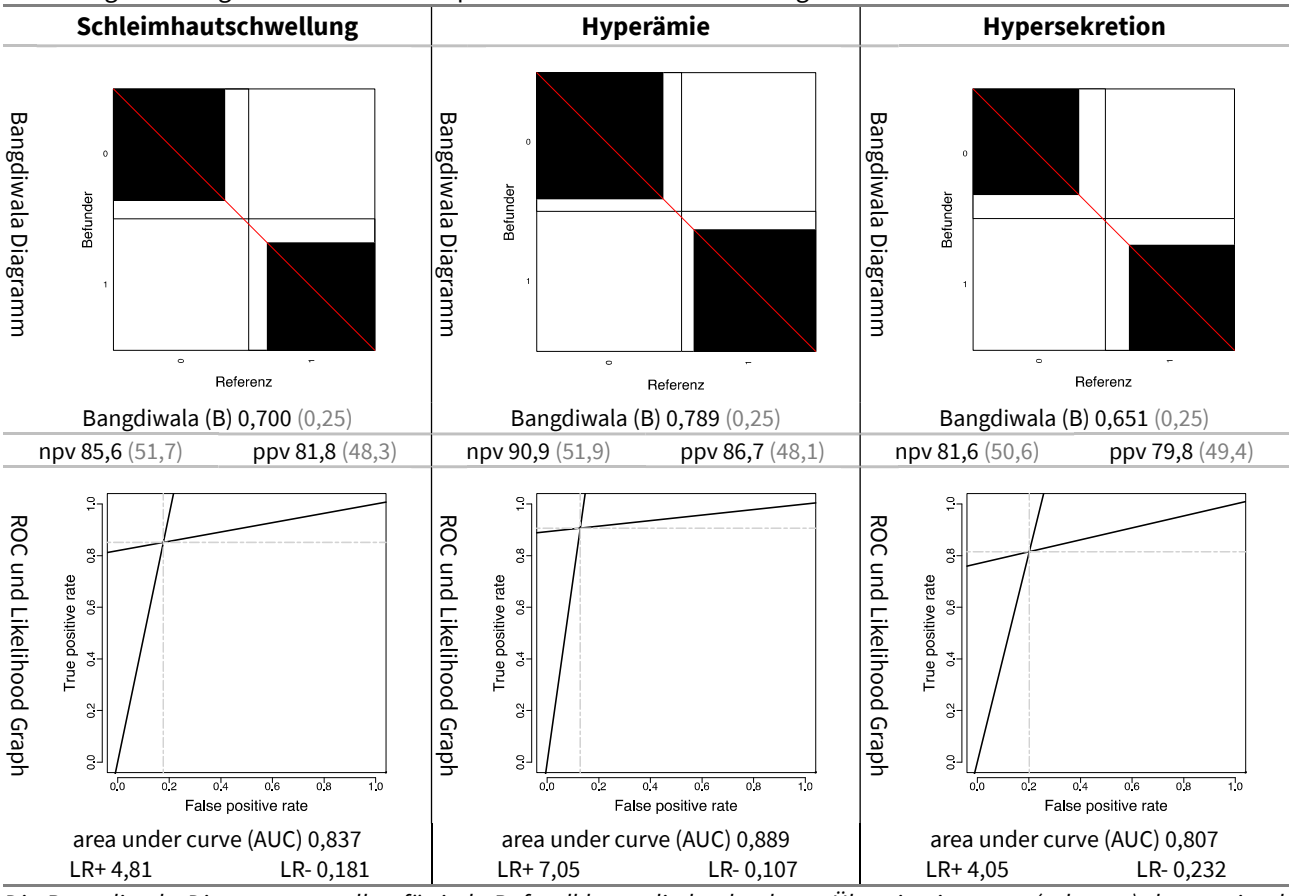
Die Vier-Felder-Tafeln und ihre Randsummen sind Berechnungsgrundlage sämtlicher Kennwerte und Diagramme.

Tabelle 4.90: Kennwerte Schleimhautphänomene versus Entzündung

Schwellung				Hyperämie				Hypersekretion			
Prävalenz		48,3		Prävalenz		48,1		Prävalenz		49,4	
	beo.	erw.	kor.		beo.	erw.	kor.		beo.	erw.	kor.
neg. Übereinst.	83,4	49,3	=κ	neg. Übereinst.	89,0	50,8	=κ	neg. Übereinst.	80,7	50,1	=κ
pos. Übereinst.	83,9	50,7	=κ	pos. Übereinst.	88,6	49,2	=κ	pos. Übereinst.	80,6	49,9	=κ
Spezifität	82,3	49,7	64,8	Spezifität	87,2	49,7	74,4	Spezifität	79,9	49,6	60,1
Sensitivität	85,1	50,3	70,1	Sensitivität	90,6	50,3	81,2	Sensitivität	81,5	50,4	62,7
Genauigkeit	83,7	50,0	67,3	Genauigkeit	88,8	50,0	77,7	Genauigkeit	80,7	50,0	61,3
odds ratio		26,6		odds ratio		65,7		odds ratio		17,5	
Yule Q / Y		0,928 / 0,675		Yule Q / Y		0,97 / 0,78		Yule Q / Y		0,892 / 0,614	

Beobachtete Kennwerte (beo.), Erwartungswerte gemäß Randsummen (erw.) & für erwartete Übereinstimmung korrigierte Werte (kor.) in Prozent. Die odds ratio und ihre Transformationen (Yule Q & Y) sind prävalenzunabhängig.

Abbildung 4.46: Diagramme Schleimhautphänomene versus Entzündung



Die Bangdiwala-Diagramme stellen für jede Befundklasse die beobachtete Übereinstimmung (schwarz) der maximal möglichen (umgebendes Rechteck) gegenüber.Die Seitenverhältnisse illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Die Receiver Operator Characteristic (ROC) trägt die Rate falsch Positiver (=1-Spezifität, fpr) gegen die Rate richtig Positiver (=Sensitivität, tpr) auf. Die Steigung der Tangenten durch den Ursprung bzw. (1,1) definieren die Likelihood ratios (LR+/LR-), die Fläche darunter die Area Under the Curve (AUC).

Tabelle 4.91: Schleimhautphänomene versus Befund Entzündung

Befund				**Befundverteilung**			**Präzision** Übereinstimmung der Befunder untereinander			**Kongruenz zu Befund Entzündung** Übereinstimmung mit Befund Entzündung innerhalb der Befunder					
										Entzündung					
				Anzahl verschiedener				Kappa Fleiss		ja		nein		keine Angabe	
Anzahl der Befunde	Schleimhautschwellung	Hyperämie	Hypersekretion	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Häufigkeit	Prozent in jeweiliger Klasse	Häufigkeit	Prozent in jeweiliger Klasse	Häufigkeit	Prozent in jeweiliger Klasse
x	**gesamt**			446 (53,0 %)	20		21	0,196	gering	282	63,3 %	96	21,5 %	68	15,2 %
0			0	394 (47,0 %)	20	40	50,4	0,326	mäßig	7	1,8 %	196	49,7 %	191	48,5 %
			1	70 (8,3 %)	19	30	19,3	0,080	kaum	20	28,6 %	30	42,9 %	20	28,6 %
1		1	0	31 (3,7 %)	15	21	13,3	0,022	keine	13	41,9 %	10	32,9 %	8	25,8 %
	1	0	0	27 (3,2 %)	14	19	10,7	0,003	keine	3	11,1 %	18	66,7 %	6	22,2 %
		1	1	34 (4 %)	15	20	13,8	0,032	keine	20	58,8 %	5	14,7 %	9	26,5 %
2	1	0	1	32 (3,8 %)	16	16	15,0	0,063	kaum	17	53,1 %	11	34,4 %	4	12,5 %
	1	1	0	55 (6,5 %)	16	27	12,4	0,010	keine	37	67,3 %	10	18,2 %	8	14,5 %
3	1	1	1	197 (23,5 %)	20	36	32,8	0,259	schwach	172	87,3 %	12	6,1 %	13	6,6 %

Die Tabelle analysiert die Befundkombinationen im Sinne eines Gesamtbefundes.

Der Befund Entzündung wurde mit 87,3 % am häufigsten beim Vollbild aus Hyperämie, Hypersekretion und Schleimhautschwellung erhoben, das in fast einem Viertel der Videos gesehen wurde (23,5 %). Beim Vollbild bestand auch die größte Einigkeit der Befunder untereinander (Kappa Fleiss 0,259), die allerdings als schwach einzustufen ist. Kombinationsbefunde aus Hypersekretion, Hyperämie und Schleimhautschwellung haben mit 71,3 % (318/446) einen wesentlich höheren Anteil als Einzelbefunde mit 28,7 % (128/446). Zweifachbefunde haben einen Anteil von 27,1 % (121/446). Einfach- und Zweifachbefunde liegen also in etwa gleichauf, Dreifachbefunde nur wenig darunter. Abseits des Vollbildes und unauffälliger Befunde besteht kaum Einigkeit innerhalb der Befunder.

Singuläre Befunde von Schwellung, Hyperämie und Hypersekretion haben nur einen schwachen positiven prädiktiven Wert hinsichtlich eines gleichzeitigen Entzündungsbefundes. Am besten schneidet, wie bei Betrachtung auf Ebene der Einzelbefunde, die Hyperämie ab (41,9 %). Dahinter folgt die Hypersekretion mit 28,6 %, Schlusslicht ist die Schleimhautschwellung mit 11,1 %. Die positiven prädiktiven Werte von Zweifachkombinationen liegen mit Werten von mindestens 50 % deutlich über dem Niveau der Einzelbefunde. Die beste Vorhersagekraft kommt mit 67,3 % der Kombination aus Hyperämie und Schleimhautschwellung zu, gefolgt von der Kombination aus Hyperämie und Hypersekretion. Das Vollbild einer Entzündung aus Hyperämie, Hypersekretion und Schwellung hat mit 87,3 % die weitaus beste Vorhersagekraft hinsichtlich eines gleichzeitigen Entzündungsbefundes.

Der wichtigste Prädiktor für das Fehlen eines Entzündungsbefundes ist mit 66,7 % die Abwesenheit einer singulären Schwellung – der mit 3,2 % seltenste Schleimhautbefund.

4.2.6.2.2 Kongruenz des Schleimhautbefundes der Untersucher zum Entzündungsbefund des Goldstandards

Neben der Konsistenz von Schleimhautbefund und Entzündungsbefund innerhalb der Untersucher, die im vorausgehenden Abschnitt untersucht wurde, wurde auch die Kongruenz des Schleimhautbefundes zum Entzündungs-Referenz-Befund des Goldstandards betrachtet.

Richtigkeit Einzelbefunde: Die positiven prädiktiven Werte von Schwellung und Hypersekretion liegen beide bei ca. 80 %, der positive prädiktive Wert von Hyperämie setzt sich mit gut 86 % erkennbar dagegen ab.

Tabelle 4.92: Schleimhautphänomene versus Befund Entzündung des Goldstandards

Befund				**Befundverteilung**				**Richtigkeit** Übereinstimmung der Befunder untereinander				**Kongruenz zu Entzündung** Übereinstimmung mit Befund Entzündung des Goldstandards			
				Anzahl verschiedener						Kappa Cohen		Entzündung			
												ja		Nein	
Anzahl der Befunde	Schleimhautschwellung	Hyperämie	Hypersekretion	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**	Häufigkeit	Prozent in jeweiliger Klasse	Häufigkeit	Prozent in jeweiliger Klasse
x	**gesamt**			**29 [580]**	**446**			**25,9**	**34,8**	**0,217**	**schwach**	**312**	**70,0 %**	**134**	**30,0 %**
0			0	13 [260]	394	20	40	78	49,5	0,368	mäßig	168	42,6 %	226	57,4 %
			1	2 [40]	70	19	30	23,1	12,9	0,110	gering	35	50,0 %	35	50,0 %
1		1	0	2 [40]	31	15	21	0	0	-0,044	keine	21	67,7 %	10	32,3 %
	1	0	0	na	27	14	19	na	na	na	na	16	59,3 %	11	40,7 %
		1	1	3 [60]	34	15	20	0	0	-0,055	keine	27	79,4 %	7	20,6 %
2	1	0	1	3 [60]	32	16	16	6,9	12,5	0,04	keine	24	75,0 %	8	25,0 %
	1	1	0	3 [60]	55	16	27	13,8	14,5	0,078	kaum	39	70,9 %	16	29,1 %
3	1	1	1	16 [320]	197	20	36	39.9	63,5	0,274	schwach	150	76,1 %	47	23,9 %

Die Tabelle analysiert die Befundkombinationen im Sinne eines Gesamtbefundes.

Eine solitäre Hypersekretion wurde genauso häufig bei Entzündung, wie bei unauffälligen Befunden, gesehen und hat somit keine Vorhersagekraft. Den besten prädiktiven Wert bei Einzelbefunden erzielte die singuläre Hyperämie mit 67,7 %. In zwei Dritteln der Fälle mit einer Hyperämie lag also auch eine Entzündung vor. Die prädiktiven Werte kombinierter Befunde liegen mit mindestens 70 % deutlich über denen von Einzelbefunden.

4.2.6.3 Entzündungsbereich

Befundverteilung Einzelbefunde: Tabelle 4.93 zeigt, dass Entzündungen vom Goldstandard in allen Bereichen erheblich häufiger gesehen wurden als von den Befundern. Die Beteiligung an der Erhebung von Entzündungen war mit mindestens 17/20 in allen anatomischen Abschnitten hoch. Entzündungen im Stenosebereich und im Bronchus werden von den Befundern in etwa dreimal so vielen Videomitschnitten gesehen, wie vom Goldstandard. In Larynx und Trachea liegt der Faktor in etwa bei 2.

Präzision Einzelbefunde: Die Übereinstimmung innerhalb der Befunder ist in Trachea und Bronchus mit einem Kappa Fleiss von 0,226 bzw. 0,222 höher ausgeprägt, als im Larynx. Bei Entzündungen im Stenosebereich gehen die Befunde am weitesten auseinander.

Tabelle 4.93: Einzelbefunde Entzündungsbereich

Befund	Befundverteilung				Präzision Übereinstimmung der Befunder untereinander			Richtigkeit Übereinstimmung mit Goldstandard als Referenz				
	Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
	Anzahl verschiedener											
	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	Klassifikation **modifiziert nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa **nach Cohen**	Klassifikation **modifiziert nach Landis**	Datenabdeckung [%]
Stenosebereich	9 [180]	125	17	32	23,6	0,124[31]	gering	16,3	23,4	0,013	keine	95,7
Larynx	17 [340]	90	18	28	20,5	0,148	gering	20,5	75,6	0,177	gering	
Trachea	13 [260]	131	20	27	28,6	0,226	schwach	34,7	66,7	0,311	mäßig	
Bronchus	6 [120]	69	19	18	24,6	0,222	schwach	38,4	63,2	0,416	moderat	

Zur besseren Vergleichbarkeit zwischen Referenz und Befundern ist in eckigen Klammern die gewichtete Anzahl der Befunde des Goldstandards angegeben (Faktor 20 im Vgl. zu den Befundern).

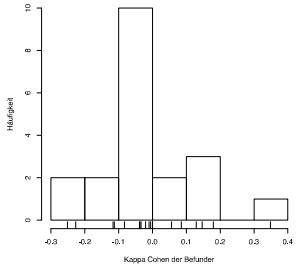
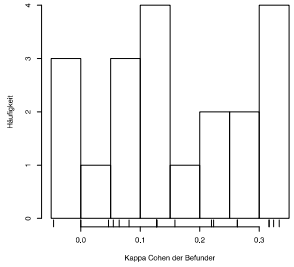
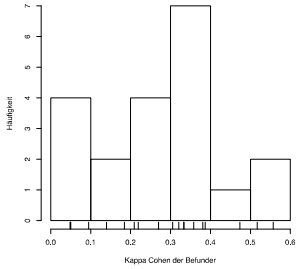
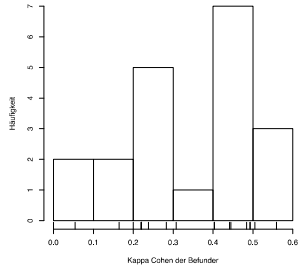
Richtigkeit Einzelbefunde: Die Richtigkeit der Befunde zur Entzündungslokalisation nimmt von proximal nach distal zu. Entzündungen im Stenosebereich schneiden, wie bei der Präzision, am schlechtesten ab. Die Sensitivität liegt im Larynx bei einem Fünftel der Befunde und verdoppelt sich über die Trachea auf knapp 40 % im Bronchus. Der positive prädiktive Wert ist hingegen proximal, im Larynx mit gut 75 % am höchsten und fällt nach distal hin auf 63,2 % im Bronchus leicht ab. Die Diagnose einer Entzündung im Stenosebereich war nur in etwas weniger als einem Viertel der Befunde richtig. Bei der Spezifität sind nur geringere Unterschiede zu erkennen. Die Werte liegen – mit Ausnahme der Entzündungen im Stenosebereich – sämtlich deutlich über 95 %. Bei den negativen prädiktiven Werten ist jedoch analog zur Sensitivität ein Anstieg von proximal (Larynx 63 %) nach distal (Bronchus 90,6 %) zu beobachten.

Am deutlichsten zeigt sich dieser Trend in den Kappawerten. Während im Stenosebereich keine Übereinstimmung erzielt wird, ist die Befundkonkordanz im Larynx mit einem Kappa Cohen von 0,177 gering, in der Trachea bei einem Kappa Cohen von 0,311 mäßig und im Bronchus mit 0,416 moderat ausgeprägt.

Hinsichtlich der odds ratio setzen sich Entzündungen im Bronchus mit 16,6 deutlich gegen die im Larynx (5,3) und der Trachea (6,3) ab, die in etwa gleich auf liegen. Die hohe odds ratio bei Entzündungen im Bronchus geht wesentlich auf eine hohe Treffsicherheit bei positiven Befunden zurück: Die LR+ ist mit 10,6 im Bronchus etwa doppelt so hoch wie in Trachea und Larynx. Die Treffsicherheit bei der negativen Befundung – gemessen an der LR- - nimmt ebenfalls nach distal leicht zu, was sich in einer abnehmenden LR- niederschlägt. Dem entsprechend ergibt sich insgesamt eine von proximal nach distal zunehmende Area under the curve. Die nahezu parallel zur Diagonalen verlaufenden Graphen bei Entzündungen im Stenosebereich illustrieren die Zufälligkeit der Befunde.

[31] *Bei dem hier angegebenen Kappa nach Fleiss handelt es sich um das Gesamt-Kappa für die 3 Kategorien 1 („ja“) 0 („nein“) und NA (Fehlwert). Die Fehlwerte entstehen durch den Korrekturalgorithmus der Rohdaten (Eliminierung der Kategorie „gesamt“) und werden hier für die Berechnung von Kappa als eigene Kategorie behandelt. Die Kappawerte der Einzelkategorien lauten: Stenosebereich* **0** *0,154;* **1** *0,104 und* **NA** *0,034. Larynx* **0** *0,167;* **1** *0,149 und* **NA** *0,034, Trachea* **0** *0,265;* **1** *0,209 und* **NA** *0,034 und Bronchus* **0** *0,273;* **1** *0,206 und* **NA** *0,034.*

Tabelle 4.94: Paarweises Kappa Cohen Entzündungsbereich

Stenosebereich				Larynx				Trachea				Bronchus			
unifiziertes Kappa Cohen				unifiziertes Kappa Cohen				unifiziertes Kappa Cohen				unifiziertes Kappa Cohen			
0,013				**0,177**				**0,311**				**0,416**			
paarweises Kappa Cohen				paarweises Kappa Cohen				paarweises Kappa Cohen				paarweises Kappa Cohen			
min.	Ø	med.	max.	min.	Ø	med.	max.	min.	Ø	med.	max.	min.	Ø	med.	max.
-0,252	**0,001**	-,008	0,349	-0,046	**0,156**	0,128	0,333	0,049	**0,275**	0,288	0,557	0	**0,335**	0,356	0,559
															

Die Tabelle vergleicht die Ergebnisse der beiden Berechnungsvarianten „vereintes“ und „paarweises“ Kappa Cohen.

Paarweises Kappa Cohen: Wird Kappa Cohen paarweise mit dem Goldstandard berechnet und der Mittelwert dieser Kappawerte gebildet, finden sich absolut gesehen, etwas niedrigere Werte für Kappa Cohen, die jedoch in der gleichen Größenordnung liegen und relativ zueinander das gleiche Bild zeichnen.

Befundverteilung Befundkombinationen: In etwas mehr als der Hälfte der Videos (54,7 %, 460/840) wurde vom Goldstandard die Lokalisation einer Entzündung angegeben. In 21,7 % (140/460) ist nur ein einziger anatomischer Abschnitt von der Entzündung betroffen. In der mit 69,5 % (320/460) überwiegenden Zahl der Videos erstreckte sich die Entzündung auf 2 oder mehr Abschnitte. Die Befunder machten in 30,4 % (255/840) Angaben zu Lage der Entzündung. Der Anteil von Entzündungen, die nur einen Abschnitt betreffen, fällt mit 60 % (155/255) erheblich größer aus als beim Goldstandard. Mehrfache Entzündungen machen dem entsprechend knapp 40 % (100/255) aus. An singulären Entzündungen beteiligt sich die Mehrheit der Befunder, mehrfache Entzündungen werden von deutlich weniger Befundern gesehen. Eine Ausnahme hiervon sind generalisierte Entzündungen, die von 14/20 Befundern angegeben wurden. In den meisten Befundklassen liegen signifikant unterschiedliche Prävalenzen vor. Dem entsprechend fällt der McNemar Test abseits bronchialer, trachealer sowie kombiniert tracheobronchialer Entzündungen signifikant aus.

Präzision Befundkombinationen: Eine, wenn auch nur geringe Übereinstimmung, wird nur bei singulären Entzündungen in Larynx und der Trachea sowie bei generalisierten Stenosen erreicht. Die Übereinstimmung negativer Befunde ist schwach bis zufällig.

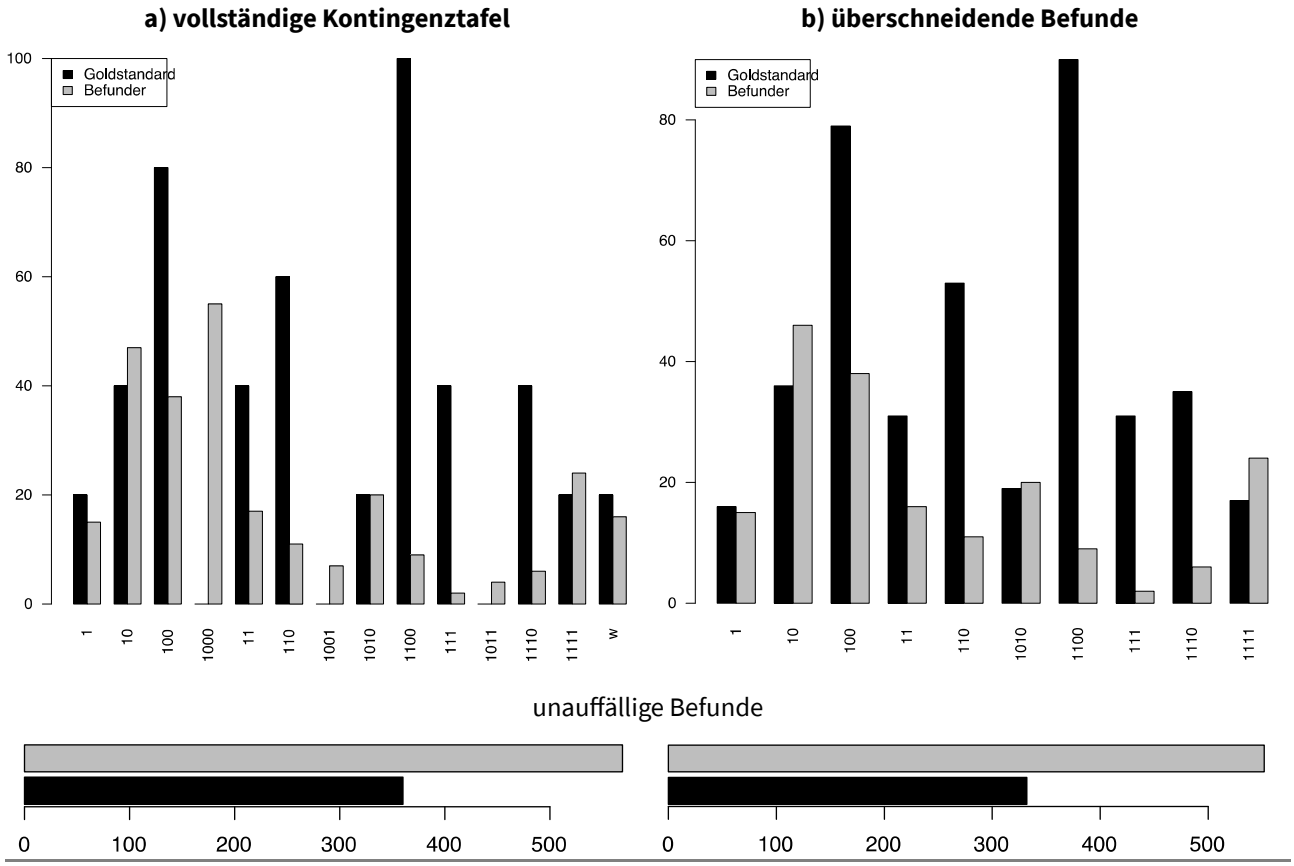
Tabelle 4.95: Inter-Beobachter-Variabilität Befundkombinationen Entzündungsbereich

Befund					**Befundverteilung**				**Präzision** Übereinstimmung der Befunder untereinander			**Richtigkeit** Übereinstimmung mit Goldstandard in den überschneidenden Befunden				
					Referenz	Befunder				Kappa Fleiss				Kappa Cohen		
					Anzahl verschiedener											
Entzündungsbereiche	Stenosebereich	Larynx	Trachea	Bronchus	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	Klassifikation **modifiziert nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	**Kappa nach Cohen**	Klassifikation **modifiziert nach Landis**	Datenabdeckung [%]
x	**gesamt**				**23 [460]**	**255**	**20**	**38**	**13,8**	**0,138**	**gering**	**9,1**	**19,8**	**0,126**	**gering**	**88,0**
0				0	18 [360]	569	20	42	67,7	0,227	schwach	84,9	51,1	0,175	gering	93,0
1				1	1 [20]	15	9	9	12,5	0,046	keine	12,5	13,3	0,110	gering	87,9
			1	0	2 [40]	47	18	19	20,6	0,119	gering	19,4	15,2	0,123	gering	93,8
		1	0	0	4 [80]	38	15	18	19,3	0,147	gering	16,5	34,2	0,164	gering	99,0
	1	0	0	0	NA	55	13	24	14,7	0,036	keine	NA	NA	NA	NA	NA
2			1	1	2 [40]	17	10	10	12,0	0,036	keine	9,7	18,7	0,102	gering	81,5
		1	1	0	3 [60]	11	7	8	12,5	0,026	keine	9,4	45,5	0,135	gering	89,4
	1	0	0	1	NA	7	6	5	10,0	0,022	keine	NA	NA	NA	NA	NA
	1	0	1	0	1 [20]	20	9	12	11,7	0,030	keine	5,3	5,0	0,026	keine	97,4
	1	1	0	0	5 [100]	9	6	7	15,0	0,025	keine	0	0	-0,023	keine	90,8
3		1	1	1	2 [40]	2	1	2	NA	-0,002	keine	3,2	50,0	0,056	kaum	78,0
	1	0	1	1	NA	4	3	3	10,0	0,022	keine	NA	NA	NA	NA	NA
	1	1	1	0	2 [40]	6	6	5	10,0	0,010	keine	5,7	33,3	0,085	kaum	88,6
4	1	1	1	1	1 [20]	24	14	9	17,5	0,174	gering	17,6	12,5	0,123	gering	92,7
n	ω				1 [20]	16	7	9	10,8	0,034	keine	NA	NA	NA	NA	NA

Die Tabelle analysiert die Befundkombinationen im Sinne eines Gesamtbefundes. Sie gliedert sich in die 3 Hauptbereiche Befundverteilung, Präzision und Richtigkeit. Details siehe Kommentar Tabelle 4.15 auf Seite 109.

Richtigkeit Befundkombinationen: Bei singulären Stenosen wurden Trachealstenosen mit einer Sensitivität von 19,4 % am besten erkannt. Vergleichbar gut wurden Larynxentzündungen mit einer Sensitivität von 16,5 % beurteilt. Die Validität trachealer Entzündungen ist mit einem positiven prädiktiven Wert (ppv) von 15,2 gering ausgeprägt. Larynxstenosen wurden mit einem ppv von 34,2 % doppelt so sicher diagnostiziert. Das Kappa Cohen laryngealer Entzündungen liegt mit 0,164 über dem trachealer Stenosen mit 0,123. Der Erwartungswert im Assoziationsdiagramm wird dem entsprechend deutlich übertroffen. Bei bronchialen Entzündungen sind Sensitivität mit 12,5 % und pvv mit 13,3 % auf niedrigem Niveau annähernd gleich. Bronchiale Entzündungen erreichen mit 7,8 die höchste odds ratio innerhalb singulärer Entzündungen, gefolgt von laryngealen (OR 5) und trachealen (OR 4,1) Entzündungen. Multiple Entzündungen führen mit dem Spitzenreiter generalisierter Entzündungen (OR 23,6) das Feld klar an.

Abbildung 4.47: Vergleich Randverteilungen Befundkombinationen Entzündungsbereich



Die Assoziationsdiagramme zeigen das Verhältnis der beobachteten Werte zu den Erwartungswerten.

Tabelle 4.96: Vierfeldertafeln Einzelbefunde Entzündungsbereich

Stenosebereich						Larynx						Trachea						Bronchus					
McN p<0,01		Referenz 0	1	ω	∑	McN p<0,01		Referenz 0	1	ω	∑	McN p<0,01		Referenz 0	1	ω	∑	McN p<0,01		Referenz 0	1	ω	∑
Befunder	0	531	149	19	699	Befunder	0	450	264	20	734	Befunder	0	513	162	18	693	Befunder	0	667	69	19	755
	1	95	29	1	125		1	22	68	0	90		1	43	86	2	131		1	25	43	1	69
	ω	14	2	0	16		ω	8	8	0	16		ω	4	12	0	16		ω	8	8	0	16
∑		640	180	20	840	∑		480	340	20	840	∑		560	260	20	840	∑		700	120	20	840

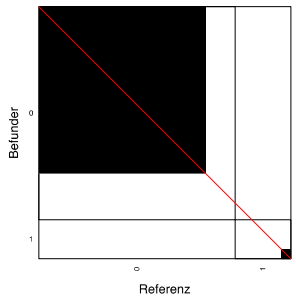
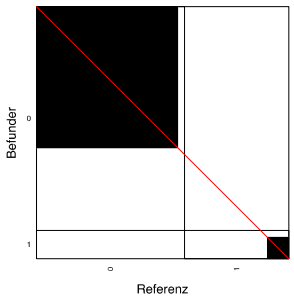
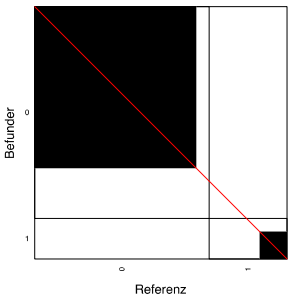
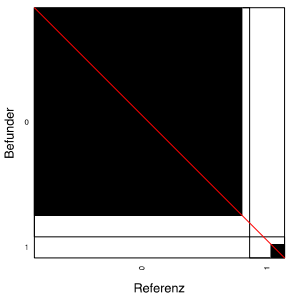
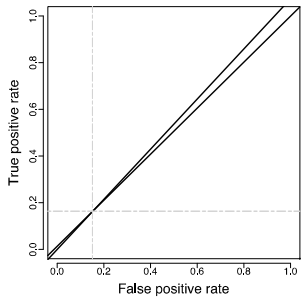
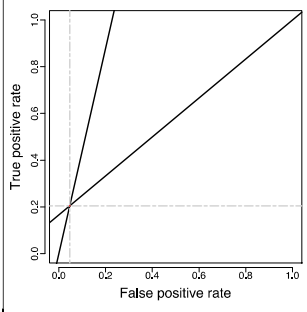
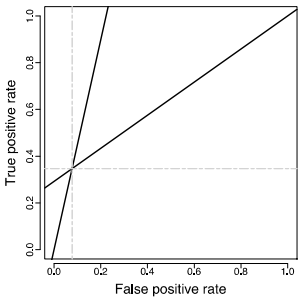
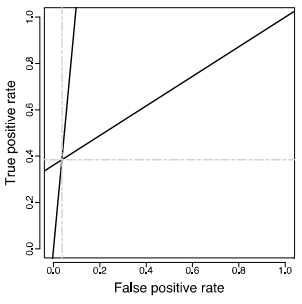
Die Vier-Felder-Tafeln und ihre Randsummen sind Berechnungsgrundlage sämtlicher Kennwerte und Diagramme.

