Kapitel 2

Methodik

2.1 Daten (Studie)

Die Daten entstammen einer prospektiven diagnostischen Studie, die von 08/1999 bis 11/2001 an dem Universitätsklinikum Charité durchgeführt wurde. 206 aufeinander folgende Patienten mit Verdacht auf Pankreaskarzinom oder Pankreasläsion unklarer Genese wurden mit sechs verschiedenen diagnostischen Verfahren untersucht: konventioneller transabdomineller Ultraschall (US), endoskopischer Ultraschall (EUS), Computertomographie unter Verwendung von Kontrastmittel in Spiraltechnik (CT), Magnet-Resonanz-Tomographie (MRT) mit Einbeziehung von MR-Cholangio-Pankreatikographie (MRCP) und MR-Angiographie (MRA), endoskopisch retrograde Cholangio-Pankreaticographie (ERCP) sowie Positronen-Emissions-Tomographie mit Fluor-markierter Desoxy-Glucose (FDG-PET). Die Studie wurde von der lokalen Ethikkomission bewilligt und alle Patienten gaben eine schriftliche Einverständniserklärung.

Patienten mit Kontraindikation für bestimmte Modalitäten wie z.B. Patien-

ten mit B-II Mägen bei EUS, Klaustrophobie bei MR oder Kontrastmittelallergie bei CT wurden von den jeweiligen Untersuchungen ausgeschlossen, ebenso wie Patienten, die sich weigerten, einzelne Untersuchungen zu absolvieren. Logistische Gründe (z.B. Gerätedefekte) waren weitere Ursachen, Verfahren bei einzelnen Patienten nicht zu durchzuführen. Technische Details zu den bildgebenden Verfahren sind publiziert unter [39].

Alle Untersuchungen an einem Patienten wurden innerhalb von zwei Wochen durchgeführt. Die Reihenfolge der Tests erfolgte dabei nach klinischer Verfügbarkeit. Die Untersuchungen wurden von unabhängigen und erfahrenen Klinikern durchgeführt, die Zugang zu allen klinischen Informationen hatten, aber gegenüber den Ergebnissen der anderen Modalitäten geblindet waren.

Kriterien für lokale Irresektabilität waren Tumorinfiltration der obere Mesenterialarterie, der Mesenterialwurzel und/oder des Mesokolons. Infiltration der Portal- oder Mesenterialvene war, ebenso wie Tumorgröße oder das Vorhandensein von vergrößerten peripankreatischen oder paraaortalen Lymphknoten kein Kriterium für Irresektabilität. Neben der lokalen Irresektabilität war das Vorliegen von Fernmetastasen ein Ausschlusskriterium für die Resektabilität. Die mittlere Zeit zwischen Diagnostik und Operation betrug für die Patienten, die operiert wurden, 31 Tage (Spannweite: 7-51 Tage).

Die endgültige Diagnose (Goldstandard) wurde durch Laparotomie (entweder Resektion oder explorative Laparotomie mit Biopsie) oder durch histologische Analyse der gewonnenen Zytologieproben gestellt. Tumore wurden außerdem als maligne klassifiziert, wenn Patienten mit nicht histologisch gesi-

cherten Metastasen innerhalb der nächsten sechs Monate an Tumorkachexie oder anderen Komplikationen verstarben. Patienten, bei denen kein Goldstandard vorlag, wurden zunächst als benigne gewertet und dann in drei- bis sechsmonatigen Intervallen nachkontrolliert. Wenn sich in innerhalb eines Jahres die Läsion nicht veränderte, wurde der Tumor endgültig als benigne klassifiziert. Patienten, die alle Kontrolluntersuchungen ablehnten, aber nach einem Jahr nach Einschluss in die Studie noch am Leben waren, wurden ebenfalls als benigne klassifiziert. Patienten mit nur einem Jahr Nachbeobachtung ohne weitere diagnostische Tests wurden ebenso wie Patienten, die innerhalb des ersten Jahres verstarben, aber für die keine detaillierte Information vorlag, als Drop-outs gewertet.

2.2 Validitätsmaße diagnostischer Tests

Ein diagnostischer Test soll Aufschluss darüber geben, ob ein Patient an einer bestimmten Krankheit leidet oder nicht. In diesem Sinne kann jede Untersuchung, die zur Aufdeckung einer bestimmten Krankheit oder eines bestimmten pathologischen Phänomens beiträgt, als diagnostischer Test aufgefasst werden. Dabei ist der Begriff der Untersuchung sehr weit gefasst und schließt z.B. Antworten auf Fragen bei der Anamnese, körperliche Befunde, Laborwerte, physikalische Messwerte oder Interpretationen von Ergebnissen bildgebender Verfahren mit ein [40].

Normalerweise werden die Ergebnisse der Untersuchung zur Einteilung der Patienten in einen von zwei Zuständen (krank - nicht krank bzw. Phänomen vorhanden - nicht vorhanden) genutzt. Eine Erweiterung auf mehr als zwei

Zustände z.B. für verschiedene Differentialdiagnosen ist möglich, aber nicht Gegenstand dieser Arbeit. Im einfachsten Fall, den binären Tests, sind auch die Ergebnisse des diagnostischen Tests dichotom, der Test wird entweder positiv oder negativ gewertet.

2.2.1 Sensitivität und Spezifität

Die Validität eines diagnostischen Tests gibt an, wie gut der Test den tatsächlichen Sachverhalt erkennt. Voraussetzung für die Beurteilung der Validität eines Tests ist, dass dieser „wahre" Sachverhalt durch ein unabhängiges und fehlerfreies Außenkriterium, den so genannten Goldstandard festgestellt werden kann.

Der Grad der Übereinstimmung zwischen Testergebnis und tatsächlichem Zustand kann bei binären Tests durch Kreuzklassifikation der beiden dichotomen Merkmale Goldstandard und Testergebnis in einer Vierfeldertafel dargestellt werden (siehe Tabelle 2.1).[1]

Die **Sensitivität** (Sens) eines diagnostischen Tests beschreibt seine Fähigkeit, tatsächlich Kranke als krank, die **Spezifität** (Spez) beschreibt seine Fähigkeit, tatsächlich Gesunde als gesund zu erkennen:

$$\text{Sens} = \text{P}(\text{T}+|\text{K}+) \qquad \text{bzw.} \qquad \text{Spez} = \text{P}(\text{T}-|\text{K}-).$$

[1]Hinweis: Für die Parameter der Population und der Stichprobe wird die gleiche Bezeichnung verwendet. Aus dem Zusammenhang ergibt sich jeweils, ob es sich um einen festen Wert der Population, einen beobachteten Wert in der Stichprobe oder allgemein um eine Zufallsgröße handelt.

		Goldstandard		
		K+	K-	
Testergebnis	T+	n_{++}	n_{+-}	$n_{+\cdot}$
	T-	n_{-+}	n_{--}	$n_{-\cdot}$
		$n_{\cdot+}$	$n_{\cdot-}$	n

Tabelle 2.1: *Der Goldstandard hat die beiden Ausprägungen krank (K+) und gesund (K-). Entsprechend steht T+ für ein positives und T- für ein negatives Testergebnis.* n_{++} *ist die Anzahl positiver,* n_{-+} *die Anzahl negativer Testergebnisse bei den Kranken,* n_{+-} *und* n_{--} *kennzeichnen die Anzahl positiver und negativer Ergebnisse bei den Gesunden.*

D.h. die Sensitivität ist die bedingte Wahrscheinlichkeit, dass bei Vorliegen der Krankheit der Test positiv ist und die Spezifität ist die bedingte Wahrscheinlichkeit, dass bei Gesunden der Test negativ ist.

Ein Test ist valide, wenn er in der Lage ist, zwischen Gesunden und Kranken zu trennen, d.h. wenn der Anteil positiver Ergebnisse bei den Kranken höher ist als bei den Gesunden[2]:

$$\text{Test valide} \Leftrightarrow \text{Sens} > 1 - \text{Spez} \Leftrightarrow \text{Sens} + \text{Spez} > 1. \tag{2.1}$$

Sensitivität und Spezifität sind dem jeweiligen diagnostischen Test innewohnende Eigenschaften, sie sind insbesondere unabhängig von der Prävalenz. Allerdings werden sie von dem betrachteten Patientenspektrum beeinflusst. Da fortgeschrittene Erkrankungen meist leichter erkannt werden, ist die Sensitivität bei Patienten im Frühstadium oft geringer als bei Patienten im

[2] $\Leftrightarrow$ bedeutet „genau dann wenn"

Spätstadium. Ebenso ist es einfacher, Kranke von völlig Gesunden zu unterscheiden, so dass die Spezifität abnimmt, wenn statt gesunder Probanden Patienten mit Differentialdiagnosen, die ähnliche Symptome hervorrufen können, untersucht werden.

Für die Schätzung von Sensitivität und Spezifität muss das Design der Studie berücksichtigt werden [41]. Bei der *prästratifizierten Erhebung* ist die Zahl der Kranken $n_{\cdot +}$ und Gesunden $n_{\cdot -}$ vorgegeben, die untersuchte Population besteht aus zwei voneinander unabhängigen Teilpopulationen Kranker und Gesunder. Dazu muss der wahre Zustand (Goldstandard) vor der Durchführung des diagnostischen Tests bekannt sein. Bei der *ungeschichteten Erhebung* ist nur die Gesamtzahl n der Patienten vorgegeben, die tatsächlich beobachtete Anzahl der Kranken bzw. der Gesunden steht nicht vorher fest, sondern ist eine von der Prävalenz abhängige Zufallsgröße.

Im Fall einer prästratifizierten Erhebung sind die Zahl der richtig Positiven und die Zahl der richtig Negativen binomialverteilt nach B($n_{\cdot +}$, Sens) bzw. B($n_{\cdot -}$, Spez) mit Erwartungswert $E(n_{++}) = n_{\cdot +}\cdot$Sens bzw. $E(n_{--}) = n_{\cdot -}\cdot$Spez und Varianz Var($n_{++}$) = $n_{\cdot +}\cdot$Sens$\cdot$(1-Sens) bzw. Var(n_{--}) = $n_{\cdot -}\cdot$ Spez$\cdot$(1-Spez).

Unverzerrte[3] Maximum-Likelihood-Schätzer für die Sensitivität und die Spe-

[3]Ein Schätzer heißt unverzerrt oder erwartungstreu, wenn sein Erwartungswert der zu schätzende Parameter ist. D.h. der Schätzer liegt „im Mittel" richtig, hat also keinen systematischen Fehler (Bias).

zifität sind die beobachteten Anteile:

$$\widehat{\text{Sens}} = \frac{n_{++}}{n_{.+}} \qquad \text{und} \qquad \widehat{\text{Spez}} = \frac{n_{--}}{n_{.-}}. \tag{2.2}$$

Die Varianzen der Schätzer ergeben sich ebenfalls aus der Binomialverteilung:

$$\text{Var}(\widehat{\text{Sens}}) = \frac{\text{Sens}(1-\text{Sens})}{n_{.+}} \qquad \text{bzw.} \qquad \text{Var}(\widehat{\text{Spez}}) = \frac{\text{Spez}(1-\text{Spez})}{n_{.-}}$$

und können geschätzt werden durch

$$\widehat{\text{Var}}(\widehat{\text{Sens}}) = \frac{\widehat{\text{Sens}}(1-\widehat{\text{Sens}})}{n_{.+}} = \frac{n_{++} \cdot n_{-+}}{n_{.+}^3}$$

bzw.

$$\widehat{\text{Var}}(\widehat{\text{Spez}}) = \frac{\widehat{\text{Spez}}(1-\widehat{\text{Spez}})}{n_{.-}} = \frac{n_{+-} \cdot n_{--}}{n_{.-}^3}.$$

Im Fall einer ungeschichteten Erhebung liegt ein multinomiales Modell vor und die Schätzer aus (2.2) sind nicht erwartungstreu [42]:

$$\begin{aligned} \text{E}(\widehat{\text{Sens}}) &= \text{Sens}(1-\delta/n) + o(n^{-1}) \\ \text{E}(\widehat{\text{Spez}}) &= \text{Spez}(1-\delta/n) + o(n^{-1}), \end{aligned}$$

wobei $\delta = \frac{1-n_{.+}/n}{n_{.+}/n}$ und o für das Landau-Symbol steht, d.h. der zweite Summand geht für große n gegen Null.