Tabelle 4.97: Kennwerte Einzelbefunde Entzündungsbereich

Stenosebereich				Larynx				Trachea				Bronchus			
Prävalenz	22,1			Prävalenz	41,3			Prävalenz	30,8			Prävalenz	13,9		
	beo.	erw.	kor.		beo.	erw.	kor.		beo.	erw.	kor.		beo.	erw.	kor.
neg. Übst.	81,3	81,1	=κ	neg. Übst.	75.9	70,7	=κ	neg. Übst.	83,3	75,8	=κ	neg. Übst.	93,4	88,7	=κ
pos. Übst.	19,2	18,2	=κ	pos. Übst.	32,2	17,6	=κ	pos. Übst.	45,6	21,1	=κ	pos. Übst.	47,8	10,7	=κ
Spezifität	84,8	84,6	1,6	Spezifität	95,3	88,8	58,4	Spezifität	92,3	84,0	51,8	Spezifität	96,4	91,5	57,3
Sensitivität	16,3	15,4	1,0	Sensitivität	20,5	11,2	10,5	Sensitivität	34,7	16,0	22,2	Sensitivität	38,4	8,5	32,7
Genauigkeit	69,7	69,3	1,3	Genauigkeit	64,4	56,8	17,7	Genauigkeit	74,5	63,0	13,1	Genauigkeit	88,3	80,0	41,6
Odds ratio	1,1			Odds ratio	5,3			Odds ratio	6,3			Odds ratio	16,6		
Yule Q / Y	0,037 / 0,021			Yule Q / Y	0,681 / 0,393			Yule Q / Y	0,727 / 0,431			Yule Q / Y	0,887 / 0,606		

Beobachtete Kennwerte (beo.), Erwartungswerte gemäß Randsummen (erw.) & für erwartete Übereinstimmung korrigierte Werte (kor.) in Prozent. Die odds ratio und ihre Transformationen (Yule Q & Y) sind prävalenzunabhängig.

Abbildung 4.48: Diagramme Einzelbefunde Entzündungsbereich

Stenosebereich	Larynx	Trachea	Bronchus
Bangdiwala Konkordanz	Bangdiwala Konkordanz	Bangdiwala Konkordanz	Bangdiwala Konkordanz
			
Bangdiwala 0,632 (0,628)	Bangdiwala 0,565 (0,483)	Bangdiwala 0,664 (0,539)	Bangdiwala 0,864 (0,776)
npv 78,1(77,9) ppv 23,4(22,1)	npv 63,0(58,7) ppv 75,6(41,3)	npv 76,0(69,2) ppv 66,7(30,8)	npv 90,6(86,1) ppv 63,2(13,9)
ROC & Likelihood Graph	ROC & Likelihood Graph	ROC & Likelihood Graph	ROC & Likelihood Graph
			
AUC 0,506	**AUC 0,579**	**AUC 0,635**	**AUC 0,674**
LR+ 1,07 LR- 0,987	**LR+ 4,39 LR- 0,834**	**LR+ 4,48 LR- 0,708**	**LR+ 10,6 LR- 0,639**

Die Bangdiwala Diagramme illustrieren horizontal Spezifität und Sensitivität, vertikal negativen und positiven prädiktiven Wert (npv & ppv). Die Receiver Operator Characteristic (ROC) trägt die Rate falsch Positiver (=1-Spezifität, fpr) gegen die Rate richtig Positiver (= Sensitivität, tpr) auf. Die Steigung der Tangenten durch den Ursprung bzw. (1,1) definieren die Likelihood ratios (LR+/LR-), die Fläche darunter die Area Under the Curve (AUC).

Tabelle 4.98: Kontingenztafel Befundkombinationen Entzündungsbereich

Bhapkar p < 0,01		Referenz 0	1			2				3		4	n	Summen	
		0	**1**	**10**	**100**	**11**	**110**	**1010**	**1100**	**111**	**1110**	**1111**	**ω**		
0	**0**	282	14	23	60	21	22	13	80	11	19	7	17	569	287
1	**1**	5	2	0	0	2	0	0	0	2	0	4	0	15	15
	10	12	0	7	1	1	17	5	1	1	1	0	1	47	47
	100	9	0	1	13	0	2	0	6	1	6	0	0	38	38
	1000	22	2	2	1	2	5	0	10	4	5	1	1		55
2	**11**	6	0	1	0	3	1	0	0	2	0	3	1	17	17
	110	0	0	0	1	0	5	0	1	1	3	0	0	11	11
	1001	0	2	0	0	3	0	1	0	1	0	0	0		7
	1010	9	0	3	1	2	3	1	0	1	0	0	0	20	20
	1100	3	0	0	2	0	0	0	0	0	4	0	0	9	9
3	**111**	1	0	0	0	0	0	0	0	1	0	0	0	2	2
	1011	2	0	0	0	2	0	0	0	0	0	0	0		4
	1110	1	0	0	1	0	0	0	2	0	2	0	0	6	6
4	**1111**	4	0	1	0	2	3	0	0	11	0	3	0	24	24
n	**ω**	4	0	2	0	2	2	0	0	4	0	2	0	16	16
Summe überl.		336	16	38	79	33	55	19	90	35	35	19	19	774	
Summe		360	20	40	80	40	60	20	100	40	40	20	20		840

Befunder

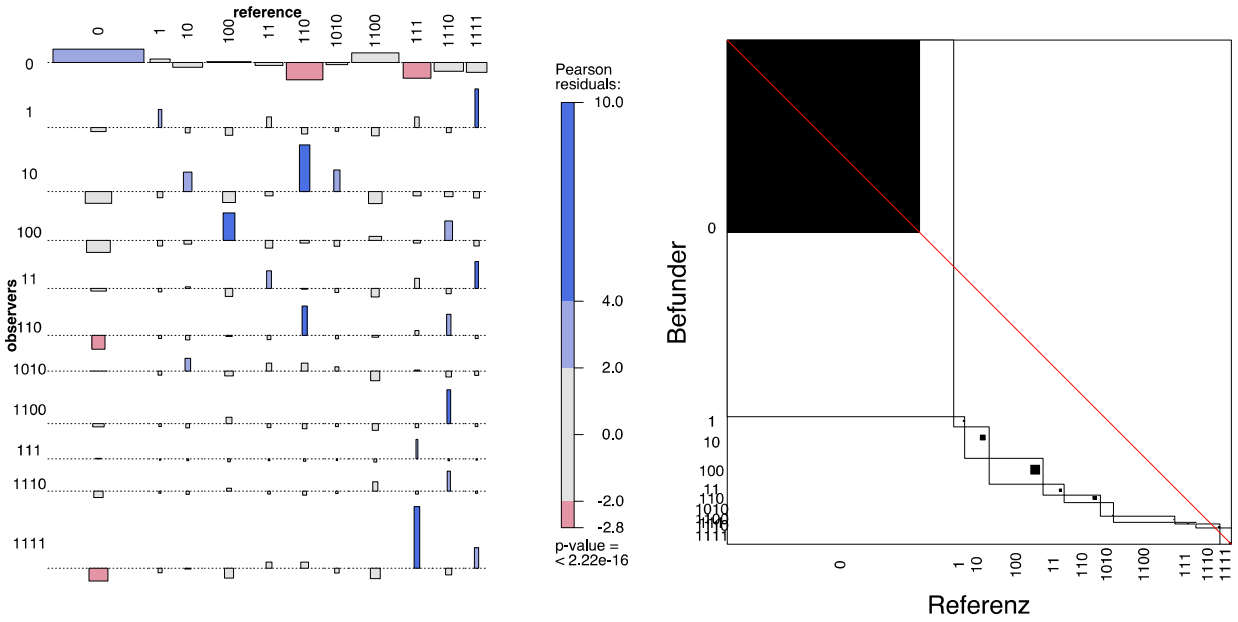
Die Kontingenztafel mit ihren Randsummen ist Berechnungsgrundlage sämtlicher Kennwerte und Diagramme. Nicht überlappende Kategorien sind grau, überlappende schwarz dargestellt.

Tabelle 4.99: Kennwerte überlappende Befundkombinationen Entzündungsbereich

Befund Anzahl	Stenosebereich	Larynx	Trachea	Bronchus	pre	McN	na	pa	spe	sen	acc	npv	ppv	B	LR-	LR+	AUC	OR	Q	Y
0	0	0	0	0	44,9	< 0,01	46,1	63,8	33,7	84,9	56,7	73,3	51,1	0,379	0,447	1,28	0,593	2,9	0,482	0,257
				1	2,2	1	98,1	12,9	98,2	12,5	96,3	98,1	13,3	0,963	0,891	6,95	0,554	7,8	0,773	0,473
1			1	0	4,9	0,275	95,1	17,1	94,5	19,4	90,8	95,8	15,2	0,902	0,853	3,5	0,569	4,1	0,609	0,339
		1	0	0	10,7	< 0,01	93,3	22,2	96,2	16,5	87,7	90,6	34,2	0,866	0,868	4,34	0,563	5	0,667	0,382
			1	1	4,2	0,029	97,1	12,8	98,2	9,68	94,5	96,1	18,7	0,943	0,92	5,27	0,539	5,7	0,703	0,411
		1	1	0	7,2	< 0,01	96,2	15,6	99,1	9,43	92,7	93,4	45,5	0,925	0,914	10,8	0,543	11,8	0,844	0,549
2	1	0	1	0	2,6	1	97,4	5,1	97,4	5,26	95	97,5	5,0	0,949	0,973	1,99	0,513	2	0,344	0,178
	1	1	0	0	12,2	< 0,01	92,8	0	98,6	0	86,6	87,7	0	0,863	1,01	0	0,493	0	-1	-1
		1	1	1	4,2	< 0,01	97,9	6,1	99,9	3,23	95,8	95,9	50	0,958	0,969	22,8	0,515	23,6	0,919	0,658
3	1	1	1	0	4,7	< 0,01	97,4	9,8	99,4	5,71	95	95,5	33,3	0,949	0,948	10,1	0,526	10,6	0,828	0,53
4	1	1	1	1	2,3	0,31	97,6	14,6	97,1	17,6	95,3	98	12,5	0,951	0,848	6,07	0,574	7,2	0,755	0,456

Prävalenz (pre), McNemar (McN) Test, negative (nag) und positive (pag) Übereinstimmung, Spezifität (spe), Sensitivität (sen), Genauigkeit (acc), negativer (npv) und positiver prädiktiver Wert (ppv), Bangdiwala (B), negative (LR-) und positive Likelihood ratio (LR+), area under the curve (AUC), odds ratio (OR), Youden Q und Y.

Abbildung 4.49: Assoziationsdiagramm



Die Assoziationsdiagramme zeigen das Verhältnis der beobachteten Werte zu den Erwartungswerten.

Zusammenfassung 4.15: Entzündung

Auf Ebene der Einzelbefunde besteht bei trachealen und bronchialen Entzündungen eine schwache Konkordanz unter den Befundern (Kappa Fleiss 0,226 und 0,222). Die Befundrichtigkeit von Entzündungen nimmt von proximal nach distal zu (Kappa Cohen Larynx 0,177, Trachea 0,311 und Bronchus 0,416). Das höchste Kappa Cohen auf der Ebene von Kombinationsbefunden haben singulär laryngeale Entzündungen. Mehrfachbefunde zeigten nur bei generalisierten Entzündungen eine geringe Präzision (Kappa Fleiss 0,174) und Richtigkeit (Kappa Cohen 0,123). Alle anderen Befundkombinationen zeigten keine Übereinstimmung untereinander und nur bei kombinierten laryngotrachealen Entzündungen und tracheobronchialen Entzündungen eine geringe Richtigkeit. Bronchiale Entzündungen werden am sichersten erkannt (Kappa Cohen & OR auf Ebene der Einzelbefunde sowie auf Ebene von Befundkombinationen).

4.3 Einflussgrößen der Befundrichtigkeit

Mit dem Arztfragebogen wurden Informationen zu Demographie, Ausbildung und Erfahrung der Befunder eingeholt. Dieser Abschnitt untersucht, inwieweit diese Variablen mit der Befundrichtigkeit, also der Übereinstimmung mit dem Referenzbefund des Goldstandards, korrelieren. So sollen Anhaltspunkte dafür gefunden werden, welche Faktoren für eine gute Ausbildung in Bronchoskopie maßgebend sein könnten.

Für die Analyse wurden lineare Modelle und Entscheidungsbäume eingesetzt. Ziel war dabei, die Wichtigkeit der einzelnen Variablen zu bestimmen. Bei linearen Modellen wurde die Variablenwichtigkeit über „*proportional marginal variance decomposition*“ (Feldman, 2005; Groemping, 2006; Grömping, 2007), bei Entscheidungsbäumen mit einem random forest errechnet (Breiman, 2001; Liaw, Wiener, 2002).

Ausschlaggebend für die Berechnung mit Entscheidungsbäumen ergänzend zu den linearen Modellen war die direkte Untersuchung des originalen Datensatzes inklusive Fehlwerte. Für die Berechnung der linearen Modelle wurden die Fehlwerte imputiert. Darüber hinaus sind Entscheidungsbäume hypothesenfrei, modellieren Interaktionen zwischen Variablen und sind anschaulich zu interpretieren.

4.3.1 Lineare Modelle

Aus methodischen Gründen können bei der Analyse mit linearen Modellen nicht alle erhobenen Variablen berücksichtigt werden. In einem mehrstufigen Verfahren wurden die folgenden Variablen zur Analyse zusammengestellt:

Tabelle 4.100: Variablen der linearen Modelle

	Variable	Ausprägungen	Abkürzung	
Hospitationen & Kurse	Alter	Alter des Untersuchers In Jahren	Alter	A
	Klinikart derHospitationen	• Universitätsklinik (U) • nicht universitäre Klinik (nU) • keine	HospitationenKlinikart	HK
	Dauer der Hospitationen	Dauer der Hospitationen in Tagen	HospitationenTage	HT
	Dauer der Kursteilnahme	Dauer der Kursteilnahme in Tagen	KursteilnahmeTage	KT
Ausbildung	Dauer der Ausbildung	Dauer der Bronchoskopieausbildung in Jahren	AusbildungJahre	AJ
	Klinikart der Ausbildung	• Universitätsklinik (U) • nicht universitäre Klinik (nU)	AusbildungKlinikart	AK
	Ausbildung in flexibler Bronchoskopie	Anzahl flexibler Bronchoskopien während der Ausbildung	AusbildungFlexibleAnzahl	AF
	Ausbildung in starrer Bronchoskopie	Anzahl starrer Bronchoskopien während der Ausbildung	AusbildungStarreAnzahl	AS
Erfahrung	Erfahrung in flexibler Bronchoskopie	Anzahl der seit der Ausbildung absolvierten flexiblen Bronchoskopien	ErfahrungFlexibleAnzahl	EF
	Erfahrung in starrer Bronchoskopie	Anzahl der seit der Ausbildung absolvierten starren Bronchoskopien	ErfahrungStarreAnzahl	ES
	Erfahrung in Interventioneller Bronchoskopie	Anzahl der seit der Ausbildung absolvierten interventionellen Bronchoskopien	ErfahrungInterventionelleAnzahl	EI
	Klinikart der Erfahrung	• Universitätsklinik (U) • nicht universitäre Klinik (nU) • keine	ErfahrungKlinikart	EK

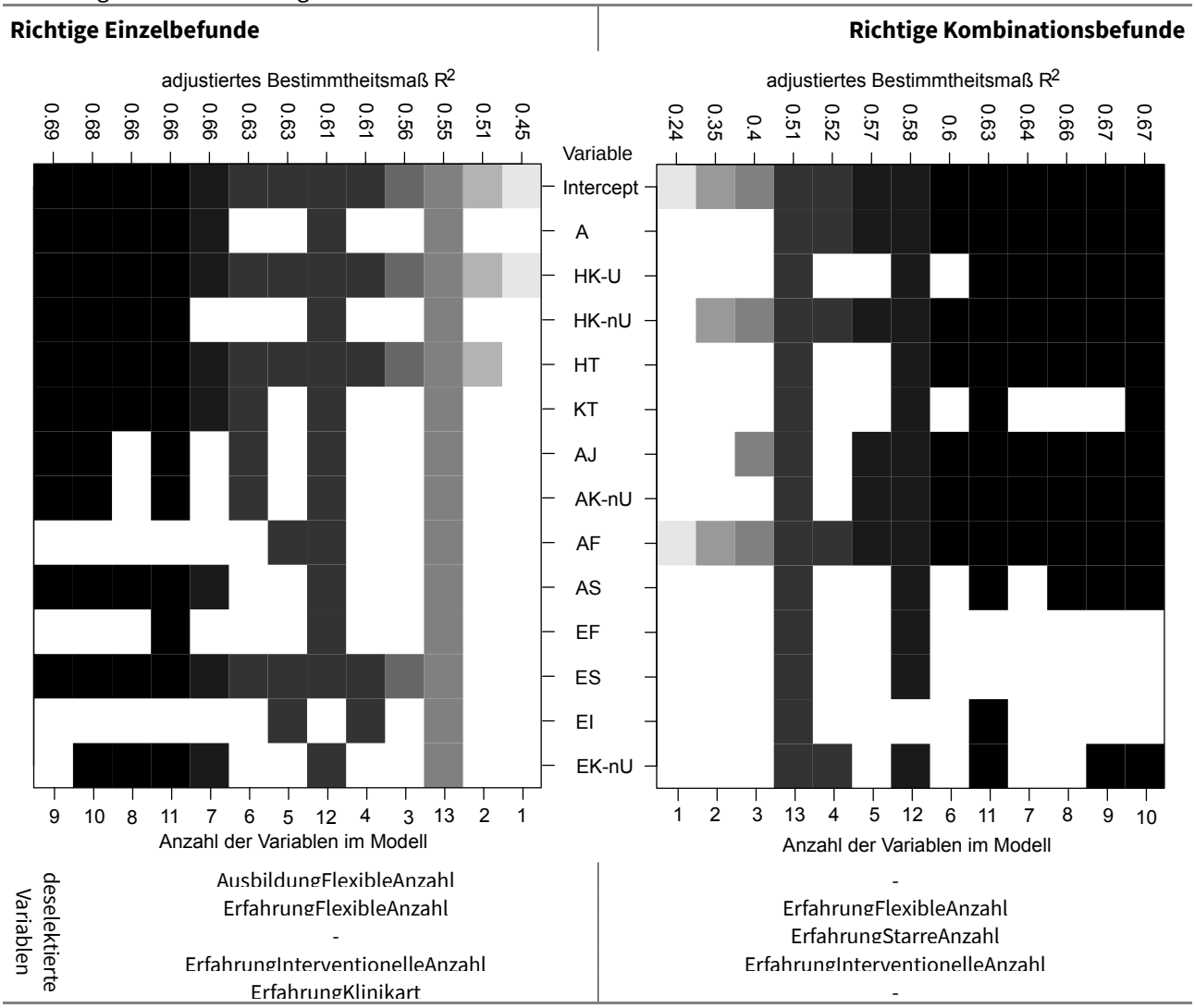
Liste der in lineare Modelle einbezogenen Variablen

Da die erklärenden Variablen des Arztfragebogens überwiegend metrisch sind, wurde als Methode der Modellbildung die multiple lineare Regression gewählt, wobei kategoriale Merkmale über Kontrastvariablen repräsentiert werden. Zunächst wurden hinsichtlich dem adjustierten Bestimmtheitsmaß (adj. R^2) möglichst optimale Modelle selektiert, die anschließend berechnet und auf ihre Kennwerte untersucht wurden. Abschließend wurde die relative Wichtigkeit der erklärenden Variablen in den Modellen bestimmt.

4.3.1.1 Modellselektion

Mit der Funktion regsubsets der R-Bibliothek leaps (Lumley, Miller, 2009; Miller, 2002) wurden in einer all subset regression die hinsichtlich des Bestimmtheitsmaßes best möglichen Modelle für alle Untergruppen (engl. subsets) von 1 bis 12 (=alle) Variablen identifiziert. Die folgende Abbildung zeigt eine horizontale Rangliste der Modelle gemäß ihrem jeweiligen adjustierten Bestimmtheitsmaß (obere horizontale Legende) und gibt an, welche Variablen Bestandteil des jeweiligen Modells sind. Das optimale Modell mit richtigen Einzelbefunden als Zielvariable ist links außen, das optimale Modell mit richtigen Befundkombinationen als Zielvariable rechts außen aufgetragen. Kategoriale Variablen werden in ihren unterschiedlichen Ausprägungen mehrfach als Variable berücksichtigt.

Abbildung 4.50: all subset Regression



Variablenabkürzungen siehe Tabelle 4.100. Ausprägungen kategorialer Variablen sind nach dem Trennstrich angegeben.

Das optimale Modell mit richtigen Einzelbefunden als Zielvariable sortiert 4 erklärende Variablen, das beste Modell mit richtigen Befundkombinationen als Zielvariable 3 erklärenden Variablen aus (Abbildung 4.50). Dem entsprechend verbleiben 9, respektive 10, erklärende Variablen in den Modellen, die jeweils gut ⅔ der beobachteten Varianz erklären.

Das Modell mit richtigen Kombinationsbefunden als Zielvariable sortiert die Erfahrung in Bronchoskopie als Variable weitgehend aus: der Anzahl flexibler, starrer und interventioneller Bronchoskopien seit der Ausbildung wird im optimalen Modell keine Bedeutung beigemessen. Bei Einzelbefunden als Zielvariable fallen Erfahrung in flexibler und interventioneller Bronchoskopie ebenfalls heraus, zusätzlich auch noch die Klinikart, an der Erfahrung gesammelt wurde. Die Anzahl starrer Bronchoskopien seit der Ausbildung ist hingegen Bestandteil der meisten Modelle bei Einzelbefunden als Zielvariable. Flexible Bronchoskopien während der Ausbildung spielen im Modell der Einzelbefund kaum eine Rolle, üben bei Befundkombinationen als Zielvariable jedoch in vielen Modellen Einfluss aus. Bei beiden Zielvariablen werden Variablen zu Hospitationen und Kursen in die meisten Modelle aufgenommen.

4.3.1.2 Multiple lineare Regression

Die selektierten Modelle wurden über die R-Funktionen lm (R Core Team, 2015) und regr (Stahel, 2013) berechnet. Dabei wurden richtige Einzelbefunde bzw. richtige Befundkombinationen als

Zielvariable gesetzt. Auf Transformationen von Variablen und die Modellierung von Interaktionen zwischen Variablen wurde verzichtet. Der Anhang zeigt Graphiken zur Evaluation der beiden Modelle.

4.3.1.2.1 Einzelbefunde multiple lineare Regression

Das optimierte Modell mit richtigen Einzelbefunden als Zielvariable (Abbildung 4.50 links außen) fällt insgesamt mit einem p-Wert von 0,006 statistisch signifikant aus. Innerhalb des Modells korrelieren die Klinikart der Hospitationen und der Ausbildung als einzige Variablen statistisch signifikant mit richtigen Einzelbefunden. Wie an den Koeffizienten abzulesen ist, erreichten Befunder, die an einer Universitätsklinik hospitiert haben, durchschnittlich knapp 30 Treffer mehr als andere Kollegen. Kollegen mit Hospitationen an nicht universitären Kliniken schneiden im Modell hingegen um ca. 15 Treffer schlechter ab.

R Ausgabe4.1: Multiple lineare Regression richtige Einzelbefunde

```
                                Coef     stcoef     signif   R2.x df p.value
(Intercept)            1555.98189000         NA 34.9688888     NA  1  0.0000
Alter                    -0.90274773 -0.3335720 -0.8321353 0.4888  1  0.0934
HospitationenKlinikart            NA         NA  1.8971621 0.5238  2  0.0010
HospitationenTage        -0.02382616 -0.2885610 -0.7131832 0.5022  1  0.1446
KursteilnahmeTage        -0.13199740 -0.3738346 -0.7103666 0.5038  1  0.0678
AusbildungJahre          -1.44353449 -0.3077710 -0.9188015 0.1755  1  0.0549
AusbildungKlinikart      22.10918276  0.6377817 -0.9750962 0.6926  1  0.0205
AusbildungStarreAnzahl    0.17784746  0.2533499  1.2336962 0.3964  1  0.1570
ErfahrungStarreAnzahl     0.05589239  0.2422176  0.6867460 0.4372  1  0.1881

Coefficients for factors:
$HospitationenKlinikart
            keine          Uniklinik nicht univ. klinik
          0.00000           29.73457          -14.67940

St.dev.error:  9.512   on 10 degrees of freedom
Multiple R^2:  0.8345    Adjusted R-squared: 0.6856
F-statistic:   5.604   on 9 and 10 d.f.,  p.value: 0.006326
```

Ausgabe der R-Funktion regr für das lineare Modell mit richtigen Einzelbefunden als Zielvariable.

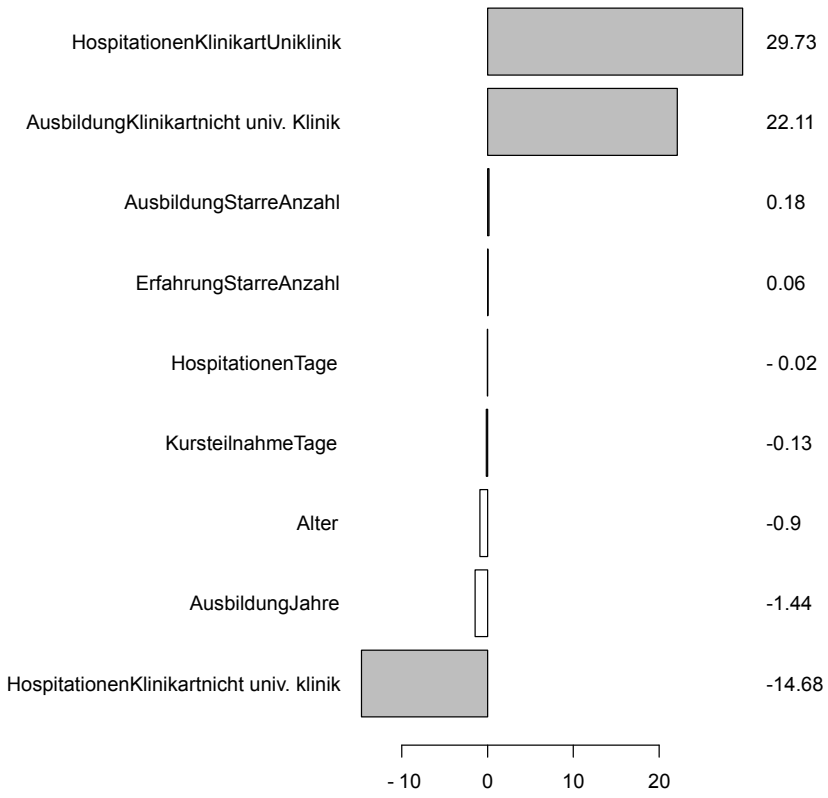
Da die Exponentialschreibweise in der Zusammenfassung des Modells schwierig zu lesen ist, zeigt Abbildung 4.51 eine Rangliste der Koeffizienten. Es wird deutlich, dass auf Symptomebene im linearen Modell im Wesentlichen 2 Faktoren einen positiven Einfluss auf die Befundrichtigkeit haben:

- **Hospitationen an einer Universitätsklinik**, und
- **Ausbildung an einer nicht universitären Klinik.**

Die beiden Ausprägungen der Variable „Klinikart der Hospitationen“ werden in der Rangliste separat aufgeführt und bilden Spitze und Schlusslicht. Einen leichten negativen Einfluss üben im Modell auch das

- **Alter** und eine
- **lange Ausbildung** aus.

Abbildung 4.51: Koeffizientenrangliste lineare Regression richtige Einzelbefunde



Variablen, die im Modell statistisch signifikante Korrelationen erreichen, sind in der Rangliste fett bzw. mit grauem Balken hervorgehoben.

4.3.1.2.2 Befundkombinationen multiple lineare Regression

Neben der Klinikart der Hospitationen findet sich im Modell mit Befundkombinationen als Zielvariable auch für das Alter und die Ausbildungsjahre eine statistisch signifikante Korrelation. Die beiden letztgenannten Faktoren wirken sich negativ aus (R-Ausgabe 4.3).
Analog zum Modell mit richtigen Einzelbefunden als Zielvariable schreibt das Modell richtiger Befundkombinationen den Variablen

- **Hospitationen an einer Universitätsklinik** und der
- **Ausbildung an einer nicht universitären Klinik**

den größten positiven Einfluss zu. Die beiden Variablen tauschen im Vergleich zu den Einzelbefunden die Plätze, ihre Regressionskoeffizienten liegen nummerisch aber noch näher beieinander. Im Gegensatz zur Klinikart der Hospitationen ist der Effekt des Kliniktyps der Ausbildung jedoch nicht statistisch signifikant.

Hospitationen an einer Universitätsklinik bewirken im Modell 20 richtige Befundkombinationen mehr, Hospitationen an nicht universitären Kliniken 31 richtige Befundkombinationen weniger. Ausbildung an einer nicht universitären Klinik verbessert die Trefferquote im Modell um knapp 20. Als dritte Variable mit positivem Einfluss kommt bei richtigen Befundkombinationen als Zielvariable

- **Erfahrung an einer nicht universitären Klinik**

hinzu, was jedoch nicht statistisch signifikant ist.

R-Ausgabe 4.2: Multiple lineare Regression richtige Befundkombinationen

```
                              coef     stcoef     signif   R2.x df p.value
(Intercept)             500.08417170         NA  8.1393689     NA  1  0.0000
Alter                    -2.57027003 -0.7372179 -1.6302728 0.5610  1  0.0050
HospitationenKlinikart           NA         NA  1.4760656 0.5693  2  0.0065
HospitationenTage        -0.03865695 -0.3607331 -1.1791573 0.5087  1  0.0886
KursteilnahmeTage        -0.09140717 -0.2001763 -0.8439416 0.5525  1  0.3384
AusbildungJahre          -3.09058908 -0.5120867 -0.4469482 0.4674  1  0.0200
AusbildungKlinikart      21.78351676  0.4863946 -1.2473287 0.8197  1  0.1533
AusbildungFlexibleAnzahl -0.05563850 -0.1957302  0.6893969 0.5303  1  0.3376
AusbildungStarreAnzahl    0.31650128  0.3489874 -0.4477297 0.4785  1  0.0895
ErfahrungKlinikart       18.78314694  0.3807492  0.8411286 0.7955  1  0.2259

Coefficients for factors:
$HospitationenKlinikart
            keine          Uniklinik nicht univ. klinik
          0.00000           19.51871          -32.96906

St.dev.error:  12.65   on 9 degrees of freedom
Multiple R^2:  0.8421    Adjusted R-squared: 0.6667
F-statistic:   4.801   on 10 and 9 d.f.,  p.value: 0.01354
```

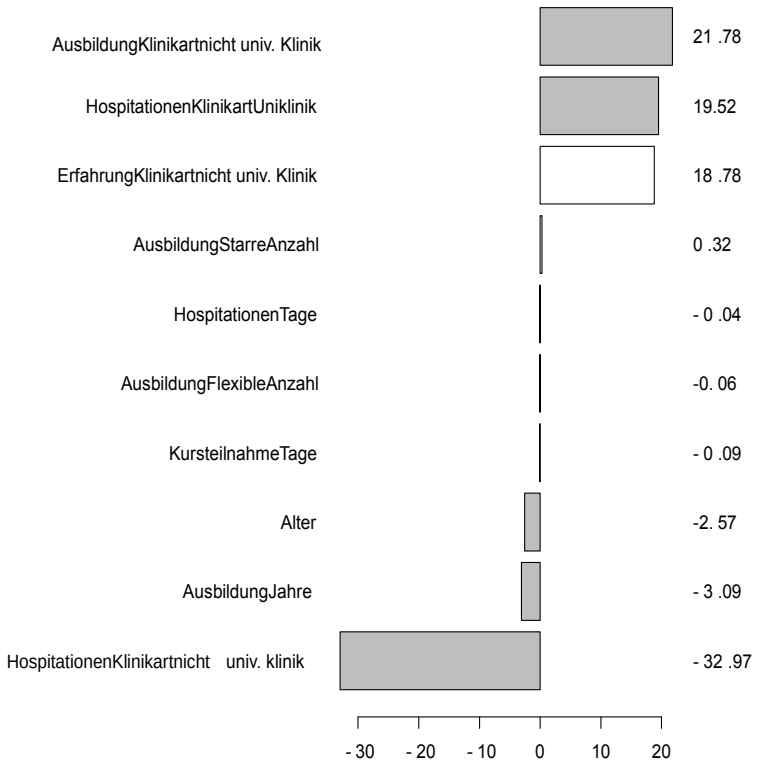
Ausgabe der R-Funktion regr für das lineare Modell mit richtigen Befundkombinationen als Zielvariable. HospitationenKlinkart und Alter erreichen als einzige Variablen statistische Signifikanz.

Im Modell mit richtigen Einzelbefunden als Zielvariable fiel diese Variable im Rahmen der Modellselektion heraus. Das negative Ende der Rangliste der Regressionskoeffizienten stimmt mit dem Modell der richtigen Einzelbefunde überein:

- **Alter** und
- **lange Ausbildung** und
- **Hospitationen an einer nicht universitären Klinik**

wirken sich im Modell negativ auf die Befundrichtigkeit aus. Der negative Einfluss dieser Variablen ist bei Befundkombinationen als Zielvariable etwas deutlicher ausgeprägt.

Abbildung 4.52: Rangliste Koeffizienten lineare Regression richtige Befundkombinationen



Variablen, die im Modell statistisch signifikante Korrelationen erreichen, sind fett hervorgehoben.

4.3.1.3 Relative Variablenwichtigkeit

Der Vergleich der Variablen und Modelle untereinander ist anhand der Regressionskoeffizienten, die an das jeweilige Skalenniveau gebunden sind nur eingeschränkt möglich. Daher wurde über die R-Bibliothek relaimpo mit proportional marginal variance decomposition die relative Wichtigkeit der Variablen in den linearen Modellen bestimmt (Grömping (2009)). Demnach ist mit richtigen Einzelbefunden als Zielvariable die

- **Klinikart der Hospitation**

die überragende Determinante, hinter der alle anderen Variablen klar zurücktreten (Abbildung 4.53). Mit richtigen Befundkombinationen als Zielvariable fällt das Ergebnis hingegen weit weniger eindeutig aus. Mit der

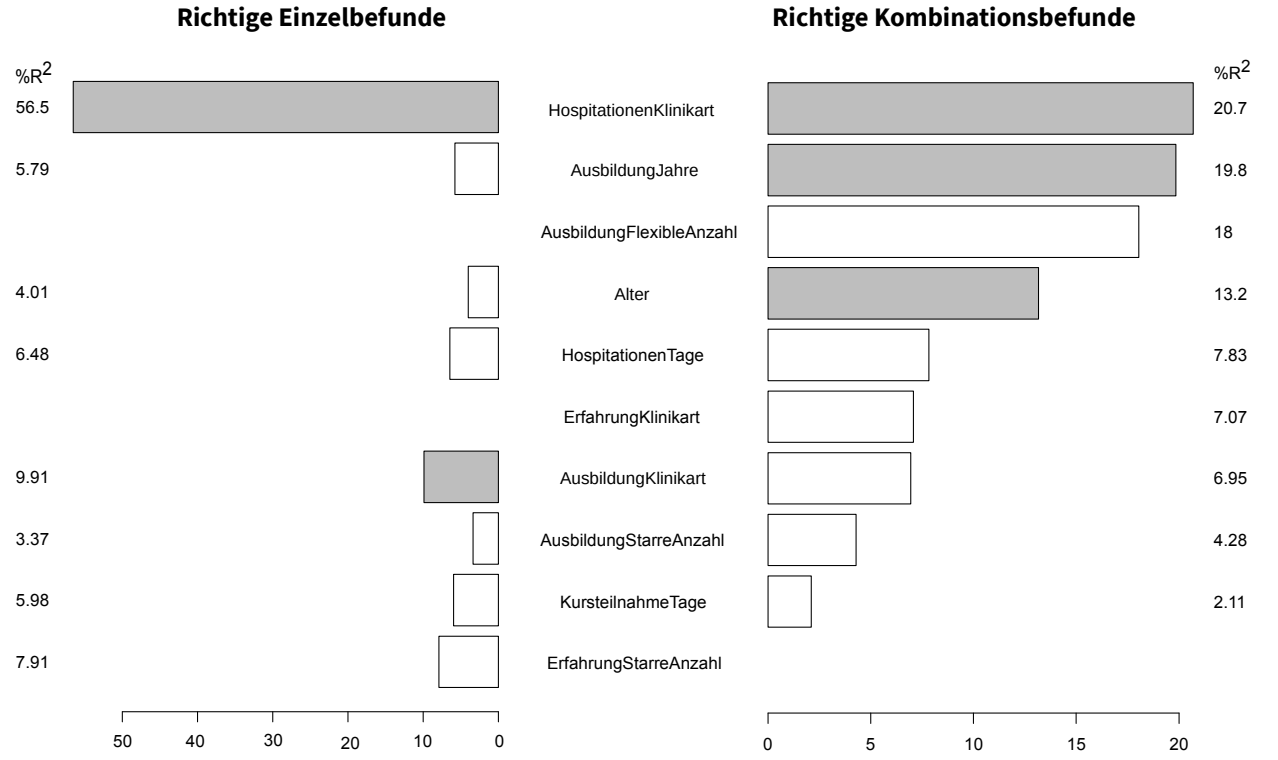
- **Dauer der Ausbildung**, der
- **Anzahl flexibler Bronchoskopien in der Ausbildung** und dem
- **Alter**

treten drei Variablen mit negativem Einfluss auf die Befundrichtigkeit in den Vordergrund, deren Regressionskoeffizienten, oberflächlich betrachtet, zunächst eine eher untergeordnete Rolle zu spielen scheinen. Der Grund hierfür ist, dass Regressionskoeffizienten die Änderung der abhängigen Variable, je Einheit der jeweiligen erklärenden Variable, wiedergeben. Während es sich bei den vermeintlichen Hauptakteuren mit großen Regressionskoeffizienten um kategoriale Merkmale handelt, deren Koeffizienten also nur einmal ins Gewicht fallen, handelt es sich bei den jetzt als mindestens ebenso wichtig identifizierten Variablen um intervallskalierte Größen, bei denen der Regressionskoeffizient **pro Einheit** zu Buche schlägt. So erzielen Befunder im linearen Modell z. B. auf Syndromebene je Jahr ihres Lebensalters 2,57 Treffer weniger. Bei einer Altersdifferenz von 10 Jahren fällt das Alter somit beispielsweise mehr ins Gewicht als die sofort ins Auge fallenden „Einzelkämpfer“-Variablen mit großen Regressionskoeffizienten. Überraschend ist, dass die Anzahl der Bronchoskopien während der Ausbildung sich nummerisch leicht negativ auswirkt und als zweiteinflussreichste Variable identifiziert wird. Dieses Ergebnis beruht u. a. auf einem Ausreißer: der Befunder mit der weitab größten Zahl an Befunden in der Ausbildung (352, Abbildung 4.6 Seite 93) ist Schlusslicht der Rangliste der Befundrichtigkeit. Ein ähnlicher Effekt besteht auch bei der Ausbildungszeit: die Befunder mit den weitaus längsten Ausbildungszeiten von 10 bzw. 15 Jahren sind unter den 3 letzten der Rangliste richtiger Befunde.

Zusammenfassung 4.16: Lineares Modell

Beide linearen Modelle sind statistisch signifikant und erklären etwa ⅔ der beobachteten Varianz. Unabhängig davon, ob richtige Einzelbefunde oder richtige Kombinationsbefunde als Zielvariable gesetzt werden, kommt Erfahrung im Vergleich zu Ausbildung eine untergeordnete Bedeutung zu. Die meisten erfahrungsbezogenen Variablen werden bereits im Rahmen der Modellselektion aussortiert. Allerdings erkennt das Modell mit richtigen Befundkombinationen als Zielvariable einen positiven Effekt von Erfahrung an nicht universitären Kliniken. Übereinstimmend wird die Klinikart der Hospitationen als einflussreichste Variable identifiziert. Die Modelle stimmen auch darin überein, dass sich die Ausbildung an einer nicht universitären Klinik günstig auswirkt. Ungünstig wirken sich in beiden Modellen Alter und eine lange Ausbildungszeit aus. Letzteres geht – genau wie ein negativer Trend für eine hohe Anzahl flexibler Bronchoskopien während der Ausbildung – jedoch zum Teil auf Ausreißer zurück.

Abbildung 4.53: Variablenwichtigkeit in den linearen Modellen



Die relative Wichtigkeit der Variablen im Modell wird über ihren Anteil an R^2 ausgedrückt.

4.3.2 Entscheidungsbäume

Ergänzende Berechnungen mittels rekursivem Partitionieren sind wesentlich von der Möglichkeit zur Analyse des originalen (nicht imputierten) Datensatzes motiviert. Dieser Datensatz wird nachfolgend als „nativ“ bezeichnet. Weitere Vorteile sind die Hypothesenfreiheit des Ansatzes, die Modellierung von Interaktionen und die verständliche Visualisierung der Modelle. Aggregiert man viele Entscheidungsbäume zu einem random forest, kann der relative Einfluss von Variablen ermittelt werden. Datengrundlage der Entscheidungsbäume war einerseits der ursprüngliche, native Datensatz des Arztfragebogens, andererseits der mittels Imputation vervollständigte Datensatz ohne Fehlwerte, der bereits für die linearen Modelle verwendet wurde.

Die Voraussetzungen für Entscheidungsbäume sind im Vergleich zu linearen Modellen deutlich laxer, weswegen mehr Variablen einbezogen werden konnten. Dabei konnten auch Dummy-Variablen, also Variablen, die aus anderen abgeleitet wurden, berücksichtigt werden.

Tabelle 4.101: Variablen der Entscheidungsbäume

	Variable	Ausprägungen	Abkürzung	
Hospitationen & Kurse	Alter	Alter des Untersuchers In Jahren	Alter	A
	Qualifikation		**Qualifikation**	
	Klinikart derHospitationen	• Universitätsklinik (U) • nicht universitäre Klinik (nU) • keine	HospitationenKlinikart	HK
	Anzahl der Hospitationen		**HospitationenAnzahl**	
	Dauer der Hospitationen	Dauer der Hospitationen in Tagen	HospitationenTage	HT
	Anzahl der Kurse		**KursteilnahmeAnzahl**	
	Dauer der Kurse	Dauer der Kursteilnahme in Tagen	KursteilnahmeTage	KT
Ausbildung	Dauer der Ausbildung	Dauer der Bronchoskopieausbildung in Jahren	AusbildungJahre	AJ
	Klinikart der Ausbildung	• Universitätsklinik (U) • nicht universitäre Klinik (nU)	AusbildungKlinikart	AK
	Ausbildung in flexibler Bronchoskopie	Anzahl flexibler Bronchoskopien während der Ausbildung	AusbildungFlexibleAnzahl	AF

	Variable	Ausprägungen	Abkürzung	
	Ausbildung in starrer Bronchoskopie	Anzahl starrer Bronchoskopien während der Ausbildung	AusbildungStarreAnzahl	AS
	Ausbildung in interventioneller Bronchoskopie	**Anzahl interventioneller Bronchoskopien während der Ausbildung**	**AusbildungInterventionelleAnzahl**	**AI**
Erfahrung	Erfahrung in flexibler Bronchoskopie	Anzahl der seit der Ausbildung absolvierten flexiblen Bronchoskopien	ErfahrungFlexibleAnzahl	EF
		Anzahl der in Ausbildung und beruflicher Tätigkeit insgesamt absolvierten flexiblen Bronchoskopien	**ErfahrungFlexibleGesamtAnzahl**	**EG**
	Erfahrung in starrer Bronchoskopie	Anzahl der seit der Ausbildung absolvierten starren Bronchoskopien	ErfahrungStarreAnzahl	ES
		Anzahl der in Ausbildung und beruflicher Tätigkeit insgesamt absolvierten starren Bronchoskopien	**ErfahrungStarreGesamtAnzahl**	**EG**
	Erfahrung in Interventioneller Bronchoskopie	Anzahl der seit der Ausbildung absolvierten interventionellen Bronchoskopien	ErfahrungInterventionelleAnzahl	EI
		Anzahl der in Ausbildung und beruflicher Tätigkeit insgesamt absolvierten interventionellen Bronchoskopien	**ErfahrungInterventionelleGesamtAnzahl**	**EG**
	Klinikart der Erfahrung	• Universitätsklinik (U) • nicht universitäre Klinik (nU) • keine	ErfahrungKlinikart	EK

Liste der in die Konstruktion von Entscheidungsbäumen einbezogenen Variablen. Variablen, die in den linearen Modellen nicht vorkommen, sind fett hervorgehoben.

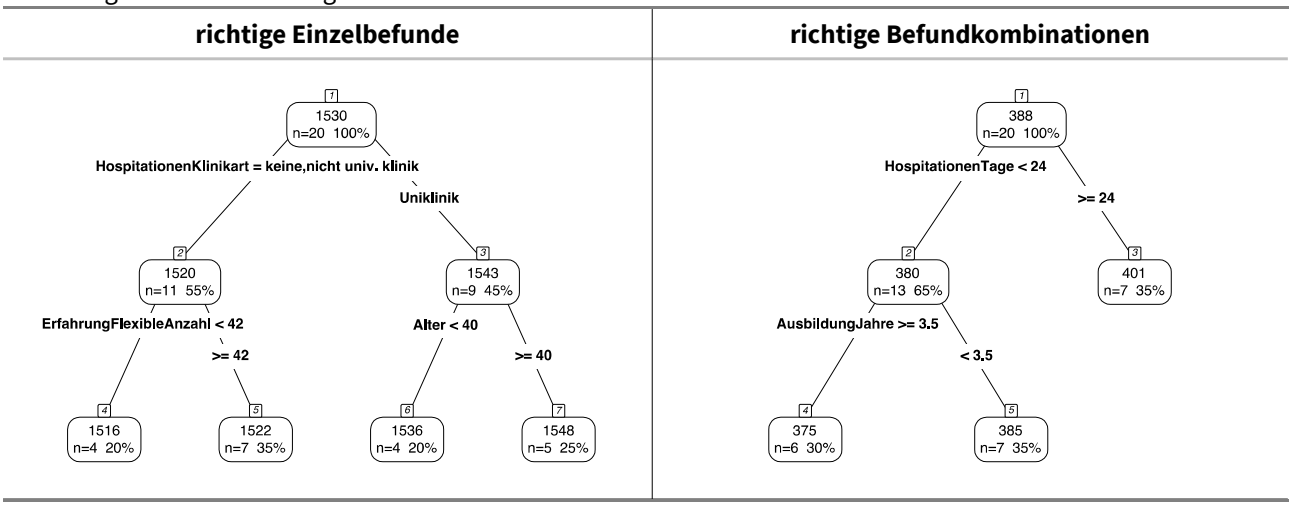
4.3.2.1 CART

Die Entscheidungsbäume wurden mit dem in der R-Bibliothek rpart (Therneau u. a., 2012) implementierten klassischen CART-Algorithmus (Breiman u. a., 1983) ohne pruning konstruiert. Die Ergebnisse der Analysen werden im Folgenden, separat nach Datensätzen (nativ/imputiert) und innerhalb der Datensätze, getrennt nach Zielvariablen (richtige Einzelbefunde/Kombinationsbefunde) dargestellt. Das einflussreichste Trennkriterium erzeugt die erste Auftrennung (engl. split). Nach unten verästeln sich die Bäume in weniger wichtige Trennkriterien. Die hinsichtlich der Zielvariable schwächere Gruppe wird dabei links aufgetragen, die stärkere rechts. CART partitioniert grundsätzlich immer in 2 Gruppen (engl. binary splits). Die Hyperparameter wurden so gewählt, dass Bäume geringer Komplexität wachsen, um eine Überanpassung (engl. overfitting) zu vermeiden. Im Gegenzug wurde auf ein nachträgliches Beschneiden der Bäume (engl. pruning) verzichtet. Die minimale Anzahl an Befundern bzw. Beobachtungen in einem weiterführenden Knoten (engl. minsplit) wurde auf 8, die minimale Anzahl in einem Endknoten (engl. minbucket) auf 4 festgelegt.

4.3.2.1.1 CART mit nativem Datensatz

Abbildung 4.54 gibt einen Überblick über die aus dem originalen Datensatz mit CART gewachsenen Bäume. Der Baum mit richtigen Einzelbefunden wählt 3, der Baum richtiger Kombinationsbefunde 2 Variablen für die Partitionierung aus, wobei sich – anders als bei den linearen Modellen – keine Überschneidungen der selektierten Variablen finden.

Abbildung 4.54: Überblick Regressionsbäume mit CART



Alle Bäume wuchsen mit den Hyperparametern minsplit = 8 und minbucket = 4. Datengrundlage war der native Arztfragebogen. Fehlwerte wurden von CART durch Surrogatvariablen ersetzt. An den Knotenpunkten ist die durchschnittliche Trefferquote aller Fälle des Knotens und darunter die Fallzahl im Knoten absolut und prozentual angegeben. Die Äste sind mit Variable und Schwellenwert der Partitionierung beschriftet.

4.3.2.1.1.1 CART Einzelbefunde nativer Datensatz

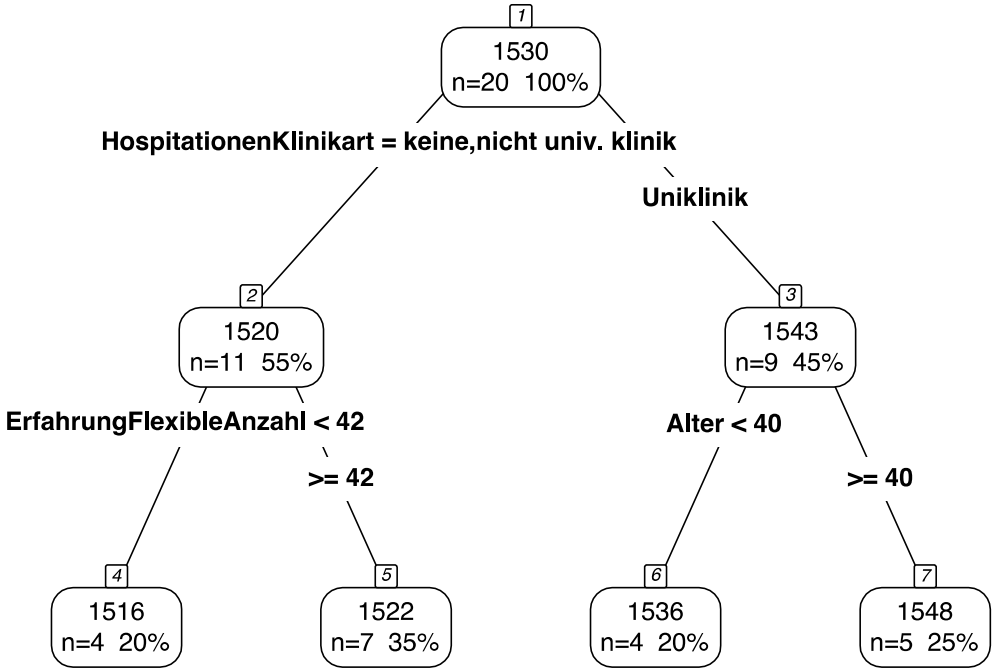
Setzt man richtige Einzelbefunden als Zielvariable, findet sich analog zum linearen Modell die Klinikart der Hospitationen als wichtigster Faktor. Dabei muss beachtet werden, dass die Variable Klinikart der Hospitationen, die eigentlich 3 Ausprägungen besitzt, von CART in eine binäre Variable mit den Klassen „keine Hospitation“ + „Hospitation an nicht universitärer Klinik“ einerseits und „Hospitation an universitärer Klinik“ andererseits, umgewandelt wird. Denn der CART-Algorithmus beherrscht ausschließlich binäre Auftrennungen. Abseits des Kliniktyps der Hospitationen kommt es gemäß dem Baum auf Alter und Erfahrung in flexibler Bronchoskopie an. Am besten schneiden Befunder ab, die Hospitationen an einer Universitätsklinik absolviert haben und über 40 Jahre alt sind, gefolgt von ihren jüngeren Kollegen. Befunder, die an einer außeruniversitären Klinik oder gar nicht hospitiert haben, können das durch Erfahrung in flexibler Bronchoskopie[32] teilweise kompensieren. Sie erreichen jedoch im Schnitt nicht mehr das Niveau der Kollegen, die an der Uniklinik hospitiert haben.

Die Kennwerte des Knotens 1 in Ausgabe 4.3 zeigen, dass die infrage kommenden primären Trennvariablen vom Thema Hospitationen bestimmt werden: Die Dauer der Hospitationen in Tagen sowie deren Anzahl liegen hinsichtlich ihres „improve“ nahe beieinander, nahe an der Klinikart der Hospitationen und setzen sich von der Anzahl flexibler Bronchoskopien während der Ausbildung deutlich ab. Dauer und Anzahl der Hospitationen wären die besten Surrogatvariablen der Klinikart der Hospitationen, auf die in Abwesenheit von Fehlwerten aber nicht zurückgegriffen werden musste. Im Knoten 2 dominieren Variablen der Erfahrung deutlich die der Ausbildung. Ein Fehlwert bei der Anzahl der flexiblen Bronchoskopien wurde durch die Gesamtzahl der in Ausbildung und beruflicher Tätigkeit absolvierten Bronchoskopien[33] ersetzt. Im Knoten 3 ist das Alter als Trennkriterium weitgehend alternativlos: Der improve sämtlicher anderen Variablen fällt deutlich geringer aus, als der des Alters. Obwohl es im Gesamtdatensatz bei der Variable Alter zwei Fehlwerte gibt, musste auch hier nicht auch Surrogatvariablen zurückgegriffen werden, da sich beide Fehlwerte in der Gruppe von Befundern mit Hospitationen außerhalb einer Universitätsklinik befinden.

[32] *mehr als 42 flexible Bronchoskopien seit der Ausbildung*

[33] *Die Gesamtzahl der in Ausbildung und beruflicher Tätigkeit absolvierten Bronchoskopien entspricht in diesem Fall der Anzahl der Bronchoskopien während der Ausbildung.*

Abbildung 4.55: CART Entscheidungsbaum richtige Einzelbefunde nativer Datensatz



An den Knotenpunkten ist die durchschnittliche Trefferquote aller Fälle des Knotens und darunter die Fallzahl im Knoten absolut und prozentual angegeben. Die Äste sind mit Variable und Schwellenwert der Partitionierung beschriftet.

R Ausgabe 4.3: Kennwerte CART richtige Einzelbefunde nativer Datensatz

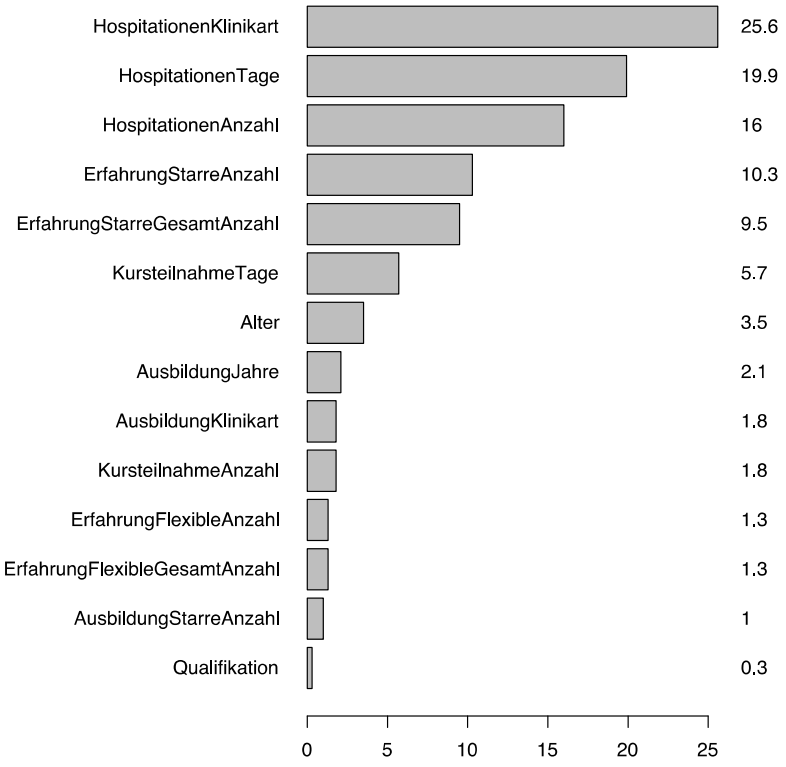
```
Node number 1: 20 observations,    complexity param=0.4793245
  mean=1529.9, MSE=273.39
  left son=2 (11 obs) right son=3 (9 obs)
  Primary splits:
      HospitationenKlinikart   splits as  LRL,       improve=0.4793245, (0 missing)
      HospitationenTage        < 24    to the left,  improve=0.4119373, (1 missing)
      HospitationenAnzahl      < 1.5   to the left,  improve=0.4037226, (1 missing)
      AusbildungFlexibleAnzahl < 15    to the right, improve=0.2244546, (5 missing)
      ErfahrungFlexibleAnzahl  < 42.5  to the left,  improve=0.2070328, (1 missing)
  Surrogate splits:
      HospitationenTage           < 24    to the left,  agree=0.90, adj=0.778, (0 split)
      HospitationenAnzahl         < 0.5   to the left,  agree=0.80, adj=0.556, (0 split)
      ErfahrungStarreAnzahl       < 150   to the left,  agree=0.70, adj=0.333, (0 split)
      ErfahrungStarreGesamtAnzahl < 160   to the left,  agree=0.70, adj=0.333, (0 split)
      KursteilnahmeTage           < 67.5  to the left,  agree=0.65, adj=0.222, (0 split)

Node number 2: 11 observations,    complexity param=0.0144506
  mean=1519.545, MSE=72.06612
  left son=4 (4 obs) right son=5 (7 obs)
  Primary splits:
      ErfahrungFlexibleAnzahl       < 42.5  to the left,  improve=0.17029820, (1 missing)
      ErfahrungFlexibleGesamtAnzahl < 105   to the left,  improve=0.09967235, (0 missing)
      AusbildungJahre               < 2.5   to the left,  improve=0.07631881, (3 missing)
      KursteilnahmeTage             < 5     to the right, improve=0.04177998, (0 missing)
      AusbildungStarreAnzahl        < 17.5  to the left,  improve=0.03165520, (0 missing)
  Surrogate splits:
      ErfahrungFlexibleGesamtAnzahl < 105   to the left,  agree=1.0, adj=1.00, (1 split)
      AusbildungStarreAnzahl        < 17.5  to the left,  agree=0.9, adj=0.75, (0 split)
      ErfahrungStarreGesamtAnzahl   < 31.5  to the left,  agree=0.9, adj=0.75, (0 split)
      Qualifikation                 splits as  LR,        agree=0.7, adj=0.25, (0 split)
      AusbildungJahre               < 1.5   to the left,  agree=0.7, adj=0.25, (0 split)
```

```
Node number 3: 9 observations,    complexity param=0.06555145
  mean=1542.556, MSE=228.2469
  left son=6 (4 obs) right son=7 (5 obs)
  Primary splits:
      Alter                       < 40.5  to the left,  improve=0.17448070, (0 missing)
      ErfahrungStarreAnzahl       < 4.5   to the left,  improve=0.05453267, (0 missing)
      ErfahrungStarreGesamtAnzahl < 8     to the left,  improve=0.05453267, (0 missing)
      HospitationenTage           < 107   to the right, improve=0.03218980, (1 missing)
      ErfahrungFlexibleAnzahl     < 155   to the right, improve=0.01006328, (0 missing)
  Surrogate splits:
      HospitationenAnzahl   < 1.5   to the left,  agree=0.778, adj=0.5, (0 split)
      KursteilnahmeAnzahl   < 0.5   to the left,  agree=0.778, adj=0.5, (0 split)
      AusbildungJahre       < 5.5   to the right, agree=0.778, adj=0.5, (0 split)
      AusbildungKlinikart   splits as  LR,        agree=0.778, adj=0.5, (0 split)
      ErfahrungStarreAnzahl < 102.5 to the left,  agree=0.778, adj=0.5, (0 split)
```

Ausgabe der Funktion rpart aus der gleichnamigen R-Bibliothek.

Die überragende Bedeutung von Hospitationen im Baummodell mit richtigen Einzelbefunden als Zielvariable drückt sich auch in der Rangliste der relativen Variablenwichtigkeit aus: Die drei ersten Plätze werden von den Variablen der Hospitation besetzt. Analog zum improve in Knoten 1 (R-Ausgabe 4.1) führt die Klinikart der Hospitationen die Rangliste mit erkennbarem Abstand an.

Abbildung 4.56: Rangliste der Variablenwichtigkeit CART richtige Einzelbefunde nativ



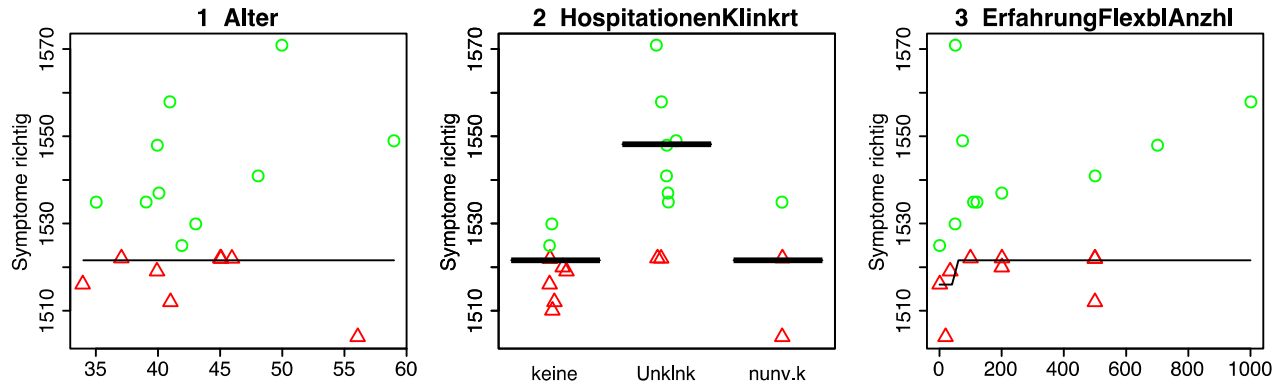
Rangliste der relativen Variablenwichtigkeit im Entscheidungsbaum: Werte summieren sich zu 100.

Die graphischen Darstellungen des Variableneinflusses in Abbildung 4.57a) macht visuell nachvollziehbar, warum die Klinikart der Hospitationen von CART als wichtigste Trennvariable ausgewählt wurde: In der Gruppe der Befunder mit Hospitationen an einer Uniklinik finden sich nur zwei Befunder der „schlechteren Hälfte“, hingegen zahlreiche exzellente Befunder. Die durchschnittliche Trefferzahl (schwarzer Balken) setzt sich hier deutlich von den beiden anderen Gruppen ab, die in etwa auf einem Niveau liegen und in denen Befunder aus der „schlechteren Hälfte“ überwiegen. Weder das Alter noch die Anzahl flexibler Bronchoskopien bewirken eine vergleichbar klare Auftrennung des Datensatzes.