Approximative Varianzen erhält man durch Taylorentwicklung der Varianz- und Kovarianzformeln der Multinomialverteilung. Die üblicherweise verwendete Formel [43] ohne Terme höherer Ordnung entspricht der Varianz bei prästratifizierten Erhebungen. Für die Schätzung der Varianz der Sensitivität ergibt sich z.B.:

$$\widehat{\text{Var}}(\widehat{\text{Sens}}) \approx \frac{n_{++}/n \cdot n_{-+}/n}{(n_{.+}/n)^3 \cdot n} = \frac{n_{++} \cdot n_{-+}}{n_{.+}^3}. \tag{2.3}$$

Wenn auch Terme zweiter Ordnung berücksichtigt werden, vergrößert sich die Varianz geringfügig [42]:

$$\mathrm{Var}(\widehat{\mathrm{Sens}}) \approx \frac{n_{++} \cdot n_{-+}}{n_{\cdot+}^3} \left(1 + \frac{n_{\cdot+}/n}{n(1 - n_{\cdot+}/n)}\right).$$

Bei der Berechnung von Konfidenzintervallen wird im ungeschichteten Fall meist die Varianz aus Formel (2.3) verwendet, so dass die Konfidenzintervalle für beide Fälle (stratifizierte und ungeschichtete Stichprobe) gleich sind.

Die Grenzen eines approximativen Konfidenzintervalls enthält man durch Normalverteilungsapproximation:

$$\widehat{\mathrm{Sens}} \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\mathrm{Sens}})} = \widehat{\mathrm{Sens}} \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{\mathrm{Sens}}(1 - \widehat{\mathrm{Sens}})}{n_{\cdot+}}}$$

bzw.

$$\widehat{\mathrm{Spez}} \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\mathrm{Spez}})} = \widehat{\mathrm{Spez}} \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{\mathrm{Spez}}(1 - \widehat{\mathrm{Spez}})}{n_{\cdot-}}}.$$

wobei $z_{1-\alpha/2}$ das $1 - \alpha/2$-Quantil der Standardnormalverteilung ist.

Diese approximativen Konfidenzintervalle, oft auch Wald-Konfidenzintervalle genannt, haben zwei Nachteile. Zum einen halten sie das Konfidenzniveau, insbesondere bei kleinen Fallzahlen und bei Wahrscheinlichkeiten nahe bei Eins oder Null, oft nicht ein. Die Überdeckungswahrscheinlichkeit liegt in einigen Fällen nur bei 60% [44]. Zum anderen ist es möglich, dass die obere Intervallgrenze Eins überschreitet bzw. die untere Intervallgrenze kleiner als Null wird (Overshot). Durch Verwendung einer Kontinuitätskorrektur erhöht

sich die Überdeckungswahrscheinlichkeit, allerdings kommt es zu mehr Overshot.

Exakte Konfidenzintervalle nach Clopper und Pearson [45] basieren auf der Invertierung des exakten Tests von Fisher. Sie garantieren die strikte Einhaltung des Niveaus, sind aber sehr konservativ[4]. Wenn der p-Wert der beobachteten Tafel nur halb gezählt wird, erhält man die so genannten mid-p-value Konfidenzintervalle. Die mittels mid-p-value-Methode gewonnenen Konfidenzintervalle für Wahrscheinlichkeiten [46] sind weniger konservativ, aber immer noch zu breit.

Newcombe [47] empfiehlt aufgrund seiner Simulationsstudien die Methode von Wilson [48]. Dessen auf den Likelihood Scores beruhenden Score-Intervalle halten im Durchschnitt das Niveau ein und zeigen keinen Overshot. Das Score-Konfidenzintervall für die Sensitivität wird berechnet mittels:

$$\frac{2 \cdot \widehat{\text{Sens}} + z^2_{1-\alpha/2} \pm z_{1-\alpha/2}\sqrt{z^2_{1-\alpha/2} + 4 \cdot n_{\cdot+} \cdot \widehat{\text{Sens}} \cdot \left(1 - \widehat{\text{Sens}}\right)}}{2\left(n_{\cdot+} + z^2_{1-\alpha/2}\right)}. \tag{2.4}$$

Durch Ersetzen von $\widehat{\text{Sens}}$ durch $\widehat{\text{Spez}}$ und von $n_{\cdot+}$ durch $n_{\cdot-}$ erhält man ein Konfidenzintervall für die Spezifität.

Die von Agresti und Coull [44] vorgeschlagenen und von Zhou [49] empfohlenen Konfidenzintervalle sind eine einfache Approximation der Score Intervalle von Wilson und können in der Form der Wald-Konfidenzintervalle dargestellt

[4]D.h. ihre Überdeckungswahrscheinlichkeit ist größer als 1-α, die Konfidenzintervalle sind somit breiter als notwendig.

werden:

$$\widetilde{\text{Sens}} \pm z_{1-\alpha/2}\sqrt{\frac{\widetilde{\text{Sens}}(1-\widetilde{\text{Sens}})}{\tilde{n}_{\cdot+}}}$$

wobei $\tilde{n}_{\cdot+} = n_{\cdot+} + z^2_{1-\alpha/2}$ und $\widetilde{\text{Sens}} = \frac{\widehat{\text{Sens}}+z^2_{1-\alpha/2}/2}{n_{\cdot+}+z^2_{1-\alpha/2}}$.
Durch Ersetzen von $\widetilde{\text{Sens}}$ durch $\widetilde{\text{Spez}}$, $\widehat{\text{Sens}}$ durch $\widehat{\text{Spez}}$ und von $n_{\cdot+}$ durch $n_{\cdot-}$ erhält man das entsprechende Konfidenzintervall für die Spezifität.

Bei beiden nach der Score Methode berechneten Konfidenzintervallen ist die strikte Einhaltung des Niveaus nicht gewährleistet. In beiden Fällen wird das Niveau aber deutlich besser eingehalten als bei der klassischen Methode mittels Normalverteilungsapproximation [47].

2.2.2 Prädiktive Werte

Sensitivität und Spezifität beziehen sich auf die Gruppe der Erkrankten bzw. der Gesunden. Da im medizinischen Alltag vorher nicht bekannt ist, ob ein Patient krank ist oder nicht, sondern aufgrund des Testergebnisses beurteilt werden soll, ob die Krankheit vorliegt, sind die Wahrscheinlichkeiten, mit denen ein diagnostischer Test zu einer korrekten Diagnose führt, von praktischer Bedeutung.

Die Wahrscheinlichkeit, dass ein Patient mit positivem Befund tatsächlich krank ist, wird als **positiver prädiktiver Wert** bezeichnet:

$$\text{ppW} = \text{P}(\text{K}+|\text{T}+).$$

Analog ist der **negative prädiktive Wert** definiert als:

$$\text{npW} = \text{P}(\text{K}-|\text{T}-).$$

Die prädiktiven Werte werden analog zu Sensitivität und Spezifität durch die beobachteten Anteile geschätzt:

$$\widehat{\text{ppW}} = \frac{n_{++}}{n_{+\cdot}} \qquad \text{und} \qquad \widehat{\text{npW}} = \frac{n_{--}}{n_{-\cdot}}. \tag{2.5}$$

Die prädiktiven Werte sind kein stabiles Charakteristikum des diagnostischen Tests, sie sind abhängig vom Anteil der Kranken in der untersuchten Population, der sogenannten Prävalenz (Präv). Wenn die Prävalenz abnimmt, verringert sich die Zahl der richtig Positiven im Verhältnis zu den falsch Positiven und der positive prädiktive Wert wird kleiner. Der Anteil richtig Negativer hingegen steigt im Verhältnis zu den falsch Negativen, der negative prädiktive Wert wird mit abnehmender Prävalenz größer. Da bei diagnostischen Studien die Prävalenz oft deutlich höher ist als im klinischen Alltag, ist Vorsicht geboten bei der Übertragung von Studienergebnissen in die Praxis. Der positive prädiktive Wert wird in der Praxis meist deutlich niedriger, der negative prädiktive Wert hingegen im Alltag höher sein als in der Studienpopulation.

Mittels des Satz von Bayes können die prädiktiven Werte in Abhängigkeit von Sensitivität, Spezifität und Prävalenz berechnet werden:

$$\text{ppW} = \frac{\text{Präv} \cdot \text{Sens}}{\text{Präv} \cdot \text{Sens} + (1 - \text{Präv}) \cdot (1 - \text{Spez})}$$

bzw.

$$\text{npW} = \frac{(1 - \text{Präv}) \cdot (\text{Spez})}{(1 - \text{Präv}) \cdot (\text{Spez}) + \text{Präv} \cdot (1 - \text{Sens})}.$$

2.2.3 Likelihood Ratios

Durch Darstellung der Wahrscheinlichkeiten als Odds (Chancenverhältnisse) können obige Formeln auch formuliert werden als:

$$\frac{\text{ppW}}{1-\text{ppW}} = \underbrace{\frac{\text{P(K+|T+)}}{\text{P(K-|T+)}}}_{\textit{post-Test Odds}} = \underbrace{\frac{\text{Präv}}{1-\text{Präv}}}_{\textit{prä-Test Odds}} \cdot \frac{\text{Sens}}{1-\text{Spez}} \tag{2.6}$$

bzw.

$$\frac{1-\text{npW}}{\text{npW}} = \underbrace{\frac{\text{P(K+|T-)}}{\text{P(K-|T-)}}}_{\textit{post-Test Odds}} = \underbrace{\frac{\text{Präv}}{1-\text{Präv}}}_{\textit{prä-Test Odds}} \cdot \frac{1-\text{Sens}}{\text{Spez}} \tag{2.7}$$

Die in (2.6) und (2.7) verwendeten Brüche, die die Veränderung der prä-Test Odds durch den Test beschreiben, werden auch Likelihood Ratios genannt. Das **positive Likelihood Ratio** ist gemäß

$$\text{LR}^+ = \frac{\text{Sens}}{1-\text{Spez}} \tag{2.8}$$

und das **negative Likelihood Ratio** entsprechend als

$$\text{LR}^- = \frac{1-\text{Sens}}{\text{Spez}} \tag{2.9}$$

definiert.

Die Likelihood Ratios werden auch Vorhersageschärfe genannt und geben an, um wie viel wahrscheinlicher ein positives bzw. negatives Ergebnis bei einem Kranken in Verhältnis zu einem Gesunden ist. Sie sind ein Maß dafür, um wie viel ein positives oder negatives Resultat die Prä-Test-Wahrscheinlichkeiten verändert bzw. um wie viel die Sicherheit oder Evidenz durch ein positives bzw. negatives Testergebnis zunimmt.

Nach Bender [50] kann man die diagnostische Leistungsfähigkeit eines Testes in Abhängigkeit von den Likelihood Ratios folgendermaßen beurteilen:

LR^+	LR^-	Bewertung
> 10	$< 0{,}1$	sehr gut
5-10	0,1-0,2	gut
2-5	0,2-0,5	mäßig
1-2	0,5-1,0	schlecht

Tabelle 2.2: *Bewertung der Likelihood Ratios*

Einen Punktschätzer für LR^+ und LR^- erhält man, indem die Schätzungen für die Sensitivität und die Spezifität in Gleichungen (2.8) bzw. (2.9) eingesetzt werden. Im Falle prästratifizierter Erhebungen sind LR^+ und LR^- Quotienten zweier unabhängiger Wahrscheinlichkeiten, die Konfidenzintervalle können analog zum relativen Risiko berechnet werden.