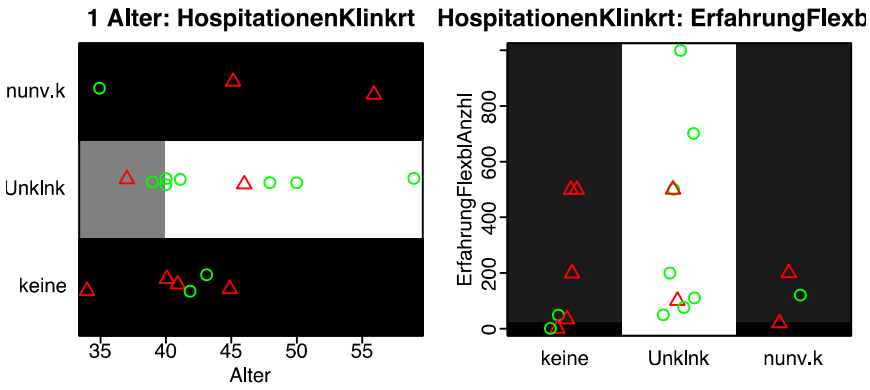
Innerhalb der Befundergruppe, die Hospitationen an einem universitären Haus absolviert haben, wirkt sich ein Alter von über 40 Jahren positiv auf die Genauigkeit aus. In der Gruppe ohne Hospitationen bzw. mit Hospitationen an außeruniversitären Häusern ist hingegen kein Effekt zu erkennen[34]. Die Erfahrung an flexiblen Bronchoskopien übt einen genau gegenläufigen Einfluss aus: für die Gruppe ohne bzw. mit außeruniversitären Hospitationen errechnet sich ein positiver Einfluss auf die Befundgenauigkeit oberhalb von 42 Bronchoskopien. Die Gruppe mit Uniklinik-Hospitationen bleibt von der Anzahl der absolvierten Bronchoskopien dagegen unbeeinflusst.

Abbildung 4.57: Interaktion im CART-Modell richtiger Einzelbefunde nativer Datensatz

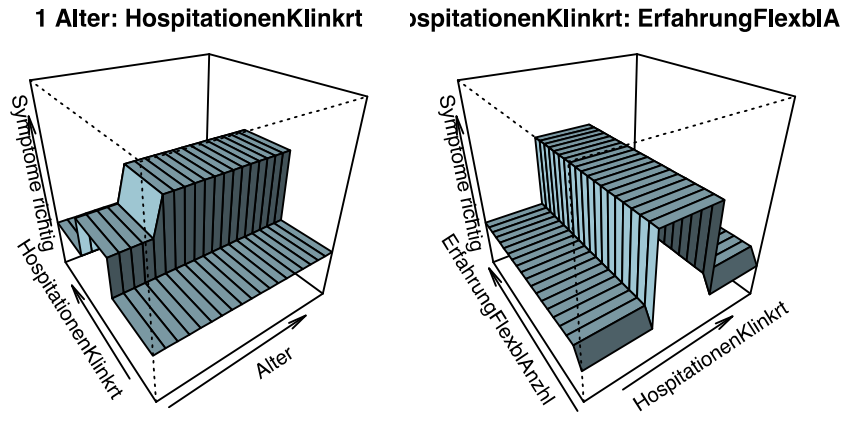
a) Einfluss der einzelnen Variablen



b) 2D Modell der Variableninteraktion



c) 3D Modell der Variableninteraktion



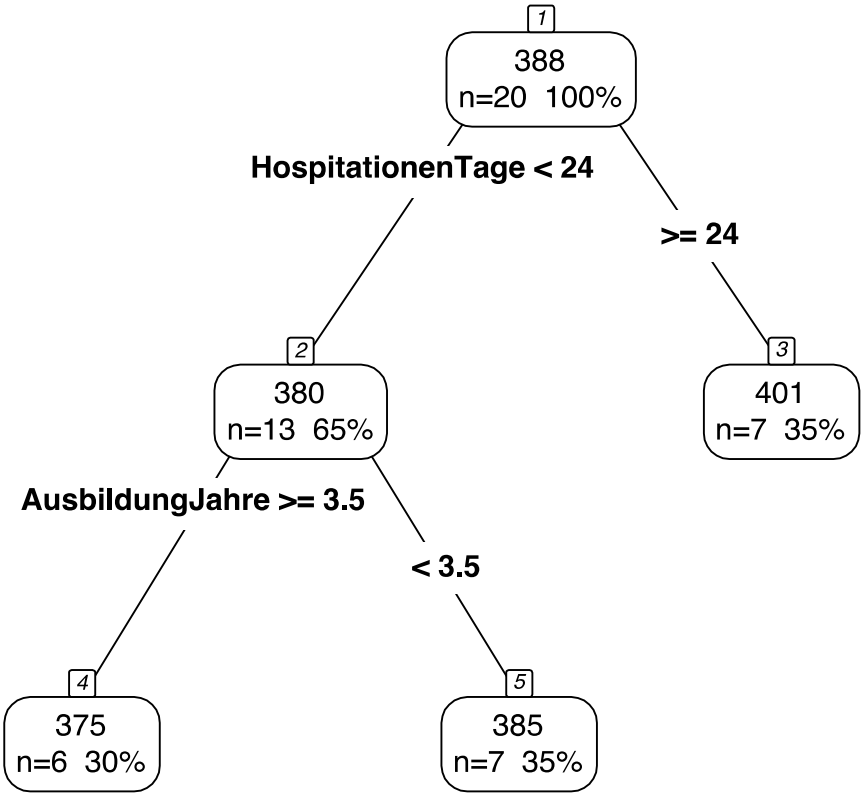
Die Diagramme der ersten Zeile illustrieren den Einfluss derjenigen Variablen aus dem Baummodell, die für eine Partitionierung herangezogen wurden, auf die insgesamt erreichten Treffer.

[34] *Abbildung 4.57c): 3D Modell linke Abbildung*

4.3.2.1.1.2 CART Befundkombinationen nativer Datensatz

Der Baum mit richtigen Befundkombinationen als Zielvariable schreibt ebenfalls Hospitationen eine zentrale Bedeutung zu, allerdings nicht der Klinikart, an der sie absolviert werden, sondern ihrer Dauer. Untersucher mit Hospitationen von mehr als 24 Tagen erzielen im Mittel 21 richtige Befunde mehr, als Ihre Kollegen. Innerhalb der Gruppe mit Hospitationen unter 24 Tagen ist die Dauer der Ausbildung ausschlaggebendes Kriterium: eine zu lange Ausbildung von über 3,5 Jahren wirkt sich negativ aus.

Abbildung 4.58: CART Entscheidungsbaum richtige Befundkombinationen



An den Knotenpunkten ist die durchschnittliche Trefferquote aller Fälle des Knotens und darunter die Fallzahl im Knoten absolut und prozentual angegeben. Die Äste sind mit Variablen und Schwellenwert der Partitionierung beschriftet.

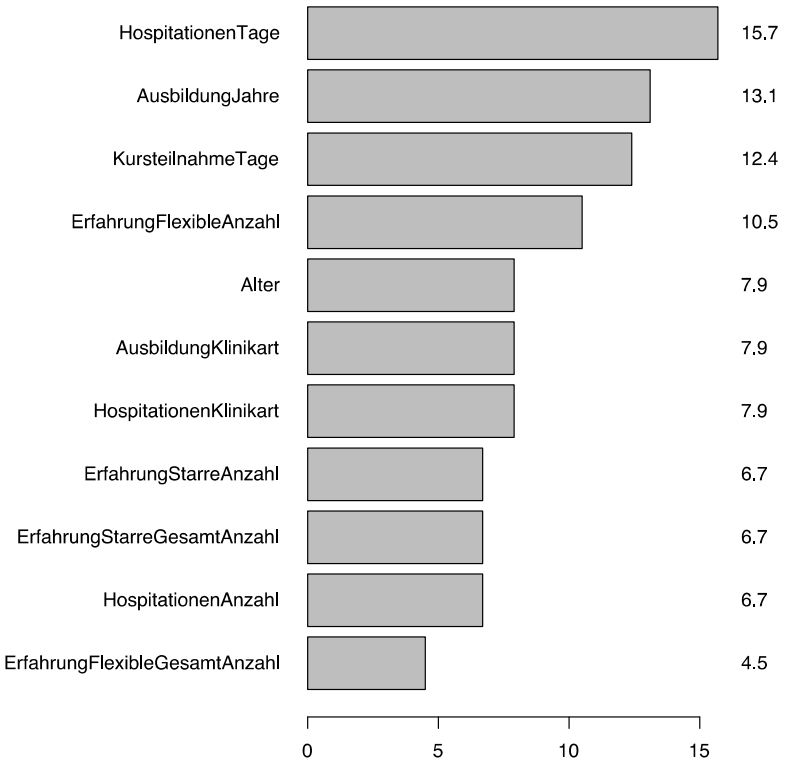
R Ausgabe 4.4: Kennwerte CART richtige Befundkombinationen nativer Datensatz

```
Node number 1: 20 observations,    complexity param=0.2132445
  mean=387.7, MSE=456.31
  left son=2 (13 obs) right son=3 (7 obs)
  Primary splits:
      HospitationenTage        < 24   to the left,  improve=0.21484870, (1 missing)
      AusbildungJahre          < 1.5  to the right, improve=0.20545240, (4 missing)
      AusbildungFlexibleAnzahl < 15   to the right, improve=0.14266770, (5 missing)
      HospitationenKlinikart   splits as  LRL,      improve=0.13710530, (0 missing)
      KursteilnahmeTage        < 2.5  to the right, improve=0.09854442, (2 missing)
  Surrogate splits:
      HospitationenAnzahl          < 0.5  to the left,  agree=0.789, adj=0.429, (0 split)
      ErfahrungStarreAnzahl        < 150  to the left,  agree=0.789, adj=0.429, (1 split)
      ErfahrungStarreGesamtAnzahl  < 160  to the left,  agree=0.789, adj=0.429, (0 split)
      KursteilnahmeTage            < 67.5 to the left,  agree=0.737, adj=0.286, (0 split)
      ErfahrungFlexibleGesamtAnzahl< 476  to the left,  agree=0.737, adj=0.286, (0 split)
```

```
Node number 2: 13 observations,    complexity param=0.03423083
  mean=380.4615, MSE=311.1716
  left son=4 (6 obs) right son=5 (7 obs)
  Primary splits:
      AusbildungJahre               < 3.5  to the right, improve=0.4050202, (3 missing)
      ErfahrungFlexibleGesamtAnzahl < 175  to the right, improve=0.3971419, (0 missing)
      Alter                         < 44   to the right, improve=0.3966584, (2 missing)
      KursteilnahmeTage             < 2.5  to the right, improve=0.3514114, (0 missing)
      KursteilnahmeAnzahl           < 0.5  to the right, improve=0.2684389, (0 missing)
  Surrogate splits:
      ErfahrungFlexibleAnzahl < 62.5 to the right, agree=0.9, adj=0.8, (2 split)
      Alter                   < 44   to the right, agree=0.8, adj=0.6, (0 split)
      HospitationenKlinikart  splits as  RLL,      agree=0.8, adj=0.6, (1 split)
      KursteilnahmeTage       < 2.5  to the right, agree=0.8, adj=0.6, (0 split)
      AusbildungKlinikart     splits as  RL,       agree=0.8, adj=0.6, (0 split)
```

Ausgabe der Funktion rpart aus der gleichnamigen R-Bibliothek.

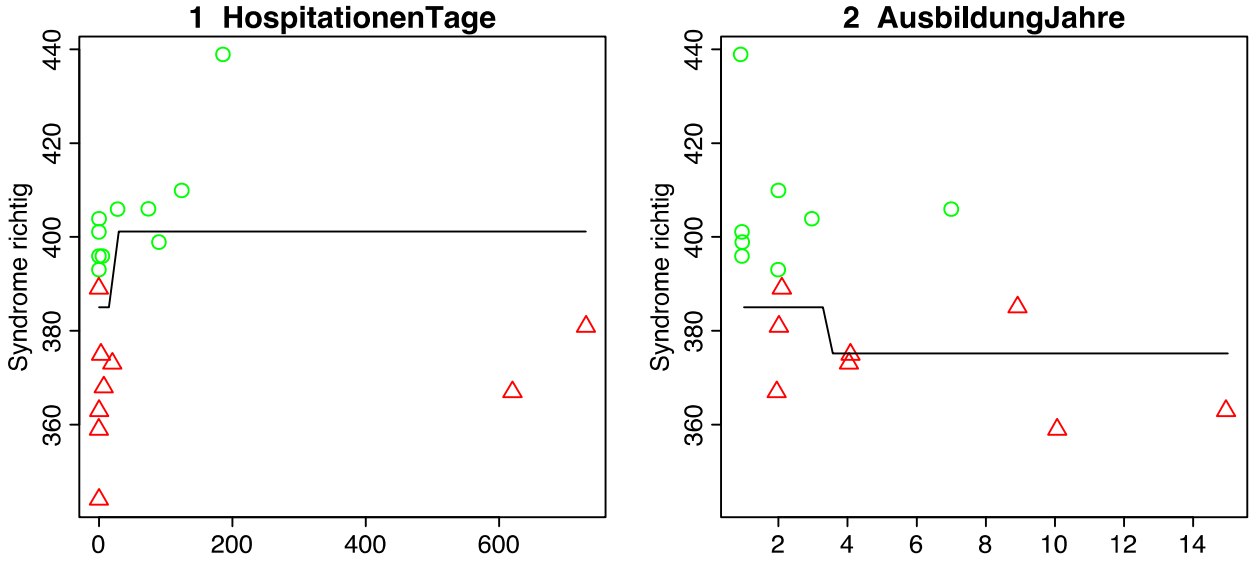
Abbildung 4.59: Rangliste Variablenwichtigkeit CART Befundkombinationen nativ



Rangliste der relativen Variablenwichtigkeit im Entscheidungsbaum: Werte summieren sich zu 100.

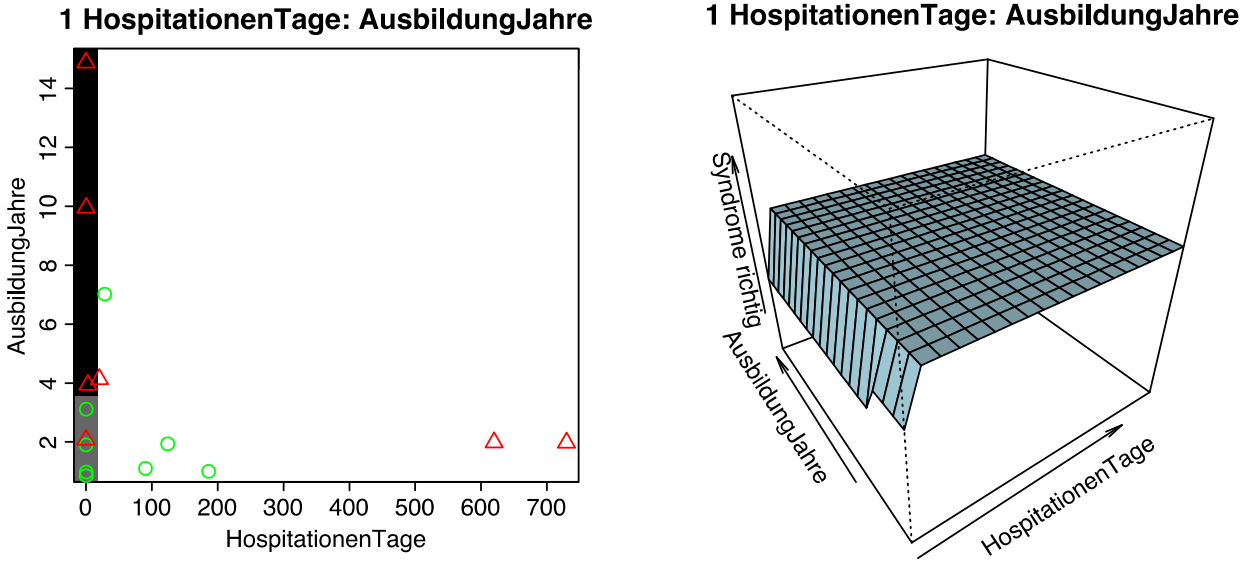
Abbildung 4.60: Variableninteraktion im CART-modell: Treffer Syndromebene

a) Einfluss der einzelnen Variablen



b) 2D Modell der Variableninteraktion

c) 3D Modell der Variableninteraktion

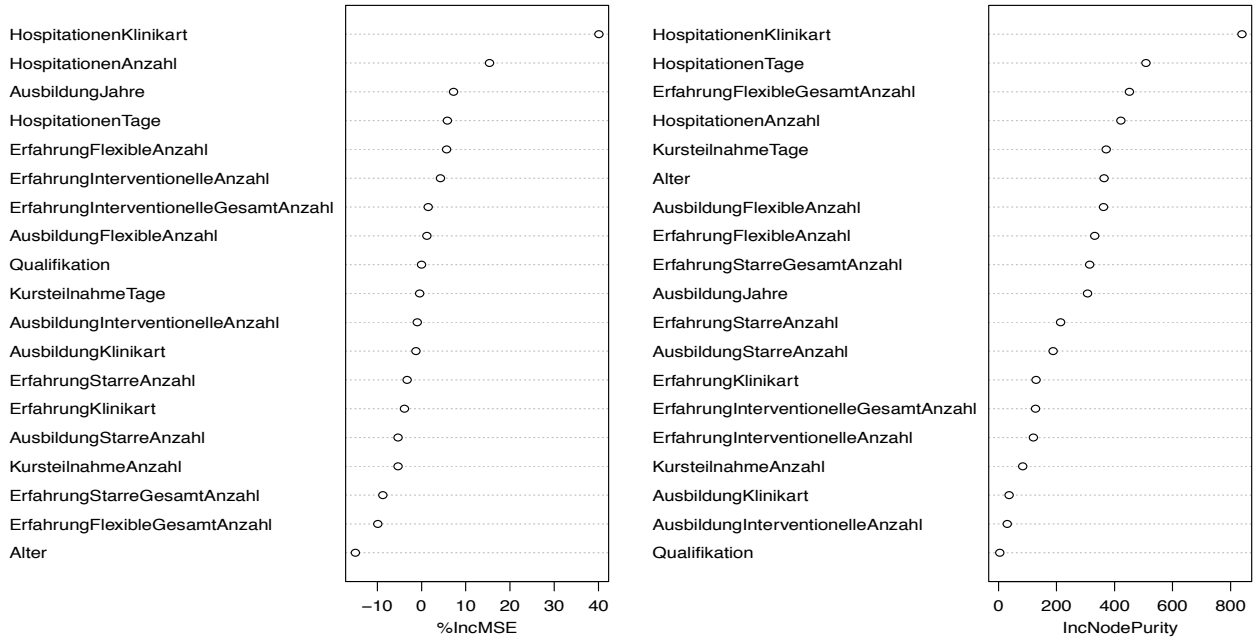


Die Diagramme der ersten Zeile illustrieren den Einfluss derjenigen Variablen aus dem Baummodell, die für eine Partitionierung herangezogen wurden. In der zweiten Reihe sind die Interaktionen dieser Variablen dargestellt.

4.3.2.2 Variablenwichtigkeit im random forest

Für eine robustere Abschätzung der relativen Wichtigkeit der Variablen, wird ihr Einfluss in einem Wald von Entscheidungsbäumen untersucht. Durch die hohe Zahl aus einer zufällig zusammengestellten Unterauswahl von Variablen konstruierten Bäumen, wird der Instabilität einzelner Bäume entgegen gewirkt. Die Wichtigkeit der Variablen wird quer über alle konstruierten Bäume bestimmt. Analog zur relativen Wichtigkeit bei linearen Modellen erhält man hierdurch Zahlenwerte für den Einfluss der einzelnen Variablen anhand derer eine Rangliste gebildet werden kann.

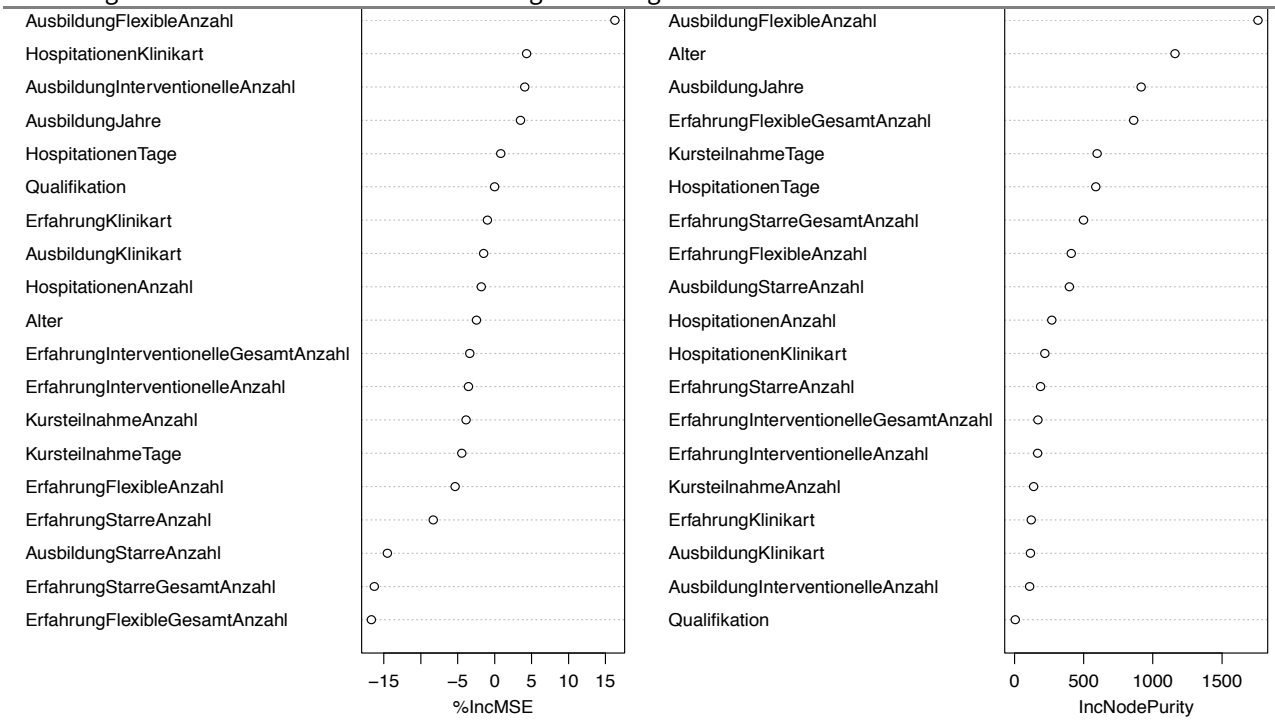
Abbildung 4.61: Random Forest Variablenwichtigkeit richtige Einzelbefunde



Relative Variablenwichtigkeit im random forest.

In Übereinstimmung mit den Ergebnissen der linearen Modelle setzt sich die Klinikart der Hospitation deutlich von sämtlichen anderen Variablen ab. Insgesamt führen die Variablen zur Hospitation das Feld klar an.

Abbildung 4.62: Random Forest Variablenwichtigkeit richtige Befundkombinationen



Relative Variablenwichtigkeit im random forest.

Zusammenfassung 4.17: Entscheidungsbäume

Mithilfe rekursiver Partitionierung konnte der ursprüngliche Datensatz, inklusive von Fehlwerten und abgeleiteten Variablen, direkt, d. h. ohne vorherige Imputation und in einem hypothesenfreien Ansatz untersucht werden. Rekursives Partitionieren modelliert Interaktionen und die Ergebnisse können in anschaulich interpretierbaren Entscheidungsbäumen dargestellt werden. Neben dem klassischen CART-Algorithmus wurden random forests erzeugt, mit denen u. a. die

relative Variablenwichtigkeit abgeschätzt werden kann. Bei richtigen Einzelbefunden als Zielvariable identifizierte CART – genau wie die linearen Modelle – die Klinikart der Hospitationen als einflussreichsten Faktor. Mit richtigen Befundkombinationen als Zielvariable waren die Tage der Hospitationen führend. Gemäß dem CART-Modell sollten Hospitationen mindestens 24 Tage dauern. Als nachgeordnete Variablen wurden bei Einzelbefunden die Erfahrung in flexibler Bronchoskopie sowie das Alter, bei Kombinationsbefunden die Ausbildungszeit ermittelt. Wendet man CART auf den über Imputation vervollständigten Datensatz an, bleibt bei richtigen Einzelbefunden als Zielvariable die Klinikart der Hospitationen erhalten, bei richtigen Kombinationsbefunden rücken die Ausbildungsjahre in den Vordergrund. Letztere werden sogar zweimal mit jeweils unterschiedlichen Grenzwerten zur Partitionierung herangezogen. Der Vergleich mit den Ergebnissen des nativen Datensatzes illustriert die Empfindlichkeit der Bäume gegenüber verhältnismäßig geringen Änderungen des Datensatzes (nur etwa 4 % wurden imputiert). Dieser Schwäche des rekursiven Partitionierens wirken random forests entgegen. Random forests zeigen, dass sich die Hospitation der Klinikart bei richtigen Einzelbefunden der Zielvariable deutlich gegen sämtliche anderen Einflussfaktoren absetzt. Anzahl und Dauer der Hospitationen folgen der Klinikart nach – je nachdem, welches Kriterium der Variablenwichtigkeit angewandt wird. Bei richtigen Befundkombinationen dominiert die Anzahl flexibler Bronchoskopien während der Ausbildung die übrigen Variablen ähnlich deutlich. Aus den Entscheidungsbäumen lassen sich ingesamt in etwa folgende Empfehlungen ableiten: Die Ausbildung in pädiatrischer Bronchoskopie sollte mindestens eine Hospitation an einer Universitätsklinik mit einer Dauer von mindestens 24 Tagen beinhalten. Eine Erfahrung von mindestens 42 eigenständig durchgeführten Bronchoskopien (nach der Ausbildung) scheint vorteilhaft. Eine zu hohe Anzahl flexibler Bronchoskopien während der Ausbildung (> 54) wirkt sich im Modell dagegen ebenso wie eine zu lange Ausbildungszeit (> 3,5 Jahre) eher negativ aus. Die Ergebnisse der Entscheidungsbäume sind weitgehend kongruent zu denen der linearen Modelle.

Literatur

Breiman, Leo (2001): „Random Forests“. In: *Machine Learning*. 45 (1), S. 5–32, DOI: 10.1023/A:1010933404324.

Breiman, Leo; Friedman, R. A.; Olshen, R. A.; u. a. (1983): *Classification and Regression Trees*. Wadsworth Publishing Co Inc. — ISBN: 0-534-98053-8

Feldman, Barry E. (2005): „Relative Importance and Value“. In: *SSRN Electronic Journal*., DOI: 10.2139/ssrn.2255827.

Groemping, Ulrike (2006): „Relative Importance for Linear Regression in R: The Package relaimpo“. In: *Journal of Statistical Software*. 17 (1), S. 1–27.

Grömping, Ulrike (2007): „Estimators of Relative Importance in Linear Regression Based on Variance Decomposition“. In: *The American Statistician*. 61 (2), S. 139–147, DOI: 10.1198/000313007X188252.

Grömping, Ulrike (2009): „Variable Importance Assessment in Regression: Linear Regression versus Random Forest“. In: *The American Statistician*. 63 (4), S. 308–319, DOI: 10.1198/tast.2009.08199.

Liaw, Andy; Wiener, Matthew (2002): „Classification and Regression by randomForest“. In: *R News*. 2 (3), S. 18–22.

Lumley, Thomas; Miller, Alan (2009): *leaps: regression subset selection*. o.V.

Miller, Alan J. (2002): *Subset selection in regression*. 2nd ed. Boca Raton: Chapman & Hall/CRC (Monographs on statistics and applied probability). — ISBN: 978-1-58488-171-1

R Core Team (2015): *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Stahel, Werner (2013): *The R-Function regr and Package regr0 for an Augmented Regression Analysis*. ETH Zürich.

Therneau, Terry M.; Atkinson, Beth; Ripley, Brian (2012): *rpart: Recursive Partitioning*. o.V.

Thompson, Austin B.; Huerta, Guillermo; Robbins, Richard A.; u. a. (1993): „The Bronchitis Index: A Semiquantitative Visual Scale for the Assessment of Airways Inflammation“. In: *Chest*. 103 (5), S. 1482–1488, DOI: 10.1378/chest.103.5.1482.

5 Diskussion

Kapitelverzeichnis

Zwanzig Ärzte beurteilten je 42 Videomitschnitte bronchoskopischer Untersuchungen, die das Spektrum der pädiatrischen Bronchoskopie weitgehend abdecken. Damit wurde erstmals in größerem Umfang die Inter-Beobachter-Variabilität in der pädiatrischen Bronchoskopie untersucht. Die berechneten Maßzahlen zur Verlässlichkeit der einzelnen bronchoskopischen Befunddomänen zeigte Stärken und Schwächen des Verfahrens auf. Mit einem Fragebogen zu Ausbildung und Erfahrung der Beobachter wurde versucht, Ursachen für die Abweichung vom Referenzbefund zu identifizieren. Die dabei ermittelten Einflussgrößen der Befundrichtigkeit können zur Gestaltung eines evidenzbasierten Ausbildungscurriculums in der pädiatrischen Bronchoskopie herangezogen werden. Der in Anlehnung an Vorarbeiten gestaltete Befundfragebogen (Caliebe, 1968; Deutsche Gesellschaft für Endoskopie, 1974; Ernst, Becker, 2001; Wunderlich, 1969) ist ein möglicher Ansatz für die Entwicklung eines zukünftig einheitlichen Befundschemas in der pädiatrischen Bronchoskopie. Eine einheitliche, gleichzeitig menschen- und maschinenlesbare („literate programming“ (Knuth, 1984)), Befunddokumentation wäre die Grundlage für

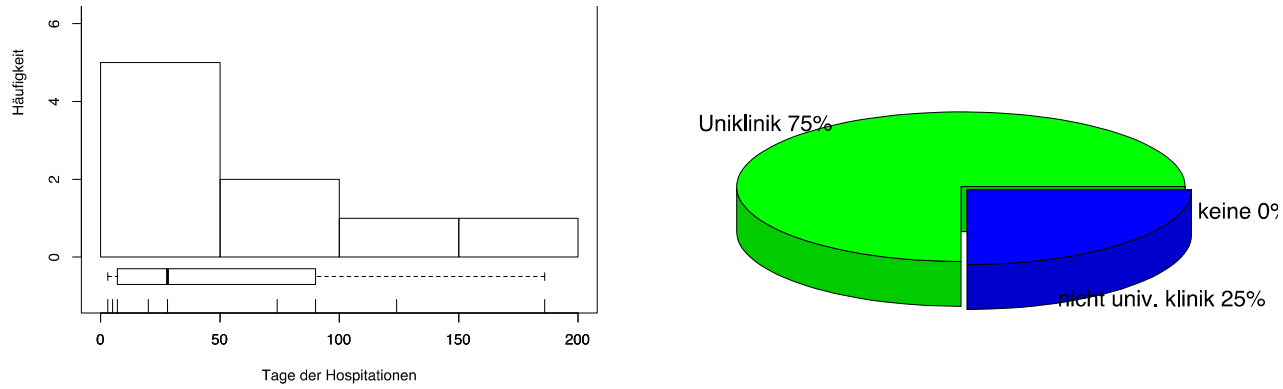
- die elektronische Archivierung in standardisierten Formaten wie z. B. DICOM SR (Hussein u. a., 2004a, 2004b),
- Qualitätsmanagement (Häussinger u. a., 2004),
- eine evidenzbasierte Ausbildung,
- epidemiologische Untersuchungen,
- telemedizinische Anwendungen und
- Expertensysteme (engl. clinical decision support, CDS) sowie
- multizentrische klinische Studien.

5.1 Untersucher

Die 20 ärztlichen Beobachter der Studie rekrutierten sich aus der Arbeitsgruppe Bronchoskopie der Gesellschaft für pädiatrische Pneumologie (GPP). Zum Zeitpunkt der Befundung waren sie bei einem Median von 42 Jahren zwischen 34 und 59 Jahre alt.

Bei 62 % der Teilnehmer waren Hospitationen Bestandteil der Ausbildung: im Median wurden 2,5 Hospitationen mit einer medianen Dauer von 28 Tagen absolviert. Drei Viertel der Hospitationen fanden an einer Universitätsklinik statt. Womöglich waren im Falle der Extremwerte eher Ausbildungsabschnitte gemeint, in denen die Bronchoskopie im Fokus stand. Folgt man dieser Interpretation, wurden im Falle der beiden extremsten Angaben hinsichtlich der Anzahl an Hospitationen (110 Hospitationen mit 730 Tagen und 20 Hospitationen mit 124 Tagen) in etwa wöchentliche Ausbildungsphasen praktiziert. Die Variabilität bei den Angaben zu Anzahl und Länge der Hospitationen spricht dafür, dass der Begriff „Hospitationen“ offenbar unterschiedlich interpretiert wird. Vielerorts scheint ein zeitlich begrenztes Curriculum von etwa einem Monat etabliert zu sein. In einigen Fällen wurden aber offenbar keine Hospitationen im Sinne eines zeitlich klar umrissenen Blockpraktikums angeboten, sondern praktische Fertigkeiten in Bronchoskopie, eher begleitend zur regulären Weiterbildung, in etwa wöchentlichen Abschnitten gelehrt. Solche Ausbildungsabschnitte wurden von den knapp 40 % Untersuchern ohne Hospitationen möglicherweise nicht im Sinne von „Hospitationen“, sondern im Sinne regulärer Ausbildung in Bronchoskopie verstanden.