Meist wird dazu die so genannte Log-Methode [51] verwendet. Da die Verteilung der Likelihood Ratios schief ist, werden LR^+ bzw. LR^- zunächst logarithmiert, um annähernd normalverteilte Größen zu erhalten. Die Varianz der logarithmierten Likelihood Ratios kann geschätzt werden durch [52]:

$$\widehat{\mathrm{Var}}\left(\ln(\widehat{\mathrm{LR}^+})\right) = \frac{1-\widehat{\mathrm{Sens}}}{n_{++}} + \frac{\widehat{\mathrm{Spez}}}{n_{+-}} \tag{2.10}$$

bzw.

$$\widehat{\mathrm{Var}}\left(\ln(\widehat{\mathrm{LR}^-})\right) = \frac{\widehat{\mathrm{Sens}}}{n_{-+}} + \frac{1-\widehat{\mathrm{Spez}}}{n_{--}} \tag{2.11}$$

Damit können Konfidenzintervalle für die logarithmierten Likelihood Ratios berechnet werden. Durch Rücktransformation erhält man anschließend die gewünschten Intervallgrenzen:

$$\exp\left(\ln\left(\frac{\widehat{\text{Sens}}}{1-\widehat{\text{Spez}}}\right) \pm z_{1-\alpha/2}\sqrt{\frac{1-\widehat{\text{Sens}}}{n_{++}} + \frac{\widehat{\text{Spez}}}{n_{+-}}}\right).$$

Falls eine der beobachteten Häufigkeiten in Tabelle 2.1 und damit Sens, 1-Sens, Spez oder 1-Spez gleich Null ist, sind die Varianzen in (2.10) bzw. (2.11) undefiniert. In diesen Fällen behilft man sich durch Addition von +0,5 zu allen Zellen.

Koopman [53] leitete ein Score-Konfidenzintervall für das Verhältnis zweier binomial verteilter Proportionen her, das im Gegensatz zu obigen mittels Log-Methode ermittelten immer konsistent zum Pearsonschen χ^2-Test ist. Sein Konfidenzintervall basiert auf der Prüfgröße für den Test der Hypothesen $H_0 : \frac{p_1}{p_2} = \theta_0$ gegen $H_1 : \frac{p_1}{p_2} \neq \theta_0$, wobei die beiden Wahrscheinlichkeiten p_1 und p_2 durch die Likelihood-Ratio-Methode mit der Restriktion $\frac{p_1}{p_2} = \theta_0$ geschätzt werden. Das Konfidenzintervall ist asymmetrisch und muss, da keine explizite Formel existiert, iterativ berechnet werden.

2.2.4 Weitere Validitätsmaße

Als weitere Validitätsmaße werden auch die **Aufdeckungsrate** n_{++}/n, der **Gesamtfehler** $= (n_{+-} + n_{-+})/n$ oder die **Treffsicherheit (accuracy)** $(n_{++} + n_{--})/n$ zur Beschreibung der Validität eines diagnostischen Tests verwendet. Diese Maßzahlen sind einfach zu berechnen und zu interpretieren, haben aber den Nachteil, dass sie von der Prävalenz abhängen.

Der **Youden-Index** (YI), definiert als

$$\text{YI} = \text{Sens} + \text{Spez} - 1$$

ist hingegen prävalenzunabhängig. Er liegt zwischen +1 und -1 und ist größer als Null, wenn ein positiver Zusammenhang zwischen Testergebnis und tatsächlichem Sachverhalt besteht. Ein Test ist diagnostisch wertlos, wenn der Youden-Index gleich Null ist (vgl. (2.1)). Falls die Werte kleiner Null sind, ist die Zuordnung krank - gesund falsch und muss umgekehrt werden.

Der Vorteil des Youden-Index, die Validität im Gegensatz zu obigen paarigen Maßzahlen mit einem Wert auszudrücken, ist gleichzeitig der größte Nachteil. Durch die Zusammenfassung gehen Informationen verloren, wobei besonders problematisch ist, dass falsch Positive und falsch Negative gleich gewichtet werden.

2.3 Vergleich diagnostischer Tests

In der medizinischen Praxis gibt es oft mehrere unterschiedliche diagnostische Tests für die gleiche Krankheit. Es liegt nahe, die diagnostische Leistungsfähigkeit der Tests zu vergleichen und zu testen, ob einer der Tests den anderen überlegen ist. Bei dichotomen Tests werden dazu meist die Sensitivitäten und Spezifitäten verglichen.

In diesem Fall lautet die Nullhypothese, dass die Sensitivität bzw. die Spezifität bei den beiden Tests A und B gleich ist, und die Alternativhypothese, dass sich die beiden Gütemaße unterscheiden:

$$\text{H}_0 : \theta_A = \theta_B \quad \text{vs.} \quad \text{H}_A : \theta_A \neq \theta_B, \qquad \text{mit } \theta \in \{\text{Sens, Spez}\}$$

Die verwendete Prüfgröße ist abhänig davon, ob beide Tests an den gleichen Patienten oder verschiedenen Patientengruppen durchgeführt wurden, d.h. ob gepaarte oder ungepaarte Stichproben vorliegen.

In dem deutlich häufigeren Fall gepaarter Stichproben lassen sich die Ergebnisse beider Tests gemäß Tabelle 2.3 mittels Kreuzklassifikation darstellen.