Abbildung 5.1: Hospitationen



Links: Balkendiagramm der Hospitationsdauer. Rechts: Anteil der Hospitationen an Universitätskliniken.

Zwei Drittel der Untersucher (67 %) hatten in Ihrer Ausbildung einen Bronchoskopiekurs von im Median 7 tägiger Dauer besucht. Die Ausbildung dauerte im Median 2 Jahre und erfolgte in knapp zwei Dritteln der Fälle (65 %) an einer Universitätsklinik. Dabei wurden im Median 40, im Mittel 76 flexible Bronchoskopien durchgeführt. Median und Mittel liegen somit deutlich unter der im Ausbildungskatalog des Schwerpunktes „Kinderpneumologie“ derzeit (2016) geforderten Mindestanzahl von 100 flexiblen Bronchoskopien.

Praktische Erfahrung in starrer Bronchoskopie sammelten während der Ausbildung zwei Drittel der Teilnehmer (65 %), die im Median 20 und im Mittel 42 starre Bronchoskopien durchführten. In interventioneller Bronchoskopie wurden nur 15 % der Teilnehmer geschult. Sie gaben an 5, 6 bzw. 25 Interventionen während der Ausbildung durchgeführt zu haben. Offenbar bleiben Interventionen meist erfahrenen Kollegen vorbehalten und sind damit selten Bestandteil der Ausbildung. Dem entsprechend sind für die Disziplinen starre bzw. interventionelle Bronchoskopie im Ausbildungskatalog des Schwerpunktes „Kinderpneumologie“ keine Richtzahlen vorgesehen.

Bis auf eine Ausnahme und einen Fehlwert verfügten sämtliche Teilnehmer über die Anerkennung des Facharztes für Kinderheilkunde. Die Erfahrung in flexibler Bronchoskopie lag bei im Median 200 und im Mittel 285 Bronchoskopien. Knapp die Hälfte der Untersucher (9 bzw. 45 %) hatte mit im Median 50 Eingriffen (Mittel 86) Erfahrung in starrer Bronchoskopie. Ein knappes Drittel (6 bzw. 30 %) hatte auch Erfahrung in interventioneller Bronchoskopie gesammelt. Die Hälfte dieser Untersucher machte keine Angaben zur Anzahl der Eingriffe, die verbliebenen 3 Untersucher gaben 7, 80 bzw. 100 interventionelle Bronchoskopien an. Drei Viertel der Befunder sammelten ihre Erfahrung an einer Universitätsklinik.

Fazit 1: Untersucher

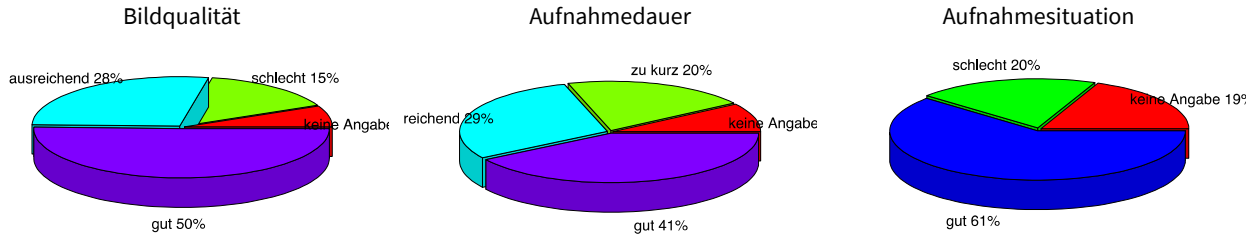
Die Befunder dieser Studie sind überwiegend Fachärzte aus dem universitären Umfeld, die bereits umfangreiche Erfahrung in der pädiatrischen Bronchoskopie gesammelt haben und als Spezialisten gelten dürfen.

5.2 Bildqualität

Die objektive Bildqualität war wegen der verschiedenen Geräte, auf denen das Videomaterial bei der Befundung dargestellt wurde, nicht bestimmbar, wurde aber im Wesentlichen durch das VHS-Videoformat als „Flaschenhals“ definiert. Aus Studien zum Einfluss von Datenkompressionsverfahren auf die Befundung, lässt sich analog schließen, dass das VHS-Format für die meisten Befundkategorien ausreichend sein sollte (Rabenstein u. a., 2002; Seidenari u. a., 2004). Eine Ausnahme hierzu bildet vermutlich die Schleimhautbeurteilung, für die eine möglichst hohe Auflösung von besonderer Bedeutung ist. So konnten wir z. B. zeigen, dass Hyperämie der wichtigs-

te prognostische Faktor des Befundes „Entzündung“ ist. Im Rahmen dieser Arbeit wurde eine subjektive Einschätzung der Videoqualität erhoben, die sich aus den Kriterien technische Bildqualität, ausreichende Aufnahmedauer und Aufnahmesituation zusammensetzt. Nach allen Kriterien wurde die Videoqualität überwiegend als „gut“ bewertet, wobei der subjektive Eindruck der Befunder erheblich variierte.

Abbildung 5.2: Videoqualität



Subjektive Bewertung von Bildqualität, Aufnahmedauer und Aufnahmesituation.

Fazit 2: Bildqualität

Da die objektive Bildqualität wegen unterschiedlicher Anzeigegeräte in dieser Studie nicht bestimmbar war, wurde eine subjektive Einschätzung erhoben, die sich aus Bildqualität, Aufnahmedauer und Aufnahmesituation zusammensetzt. Die Mitschnitte wurden zwar durchaus variabel, aber in allen Qualitäten überwiegend als „gut“ bewertet. In Zusammenschau mit Arbeiten, die keinen wesentlichen Einfluss von Kompressionsverfahren auf die Befundung zeigen konnten, nehmen wir an, dass die Bildqualität als limitierender Faktor der Befundübereinstimmung weitgehend ausgeschlossen werden kann. Ausnahmen hiervon sind womöglich Befunde zu Schleimhautbeschaffenheit, bei denen der Auflösung besondere Bedeutung zukommt.

5.3 Auswertung

Trotz der hohen Relevanz für die medizinische Versorgung und klinische Forschung finden sich Analysen der Übereinstimmung einer größeren Anzahl von Befundern in der medizinischen Literatur vergleichsweise selten. Meist werden nur 2, selten mehr als 5 Beurteiler miteinander verglichen. Es ist anzunehmen, dass hierfür mehrere Faktoren ausschlaggebend sind. Trotz der zunehmenden Durchdringung der Medizin mit elektronischer Datenverarbeitung beschränkt sich die maschinenlesbare Kodierung klinischer Informationen immer noch vorwiegend auf Diagnosen (ICD-10) und Abrechnungsdaten (DRG). Klinische Befunde werden weiterhin in unstrukturiertem Freitext erfasst, wobei weder eine einheitliche, definierte Terminologie, noch standardisierte Abkürzungen Verwendung finden. Abseits des DICOM-Formates für Medizinische Bildgebung sind kaum Datenformate für den elektronischen Austausch etabliert. Klinische Befunde sind systematischen oder gar zentrenübergreifenden Auswertungen damit weitgehend unzugänglich. In der Folge beschränken sich Analysen meist auf zusätzlich zur klinischen Routine im Rahmen von Studien erhobene Daten. Dabei wird meist auf Scores zurückgegriffen, die auf einzelne Phänomene abzielen. Klassifikationssysteme, welche die standardisierte Dokumentation einer umfangreicheren Untersuchung erlauben, fehlen. Die einzige uns bekannte Ausnahme bildet derzeit das AMDP-System in der Psychiatrie. Viele der für die statistische Auswertung der Inter-Untersucher-Variabilität geeigneten Maßzahlen, wie z. B. das Kappa nach Fleiss, sind in gängiger Statistiksoftware nicht implementiert bzw. nur umständlich oder eingeschränkt über die graphische Benutzeroberfläche zugänglich. Zwar gibt es Spezialanwendungen und auch Makros für die Tabellenkalkulation Excel (Gwet, 2011; Mackinnon, 2000), die eine komfortable Berechnung der Übereinstimmung einzelner Untersucherpaare ermöglichen. Varianten des klassischen Kappa Cohen, welche die Übereinstimmung zwischen mehreren Untersuchern erlauben, sind aber wenig ver-

breitet. Umfangreichere Analysen (z. B. paarweises Kappa bei mehreren Untersuchern, Kappa je Kategorie etc.) erreichen rasch eine Komplexität, die praktisch nur noch mithilfe einer Datenbank und maßgeschneidertem Programmcode bewältigt werden kann. Klare Kriterien zur Beurteilung von Kappawerten gibt es nicht. Die Vergleichbarkeit zwischen verschiedenen Studien ist problematisch, da Kappawerte von Prävalenz und dem jeweiligen Klassifikationssystem beeinflusst werden. Letzteres ist ein wichtiger Grund, warum einheitliche Befundsysteme einen entscheidenden Fortschritt für die Untersuchung der Inter-Beobachter-Variabilität und diagnostischen Genauigkeit bedeuten: da Anzahl und Ausprägung der Befundvariablen klar definiert sind, werden Kappawerte zentrenübergreifend vergleichbarer.

Tabelle 5.1: Treffer einer Pubmed-Recherche zu Multi-Rater-Kappas

	Kappa Fleiss	**Kappa Light**	**Kappa Conger**
+ „inter rater“	95	29	0
+ „inter observer“	83	78	0
+ „agreement“	330	483	1

Gemäß Rechercheergebnis im Herbst 2016.

Um die Qualität der Publikationen von Studien zur diagnostischen Genauigkeit zu verbessern wurden „Standards for Reporting of Diagnostic Accuracy“ (STARD) vorgeschlagen, an denen sich auch diese Studie orientiert.

5.3.1 Syndromale Übereinstimmung

„Das Ganze ist mehr als die Summe seiner Teile“ wird Aristoteles sinngemäß in seiner Metaphysik zitiert. Diese alte Weisheit gilt in besonderem Maße für klinische Befunde, bei denen mehrere Symptome ein Syndrom bilden. So sind die Einzelbefunde Erythem, Schwellung und Hypersekretion für sich genommen, womöglich noch bedeutungslos. Zusammen sind sie jedoch pathognomisch für eine Entzündung. Kombinierte bzw. mehrfache Stenosen sind therapeutisch anders zu bewerten, als ein isolierter lokaler Befund. Syndrome können jedoch bei herkömmlichen Auswertungsmethoden auf Ebene einzelner Variablen nicht untersucht werden. Im Rahmen dieser Studie wurde daher ein Verfahren zur Auswertung der syndromalen Übereinstimmung erprobt und mit den Ergebnissen der Einzelbefunde verglichen. Hierzu wurden Einzelbefunde zu Mehrfachbefunden aggregiert, indem die binären Werte der Einzelbefunde zu mehrstelligen Zahlen zusammengefügt wurden. Über diese mehrstelligen Kombinationen der Einzelbefunde entsteht für jedes beobachtete Syndrom ein eindeutiger Identifikator, mit dessen Hilfe die Verteilung der Syndrome konstruiert werden kann. So werden Befundkombinationen der statistischen Auswertung zugänglich. Hierdurch ergibt sich ein wesentlich differenzierteres Bild der Befundung als auf Ebene der Einzelbefunde. Auf diese Weise konnten aus Einzelbefunden auch Kappa-Werte für übergeordnete anatomische Abschnitte abgeleitet werden.

5.3.2 Kappa bei mehreren Befundern und Goldstandard

Das klassische Kappa nach Cohen ist nur für die Übereinstimmung zweier (gleichwertiger) Befunder definiert. Für die Übereinstimmung mehrerer Befunder wurden Varianten von Kappa entwickelt. Prominentester Vertreter ist das Kappa Fleiss (Fleiss, 1971), weitere Vertreter sind das Kappa Light (Light, 1971) und das Kappa Conger (Conger, 1980). Wie Kappa Cohen untersuchen diese Maßzahlen jedoch die Übereinstimmung zwischen gleichberechtigten Befundern. Uns ist keine Variante von Kappa bekannt, die den Vergleich mehrerer Befunder zu einer Referenz erlaubt. Vereinzelt wurden hierfür zwar Maßzahlen vorgeschlagen, wie z. B. der G-Index. Der G-Index ist jedoch derzeit in keiner gängigen Statistiksoftware verfügbar und wurde nach unserer Kenntnis in keiner medizinischen Publikation zur Beobachterübereinstimmung jüngeren Datums praktisch angewandt. Ein mögliches Vorgehen, um trotz der Beschränkung auf zwei Untersucher,

mit Kappa Cohen eine Gruppe von Befundern zu einer Referenz zu vergleichen, ist die paarweise Bestimmung von Kappa Cohen für jeden einzelnen Befunder und den Goldstandard mit anschließender Bildung des Mittelwertes. In dieser Arbeit wurde ein anderer Ansatz verfolgt, den wir als vereintes Kappa Cohen (engl. unified) bezeichnen. Beim vereinten Kappa Cohen werden die Befunder zu einem virtuellen Befunder vereint. Der virtuelle Befunder wird dann mit dem Goldstandard verglichen. Wie beim paarweisen Kappa Cohen wird der Befund der Referenz dabei, entsprechend der Anzahl der zu vergleichenden Befunder, mehrfach gewichtet. Das Modell des vereinten Kappa betont, dass es um eine Analyse der Gesamtheit der Befunder im Vergleich zum Goldstandard geht, weniger um einen Vergleich einzelner Befunder. Vorteil dieses Verfahrens ist u. a., dass Befunder und Referenz in einer einzigen Kontingenztafel verglichen werden können (anstatt in vielen separaten). Der Vergleich in einer einzigen Kontingenztafel ist insbesondere hinsichtlich der syndromalen Übereinstimmung von Bedeutung. Die Kontingenztafel des vereinten Kappas stellt das gesamte Spektrum aller gewählten Befundkombinationen dar. Beim paarweisen Kappa können die Befundkategorien in jeder einzelnen paarweisen Kontingenztafel variieren. In die Berechnung von Kappa Cohen können jedoch nur die zwischen Befunder(n) und Referenz überlappenden Befundkategorien einbezogen werden. Nicht überlappende Kategorien werden zu Fehlwerten, deren Anteil über die „Datenabdeckung“ angegeben wurde. Gemäß den Erfahrungen dieser Arbeit liegt das vereinte Kappa tendenziell leicht über dem paarweisen Kappa, aber in der gleichen Größenordnung.

Fazit 3: Auswertung

Im Rahmen dieser Studie wurden Varianten bzw. Erweiterungen klassischer Methoden der Auswertung der Inter-Beobachter-Variabilität etabliert. Als Maßzahl für die referenzunabhängige Übereinstimmung von Befundern in positiven Befunden wurde die „durchschnittliche positive Übereinstimmung“ vorgeschlagen. Durch Aggregation von Einzelbefunden zu Mehrfachbefunden wurde die Abstraktion vom Symptom zum Syndrom nachgebildet. Hierdurch können die Verteilungen von Befundkombinationen bzw. Syndromen untersucht und auf ihre Konkordanz innerhalb der Befunder und zum Goldstandard überprüft werden. So kann z. B. zwischen einfachen und mehrfachen Stenosen differenziert werden. Die Methode erlaubt auch aus Einzelbefunden Gesamtbefunde für übergeordnete Abschnitte zu konstruieren. Aus den Befunden zur supraglottischen, glottischen und subglottischen Stenoselokalisation kann beispielsweise ein Gesamtbefund zur Stenoselokalisation im Larynx abgeleitet werden. Diese Abstraktion ist über mehrere Ebenen möglich, sodass die Befunde aus Larynx, Trachea und Bronchien in einem zweiten Schritt zu einem Gesamtbefund für die unteren Atemwege fusioniert werden könnten. Das Kappa nach Cohen ist nur für zwei gleichberechtigte Beurteiler definiert. In dieser Arbeit wurde ein Berechnungsmodus erprobt, mit dem Kappa Cohen zum Vergleich einer Gruppe von Befundern mit einer Referenz herangezogen werden kann. Die Befunder werden dabei zu einem virtuellen Befunder vereint, der anschließend mit dem Goldstandard verglichen wird. Die Methode wird daher als „vereintes“ Kappa Cohen bezeichnet. Die Ergebnisse dieses Verfahren wurden mit dem Mittelwert aller paarweisen Kappawerte zwischen Befundern und Referenz verglichen, wobei beide Kappa-Werte stets in der gleichen Größenordnung lagen. Meist lagen die vereinten Kappawerte etwas über den paarweisen. Diese Verfahren sind als Vorschläge zu verstehen, wie methodische Lücken bei der Anwendung von Kappa zur Analyse von Syndromen und zum Vergleich einer Gruppe von Befundern mit einer Referenz geschlossen werden könnten.

5.4 Befundübereinstimmung

Mit dem Befundbogen wurden von den 20 teilnehmenden Untersuchern im multiple-choice-Verfahren standardisierte Befunde zu einer Bibliothek aus 42 Videomitschnitten erhoben, die den Großteil des in der Praxis relevanten Befundspektrums pädiatrischer Bronchoskopie repräsentiert. Der Befundbogen ging dabei auf Stenosen, die Schleimhautbeschaffenheit, Entzündung, Malazie, Pulsationen und Kompressionen ein. Die Übereinstimmung zum Goldstandard wurde nach den Kriterien Deskription, Präzision und Richtigkeit analysiert. Die Deskription beschreibt die Verteilung der Befunde bei den Untersuchern im Vergleich zum Goldstandard. Die Präzision untersucht die Übereinstimmung der Untersucher untereinander (Kenngröße Kappa Fleiss). Die Richtigkeit (Kenngröße Kappa Cohen) stellt die Übereinstimmung der Untersucher mit dem Goldstandard dar.

5.4.1 Hauptdiagnose

Die Hauptdiagnose wurde im Freitext angegeben und sekundär in die Kategorien „falsch", „richtig" und „ähnlich" klassifiziert. Die Kategorie „ähnlich" soll dabei der unterschiedlichen Nomenklatur der Freitextangaben gerecht werden und ist am ehesten im Sinne von „synonym" zu verstehen. Mit 45,2 % stimmte knapp die Hälfte der Diagnosen mit dem Referenzbefund des Goldstandards überein, 15,5 % der Diagnosen waren ähnlich. Insgesamt wurde also bei gut 60 % der Videos eine korrekte Diagnose gestellt. In 28,8 % wich die Diagnose erheblich ab, sodass sie im Kontext dieser Studie als falsch angesehen wurde. In 11 % der Fälle wurde keine Diagnose angegeben.

Die Verlässlichkeit von Diagnosen ist vergleichsweise schlecht untersucht und wird gerne überschätzt (Meyer AD u. a., 2013). Trotz einer zunehmend hoch technisierten Medizin trägt die Anamnese in der klinischen Praxis unverändert den größten Anteil zur Diagnosefindung bei (Hampton u. a., 1975; Peterson u. a., 1992). In einer Studie zur Differentialdiagnose bei primären Lungentumoren stellten 6 erfahrene Untersucher die abschließende Diagnose eines Malignoms (dichotome Skala) zu 57 - 77 % (Mittel 64,6 %) alleine aufgrund klinischer Angaben (Segnan u. a., 1992). Zusätzliche Röntgenbilder veränderten den Prozentsatz auf 78 - 92 % (Mittel 84,5 %), ein zusätzlicher bronchoskopischer Befund auf 70 - 94 % (Mittel 82,9 %) und histologische Befunde auf 90 - 99 % (Mittel 95,5 %). In der Literatur zu Bronchoskopien bei Kindern und auch Erwachsenen sind uns gegenwärtig (2016) keine Studien bekannt, die zum direkten Vergleich der Verlässlichkeit von Diagnosen herangezogen werden könnten. Am ehesten vergleichbar sind Untersuchungen aus der gastroenterologischen Endoskopie bei Erwachsenen. Diese Untersuchungen sind jedoch auf einzelne Diagnosen wie z. B. Polypen beschränkt und beziehen eine erheblich kleinere Anzahl von Untersuchern ein. Eine Studie, die umfangreiche endoskopische Befunde vergleicht, haben wir im Rahmen der ausführlichen Literaturrecherche nicht gefunden[35]. Eine der ausführlichsten Studien in der Endoskopie untersuchte die Klassifikation kleiner kolorektaler Polypen (< 1 cm) in Adenome bzw. Hyperplasien, wobei histologische Befunde als Goldstandard herangezogen wurden (Repici u. a., 2016). Eine Videobibliothek von 55 Polypen wurde von 6 Untersuchern gemäß der NICE[36] Klassifikation mit Unterstützung von FICE[37], einer virtuellen Farbdarstellung der Neoangiogenese und des Schleimhautreliefs, befundet. Die diagnostische Genauigkeit lag insgesamt bei 77 %. In der Subgruppe der subjektive unsicheren Diagnosen bei 64 %.

[35]Recherche über PubMed u. a. mit den Suchbegriffen „inter observer", „inter rater" und „diagnosis agreement"+ „bronchoscopy" bzw. „endoscopy"

[36]Narrow-band Imaging International Colorectal Endoscopic (NICE)

[37]Fujinon Spectral Imaging Color Enhancement (FICE) ist eine Alternative zum Narrow band imaging (NBI)

Die Übereinstimmung in der Hauptdiagnose von gut 60 % in der vorliegenden Studie erscheint somit erfreulich hoch. Insbesondere wenn man berücksichtigt, dass ohne jegliche Hintergrundinformation zu den Patienten befundet wurde, keine unterstützende bildgebende Techniken eingesetzt wurden (keine FICE) und Freitextdiagnosen bei breitem Befundspektrum angegeben wurden. Gleichzeitig wird deutlich, dass klinische Angaben zu Vorgeschichte und Beschwerden auch für die Bronchoskopie von entscheidender Bedeutung sind und eine isolierte Beurteilung ausschließlich auf Grundlage von Bildmaterial oft problematisch ist.

5.4.2 Stenosen

Die Beurteilung und Behandlung von Stenosen zählt zu den wichtigsten Domänen der pädiatrischen Bronchoskopie. Dabei geht es insbesondere um die Einschätzung von Grad und Lage der Stenose. Ergänzend wurden die morphologischen Kriterien kurzstreckig/langstreckig, membranös und ringförmig erfragt, Merkmale, die in der Differentialdiagnose von pneumologischen Krankheitsbildern hilfreich sind.

5.4.2.1 Stenosegrad

Der Stenosegrad wurde zunächst als freie Prozentzahl angegeben und sekundär leicht modifiziert (siehe Einleitung) nach Myer-Cotton klassifiziert. Die Untersucher sahen 441 Stenosen in 40 der 42 Videos, der Goldstandard 740 Stenosen in 36 Videos. Überraschend wurden trotz des relativ groben, vierklassigen Myer-Cotton-Schemas bei der Beurteilung des maximalen Stenosegrades nur eine geringe Übereinstimmung erzielt und das selbst bei höchstgradigen Stenosen. Das Kappa nach Fleiss als Maß der Übereinstimmung innerhalb der Befunder erreicht bei hochgradigen Stenosen Grad III und IV mit 0,264 bzw. 0,275 zwar die besten Werte, ist aber absolut gesehen im Sinne einer lediglich schwachen Übereinstimmung zu bewerten. Im Vergleich zum Goldstandard wurde in der Klasse III mit einem Kappa Cohen von 0,232 die beste Übereinstimmung erzielt, die ebenfalls schwach ist. Ausgerechnet bei höchstgradigen Stenosen findet sich hingegen kaum eine Übereinstimmung mit dem Goldstandard.

Tabelle 5.2: maximaler Stenosegrad (modifizierte Mayr-Cotton Klassifikation)

Befund	**Befundverteilung**				**Präzision** Übereinstimmung der Befunder			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz			
	Referenz	Befunder				Kappa Fleiss				Kappa Cohen	
	Anzahl verschiedener										
Stenosegrad Myer-Cotton Klassifikation	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**
gesamt	42 [840]	461	20	40	25,0	0,201	schwach	33,9	34,5	0,139	gering
keine Angabe	0	379	20	42	46,1	0,272	schwach	NA	NA	NA	NA
keine Stenose	5 [100]	20	4	14	11,0	0,013	keine	69,2	45,0	0,529	gut
Grad I	2 [40]	146	20	36	23,3	0,111	gering	60,0	4,1	0,038	keine
Grad II	8 [160]	100	19	32	20,5	0,078	kaum	29,6	29,0	0,1	kaum
Grad III	16 [320]	162	20	28	32,9	0,264	schwach	46,9	65,4	0,232	schwach
Grad IV	11 [220]	33	14	11	23,3	0,275	schwach	9,7	33,3	0,043	kaum

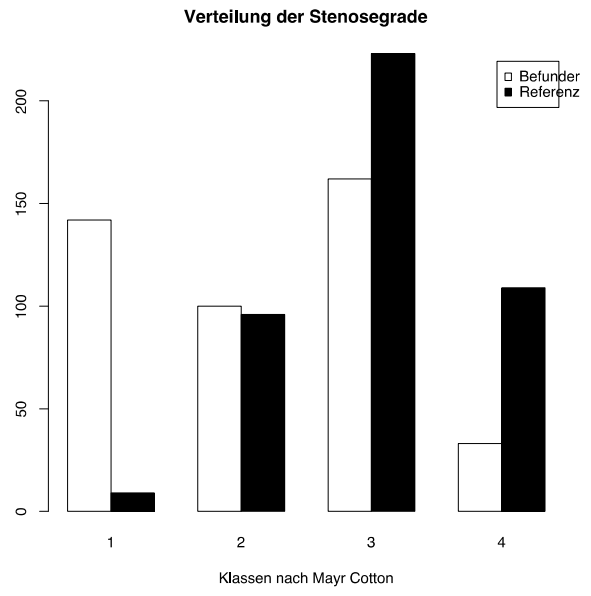
Zur besseren Vergleichbarkeit zwischen Referenz und Befundern ist in eckigen Klammern die Gewichtung der Befunde des Goldstandards angegeben. Da jeder Befund mit dem Goldstandard verglichen wird, ergibt sich bei 20 Befundern für die Gewichtung der Befunde des Goldstandards der Faktor 20.

Eine Erklärung für dieses kontraintuitive Ergebnis findet sich nicht nur in den unterschiedlichen absoluten Häufigkeiten von Stenosebefunden (441 vs. 740), sondern auch in unterschiedlichen Häufigkeiten innerhalb der vier Befundklassen der (leicht modifizierten) Myer-Cotton-Klassifkation. Das paarige Histogramm der Befundhäufigkeiten (Abbildung 5.3a) zeigt deutlich divergente Prävalenzen zwischen Untersuchern und dem Goldstandard in den einzelnen Befundklassen: während die Untersucher viele geringradige Stenosen der Klasse I sahen, aber nur wenig höchstgradige der Klasse IV, ist es beim Goldstandard genau umgekehrt. Nur in den beiden mittleren Klassen (II und III) gibt es in etwa vergleichbare Prävalenzen. Der McNemar Test – als Indikator erheblich abweichender Randsummen –fällt nur in Klasse II nicht signifikant aus. Die Übereinstimmung zwischen Untersuchern und dem Goldstandard bei höchstgradigen Stenosen ist also schon alleine aufgrund unterschiedlicher Prävalenzen – absolut, wie auch innerhalb der Stenosegrade – gering. Auch Maßzahlen wie z. B. Kappa werden erheblich von unterschiedlichen Randsummen beeinträchtigt (Gwet, 2002).

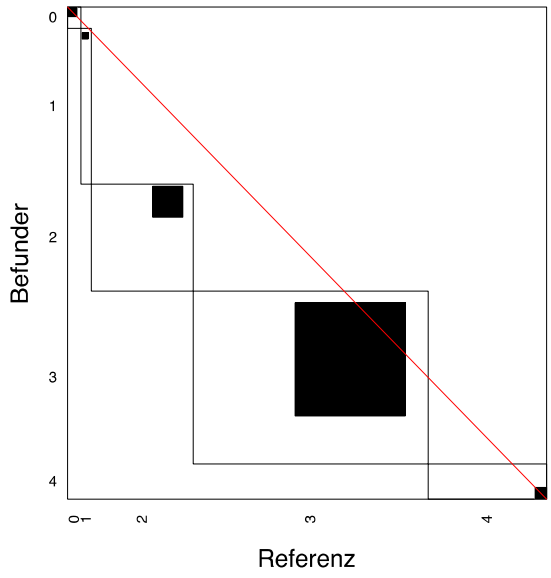
Die Fehlklassifikationen im Vergleich zum Goldstandard sind jedoch auch in den Klassen 2 und 3 nach Myer-Cotton mit vergleichbaren Befundhäufigkeiten erheblich. Die unterschiedlichen Befundprävalenzen können also nur einen Teil der geringen Übereinstimmung erklären. Wie aus der Kontingenztafel (Tabelle 5.3) hervorgeht und in Abbildung 5.4 illustriert wird, werden die Stenosen keineswegs nur in benachbarte Klassen eingestuft. Höchstgradige Stenosen der Klasse 4 werden nicht nur in Klasse 3, sondern häufig auch in Klasse 2 und sogar Klasse 1 eingestuft. Einige Untersucher sahen sogar gar keine Stenose. Stenosen der Klasse 3 werden fast ebenso häufig als zweitgradig wie auch als erstgradig bewertet. Mittelgradige Stenosen der Klasse 2 werden meist als erstgradig unterschätzt. Nur geringradige Stenosen der Klasse 1 zeigen die erwarteten, steil abfallenden Säulen in benachbarten Befundklassen.

Abbildung 5.3: Vergleich Befundung Stenosegrade nach Myer-Cotton (modifiziert)

a) gepaartes Histogramm



b) Bangdiwala-Diagramm



a) Die paarigen Balkendiagramme stellen die Befundhäufigkeiten der Stenosegrade gemäß der (modifizierten) Myer-Cotton-Klassifikation bei den Untersuchern denen des Goldstandards gegenüber. Die Befundhäufigkeiten sind in der Kontingenztafel (Tabelle 5.3) als Randsummen repräsentiert. b) Das Bangdiwala-Diagramm vergleicht die bei Abwesenheit von Fehlklassierungen gemäß Randsummen zu erwartenden Häufigkeiten (Rechtecke) mit den beobachteten Häufigkeiten (schwarze Quadrate) für die von Untersuchern und Goldstandard einheitlich klassierten Fälle. Bei gleichen Prävalenzen zwischen Untersuchern und Goldstandard werden die Rechtecke zu Quadraten (siehe Klasse 2 nach Mayr-Cotton). Bei gleichen Prävalenzen in allen Klassen liegen die Schnittpunkte der Rechtecke auf der Diagonalen.

In den Klassen 2 - 4 nach Myer-Cotton ist eine klare Tendenz zur Unterschätzung des Stenosegrades erkennbar. Die Untersucher schätzen die in den Videos gezeigten Stenosen also offenbar generell etwas weniger ausgeprägt ein als der Goldstandard. Studien zur Größenbeurteilung in der Endoskopie beschrieben ebenfalls eine systematische Unterschätzung der Größe von Objekten (Fennerty u. a., 1993; Margulies u. a., 1994). Im Unterschied zur hier beobachteten Tendenz war die Bezugsgröße jedoch die tatsächliche Objektgröße, was z. B. durch Distorison der Linsen des Bronchoskops (Masters u. a., 2005) erklärt werden kann.