	Kranke		Gesunde	
	T_B^+	T_B^-	T_B^+	T_B^-
T_A^+	n_{++}^K	n_{+-}^K	n_{++}^G	n_{+-}^G
T_A^-	n_{-+}^K	n_{--}^K	n_{-+}^G	n_{--}^G

Tabelle 2.3: *Kreuzklassifikation der Ergebnisse zweier Tests bei gepaarten Stichproben. T_A^+ bezeichnet ein positives Ergebnis bei Test A, T_A^- ein negatives. Analog bezeichnen T_B^+ bzw. T_B^- ein positives bzw. negatives Ergebnis bei Test B.*

Für den Vergleich der Sensitivitäten bzw. Spezifitäten lauten die asymptotisch χ_1^2-verteilten Prüfgrößen:

$$\chi^2_{\mathrm{Sens}} = \frac{(n_{+-}^K - n_{-+}^K)^2}{(n_{+-}^K - n_{-+}^K)} \qquad \text{bzw.} \qquad \chi^2_{\mathrm{Spez}} = \frac{(n_{+-}^G - n_{-+}^G)^2}{(n_{+-}^G - n_{-+}^G)}.$$

Dieser Test entspricht dem McNemar-Test innerhalb der Subpopulation der Kranken bzw. Gesunden. Falls die Anzahl der diskordanten Paare $(n_{+-}^K + n_{-+}^K)$ bzw. $(n_{+-}^G + n_{-+}^G)$ kleiner als 20 ist, sollte eine exakte Version [54] oder der Binomialtest mit $\pi = 1/2$ und $n = n_{+-}^K + n_{-+}^K$ bzw. $n = n_{+-}^G + n_{-+}^G$ verwendet werden.

Es existieren verschiedene Methoden zur Berechnung von Konfidenzintervallen für die Differenz zweier Proportionen bei gepaarten Stichproben. Newcombe [55] zeigt, dass asymptotische Methoden zu Overshot (vgl. Seite 20) neigen und ebenso wie die exakte Methode eine schlechte Überdeckungswahrscheinlichkeit haben. Bessere Ergebnisse erzielen die auf der Profile Likelihood[5] basierenden Methoden. Diese sind jedoch nur iterativ zu berechnen.

Newcombe [55] empfiehlt aufgrund seiner Simulationsstudien einfacher zu berechnende, auf den Score-Intervallen von Wilson basierenden Konfidenzintervalle. Dazu werden zunächst nach Formel (2.4) Konfidenzintervalle für p_1 und p_2 berechnet, die dann mittels des ϕ-Koeffizeinten für Abhängigkeit korrigiert werden.

	Kranke			Gesunde		
	+	-	Σ	+	-	Σ
T_A	n^K_{A+}	n^K_{A-}	$n^K_{A\cdot}$	n^G_{A+}	n^G_{A-}	$n^G_{A\cdot}$
T_B	n^K_{B+}	n^K_{B-}	$n^K_{B\cdot}$	n^G_{B+}	n^G_{B-}	$n^G_{B\cdot}$
	$n^K_{\cdot+}$	$n^K_{\cdot-}$	$n^K_{\cdot\cdot}$	$n^G_{\cdot+}$	$n^G_{\cdot-}$	$n^G_{\cdot\cdot}$

Tabelle 2.4: *Kreuzklassifikation der Ergebnisse zweier Tests bei ungepaarten Stichproben. Mit T_A werden die Ergebnisse von Test A, mit T_B die Ergebnisse von Test B bezeichnet. + kennzeichnet ein positives, - ein negatives Testergebnis.*

[5] Die Profile Likelihood (L_P) für einen interessierenden Parameter A erhält man aus der üblichen Likelihoodfunktion L durch Maximieren über den „störenden" Parameter (Nuissanceparameter) B, d.h. $L_P(A) = \max A\ L(A, B)$

Im selteneren Fall ungepaarter Stichproben lassen sich die Ergebnisse in einer Kreuzklassifikation gemäß Tabelle 2.4 darstellen. Da unabhängige Wahrscheinlichkeiten verglichen werden, können die Hypothesen mit dem χ^2-Test für Vierfeldertafeln getestet werden. Die asymptotisch χ^2-verteilten Prüfgrößen für den Vergleich der Sensitivitäten bzw. Spezifitäten lauten:

$$\begin{aligned}
\chi^2_{\text{Sens}} &= \sum \frac{(\text{beobachtet} - \text{erwartet})^2}{\text{erwartet}} \\
&= \frac{(n^K_{A+} - (n^K_{\cdot+} \cdot n^K_{A\cdot})/n^K_{\cdot\cdot})^2}{(n^K_{\cdot+} \cdot n^K_{A\cdot})/n^K_{\cdot\cdot}} + \frac{(n^K_{A-} - (n^K_{\cdot-} \cdot n^K_{A\cdot})/n^K_{\cdot\cdot})^2}{(n^K_{\cdot-} \cdot n^K_{A\cdot})/n^K_{\cdot\cdot}} + \\
&\quad \frac{(n^K_{B+} - (n^K_{\cdot+} \cdot n^K_{B\cdot})/n^K_{\cdot\cdot})^2}{(n^K_{\cdot+} \cdot n^K_{B\cdot})/n^K_{\cdot\cdot}} + \frac{(n^K_{B-} - (n^K_{\cdot-} \cdot n^K_{B\cdot})/n^K_{\cdot\cdot})^2}{(n^K_{\cdot-} \cdot n^K_{B\cdot})/n^K_{\cdot\cdot}}
\end{aligned}$$

bzw.

$$\begin{aligned}
\chi^2_{\text{Spez}} &= \sum \frac{(\text{beobachtet} - \text{erwartet})^2}{\text{erwartet}} \\
&= \frac{(n^G_{A+} - (n^G_{\cdot+} \cdot n^G_{A\cdot})/n^G_{\cdot\cdot})^2}{(n^G_{\cdot+} \cdot n^G_{A\cdot})/n^G_{\cdot\cdot}} + \frac{(n^G_{A-} - (n^G_{\cdot-} \cdot n^G_{A\cdot})/n^G_{\cdot\cdot})^2}{(n^G_{\cdot-} \cdot n^G_{A\cdot})/n^G_{\cdot\cdot}} + \\
&\quad \frac{(n^G_{B+} - (n^G_{\cdot+} \cdot n^G_{B\cdot})/n^G_{\cdot\cdot})^2}{(n^G_{\cdot+} \cdot n^G_{B\cdot})/n^G_{\cdot\cdot}} + \frac{(n^G_{B-} - (n^G_{\cdot-} \cdot n^G_{B\cdot})/n^G_{\cdot\cdot})^2}{(n^G_{\cdot-} \cdot n^G_{B\cdot})/n^G_{\cdot\cdot}}.
\end{aligned}$$

Eine exakte Alternative zum χ^2-Test stellt der exakte Test von Fisher dar. Meist wird empfohlen, den exakten Test zu verwenden, falls die Anzahl der erwarteten Fälle in einer Zelle kleiner ist als fünf. Da aber diese Erwartungswertbedingung sehr vage ist und Permutationstests für moderne Computer kein Problem mehr darstellen, sollte der exakte Test von Fisher stets verwendet werden.

Die Eigenschaften der verschiedenenen Methoden zur Berechnung von Konfidenzintervallen für die Differenz zweier Proportionen im ungepaarten Fall ähneln denen für gepaarte Stichproben: Asymptotische Methoden neigen zu Overshot (vgl. Seite 20) und haben ebenso wie die exakte Methode eine schlechte Überdeckungswahrscheinlichkeit [56]. Auch hier erzielen die auf den Score-Intervallen von Wilson basierenden Konfidenzintervalle gute Ergebnisse.

Wenn sowohl die Sensitivität als auch die Spezifität eines Testes höher ist als die eines anderen, ist offensichtlich, dass dieser Test besser ist. Falls jedoch nur eines der beiden Maße größer, das andere aber kleiner ist, ist nicht sofort klar, ob einer der beiden Test besser ist als der andere. Es ist z.B. möglich, dass die Abnahme der Spezifität durch die Zunahme an Sensitivität aufgefangen wird, so dass trotz geringerer Spezifität die prädiktiven Werte höher sind bzw. das positive Likelihood Ratio größer und das negative Likelihood Ratio kleiner ist als bei dem anderen Test.

Da die prädiktiven Werte prävalenzabhängig sind, muss man bei ihrem Vergleich vorsichtig sein. Da aber bei fester Prävalenz gilt [41]:

$$\mathrm{ppW}_A \geq \mathrm{ppW}_B \Leftrightarrow \mathrm{LR}_A^+ \geq \mathrm{LR}_B^+ \tag{2.12}$$

bzw.

$$\mathrm{npW}_A \geq \mathrm{npW}_B \Leftrightarrow \mathrm{LR}_A^- \leq \mathrm{LR}_B^- \tag{2.13}$$

genügt es, die Likelihood Ratios zu vergleichen. Ein diagnostischer Test T_A, der die Bedingung (2.12) erfüllt, ist zur Bestätigung einer Krankheit besser geeignet, als ein Vergleichstest T_B. Falls der Test T_A die Bedingung (2.13) erfüllt, ist er besser zum Ausschluss geeignet als Test T_B. Erfüllt ein Test T_A

beide Bedingungen, ist er dem Test T_B insgesamt überlegen.

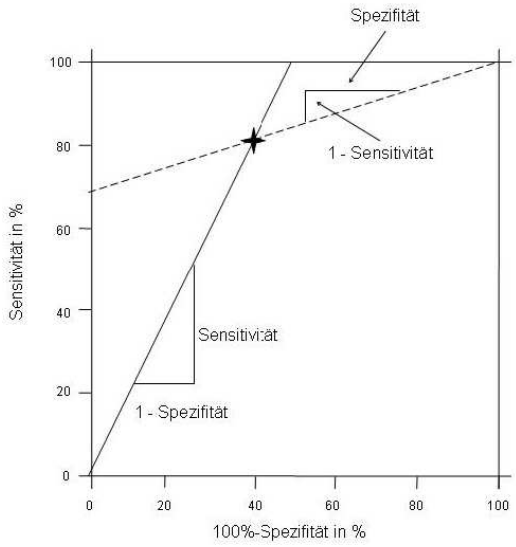

Abbildung 2.1: *Likelihood-Ratio-Grafik nach Biggerstaff: Darstellung von Sensitivität, Spezifität, positivem Likelihood Ratio und negativem Likelihood Ratio eines Tests X.*

Biggerstaff [57] hat eine graphische Methode zum Vergleich der Likelihood Ratios entwickelt (siehe Abbildung 2.1). Auf der x-Achse wird der Anteil der falsch Positiven $n_{+-}/n_{\cdot-}$, d.h. 1 - Spezifität aufgetragen, auf der y-Achse der Anteil der richtig Positiven $n_{++}/n_{\cdot+}$, d.h. die Sensitivität. Durch ein Kreuz x wird der erste Test markiert. Die Steigung der Gerade von (0;0) durch x enspricht dem positiven Likelihood Ratio des Tests, die Steigung der Geraden von (1;1) durch x entspricht dem negativen Likelihood Ratio.

Tests, die links der Gerade LR^+ und oberhalb der Gerade LR^- liegen, haben ein größeres positives Likelihood Ratio und ein kleineres negatives Likelihood Ratio und sind besser als der Ausgangstest (siehe Abbildung 2.2). Tests, die

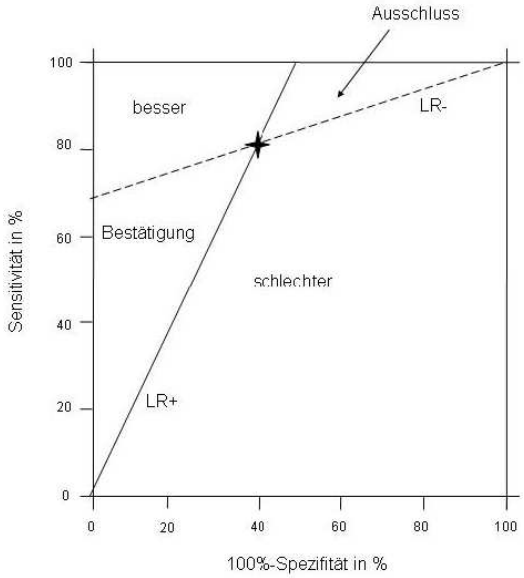

Abbildung 2.2: *Likelihood-Ratio-Grafik nach Biggerstaff: Bereiche in denen ein Vergleichstest insgesamt besser, besser zum Ausschluss bzw. besser zur Bestätigung einer Krankeit geeignet, oder schlechter als ein Ausgangstest X ist.*

links der Gerade LR^+ liegen, aber unterhalb der Gerade LR^-, erfüllen Bedingung (2.12) und sind besser zur Bestätigung einer Krankheit geeignet als der Ausgangstest. Analog erfüllen Tests, die rechts der Gerade LR^+, aber oberhalb der Gerade LR^- liegen, Bedingung (2.13) und sind besser zum Ausschluss der Krankeit geeignet. Tests, die sowohl rechts der Gerade LR^+ als auch unterhalb der Gerade LR^- liegen, haben ein kleineres positives Likelihood Ratio und ein größeres negatives Likelihood Ratio und sind dem Ausgangstest unterlegen.