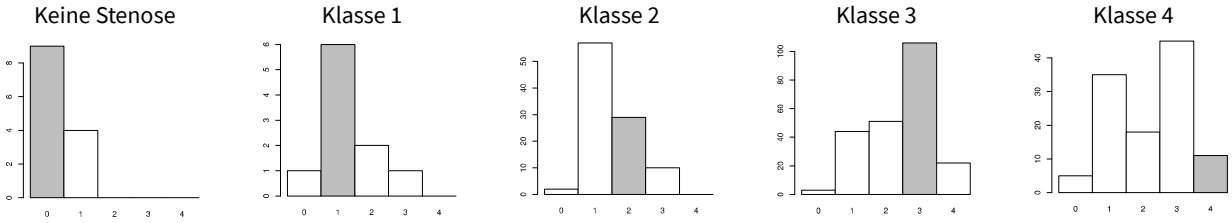
Möglicherweise spielen auch psychophysikalische Effekte bei der Beurteilung des Stenosegrades eine Rolle. Untersuchungen aus der Kartographie (Chang, 1977; Cox, 1976; Ekman, Junge, 1961; Flannery, 1971; Meihoefer, 1973), Statistik (Cleveland u. a., 1982; Croxton, 1932) und Psychologie (Teghtsoonian, 1965) zeigen eine tendenzielle Unterschätzung der Verhältnisse von Kreisflächen. Die Verhältnisse von Balken zueinander werden verlässlicher geschätzt, als die von Kreisen (Croxton, 1932: S. 55–56): Kreisflächen mit Faktor 2 werden im Durchschnitt als 1,7 fach eingeschätzt, Kreisflächen, die sich um den Faktor 10 unterscheiden, dagegen nur als um Faktor 6 differierend empfunden (Teghtsoonian, 1965: S. 394). Das Webersche Gesetz sagt voraus, dass der kleinste wahrnehmbare Unterschied mit zunehmender Größe des Reizes abnimmt. Demnach wären Unterschiede bei höhergradigen Stenosen gröber beurteilbar als bei niedriggradigen – vielleicht ein Grund für die bei höhergradigen Stenosen beobachtete deutlich höhere Variabilität im Vergleich zu niedriggradigen Stenosen (Abbildung 5.4).

Tabelle 5.3: Kontingenztafel maximaler Stenosegrad

Mc Nemar <0,01		**Goldstandard** keine Stenose	Grad I (<50%)	Grad II (51-70%)	Grad III (71-90%)	Grad IV (>90%)	Randsummen
	keine Angabe	87	30	62	94	106	379
Befunder	keine Stenose	9	1	2	3	5	20
	Grad I < 50%	4	6	57	44	35	146
	Grad II 51-70%	0	2	29	51	18	100
	Grad III 71-90%	0	1	10	106	45	162
	Grad IV >90%	0	0	0	22	11	33
		13	10	98	226	114	461
	Randsummen	100	40	160	320	220	840

Der McNemar Test prüft die Homogenität der Randsummen. Das signifikante Testergebnis belegt erheblich divergierende Befundverteilungen zwischen den Befundern und dem Goldstandard.

Abbildung 5.4: Befundverteilungen der Untersucher in der Myer-Cotton-Klassfikation



Die Säulendiagramme illustrieren die Befundverteilungen der Untersucher in den einzelnen Klassen der Myer-Cotton Klassifikation. Die Säule der (gemäß dem Goldstandard) jeweils korrekten Befundklasse ist grau hervorgehoben.

Insgesamt suggerieren die Daten, dass die Beurteilung des Stenosegrades rein auf Grundlage der Videomitschnitte unzuverlässig ist. Eine mögliche Erklärung hierfür könnten klinische Angaben und Voruntersuchungen zu den Beschwerdebildern der Patienten sein, die zwar dem Goldstandard, nicht aber den Untersuchern bekannt waren. So wird die Indikation zu einer Bronchoskopie z. B. häufig aufgrund einer klinisch apparenten Dyspnoe bzw. eines Stridors gestellt, was sicherlich einen Beobachtungsbias hinsichtlich des zu erwartenden Stenosegrades erzeugt.

5.4.2.2 Stenoselokalisation

Die Stenoselokalisation konnte mit einem auf die Anwendung in der Bronchoskopie zugeschnittenem vereinfachten Schema der Atemwege beurteilt werden, das den Larynx in supraglottisch, glottisch und subglottisch differenzierte, die Trachea in ein proximales, mittleres und distales Drittel unterteilte und den Hauptbronchus in rechts und links. Die Lappenbronchien wurden rechts in Ober-, Mittel- und Unterlappen, links in Oberlappen, Lingula und Unterlappen eingeteilt. Befundverteilung, Präzision und Kappa wurden sowohl auf Ebene der Einzelbefunde, als auch syndromal auf Ebene der jeweils übergeordneten anatomischen Abschnitte Larynx, Trachea, Hauptbronchus und Lappenbronchien analysiert. Abschließend wurden die Stenoselokalisationen auch abschnittsübergreifend im Sinne eines Gesamtbefundes der Stenoseloklisation interpretiert.

5.4.2.2.1 Larynx

Supraglottische und glottische Stenosen wurden nur in etwa einem Viertel der Fälle erkannt, subglottische Stenosen hingegen in fast 60 % der Fälle. Der Befund einer supraglottischen Stenose war mit einem positiven prädiktiven Wert von ca. 60 % am unzuverlässigsten. Wurden glottische Stenosen erkannt, war dieser Befund mit einem positiven prädiktiven Wert von ca. 75 % genauso verlässlich wie der Befund subglottischer Stenosen. Stenosen im Larynx werden sowohl innerhalb der Befunder untereinander (Kappa Fleiss) als auch im Vergleich zum Goldstandard (Kappa Cohen) mit von proximal nach distal zunehmender Verlässlichkeit erkannt. Das gilt auch für die prävalenzunabhängigen Maßzahlen Odds ratio (OR) und Area Under Curve (AUC). Die Kontingenztafel und das Assoziationsdiagramm der Kombinationsbefunde zeigen, dass singuläre Stenoselokalisationen im Larynx vergleichsweise sicher erkannt werden, bei Kombinationen von Stenosen in mehreren Abschnitten gibt es so gut wie keine Übereinstimmung. Der überwiegende Anteil der Fehlklassifikationen geht auf falsch negative Befunde zurück. Die Uneinigkeit zwischen dem Goldstandard und den Untersuchern besteht also weniger in der unterschiedlichen Beurteilung der Lage von Stenosen als in der prinzipiellen Frage ihrer Existenz. Eine Erklärung dafür, dass die Lage subglottischer Stenosen mit Abstand am zuverlässigsten beurteilt wird, könnte darin liegen, dass die Prävalenz subglottischer Stenosen im Klinikalltag vermutlich am höchsten ist und die Untersucher hierdurch im Erkennen dieses Befundes besonders erfahren sind.

5.4.2.2.2 Trachea

Stenosen im mittleren Drittel der Trachea werden mit einer Sensitivität von knapp 50 % am sichersten erkannt. Ihr positiver prädiktiver Wert ist mit ebenfalls knapp über 50 % aber am niedrigsten. Der positive prädiktive Wert von Stenosen im proximalen und distalen Drittel liegt mit ca. 70 % deutlich darüber. Die Analyse der Kombinationsbefunde zeigt, dass nennenswerte Übereinstimmungen sowohl zwischen den Befundern, als auch zum Goldstandard nur bei singulären Stenoselokalisationen zu beobachten sind, wobei der Befund einer singulären mittleren Trachealstenose gemäß dem Goldstandard nicht vorhanden war. Wie im Larynx ist auch in der Trachea bei den Einzelbefunden eine Zunahme der Präzision und Richtigkeit von proximal nach distal zu beobachten.

5.4.2.2.3 Hauptbronchus

Auf Ebene der Einzelbefunde wird die Stenoselokalisation in beiden Hauptbronchien in der Hälfte der Fälle erkannt. Der prädiktive Wert von Stenosen im rechten Hauptbronchus liegt mit 66,7 über dem von Stenosen im linken Hauptbronchus mit 61,2. Beidseitigen Stenosen wurden wie linksseitige Stenosen mit einer Sensitivität von gut 30 % erkannt. Das Kappa Cohen für beidseitige Stenosen liegt mit 0,392 zwischen dem der linksseitigen und rechtsseitigen Trachealstenosen.

Sowohl auf Ebene der Einzelbefunde als auch auf Ebene der Kombinationsbefunde zeigte sich im linken Hauptbronchus eine im Vergleich zu rechts einheitlichere Befundung der Untersucher untereinander (Kappa Fleiss). Die Übereinstimmung mit dem Goldstandard ist hingegen bei Einzelbefunden wie auch Kombinationsbefunden rechtsseitig besser. Falsch negative Befunde machen bei linksseitigen wie auch bei rechtsseitigen Stenosen des Hauptbronchus einen erheblichen Anteil der Fehlklassifikationen aus, sodass die Kappawerte in der Subgruppenanalyse positiver Befunde deutlich ansteigen.

5.4.2.2.4 Lappenbronchien

Insgesamt gesehen, sind die für die rechten Lappenbronchien erhobenen Zahlen aufgrund der geringen Prävalenz als nicht repräsentativ einzustufen. Interessant ist jedoch, dass sich der wesentliche Anteil der Differenzen auch hier aus falsch-negativen Befunden speist. Die Aussagekraft der Studie zu Stenosen im linken Lappenbronchus ist eingeschränkt, da sie gemäß dem Referenzbefund des Goldstandards nur im linken Unterlappen zu sehen waren. Dieser Befund erreicht allerdings sowohl innerhalb der Untersucher mit einem Kappa Fleiss von 0,375, als auch im Vergleich zum Goldstandard mit einem Kappa Cohen von 0,615 Spitzenwerte. Mit einem prädiktiven Wert von 100 % sind Stenosen im linken Unterlappen der valideste Befund der Studie. Abseits von Stenosen im linken Unterlappen bewegt sich die Konkordanz der Untersucher auf Zufallsniveau.

5.4.2.2.5 Vergleich der anatomischen Abschnitte der Stenoselokalisation

Beim Vergleich der anatomischen Abschnitte werden Stenoselokalisationen im Hauptbronchus am präzisesten und abgesehen von Stenosen im linken Lappenbronchus auch am genauesten erkannt. Ähnlich einheitlich werden Larynxstenosen von den Untersuchern beurteilt, Trachealstenosen fallen dem gegenüber leicht ab. Deutlich darunter liegen in der Präzision Stenosen in den Lappenbronchien, wobei bei linksseitigen Stenosen erheblich mehr Befundeinigkeit besteht als bei rechtsseitigen. Die Übereinstimmung mit dem Goldstandard ist in Larynx und Trachea mäßig, im Hauptbronchus moderat. Überraschend ist die Richtigkeit bei Stenosen im linksseitigen Lappenbronchus, die auf Stenosen im linken Unterlappen basieren. Dieser Befund wurde trotz niedriger Prävalenz am genauesten erkannt.

Die Überlegenheit bei der Beurteilung des Hauptbronchus und des Larynx hängt vermutlich mit den visuell eindeutig bestimmbaren anatomischen Lokalisationen zusammen. In diesem Kontext ist es überraschend, dass trotz der vagen Untergliederung der Trachea in proximales, mittleres und distales Drittel eine verhältnismäßig einheitliche Beurteilung stattfindet.

5.4.2.2.6 Kombinationsbefunde der Stenoselokalisation quer über Abschnitte

Ergänzend wurden auch abschnittsübergreifende Kombinationsbefunde von Stenoselokalisationen untersucht. Dafür wurde die Befundung nachträglich so vereinfacht, dass positive Befunde in einem beliebigen anatomischen Unterabschnitt als positiver Befund des übergeordneten Abschnittes gewertet wurden. Hierdurch kann – analog zur syndromalen Analyse innerhalb der anatomischen Abschnitte – zwischen Einfach- und Mehrfachstenosen differenziert werden. Zudem ist es möglich auf diese Weise Präzision und Richtigkeit für die Stenoselokalisation insgesamt anzugeben.

Präzision und Richtigkeit sind bei Stenosen in nur einem anatomischen Abschnitt am besten. Dabei liegt die Präzision in allen Abschnitten – mit Ausnahme des rechten Lappenbronchus - gleich auf. Die Übereinstimmung mit dem Goldstandard ist bei Trachealstenosen am größten, gefolgt von Larynxstenosen und Stenosen im Hauptbronchus. Im Lappenbronchus findet sich hingegen keine Übereinstimmung. Während es bei zweifachen Stenosen noch durchaus erkennbare

Übereinstimmungen innerhalb der Untersucher wie auch mit dem Goldstandard gibt, ist die Übereinstimmung bei Dreifachstenosen nicht mehr messbar. Bei den Zweifachstenosen erreichen kombinierte Stenosen in Trachea und Hauptbronchus die größte Präzision und Richtigkeit.

Innerhalb der Untersucher, wie auch im Vergleich zum Goldstandard wurde quer über alle anatomischen Abschnitte eine mäßige Übereinstimmung erzielt.

Tabelle 5.4: Stenoselokalisation

Befund		**Befundverteilung**				**Präzision** Übereinstimmung der Befunder			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz			
		Referenz	Befunder				Kappa Fleiss				Kappa Cohen	
		Anzahl verschiedener										
Typ der Untersuchung		Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**
syn.	Larynx	23 [460]	278	20	32	26,0	0,418	moderat	29,8	49,3	0,357	mäßig
einzeln	supraglottisch	8 [160]	62	17	17	33,1	0,261	schwach	23,8	61,3	0,264	schwach
einzeln	glottisch	11 [220]	80	19	18	33,2	0,250	schwach	27,7	76,3	0,310	mäßig
einzeln	subglottisch	11 [220]	170	20	20	46,1	0,569	gut	58,6	75,8	0,561	gut
syn.	Trachea	17 [340]	227	20	28	21,0	0,324	mäßig	30,3	49,4	0,385	mäßig
einzeln	proximales Drittel	10 [200]	75	19	20	37,8	0,150	gering	26,0	69,3	0,285	schwach
einzeln	mittleres Drittel	5 [100]	90	20	17	35,4	0,299	schwach	48,0	53,3	0,442	moderat
einzeln	distales Drittel	9 [180]	108	20	16	54,2	0,393	mäßig	43,3	72,2	0,454	moderat
syn.	Hauptbronchus	8 [160]	135	20	16	30,6	0,487	moderat	29,8	49,3	0,452	moderat
einzeln	rechts	4 [80]	60	20	10	36,9	0,443	moderat	50,0	66,7	0,533	gut
einzeln	links	6 [120]	98	20	14	40,4	0,509	gut	50,0	61,2	0,484	moderat
syn.	Lappenb. rechts	1 [20]	21	11	10	13,8	0,068	schwach	29,8	49,3	0,159	schwach
einzeln	Oberlappen re.	na	13	8	6	16,7	0,099	kaum	na	na	na	na
einzeln	Mittellappen	na	6	6	5	10,0	0,011	keine	na	na	na	na
einzeln	Unterlappen re.	1 [20]	7	6	5	10,0	0,022	keine	10,0	29,0	0,137	gering
syn.	Lappenb. links	2 [40]	35	17	8	16,5	0,297	schwach	45,5	100	0,615	stark
einzeln	Oberlappen links	na	14	9	7	13,8	0,067	schwach	na	na	na	na
einzeln	Unterlappen links	2 [40]	27	16	6	40,0	0,418	beachtlich	50,0	74,1	0,581	beachtlich
einzeln	Lingula	na	13	9	7	20,0	0,091	schwach	na	na	na	na

Die Tabelle betrachtet die Übereinstimmung der Einzelbefunde sowie der jeweils übergeordneten Befundkombinationen (grau hinterlegt) und ist in die Kriterien Befundverteilung, Präzision und Richtigkeit gegliedert.

5.4.2.3 Stenoseform

Auch wenn die Aussagekraft durch eine geringe Prävalenz eingeschränkt wird, scheinen die morphologischen Befunde „membranös“ und „ringförmig“ zuverlässiger, als die Einschätzung der Länge der Stenose, selbst wenn letztere nur durch die einfachst mögliche Einteilung in „langstreckig“ und „kurzstreckig“ erfolgt. Befundkombinationen aus mehreren Merkmalen spielen mit einem Anteil von nur 4,3 % an allen Befunden so gut wie keine Rolle. Wie schon bei der Lage der Stenosen haben falsch negative Befunde einen erheblichen Anteil an der Fehlklassifikation.

Tabelle 5.5: Stenoseform

Befund	Befundverteilung				Präzision Übereinstimmung der Befunder			Richtigkeit Übereinstimmung mit Goldstandard als Referenz		
	Referenz	Befunder Anzahl verschiedener				Kappa Fleiss				Kappa Cohen
	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa **nach Cohen**
Stenoseform	42 [840]	840	20	42	NA	0,202	schwach	21,3	24,4	0,245
kurzstreckig	11 [220]	157	18	32	26,7	0,119	gering	31,4	43,9	0,189
langstreckig	5 [100]	109	19	23	30,6	0,230	schwach	33,0	30,3	0,219
membranös	3 [60]	32	14	12	38,3	0,278	schwach	31,7	59,4	0,382
ringförmig	6 [120]	49	16	12	35,8	0,269	schwach	26,7	65,3	0,323

Befundübereinstimmung der einzelnen Stenoseformen und des Globalbefundes (grau hinterlegt).

Fazit 4: Stenosen

Der Stenosegrad wird selbst bei höchstgradigen Stenosen sehr variabel beurteilt. Dabei streut die Fehlklassifikation nicht nur in benachbarte Befundklassen, sondern über fast alle Grade der Myer-Cotton-Klassifikation. Ein wesentlicher Anteil der Fehlklassifikation geht auf falsch negative Befunde zurück, also auf die Frage, ob überhaupt eine Stenose vorliegt. Die Befunder schätzten den Stenosegrad ingesamt deutlich geringer ein, als der Goldstandard, was womöglich teilweise durch psychophysikalische Phänomene erklärbar sein könnte. Die Variabilität lässt sich nur bedingt aus den zwischen Befundern und Goldstandard erheblich divergierenden Befundprävalenzen erklären. Die Ergebnisse legen insgesamt nahe, dass der Stenosegrad anhand des Bildmaterials nur unzuverlässig bestimmt werden kann – selbst bei Vereinfachung auf die vier Befundklassen der Mayr-Cotton-Klassifikation. Klinische Hintergrundinformationen wie Dyspnoe oder Stridor erzeugen vermutlich einen erheblichen Beobachtungsbias. Bei der Bestimmung der Stenoselokalisation fällt auf, dass Präzision und Richtigkeit in Larynx und Trachea von proximal nach distal zunehmen. Die Analyse der Kombinationsbefunde zeigt, dass sich verlässliche Übereinstimmungen bei der Bestimmung der Stenoselokalisation innerhalb der Befunder, wie auch beim Vergleich mit dem Goldstandard, fast nur bei singulären Stenosen finden. Bei Mehrfachstenosen ist dem gegenüber kaum eine nennenswerte Übereinstimmung feststellbar. Wie beim Stenosegrad ist ein erheblicher Anteil der Fehlbefundung durch falsch-negative Befunde zu erklären. Uneinigkeit besteht also in erster Linie darin, ob eine Stenose vorhanden ist, oder nicht, weniger in ihrer Graduierung oder Lage. Ein Vergleich zwischen den anatomischen Abschnitten zeigt die präziseste und gleichzeitig genaueste Befundung bei Stenoselokalisationen im Hauptbronchus. Hinsichtlich Präzision folgen Larynx und Trachea. Die syndromale Analyse der abschnittsübergreifenden Stenosen zeigt die beste Übereinstimmung bei singulären Stenosen. Mit Ausnahme von Stenosen im rechten Lappenbronchus liegt die Präzision in allen Abschnitten hier gleich auf. In Bezug auf die Richtigkeit führen bei singulären Stenosen Trachealstenosen und unauffällige Befunde. Dahinter reihen sich Stenosen im Larynx und Hauptbronchus ein. Die Übereinstimmung bei Doppeltstenosen ist gering, bei Dreifachstenosen fehlend. Insgesamt sind Präzision und Richtigkeit der Stenoselokalisation mit einem Kappa Fleiss von 0,328 bzw. einem Kappa Cohen von 0,346 mäßig. Bei der Stenoseform ist die Beurteilung der Merkmale membranös und ringförmig zuverlässiger, als die Ausdehnung der Stenose (kurzstreckig versus langstreckig).

5.4.3 Spezielle Stenosen

Malazie, Pulsationen und Kompressionen wurden als Sonderformen von Atemwegsstenosen aufgefasst: Malazien als „dynamische“ Stenosen, Pulsationen als direkt visuell erkennbare gefäßbedingte Ursache von Stenosen und Kompressionen als Stenosen durch Raumforderungen von außen.

5.4.3.1 Malazie

Auf Ebene der Einzelbefunde wurden tracheale Malazien mit einem Kappa nach Fleiss von 0,241 am einheitlichsten befundet und erreichten bei einem Kappa Cohen von 0,303 die beste Übereinstimmung mit dem Goldstandard. In der Rangliste folgen sowohl hinsichtlich Präzision wie auch der Richtigkeit Larynxmalazien. Bronchusmalazien bilden das Schlusslicht. Auch bei syndromaler Betrachtung zeigten singuläre Malazien im Bereich der Trachea in Präzision und Richtigkeit die höchste Übereinstimmung und werden auch hier von laryngealen Malazien gefolgt. Bei kombinierte Malazien wurden weder innerhalb der Untersucher noch im Vergleich zum Goldstandard eine erkennbare Übereinstimmung erzielt. Dabei wurde nur jede dritte singuläre Tracheomalazie als solche erkannt und nur jede 10. Laryngomalazie. Der Befund einer singulären Tracheomalazie war in 50 % der Fälle, der Befund der Laryngomalazie in ⅓ der Fälle korrekt.

5.4.3.2 Pulsationen

Auf Ebene der Einzelbefunde wurde die häufigste Diagnose, die Pulsation im Stenosebereich, sowohl in der Übereinstimmung untereinander (Kappa Fleiss 0,324) als auch zum Goldstandard (Kappa Cohen 0,358) am besten bewertet. Die weitere Rangliste folgt, sowohl hinsichtlich Präzision, wie auch hinsichtlich der Richtigkeit den Häufigkeiten der Befunde, bei den Untersuchern: Pulsationen in der Trachea (Kappa Fleiss 0,273, Kappa Cohen 0,355), Pulsationen im Bronchus (Kappa Fleiss 0,172; Kappa Cohen 0,329) und zuletzt Pulsationen im Larynx (Kappa Fleiss 0,370), die vom Goldstandard nicht gesehen wurden (Kappa Cohen daher nicht erhebbar). Auf Ebene der Befundkombinationen waren Übereinstimmungen nur bei singulären Pulsationen erkennbar. Die beste Konkordanz der Befunder findet sich bei Pulsationen im Stenosebereich mit einem schwachen Kappa Fleiss von 0,212 – einem Befund, der vom Goldstandard nicht erhoben wurde. Eine schwache Übereinstimmung mit dem Goldstandard wurde nur bei Pulsationen im Bronchus beobachtet. Allerdings gingen bei der Berechnung der Richtigkeit wegen der geringen Überschneidung der Befundkategorien etwa ⅓ der Daten verloren.

5.4.3.3 Kompressionen

Bei Betrachtung der Einzelbefunde wird die Präzision von Kompressionen in der Trachea (Kappa Fleiss 0,174) dicht gefolgt von Kompressionen im Stenosebereich (Kappa Fleiss 0,165). Auch hinsichtlich der Richtigkeit führen Kompressionen in der Trachea (Kappa Cohen 0,284), hier allerdings gefolgt von Kompressionen im Bronchus (Kappa Cohen 0,249). Laryngeale Kompressionen wurden im Gegensatz zu den Befundern vom Goldstandard nicht gesehen. Auf der Ebene der Befundkombinationen sieht der Goldstandard überwiegen Kombinationsbefunde der Kompressionen, während bei den Untersuchern Einzelbefunde überwiegen. Die bei den Befundern häufigste Befundklasse – Kompressionen im Stenosebereich – bei der zugleich mit 0,131 die höchste Präzision besteht, wurde vom Goldstandard nicht gesehen. Gemäß dem Goldstandard treten Kompressionen hauptsächlich in Assoziationen mit Kompressionen im Stenosebereich auf (7/9 Befunden). Eine Übereinstimmung mit dem Goldstandard ist nur bei singulären Kompressionen im Bronchus erkennbar (Kappa Cohen 0,153), wobei die Datenabdeckung nur bei 75,9 % liegt.

Tabelle 5.6: Inter-Beobachter-Variabilität spezielle Stenosen

Befund	Befundverteilung				Präzision Übereinstimmung der Befunder untereinander			Richtigkeit Übereinstimmung mit Goldstandard als Referenz			
	Referenz	Befunder				Kappa Fleiss				Kappa Cohen	
	Anzahl verschiedener										
	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)	Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	**modifizierte** Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa **nach Cohen**	**modifizierte** Klassifikation **nach Landis**
Malazie	**13 [260]**	**161**	**20**	**28**	**16,1**	**0,171**	**gering**	**13,6**	**27,8**	**0,221**	**schwach**
Stenosebereich	4 [80]	57	16	18	18,9	0,106	gering	12,5	17,5	0,072	kaum
Larynx	5 [100]	32	13	13	18,6	0,131	gering	17,2	53,1	0,214	schwach
Trachea	8 [160]	73	17	15	29,2	0,241	schwach	28,1	61,6	0,303	mäßig
Bronchus	2 [40]	46	14	15	17,9	0,110	gering	17,5	15,2	0,118	gering
Pulsationen	**11 [220]**	**108**	**19**	**15**	**NA**	**0,317**	**mäßig**	**9,2**	**30,6**	**0,248**	**schwach**
Stenosebereich	7 [140]	74	18	12	33,2	0,324	mäßig	33,1	62,2	0,358	mäßig
Larynx	NA	9	6	6	12,5	0,037	kaum	NA	NA	NA	NA
Trachea	7 [140]	47	16	11	30,7	0,273	schwach	27,3	80,9	0,355	mäßig
Bronchus	5 [100]	39	16	10	23,1	0,172	gering	26,0	66,7	0,329	mäßig
Kompressionen	**9 [180]**	**124**	**20**	**19**	**NA**	**0,193**	**gering**	**7,6**	**34,5**	**0,181**	**gering**
Stenosebereich	7 [140]	75	18	17	24,3	0,165	gering	21,4	40,0	0,184	gering
Larynx	NA	8	6	5	12,5	0,044	kaum	NA	NA	NA	NA
Trachea	4 [80]	46	16	12	22,0	0,174	gering	26,2	45,7	0,284	schwach
Bronchus	6 [120]	26	14	10	16,4	0,102	gering	17,5	80,8	0,249	schwach

Die Tabelle betrachtet die Übereinstimmung der Einzelbefunde sowie der jeweils übergeordneten Befundkombinationen (grau hinterlegt) und ist in die Kriterien Befundverteilung, Präzision und Richtigkeit gegliedert.

Fazit 5: Spezielle Stenosen

Pulsationen werden mit einem mäßigen Kappa Fleiss von 0,317 am präzisesten erkannt. Kompressionen und Malazien mit einem Kappa Fleiss von 0,193 bzw. 0,171 deutlich schlechter. Bei Pulsationen besteht gleichzeitig auch die beste Übereinstimmung mit dem Goldstandard, gefolgt von Malazien und Kompressionen. Pulsationen werden am verlässlichsten in Assoziation mit Stenosen erkannt, Malazien und Kompressionen in der Trachea.

5.4.4 Schleimhaut

Die Schleimhaut wurde zunächst separat nach den Kriterien Schwellung, Hyperämie und Hypersekretion beurteilt. In einem zweiten Schritt wurden diese Befunde entsprechend dem Konzept des Bronchitis Index (BI) (Thompson u. a., 1993) zu einem syndromalen Entzündungsbefund zusammengeführt. Dieser syndromale Entzündungsbefund wurde auf seine Übereinstimmung mit einem pauschal erhobenen Entzündungsbefund verglichen.

5.4.4.1 Hyperämie, Schwellung & Hypersekretion

Bei Betrachtung der Einzelbefunde wird Hypersekretion am sichersten erkannt – sowohl hinsichtlich der Präzision, als auch der Richtigkeit. Etwas weniger präzise wird Hyperämie, am we-

nigsten einheitlich Schwellung erkannt. Die Übereinstimmung mit dem Goldstandard ist bei Schwellung und Hyperämie schwach, bei Hypersekretion mäßig.

Tabelle 5.7: Schleimhaut Einzelbefunde

Befund	**Befundverteilung**				**Präzision** Übereinstimmung der Befunder			**Richtigkeit** Übereinstimmung mit Goldstandard als Referenz			
	Referenz	Befunder			Ø positive Übereinstimmung [%]	Kappa Fleiss		Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa Cohen	
	Anzahl verschiedener					Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**			Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**
	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)							
Schwellung	22 [440]	311	20	39	NA	0,153	gering	63,9	69,5	0,283	schwach
Hyperämie	24 [480]	317	20	39	NA	0,175	gering	62,8	75,1	0,269	schwach
Hypersekretion	24 [280]	333	20	38	NA	0,211	mäßig	67,8	81,4	0,384	mäßig

Übereinstimmung bei isolierter Betrachtung der einzelnen Befundkategorien. Es wurden nur aktiv als vorhanden oder abwesend erhobene Befunde berücksichtigt. Die Datenabdeckung liegt bei allen drei Befundkategorien über 70 %. Die Ergebnisse bei Wertung fehlender Befunde als negativ finden sich im Ergebnisteil.

Die Betrachtung der Kombinationsbefunde (siehe Ergebnisteil) zeigt, dass es bei singulären Schleimhautbefunden wie auch Befundkombinationen kaum Übereinstimmung gibt: weder unter den Befundern untereinander, noch im Vergleich zur Referenz. Singuläre Schleimhautbefunde wurden auch nicht sicherer erkannt als Befundkombinationen. Nur bei unauffälligen Schleimhautbefunden so wie dem anderen Extrem, dem Vollbild aus Schwellung, Hyperämie und Sekretion findet sich eine mäßige respektive schwache Präzision und Übereinstimmung.

5.4.4.2 Entzündung

Entzündung wurde einerseits als pauschaler Befund erhoben, andererseits als aus den Befunden Schwellung, Hyperämie und Hypersekretion abgeleitetes Syndrom. Letzteres entspricht dem Konzept des Bronchitisindex (BI). Diese beiden Erhebungsmethoden wurden auf ihre Kongruenz zueinander überprüft.

5.4.4.2.1 Entzündung als pauschaler Befund

Der Goldstandard beurteilte die Schleimhaut in 24 von 42 Videos als entzündet, was 480 Befunde bei den Untersuchern erwarten lässt, die tatsächlich aber nur 289 Befunde erhoben. In gut 30 % der Videos erhoben die Untersucher keinen Entzündungsbefund. Die Einheitlichkeit der Befundung innerhalb der Untersucher war bei einem Kappa Fleiss von 0,155 gering. Die Übereinstimmung mit dem Goldstandard bei einem Kappa Cohen von 0,268 immerhin schwach.

5.4.4.2.2 Entzündung als Syndrom der Schleimhautbefunde

Die Kongruenz von Schleimhautbefunden wurde sowohl im Vergleich zum von den Untersuchern pauschal erhobenen Entzündungsbefund (Präzision), als auch zu dem des Goldstandards (Richtigkeit) untersucht. Die beste Kongruenz zu Entzündungsbefunden der Untersucher liegt mit 87,3 % beim Vollbild aus Hypersekretion, Hyperämie und Schleimhautschwellung vor. Die Kombination aus Schleimhautschwellung und Hyperämie ist mit einem positiven prädiktiven Wert von 67,3 % der Kombination aus Hypersekretion und Hyperämie (ppv 58,8 %) sowie der Kombination aus Schleimhautschwellung und Hypersekretion (ppv 53,1 %) überlegen. Bei Ein-

zelbefunden kommt der Hyperämie mit 41,9 % die höchste Vorhersagekraft hinsichtlich eines gleichzeitigen Entzündungsbefundes zu. Einer isolierten Hypersekretion (ppv 28,6%) oder Schleimhautschwellung (ppv 11,1 %) hat nur eine geringe Vorhersagekraft in Bezug auf eine etwaige Entzündung.

Auch in Bezug zum Entzündungsbefund der Referenz kommt dem Vollbefund aus Hypersekretion, Hyperämie und Schleimhautschwellung eine hohe Vorhersagekraft (ppv 76,1 %) zu. Zweifachbefunde haben im Vergleich zur Referenz aber eine ähnliche Vorhersagekraft, die für die Befundkombination aus Hypersekretion und Hyperämie mit 79,4 % sogar den Spitzenwert erreicht. Bestes Einzelmerkmal für eine Entzündung ist auch hier die Hyperämie.

Tabelle 5.8: Schleimhautphänomene versus Global-Befund „Entzündung“

Befund				**Befundverteilung**				**Kappa**				**Kongruenz zu Entzündungsbefund der**			
				Anzahl verschiedener				**Präzision** Kappa Fleiss		**Richtigkeit** Kappa Cohen		Befunder		Referenz	
Anzahl der Befunde	Schwellung	Hyperämie	Hypersekretion	Befunde (max. 840)		Befunder (max. 20)	Videos (max. 42)	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Kappa **nach Cohen**	modifizierte Klassifikation **nach Landis**	Häufigkeit	Prozent in jeweiliger Klasse	Häufigkeit	Prozent in jeweiliger Klasse
x	**gesamt**			**29**	**446**			**0,196**	**gering**	**0,217**	**schwach**	282	63,3 %	312	70,0 %
0			0	13	394	20	40	0,326	mäßig	0,368	mäßig	7	1,8 %	168	42,6 %
			1	2	70	19	30	0,080	kaum	0,110	gering	20	28,6 %	35	50,0 %
1		1	0	2	31	15	21	0,022	keine	-0,044	keine	13	41,9 %	21	67,7 %
	1	0	0	na	27	14	19	0,003	keine	na	na	3	11,1 %	16	59,3 %
		1	1	3	34	15	20	0,032	keine	-0,055	keine	20	58,8 %	27	79,4 %
2	1	0	1	3	32	16	16	0,063	kaum	0,04	keine	17	53,1 %	24	75,0 %
	1	1	0	3	55	16	27	0,010	keine	0,078	kaum	37	67,3 %	39	70,9 %
3	1	1	1	16	197	20	36	0,259	schwach	0,274	schwach	172	87,3 %	150	76,1 %

Die Tabelle analysiert die Befundkombinationen der einzelnen Schleimhautphänomene im Vergleich zum unabhängig davon erhobenen pauschalen Befund „Entzündung“. Referenz ist dabei einerseits der Global-Befund „Entzündung“ der Befunder, andererseits der Global-Befund „Entzündung“ des Goldstandards. Kappa-Fleiss misst die relative Übereinstimmung der jeweiligen Befundkombination innerhalb der Befunder. Kappa Cohen misst die Übereinstimmung der Befundkombination mit der durch den Referenzbefund des Goldstandards vorgegebenen. Die Prozentzahl in der Spalte „Kongruenz zu Entzündungsbefund“ gibt an, wie groß bei der jeweiligen Befundkombination der Anteil positiver Entzündungsbefunde bei den Befundern und beim Goldstandard ist.

5.4.4.2.3 Entzündungsbereich

Bei der Lokalisation des Entzündungsbereiches findet sich nur in Trachea und Bronchus eine nennenswerte Präzision und Richtigkeit. Die Übereinstimmung mit dem Goldstandard ist bei Entzündungen im Bronchus mit einem Kappa Cohen von 0,416 am ausgeprägtesten und fällt nach proximal über die Trachea (Kappa Cohen 0,311) zum Larynx auf ein Kappa Cohen von 0,177 ab. Die Präzision in Trachea und Bronchus ist bei einem Kappa Fleiss von 0,226 bzw. 0,222 ähnlich, aber als schwach einzustufen.

Tabelle 5.9: Einzelbefunde Entzündungsbereich

Befund	Befundverteilung				Präzision Übereinstimmung der Befunder untereinander			Richtigkeit Übereinstimmung mit Goldstandard als Referenz			
	Referenz	Befunder Anzahl verschiedener			Ø positive Übereinstimmung [%]	Kappa Fleiss		Sensitivität [%]	positiver prädiktiver Wert [%]	Kappa Cohen	
	Befunde (& Videos) (max. 42)	Befunde (max. 840)	Befunder (max. 20)	Videos (max. 42)		Kappa **nach Fleiss**	Klassifikation **modifiziert nach Landis**			Kappa **nach Cohen**	Klassifikation **modifiziert nach Landis**
Stenosebereich	9 [180]	125	17	32	23,6	0,124	gering	16,3	23,4	0,013	keine
Larynx	17 [340]	90	18	28	20,5	0,148	gering	20,5	75,6	0,177	gering
Trachea	13 [260]	131	20	27	28,6	0,226	schwach	34,7	66,7	0,311	mäßig
Bronchus	6 [120]	69	19	18	24,6	0,222	schwach	38,4	63,2	0,416	moderat

Zur besseren Vergleichbarkeit zwischen Referenz und Befundern ist in eckigen Klammern die gewichtete Anzahl der Befunde des Goldstandards angegeben (Faktor 20 im Vgl. zu den Befundern)

Fazit 6: Schleimhaut und Entzündung

Die Schleimhautmerkmale Schwellung, Hyperämie und Sekretion werden innerhalb der Untersucher uneinheitlich befundet. Ihre Richtigkeit relativ zum Goldstandard liegt über der Präzision. Die Betrachtung der Kombinationsbefunde zeigt, dass bei Einfach- und Zweifachbefunden kaum Übereinstimmung besteht – weder innerhalb der Befunder, noch zum Goldstandard. Beim Vollbild aus Hyperämie, Hypersekretion und Schwellung sowie bei unauffälligen Befunden besteht hingegen eine schwache bis mäßige Übereinstimmung, sowohl bei Präzision als auch bei der Richtigkeit. Die Kombination der drei Befunde Hyperämie, Hypersekretion und Schwellung zeigt mit einem prädiktiven Wert von 87,3 % auch die beste Kongruenz zum pauschalen Befund einer Entzündung. Zweifachkombinationen erreichen prädiktive Werte zwischen 50 und 70 %, Einzelbefunde zwischen ca. 10 und 40 %. Auch im Vergleich zum Goldstandard ist die Vorhersagekraft von Dreifachbefunden mit 76 % stark, setzt sich aber weniger deutlich gegen Zweifachbefunde ab, die hier prädiktive Werte zwischen 70 und 80 % erreichen. Bester Einzelprädiktor für Entzündung ist die Hyperämie, sowohl in Bezug auf die Präzision wie auch die Richtigkeit. Hinsichtlich der Präzision ist die Hyperämie, zusammen mit der Schwellung (ppv 67,3 %), hinsichtlich der Richtigkeit, zusammen mit der Hypersekretion (ppv 79,4 %) am aussagekräftigsten.

5.4.5 Empfehlungen für die Gestaltung eines Befundbogens

Ein einheitlicher Befundbogen für die pädiatrische Bronchoskopie wäre ein entscheidender Beitrag zur Erhebung epidemiologischer Daten, zur Qualitätssicherung und zu multizentrischen klinischen Studien. Ausgehend von dem in dieser Studie angewendeten Befundbogen können Empfehlungen für die Weiterentwicklung eines solchen standardisierten Befundbogens ausgesprochen werden.

Der Stenosegrad sollte weiterhin primär als Prozentzahl erhoben und ggf. sekundär nach der Mayr-Cotton-Klassifikation kategorisiert werden, auch wenn in dieser Studie selbst nach Reduktion auf vier Befundklassen nur eine geringe Übereinstimmung erzielt wurde. Die Konkordanz von Prozentangaben kann so nicht nur über Kappa, sondern auch den sogenannten Intra-Klassen-Koeffizienten (engl. Intra-Class-Coeficient, kurz ICC) bestimmt werden. Anstatt des generalisierten maximalen Stenosegrades sollte für jede Stenose separat der maximale Stenosegrad erfragt

werden, damit einzelne Stenosen gezielter verglichen werden können bzw. bei Mehrfachstenosen der Bezug des Stenosegrades zur Stenose eindeutig bleibt.

Die Gliederung des Bronchialbaumes bis zur Ebene der Lappenbronchien hat sich als anatomisches Befundschema bewährt. Die Hauptabschnitte sind einheitlich in 3 Unterabschnitte gegliedert und die anatomischen Grenzen zwischen den Abschnitten relativ klar definierbar. Die Gliederung scheint übersichtlich genug, um im klinischen Alltag Akzeptanz zu finden und ist gleichzeitig differenziert genug, um Unterschiede bei der Befundung im Detail zu erfassen. Bei Bedarf kann das Befundschema rechnerisch auf die übergeordneten anatomischen Abschnitte Larynx, Trachea, Hauptbronchien und Lappenbronchien vereinfacht werden. Eine weitere Differenzierung bis hin zu den Segmentbronchien scheint für den klinischen Alltag wenig sinnvoll, denn die Komplexität des im Rahmen dieser Arbeit vorgeschlagenen Schemas übersteigt bereits klar die von qualifizierten Untersuchern leistbare Befundpräzison. Relative Lokalisationen wie „im Stenosebereich“ sollten wegen unklarer Bezüge bei Mehrfachbefunden vermieden werden. Das identische anatomische Schema sollte einheitlich auch auf andere Befunde, für die Lokalisationsangaben relevant sind, angewandt werden. Hierzu zählen beispielsweise die Schleimhautbefunde Schwellung, Hyperämie und Hypersekretion sowie die speziellen Stenosen Malazie, Kompression und Pulsationen. Hierdurch können Korrelationen zwischen den Befunden in zukünftigen Arbeiten besser analysiert werden.

Der Befundbogen sollte in einem maschinenlesbaren, automatisch auswertbaren Format gestaltet sein. Alternativ zu einem Befundbogen im multiple-choice-Format sollte auch eine äquivalente Kurzschrift etabliert werden, die z. B. in Klinikinformationssysteme eingepflegt und ebenfalls automatisch extrahiert werden kann.

5.5 Evidenzbasierte Ausbildung

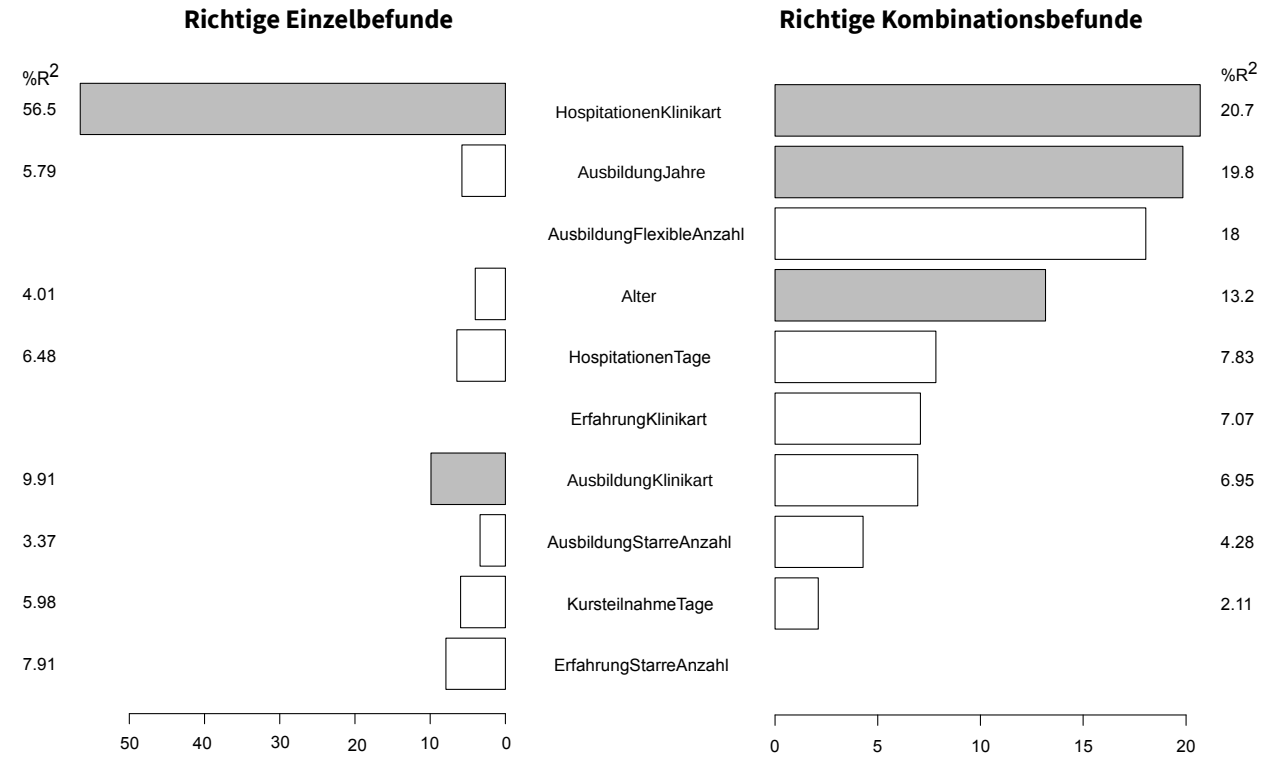
Eine Evaluierung der praktischen Fertigkeiten angehender Kinderpneumologen in Bronchoskopie findet in Deutschland derzeit nicht statt. International gibt es Bemühungen die Kompetenz im Umgang mit dem Bronchoskop in Simulatoren zu trainieren und zu testen, anstatt auf die bloße Erfüllung eines Katalogs von Eingriffen zu vertrauen (Baker u. a., 2016; Kastelik u. a., 2013). Damit verbunden ist ein genereller Trend, verstärkt auf die Prüfung von Wissen und praktischen Fertigkeiten zu setzen, anstatt Ausbildungszeit oder die Anzahl durchgeführter Untersuchungen zum Maßstab der Qualifikation zu machen (Ernst u. a., 2015). Ausbildungscurricula basieren auf Empfehlungen von Experten bzw. Fachgesellschaften. Die geforderten Inhalte variieren und werden in der Praxis unterschiedlich umgesetzt (Haponik u. a., 2000). Die Anzahl der geforderten Untersuchungen wird in der Praxis häufig nicht erreicht (Pastis u. a., 2005). Die HERMES-Initiative tritt an, um einen europäischen Konsens hinsichtlich der Anforderungen in der Ausbildung zu erzielen (Gappa u. a., 2009). Experten halten oft ein Minimum von 50 flexiblen Bronchoskopien während der Ausbildung für notwendig (Torrington, 2000), eine Zahl, die sich durch eine Umfrage in den Vereinigten Staaten belegen lässt (Leong u. a., 2014).

Über die Faktoren, die tatsächlich ausschlaggebend für eine gute Ausbildung in pädiatrische Bronchoskopie sind, ist bislang wenig bekannt. Mithilfe eines Arztfragebogens, der Informationen zu Demographie, Ausbildung und Erfahrung der Untersucher erhob, fanden wir Anhaltspunkte dafür, welche Faktoren für eine effektive Ausbildung entscheidend sein könnten. Mit linearen Modellen und Entscheidungsbäumen wurde untersucht, welche Variablen aus Ausbildung und Erfahrung Einfluss auf die Befundrichtigkeit ausüben. Dabei wurden richtige Einzelbefunde und richtige Befundkombinationen als Zielvariable gesetzt.

5.5.1 Lineares Modell

Über eine all subset regression wurde zunächst eine Vorauswahl an Variablen getroffen. Sowohl das lineare Modell mit richtigen Einzelbefunden, als auch das Modell mit richtigen Befundkombinationen fielen statistisch signifikant aus und erklären etwa ⅔ der beobachteten Varianz. Beide Modelle schreiben Erfahrung im Vergleich zu Ausbildung eine untergeordnete Bedeutung zu. Die meisten erfahrungsbezogenen Variablen werden bereits im Rahmen der Modellselektion aussortiert. Allerdings erkennt das Modell mit richtigen Befundkombinationen als Zielvariable einen positiven Effekt von Erfahrung an nicht universitären Kliniken. Übereinstimmend wird die Klinikart der Hospitationen als einflussreichste Variable identifiziert: Hospitationen an einer Universitätsklinik wirken sich positiv aus. Die Modelle stimmen aber auch darin überein, dass sich die Ausbildung an einer nicht universitären Klinik günstig auswirkt. Ungünstig wirken sich in beiden Modellen Alter und eine lange Ausbildungszeit aus. Letzteres geht – genau wie ein negativer Trend für eine hohe Anzahl flexibler Bronchoskopien während der Ausbildung – jedoch zum Teil auf Ausreißer zurück.

Abbildung 5.5: Variablenwichtigkeit in den linearen Modellen



Die relative Wichtigkeit der Variablen im Modell wird über ihren Anteil am Bestimmtheitsmaß R^2 ausgedrückt. Bei statistisch signifikanten Ergebnissen sind die jeweiligen Balken im Diagramm grau hinterlegt.

5.5.2 Entscheidungsbäume

Für die Berechnung der linearen Modelle musste der durch fehlende Angaben der Untersucher teilweise lückenhafte Datensatz mittels informierter Schätzung (engl. „informed guess“) und anschließender Imputation[38] komplettiert werden. Im Hinblick auf Kollinearität wurde die Anzahl der Variablen reduziert und auf abgeleitete Variablen (engl. „dummy variables“) ganz verzichtet. Die Modelle gehen davon aus, dass eine Normalverteilung gegeben ist.

Rekursives Partitionieren ist ein hypothesenfreier Ansatz, mit dem der originale Datensatz inklusive Fehlwerte direkt untersucht werden kann. Das Verfahren ist hinsichtlich Anzahl und Beschaffenheit der Variablen robust und modelliert Interaktionen. Die Ergebnisse können anschau-

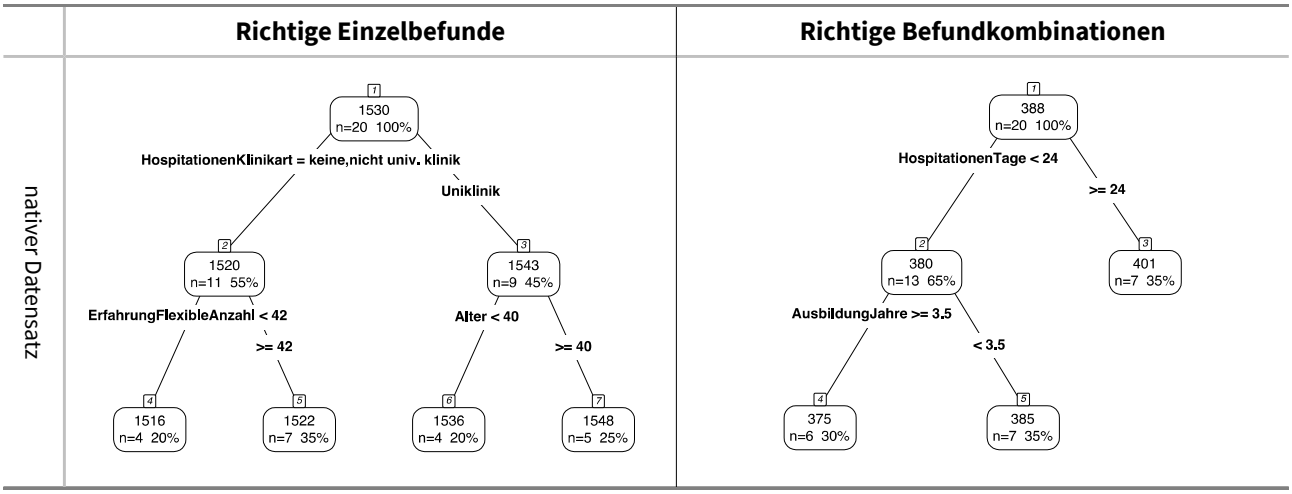
[38] *Die Imputation wurde mit random forests durchgeführt.*

lich in einem Entscheidungsbaum dargestellt werden. Rekursives Partitionieren ist in mehreren verschiedenen Algorithmen implementiert. Für diese Arbeit wurde der klassische CART-Algorithmus gewählt. Als Gegenstück zur relativen Variablenwichtigkeit der linearen Modelle wurde der Einfluss der einzelnen Variablen über random forests abgeschätzt. Random forests sind Wälder aus vielen einzelnen Entscheidungsbäumen, wobei jeder Baum im Wald nur eine zufällige Auswahl der Variablen einbezieht. So werden Ergebnisse erzielt, die weniger anfällig für overfitting und den verdeckenden Einfluss dominanter Variablen sind, als einzelne Bäume.

5.5.2.1 CART

Die Entscheidungsbäume selektieren bei richtigen Einzelbefunden die Klinikart etwaiger Hospitation als wichtigste Variable, bei richtigen Befundkombinationen als Zielvariable die Dauer der Hospitationen in Tagen. Hospitationen sind also bei beiden Zielvariablen das beherrschende Thema. Bei richtigen Einzelbefunden als Zielvariable identifiziert CART eine Erfahrung von mehr als 42 flexiblen Bronchoskopien sowie ein Alter über 40 Jahren als nachgeordnete Einflussvariablen. Gemäß dem Baum richtiger Befundkombinationen sollte die Ausbildung nicht länger als 3,5 Jahre dauern.

Abbildung 5.6: CART Entscheidungsbäume

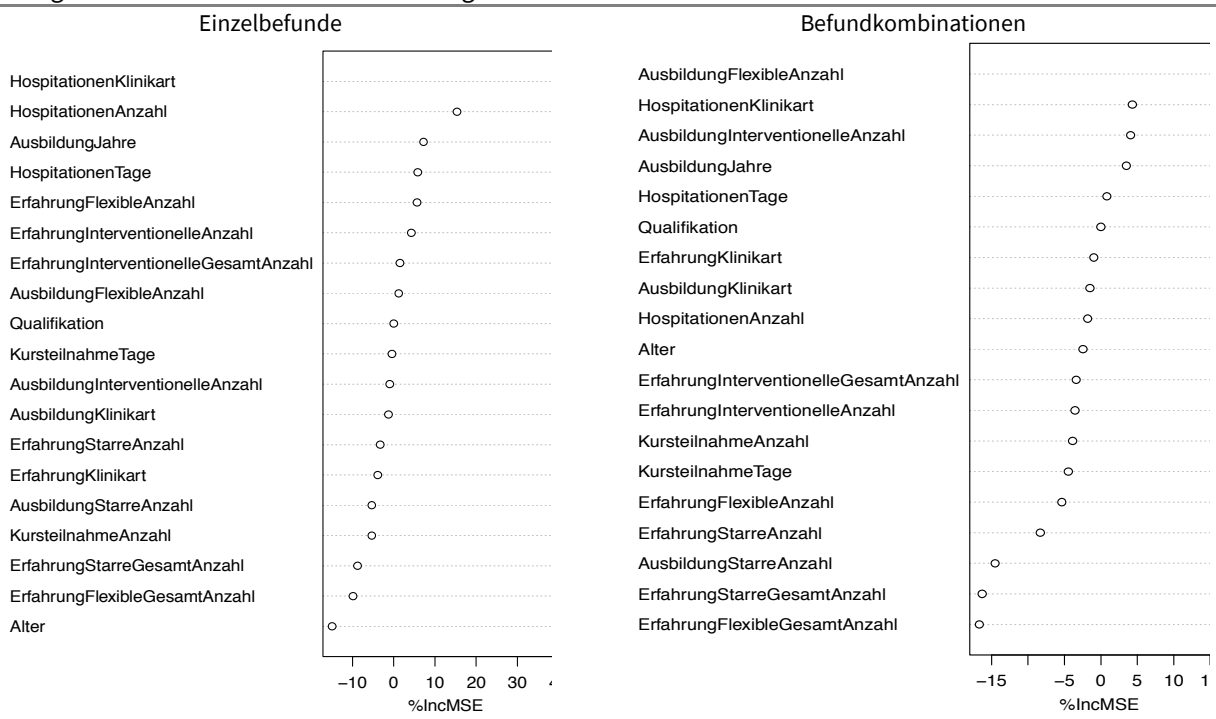


Vergleich der mit CART erzeugten Entscheidungsbäume

5.5.2.2 Random forests

Bereits geringfügige Änderung des Datensatzes können die Konfiguration einzelner Bäume erheblich verändern. Wälder aus zahlreichen, zufällig aus den Variablen zusammengesetzten Bäumen, erzielen wesentlich robustere Ergebnisse. Mit den sogenannten random forests kann darüber hinaus auch die relative Wichtigkeit von Variablen abgeschätzt werden. Mit richtigen Einzelbefunden als Zielvariable werden Klinikart und Anzahl der Hospitationen als wichtigster Prädiktor selektiert, mit richtigen Befundkombinationen als Zielvariable die Anzahl flexibler Bronchoskopien während der Ausbildung und die Klinikart der Hospitationen. Bei richtigen Befundkombinationen als Zielvariable führt die Anzahl der flexiblen Bronchoskopien während der Ausbildung.

Abbildung 5.7: Random Forest Variablenwichtigkeit



Die Abbildungen illustrieren die Variablenwichtigkeit im Random Forest. Links bei richtigen Einzelbefunden als Zielvariable, rechts bei richtigen Befundkombinationen als Zielvariable.

Fazit 7: Evidenzbasierte Ausbildung

Als wichtigsten Faktor für eine korrekte Befundung werden in fast allen Modellen Hospitationen an einer Universitätsklinik identifiziert. Auch der Ausbildung an einer nicht universitären Klinik schreiben die linearen Modelle einen positiven Effekt zu. Ein positiver Trend ist ebenfalls für Erfahrung an einer nicht universitären Klinik zu verzeichnen. Negativ wirkt sich hingegen eine zu lange Ausbildungszeit aus. Gemäß den Entscheidungsbäumen sollten Hospitationen mindestens 24 Tage dauern und eine Erfahrung von mindestens 42 flexiblen Bronchoskopien bestehen. Wie im linearen Modell kommt einer langen Ausbildung ein negativer Effekt zu: die Ausbildung sollte weniger als 3,5 Jahre dauern. Der Entscheidungswald mit richtigen Einzelbefunden als Zielvariable bestätigt die Spitzenposition der Klinikart der Hospitationen. Bei richtigen Befundkombinationen als Zielvariable führt die Anzahl flexibler Bronchoskopien während der Ausbildung die Rangliste an.

5.6 Zusammenfassung

Diese Studie untersuchte erstmals in größerem Umfang die Befundübereinstimmung in der pädiatrischen Bronchoskopie. Dafür wurde das Konzept von Kappa Cohen als etablierte Maßzahl der Befundübereinstimmung so erweitert, dass es auf Befundkombinationen und den Vergleich einer Gruppe zu einer Referenz anwendbar ist. Die Ergebnisse dieses sogenannten „vereinten“ Kappa Cohen waren dem Mittelwert des paarweisen Kappa Cohens vergleichbar, einer der wenigen in der Literatur beschriebenen Behelfsmethoden für den Vergleich einer Gruppe zu einer Referenz. Das vereinte Kappa hat gegenüber dem Mittelwert des paarweisen Kappa Cohen u. a. Vorteile bei der Betrachtung von Befundkombinationen. Mit der Analyse von aus Einzelbefunden zusammengesetzten Befundkombinationen konnte das Konzept von Symptom und Syndrom nachgebildet werden und die erhobenen Befunde genauer differenziert werden: z. B. in einfache und mehrfache Stenosen. Darüber hinaus können über die Verteilung von Befundkombinationen aus Einzelbefunden Kappa Werte für übergeordnete Abschnitte abgeleitet werden.

Im Vergleich der Befundkategorien (Tabelle 5.10) findet sich die größte Übereinstimmung bei der Lokalisation von Stenosen sowohl innerhalb der Untersucher, als auch im Vergleich zum Goldstandard. Stenosegrad und Stenoseform werden ähnlich präzise beurteilt. Der Stenosegrad gehört zu den Befunden bei denen die Übereinstimmung mit dem Goldstandard am geringsten ist – selbst bei hochgradigen Stenosen. Unter den Spezialformen der Stenosen werden Pulsationen am präzisesten und auch genausten befundet. Entzündungen werden zwar innerhalb der Untersucher wenig einheitlich, aber mit der zweitbesten Übereinstimmung zur Referenz bewertet. Quer über die Befunddomänen geht ein wesentlicher Anteil der Fehlbefundung auf falsch negative Befunde zurück, also die Frage, ob das entsprechende Merkmal überhaupt vorhanden ist. Bei der Graduierung der Merkmale ist die Befundvariabilität dem gegenüber erkennbar geringer. Bestes Einzelmerkmal für eine Entzündung ist die Hyperämie. Wirklich verlässliche Entzündungsbefunde finden sich allerdings nur beim Vollbild aus Schwellung, Hyperämie und Hypersekretion.

Obwohl die Videos von erfahrenen Untersuchern mit einem vergleichsweise einfachen Befundschema beurteilt wurden, waren die erzielten Übereinstimmungen verhältnismäßig gering. Das ist u. a. durch das ungewöhnlich große Testfeld bedingt: 20 Untersucher beurteilten je 42 Videos. Viele Studien beschränken sich auf wenige Untersucher (typischerweise 2–5) und Befunde, vielleicht auch deshalb, weil Methoden zum Vergleich mehrerer Untersucher mit einer Referenz bislang wenig etabliert sind. Durch die Berücksichtigung von Befundkombinationen entsteht zudem eine Vielzahl von Befundklassen. Ein direkter Vergleich mit anderen Arbeiten ist daher nur bedingt möglich. Teilweise ist die divergente Befundung auch auf ungleiche Verteilung der Befundprävalenzen, respektive Randsummen, zurückzuführen.

Dennoch bleibt eine erhebliche Befundvarianz bestehen die durch die genannten Faktoren nicht hinreichend erklärbar ist. Als Ursache nehmen wir insbesondere die bewusst vorenthaltenen klinischen Hintergrundinformationen an, die mutmaßlich einen erheblichen positiven Beobachtungsbias erzeugen und einen größeren Einfluss auf die Befundung haben dürften als bislang angenommen. Eine Befundung ohne sie – rein auf Grundlage des Bildmaterials – scheint schwierig.

Zieht man den Vergleich zu anderen diagnostischen Verfahren, wird deutlich, dass verhältnismäßig niedrige Kappa-Werte kein Alleinstellungsmerkmal der Bronchoskopie sind. Selbst vermeintlich objektivere technische Verfahren sind mit einer erheblichen diagnostischen Variabilität behaftet und subjektiver klinischer Diagnostik keineswegs immer überlegen. Die Verlässlichkeit und Aussagekraft diagnostischer Verfahren ist oftmals unzureichend untersucht und wird gerne überschätzt.

Über einen Fragebogen zu Demographie, Ausbildung und Erfahrung der Untersucher wurden erstmals empirisch mögliche Einflussfaktoren einer effektiven Ausbildung in pädiatrischer Bronchoskopie ermittelt. Der Ausbildung scheint eine mindestens ebenso große Bedeutung wie der Erfahrung zuzukommen. Günstig scheinen sich u. a. Hospitationen an einer Universitätsklinik aber auch Erfahrung an einer nicht universitären Klinik auszuwirken. Die von Experten vorgeschlagene Erfahrung von etwa 50 flexiblen Bronchoskopien korreliert gut mit dem in dieser Studie empirisch gefundenen Wert von 42 Bronchoskopien im Entscheidungsbaum mit richtigen Einzelbefunden als Zielvariable.

Tabelle 5.10: Vergleich der Inter-Beobachter-Variabilität verschiedener Befunddomänen

Kategorie	Befund	Präzision Übereinstimmung der Befunder untereinander			Richtigkeit Übereinstimmung mit dem Goldstandard in den überschneidenden Befunden			
			Kappa Fleiss				Kappa Cohen	
		Ø positive Übereinstimmung [%]	Kappa **nach Fleiss**	modifizierte Klassifikation **nach Landis**	Sensitivität [%]	positiver prädiktiver Wert [%]	**Kappa nach Cohen**	modifizierte Klassifikation **nach Landis**
Stenosen	Stenosegrad	25,0	0,201	schwach	33,9	34,5	0,139	gering
	Stenoselokalisation	NA	0,328	mäßig	38,5	51,7	0,346	mäßig
	Stenoseform	24,7	0,202	schwach	21,3	24,4	0,245	schwach
spezielle Stenosen	Malazie	16,1	0,171	gering	13,6	27,8	0,221	schwach
	Pulsationen	24,2	0,317	mäßig	9,2	30,6	0,248	schwach
	Kompressionen	13,4	0,193	gering	7,6	34,5	0,181	gering
Schleimhaut	Schleimhautmerkmale	20,9	0,196	gering	25,9	34,8	0,217	schwach
	Entzündung	39,7	0,155	gering	60,7	74,4	0,268	schwach
	Entzündungsbereich	13,8	0,138	gering	9,1	19,8	0,126	gering

Die Tabelle vergleicht Präzision (Kappa Fleiss) und Richtigkeit (Kappa Cohen) der Beurteilung wichtiger Merkmale. Die angegebenen Maßzahlen sind globale Werte für sämtliche Befundkombinationen des jeweiligen Merkmals.

5.7 Ausblick

Der Befundbogen dieser Studie kann als Ausgangspunkt zur Entwicklung eines einheitlichen Befundsystems in der pädiatrischen Bronchoskopie herangezogen werden. Ein solches standardisiertes Befundsystem würde die Erhebung epidemiologischer Daten, Qualitätsmanagement und zentrenübergreifende klinische Studien erheblich erleichtern bzw. überhaupt erst ermöglichen.

Die Ergebnisse dieser Studie sind in ihrer Aussagekraft u. a. dadurch eingeschränkt, dass es sich bei den Untersuchern um erfahrene Spezialisten handelte. Zukünftige Studien sollten ein breiteres, für die klinische Praxis repräsentativeres Spektrum an Befundern einbeziehen. Zur Stärkung des Referenzbefundes könnten sowohl der Konsens einer Expertengruppe herangezogen werden, als auch ergänzende morphometrische Messungen des vorgelegten Bildmaterials. Mittels computergestützter Erhebung können Fehlwerte vermieden und Beurteilungen gemäß dem intendierten Schema forciert werden.

Neben der Inter-Beobachter-Variabilität sollte auch die Intra-Beobachter-Variabilität untersucht werden. Mit ihrer Hilfe kann auch die Effektivität der Ausbildung noch genauer untersucht werden: z. B. durch Beurteilung von Videos vor und nach Lehrveranstaltungen.

Literatur

Baker, P. A.; Weller, J. M.; Baker, M. J.; u. a. (2016): „Evaluating the ORSIM® simulator for assessment of anaesthetists' skills in flexible bronchoscopy: aspects of validity and reliability". In: *British Journal of Anaesthesia*., S. aew059, DOI: 10.1093/bja/aew059.

Caliebe, W. (1968): „Dokumentationsgerechte Befunderhebung bei Bronchoskopien". In: *European Archives of Oto-Rhino-Laryngology*. 191 (2), S. 628–631, DOI: 10.1007/BF00492143.

Chang, KANG-TSUNG (1977): „Visual Estimation of Graduated Circles". In: *Cartographica: The International Journal for Geographic Information and Geovisualization*. 14 (2), S. 130–138, DOI: 10.3138/V35N-2387-G7M8-P527.

Cleveland, William S.; Harris, Charles S.; McGill, Robert (1982): „Judgments of Circle Sizes on Statistical Maps". In: *Journal of the American Statistical Association*. 77 (379), S. 541–547.

Conger, Anthony J. (1980): „Integration and generalization of kappas for multiple raters.". In: *Psychological Bulletin*. 88 (2), S. 322–328.

Cox, Carleton W. (1976): „Anchor Effects and the Estimation of Graduated Circles and Squares". In: *Cartography and Geographic Information Science*. 3 , S. 65–74, DOI: 10.1559/152304076784080195.

Croxton, Frederick E. (1932): „Graphic Comparisons by Bars, Squares, Circles, and Cubes". In: *Journal of the American Statistical Association*. 27 (177), S. 54–60.

Davoudi, Mohsen; Quadrelli, Silvia; Osann, Kathryn; u. a. (2008): „A competency-based test of bronchoscopic knowledge using the Essential Bronchoscopist: An initial concept study". In: *Respirology (Carlton, Vic.)*. 13 (5), S. 736–743, DOI: 10.1111/j.1440-1843.2008.01320.x.

Deutsche Gesellschaft für Endoskopie (1974): *Fortschritte der Endoskopie. Verhandlungsbericht; mit 56 Tab. Bd. 5. Bd. 5.* Stuttgart, New York: Schattauer. — ISBN: 978-3-7945-0425-1

Ekman, Gosta; Junge, Kenneth (1961): „Psychological Relations in the Perception of Length, Area, and Volume". In: *Scandinavian Journal of Psychology*. 2 , S. 1–10, DOI: DOI: 10.1111/j.1467-9450.1961.tb01215.x.

Ernst, Armin; Becker, Heinrich D. (2001): „Documentation in Bronchology". In: *Clinics in Chest Medicine*. 22 (2), S. 373–379.

Ernst, Armin; Wahidi, Momen M.; Read, Charles A.; u. a. (2015): „Adult Bronchoscopy Training: Current State and Suggestions for the Future: CHEST Expert Panel Report". In: *Chest*. 148 (2), S. 321–332, DOI: 10.1378/chest.14-0678.

Fennerty, M B; Davidson, J; Emerson, S S; u. a. (1993): „Are endoscopic measurements of colonic polyps reliable?". In: *The American Journal of Gastroenterology*. 88 (4), S. 496–500.

Flannery, James John (1971): „The Relative Effectiveness of Some Common Graduated Point Symbols in the Presentation of Quantitative Data". In: *Cartographica: The International Journal for Geographic Information and Geovisualization*. 8 (2), S. 96–109, DOI: 10.3138/J647-1776-745H-3667.

Fleiss, J.L. (1971): „Measuring nominal scale agreement among many raters". In: *Psychological Bulletin*. (76), S. 378–382.

Gappa, M.; Paton, J.; Baraldi, E.; u. a. (2009): „Paediatric HERMES: update of the European Training Syllabus for Paediatric Respiratory Medicine". In: *European Respiratory Journal*. 33 (3), S. 464–465, DOI: 10.1183/09031936.00001209.

Gwet, Kilem (2011): *AgreeStat*. o.V.

Gwet, Kilem (2002): „Inter-rater reliability: dependency on trait prevalence and marginal homogeneity". In: *Statistical Methods for Inter-Rater Reliability Assessment Series*. 2 , S. 1–9.

Hampton, J R; Harrison, M J; Mitchell, J R; u. a. (1975): „Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients.". In: *British Medical Journal*. 2 (5969), S. 486–489.

Haponik, Edward F.; Russell, Gregory B.; Beamis, John F.; u. a. (2000): „Bronchoscopy Training: Current Fellows' Experiences and Some Concerns for the Future". In: *Chest*. 118 (3), S. 625–630, DOI: 10.1378/chest.118.3.625.

Häussinger, K; Ballin, A; Becker, H D; u. a. (2004): „[Recommendations for quality standards in bronchoscopy]". In: *Pneumologie (Stuttgart, Germany)*. 58 (5), S. 344–56, DOI: 15162262.

Hussein, Rada; Engelmann, Uwe; Schroeter, Andre; u. a. (2004a): „DICOM Structured Reporting: Part 1. Overview and Characteristics“. In: *Radiographics*. 24 (3), S. 891–896, DOI: 10.1148/rg.243035710.

Hussein, Rada; Engelmann, Uwe; Schroeter, Andre; u. a. (2004b): „DICOM Structured Reporting: Part 2. Problems and Challenges in Implementation for PACS Workstations“. In: *Radiographics*. 24 (3), S. 897–909, DOI: 10.1148/rg.243035722.

Kastelik, Jack A.; Chowdhury, Faiza; Arnold, Anthony (2013): „Simulation-Based Bronchoscopy Training“. In: *Chest*. 144 (2), S. 718–719, DOI: 10.1378/chest.13-0880.

Knuth, Donald Ervin (1984): „Literate programming“. In: *The Computer Journal*. 27 (2), S. 97–111.

Leong, A.b.; Green, C.g.; Kurland, G.; u. a. (2014): „A survey of training in pediatric flexible bronchoscopy“. In: *Pediatric Pulmonology*. 49 (6), S. 605–610, DOI: 10.1002/ppul.22872.

Light, Richard J. (1971): „Measures of response agreement for qualitative data: Some generalizations and alternatives.“. In: *Psychological Bulletin*. 76 (5), S. 365–377.

Mackinnon, A (2000): „A spreadsheet for the calculation of comprehensive statistics for the assessment of diagnostic tests and inter-rater agreement“. In: *Computers in Biology and Medicine*. 30 (3), S. 127–134.

Margulies, C; Krevsky, B; Catalano, M F (1994): „How accurate are endoscopic estimates of size?“. In: *Gastrointestinal Endoscopy*. 40 (2 Pt 1), S. 174–177.

Masters, I. B.; Eastburn, M. M.; Francis, P. W.; u. a. (2005): „Quantification of the magnification and distortion effects of a pediatric flexible video-bronchoscope“. In: *Respiratory Research*. 6 (1), S. 16, DOI: 10.1186/1465-9921-6-16.

Meihoefer, Hans-Joachim (1973): „The Visual Perception of the Circle in Thematic Maps/Experimental Results“. In: *Cartographica: The International Journal for Geographic Information and Geovisualization*. 10 (1), S. 63–84, DOI: 10.3138/2771-5577-5417-369T.

Meyer AD; Payne VL; Meeks DW; u. a. (2013): „Physicians' diagnostic accuracy, confidence, and resource requests: A vignette study“. In: *JAMA Internal Medicine*. 173 (21), S. 1952–1958, DOI: 10.1001/jamainternmed.2013.10081.

Pastis, Nicholas J.; Nietert, Paul J.; Silvestri, Gerard A. (2005): „Variation in Training for Interventional Pulmonary Procedures Among US Pulmonary/Critical Care Fellowships: A Survey of Fellowship Directors“. In: *Chest*. 127 (5), S. 1614–1621, DOI: 10.1378/chest.127.5.1614.

Peterson, M. C.; Holbrook, J. H.; Von Hales, D.; u. a. (1992): „Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses.“. In: *Western Journal of Medicine*. 156 (2), S. 163–165.

Rabenstein, T; Maiss, J; Naegele-Jackson, S; u. a. (2002): „Tele-endoscopy: influence of data compression, bandwidth and simulated impairments on the usability of real-time digital video endoscopy transmissions for medical diagnoses“. In: *Endoscopy*. 34 (9), S. 703–710, DOI: 10.1055/s-2002-33568.

Repici, Alessandro; Ciscato, Camilla; Correale, Loredana; u. a. (2016): „Narrow-band imaging international colorectal endoscopic classification to predict polyp histology: REDEFINE (with videos)“. In: *Gastrointestinal Endoscopy*., DOI: 10.1016/j.gie.2016.02.020.

Segnan, Nereo; Bugiani, Massimo; Ronco, Guglielmo; u. a. (1992): „The differential diagnosis of primary lung cancer: Inter-observer agreement and contribution of specific diagnostic procedures“. In: *Journal of Clinical Epidemiology*. 45 (8), S. 827–833, DOI: 10.1016/0895-4356(92)90065-U.

Seidenari, Stefania; Pellacani, Giovanni; Righi, Elena; u. a. (2004): „Is JPEG compression of videomicroscopic images compatible with telediagnosis? Comparison between diagnostic performance and pattern recognition on uncompressed TIFF images and JPEG compressed ones“. In: *Telemedicine journal and e-health: the official journal of the American Telemedicine Association*. 10 (3), S. 294–303, DOI: 10.1089/tmj.2004.10.294.

Teghtsoonian, Martha (1965): „The Judgment of Size“. In: *The American Journal of Psychology*. 78 (3), S. 392–402.

Thompson, Austin B.; Huerta, Guillermo; Robbins, Richard A.; u. a. (1993): „The Bronchitis Index: A Semiquantitative Visual Scale for the Assessment of Airways Inflammation“. In: *Chest*. 103 (5), S. 1482–1488, DOI: 10.1378/chest.103.5.1482.

Torrington, Kenneth G. (2000): „Bronchoscopy Training and Competency: How Many Are Enough?“. In: *Chest*. 118 (3), S. 572–573, DOI: 10.1378/chest.118.3.572.

Wunderlich, P (1969): „EIn dokumentationsgerechter Befundbericht für die Kinderbronchologie“. In: *Zeitschrift für Erkrankungen der Atmungsorgane mit Folia Bronchologica*. 131 , S. 123–130.

6 Anhang

KAPITELVERZEICHNIS

6.1 Reproducible Research

Diese Studie wurde im Sinne von „reproducible research“ durchgeführt. Soweit wie möglich wurde auf quelloffene Software zurückgegriffen. Der den Berechnungen zugrunde liegende Datensatz kann bei Interesse über die Autorin bezogen werden. Sämtliche Berechnungen und Graphiken basieren auf Quellcode mit dem die Ergebnisse jederzeit reproduziert und nachvollzogen werden können.

6.2 Veröffentlichung

Teile dieser Arbeit wurden im Rahmen der Jahrestagungen der Gesellschaft für Pädiatrische Bronchoskopie in Hamburg (Nicolai u. a., 2005) und Dresden (Müller-Sarnowski, 2016) vorgestellt.

Müller-Sarnowski, Ann-Luise (2016): „Inter-Beobachter-Variabilität in der pädiatrischen Bronchoskopie“. Dresden 11.3.2016.

Nicolai, Thomas; Kirsten, Ann-Luise; Gerstlauer, Michael (2005): „Inter-Observer-Variabilität bei der Beurteilung von Kinderbronchoskopien“. In: *Zeitschrift der Gesellschaft für Pädiatrische Pneumologie*. (8), S. 26–29.

6.3 Bibliographie

Die bibliographischen Angaben dieser Arbeit wurden den einzelnen Kapiteln nachgestellt.

6.4 Fragebögen

Auf den beiden folgenden Seiten sind der Befundfragebogen und der Arztfragebogen zusammen mit der Kodierung der einzelnen Antwortoptionen, die zur Übertragung in die Tabellenkalkulation genutzt wurde aufgeführt.

Befundfragebogen

Videoqualität

Bildqualität:	b0	
schlecht	1	☐
ausreichend	2	☐
gut	3	☐
Aufnahmedauer:	b0	
zu kurz	1	☐
ausreichend	2	☐
gut	3	☐
Aufnahmesituation:	b0	
schlecht	1	☐
gut	2	☐

Diagnose

Diagnose [Freitext] b04 (bzw. b04a-c):

Stenose

Maximaler Stenosegrad [%]: b05

Grad

sek. Codierung gemäß Myer-Cotton-Klassifkation (b05a-d)

Lokalisation

Stenoselokalisation:		
Larynx:	*b06&*	
supraglottisch	b06a	☐
glottisch	b06b	☐
subglottisch	b06c	☐
Trachea:	*b07&*	
proximales Drittel	b07a	☐
mittleres Drittel	b07b	☐
distales Drittel	b07c	☐
Hauptbronchus:	*b08&*	
rechts	b08a	☐
links	b08b	☐
Lappenbronchus rechts:	*b09&*	
rechts Oberlappen	b09a	☐
rechts Mittellappen	b09b	☐
rechts Unterlappen	b09c	☐
Lappenbronchus links:	*b10&*	
links Oberlappen	b10a	☐
links Unterlappen	b10b	☐
links Lingula	b10c	☐

Form

Stenoseform:	*b11&*	
kurzstreckig	b11a	☐
langstreckig	b11b	☐
membranös	b11c	☐
ringförmig	b11d	☐

Schleimhaut

Schwellung:	b1	
ja	1	☐
nein	2	☐
Hyperämie:	b1	
ja	1	☐
nein	2	☐
Hypersekretion:	b1	
ja	1	☐
nein	2	☐

Entzündung

Entzündung:	b1	
ja	1	☐
nein	2	☐
Entzündungsbereich:	*b16&*	
Stenosebereich	b16a	☐
Larynx	b16b	☐
Trachea	b16c	☐
Bronchus	b16d	☐
generalisiert	b16e	☐

Malazie

Malazie:	b17&	
Stenosebereich	b17a	☐
Larynx	b17b	☐
Trachea	b17c	☐
Bronchus	b17d	☐
generalisiert	b17e	☐

Pulsationen

Pulsationen:	b18&	
Stenosebereich	b18a	☐
Larynx	b18b	☐
Trachea	b18c	☐
Bronchus	b18d	☐
generalisiert	b18e	☐

Kompressionen

Kompressionen:	b19&	
Stenosebereich	b19a	☐
Larynx	b19b	☐
Trachea	b19c	☐
Bronchus	b19d	☐
generalisiert	b19e	☐

Befund

Video Nr.:

Name oder Kürzel:

Datum:

Arztfragebogen

Status

Qualifikation:	a0	
Assistent	1	☐
Facharzt	2	☐

Ausbildung

Kurse

Kursteilnahme:	a02	
ja	1	☐
nein	2	☐
Anzahl Kurse	a03	
Gesamtdauer [Tage]	a04	

Hospitationen

Hospitationen:	a0	
an Uniklinik	1	☐
an sonstigem Krankenhaus	2	☐
keine	3	☐
Anzahl Hospitationen	a06	
Gesamtdauer [Tage]	a07	

Dauer

Dauer Ausbildung		
[Jahre]	a08	

Ort

Ort der Ausbildung	a0	
Uniklinik	1	☐
Sonstiges Krankenhaus	2	☐
War diese Klinik Ort der übrigen Ausbildung in Pädiatrie?	a10	
ja	1	☐
nein	2	☐

Disziplinen

Disziplinen: Ausbildung in ...		
flexibler Bronchoskopie	a11	
ja	1	☐
nein	2	☐
Wenn ja: Anzahl der Bronchoskopien in Ausbildung	a12	
starrer Bronchoskopie	a13	
ja	1	☐
nein	2	☐
Wenn ja: Anzahl der Bronchoskopien in Ausbildung	a14	
interventioneller Bronchosk.	a15	
ja	1	☐
nein	2	☐
Wenn ja: Anzahl der Bronchoskopien in Ausbildung	a16	

Erfahrung

Ort

Ort der Erfahrung	a1	
Uniklinik	1	☐
Sonstiges Krankenhaus	2	☐
keine	3	☐
War diese Klinik Ort der Anstellung?	a18	
ja	1	☐
nein	2	☐

Disziplinen

Disziplinen: Erfahrung in ...		
flexibler Bronchoskopie	a19	
ja	1	☐
nein	2	☐
Wenn ja: Anzahl der Bronchoskopien seit Ausbildung	a20	
starrer Bronchoskopie	a21	
ja	1	☐
nein	2	☐
Wenn ja: Anzahl der Bronchoskopien seit Ausbildung	a22	
interventioneller Bronchosk.	a23	
ja	1	☐
nein	2	☐
Wenn ja: Anzahl der Bronchoskopien seit Ausbildung	a24	

Arzt

Name oder Kürzel:

Alter:

Datum:

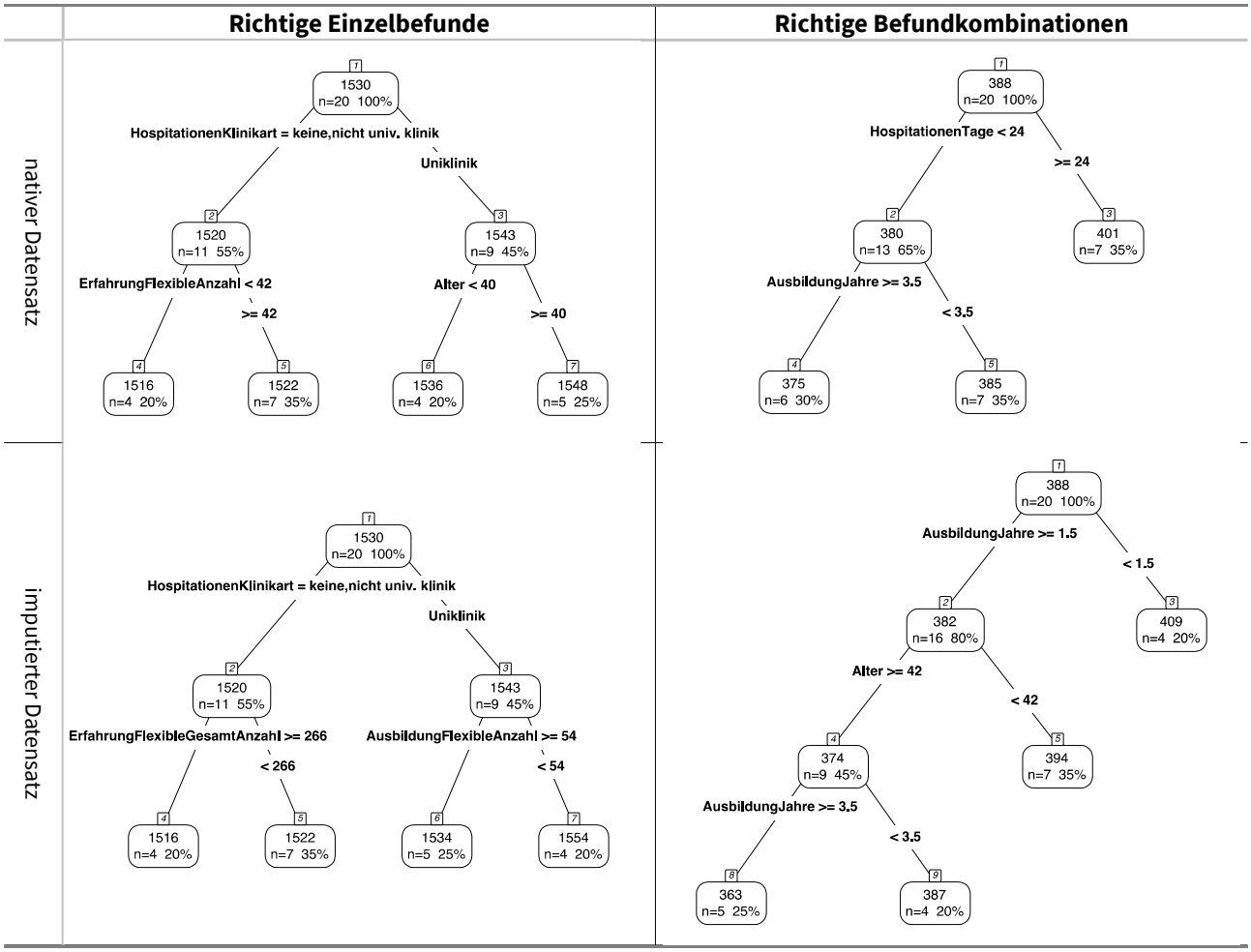
6.5 Ergänzende Berechnungen

Dieser Abschnitt zeigt einen Teil der Analysen, die ergänzend zu denen im Ergebnisteil dargestellten Berechnungen durchgeführt wurden.

6.5.1 CART mit imputiertem Datensatz

Um einen direkteren Vergleich zu den Ergebnissen der linearen Modelle ziehen und den Einfluss der Imputation fehlender Daten auf das Ergebnis abschätzen zu können, wurde ergänzend der imputierte Datensatz untersucht. Im Vergleich zum teilkomplettierten Datensatz erweitert sich der Pool zur Partitionierung herangezogener Variablen im lückenlos komplettierten Datensatz im Prinzip nur um die Anzahl der in der Ausbildung absolvierten Bronchoskopien. Die Gesamtzahl der absolvierten Bronchoskopien (Ausbildung + Bronchoskopien seit Ausbildung) ist gegenüber den Bronchoskopien seit Ausbildung alleine keine wirklich nennenswerte Veränderung. Beide Variablen werden vom Baum der Einzelbefunde gewählt. Der Baum der Befundkombinationen behält die Dauer der Ausbildung als Variable bei und ersetzt die Dauer der Hospitationen als 2. variable durch das Alter. Die Dauer der Ausbildung wird doppelt als Trennkriterium herangezogen.
Paradox ist, dass sich sowohl die Anzahl der Bronchoskopien in der Ausbildung, als auch die Gesamtzahl der absolvierten Bronchoskopien im Modell negativ auf die erzielte Befundrichtigkeit auswirken.

Abbildung 6.1: Vergleich von CART auf Basis von nativem und imputiertem Datensatz



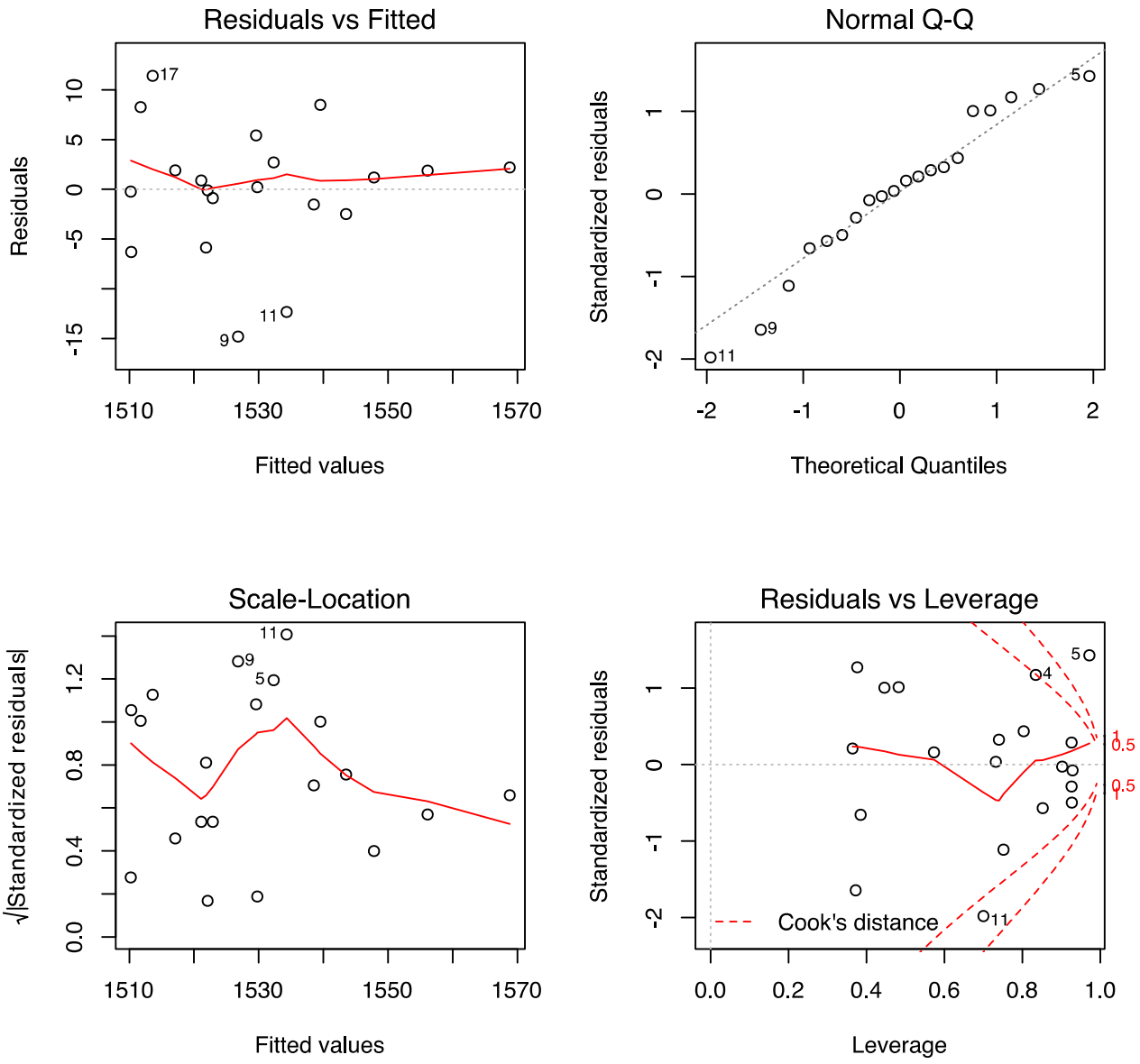
Vergleich der mit CART erzeugten Entscheidungsbäume

Der Vergleich der Bäume zeigt, dass verhältnismäßig geringfügige Änderungen des Datensatzes (nur ca. 4 % der Werte wurden imputiert!) erhebliche Änderungen an der Struktur der Bäume nach sich ziehen. Das gilt insbesondere für den Baum mit Befundkombinationen als Zielvariable.

6.5.2 Qualität multiple lineare Regressionen

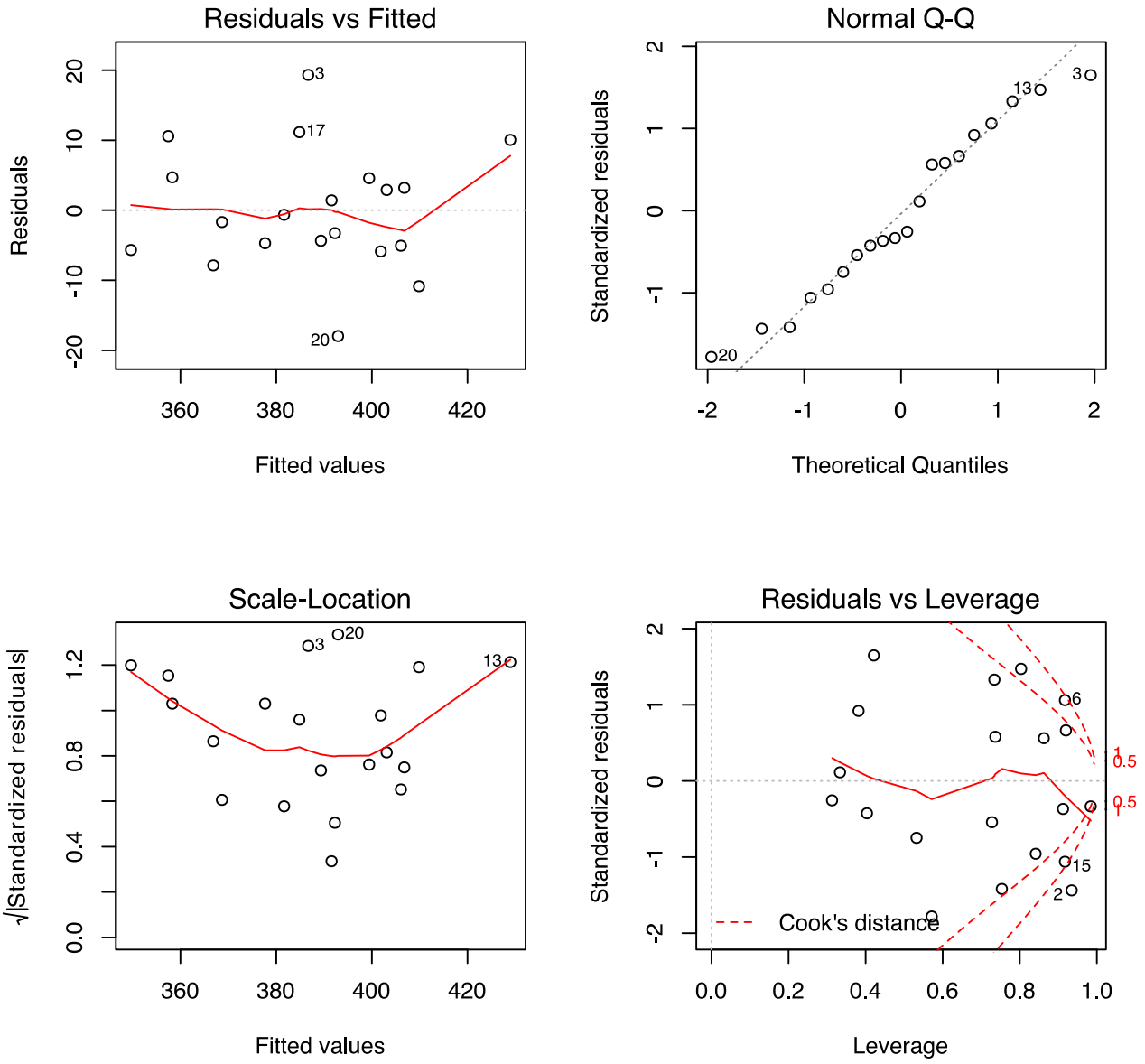
Anhand der folgenden Graphiken kann die Qualität der linearen Regressionsmodelle nachvollzogen werden.

Abbildung 6.2: Multiple lineare Regression Einzelbefunde



Das Modell hält der Testbatterie gvlma stand. Die Residuen des Modells nehmen mit zunehmender Trefferquote trichterförmig ab, was insbesondere auf die Outlier 9, 11 und 17 zurückzuführen ist. Im Q-Q Plot finden sich insbesondere an den Rändern erkennbare Abweichungen von einer Normalverteilung.

Abbildung 6.3: Multiple lineare Regression Befundkombinationen



Die Residuen scheinen gleichmäßiger über den gesamten Wertebereich zu streuen, als im Modell auf Symptomebene. Im Q-Q-Plot Fallen die Abweichungen von der Normalverteilung an den Rändern etwas milder aus. Hingegen gibt es im Zentrum eine deutlich erkennbare s-förmige Abweichung.

6.6 Verzeichnisse

6.6.1 Abbildungsverzeichnis

6.6.2 Tabellenverzeichnis

6.6.3 Formelverzeichnis

6.6.4 Verzeichnis R-Ausgaben

6.6.5 Verzeichnis der Exkurse

6.6.6 Verzeichnis der Zusammenfassungen

6.6.7 Verzeichnis der Fazits

Eidesstattliche Versicherung

Müller-Sarnowski, Ann-Luise

Ich erkläre hiermit an Eides statt,
dass ich die vorliegende Dissertation mit dem Thema
Inter-Beobachter-Variabiliät in der pädiatrischen Bronchoskopie

Selbstständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

München 27.11.2017

Ann-Luise Müller-Sarnowski