2.4 Kombination diagnostischer Tests

Nachdem untersucht wurde, welcher von mehreren Tests der beste ist, stellt sich oftmals die Frage, ob durch Kombination verschiedener Tests die diagnostische Treffsicherheit erhöht werden kann. Für die Kombination zweier Tests muss zunächst geklärt werden, wie diskordante Ergebnisse bewertet werden, ob gleichzeitig (parallel) oder nacheinander (sequentiell) getestet wird und welcher Test bei sequentieller Vorgehensweise zuerst durchgeführt werden soll.

2.4.1 Technische Fragen

Beim *parallelen Testen* werden beide Tests gleichzeitig und unabhängig voneinander durchgeführt. Diese Vorgehensweise ist sinnvoll, wenn das Ergebnis rasch vorliegen soll oder gleichzeitiges Testen, wie z.B. bei Laborparametern, unproblematisch ist.

Beim *sequentiellen Testen* werden die Tests nacheinander durchgeführt. Ob der zweite Test noch notwendig ist und durchgeführt wird, hängt vom Ergebnis des ersten Tests ab. Durch sequentielles Testen kann somit die Anzahl der durchzuführenden zweiten Tests verringert werden. Dies ist besonders dann sinnvoll, wenn der zweite Test aufwändig, teuer oder riskant ist.

Parallele oder serielle Testung führt zur gleichen Sensitivität und Spezifität, auch die Reihenfolge der Tests beim sequentiellen Testen hat keinen Einfluss auf die beiden Parameter [58]. Bei der Frage, ob parallel oder sequentiell getestet werden soll, muss man sich somit entscheiden, ob ein rasches Testergebnis wichtig ist oder es von Interesse ist, einigen Patienten den zweiten

Test und die damit verbundenen Belastungen zu ersparen. Bei der Frage der Reihenfolge bei sequentieller Durchführung sind Invasivität, Kosten und Schnelligkeit der beiden Verfahren zu berücksichtigen und der aufwändigere Test als zweiter durchzuführen.

Es gibt zwei Möglichkeiten, diskordante Ergebnisse beider Tests zu bewerten. Bei der „**b**elieve-**t**he-**p**ositive"-Regel (BTP) wird die Kombination positiv gewertet, sobald einer der beiden Einzeltests positiv ist. Bei Patienten mit positivem ersten Test ist somit bei sequentieller Durchführung der zweite Test nicht mehr notwendig, nur Patienten mit negativem ersten Test werden mit dem zweiten Verfahren untersucht.

Bei der „**b**elieve-**t**he-**n**egative"-Regel (BTN) wird bei paralleler Durchführung die Kombination nur dann als positiv gewertet, wenn beide Einzeltests positiv sind bzw. müssen bei sequentieller Durchführung Patienten mit negativem Ergebnis im ersten Test nicht weiter untersucht werden.

2.4.2 Abhängigkeit

Bei der Kombination diagnostischer Tests kann man drei Arten von Abhängigkeit unterscheiden [41]. Die erste Möglichkeit ist die (unbedingte) Unabhängigkeit:

$$\mathrm{P}(T_A^+ \cap T_B^+) = \mathrm{P}(T_A^+) \cdot \mathrm{P}(T_B^+)$$

Die Annahme der (unbedingten) Unabhängigkeit ist realitätsfern, da in der Praxis ein Zusammenhang über die Krankheit gegeben sein sollte, d.h. bei Vorliegen der Krankheit werden beide Tests mit hoher Wahrscheinlichkeit positiv und bei Nichtvorliegen der Krankhet negativ sein. Bei unabhängigen

Tests wäre somit zumindest einer der beiden diagnostisch unbrauchbar.

Die zweite Möglichkeit ist die bedingte Unabhängigkeit:

$$\mathrm{P}(T_A^+ \cap T_B^+|\mathrm{K}^+) = \mathrm{P}(T_A^+|\mathrm{K}^+) \cdot \mathrm{P}(T_B^+|\mathrm{K}^+)$$

bzw.

$$\mathrm{P}(T_A^+ \cap T_B^+|\mathrm{K}^-) = \mathrm{P}(T_A^+|\mathrm{K}^-) \cdot \mathrm{P}(T_B^+|\mathrm{K}^-).$$

Aus der bedingten Unabhängigkeit folgt nicht die unbedingte Unabhängigkeit und umgekehrt. Außerdem können zwei Tests bedingt unabhängig, und trotzdem beide valide Tests mit hoher Sensitivität und Spezifität sein [41].

Zwei Tests sind oft bedingt unabhängig, wenn sie unterschiedliche Aspekte der zu diagnostizierenden Krankheit untersuchen. Beispielsweise erfolgt die Tumordetektion beim CT morphologisch, das PET hingegen liefert funktionelle Informationen. Es ist somit anzunehmen, dass beide Verfahren voneinander bedingt unabhängig sind und sich gegenseitig ergänzen. Da aber oft beide Tests auf denselben anatomischen, morphologischen oder biochemischen Hintergründen basieren, ist die Annahme der bedingten Unabhängigkeit in der Praxis nicht immer erfüllt. Deswegen sollte sie nicht ungeprüft vorausgesetzt werden, sondern mittels χ^2-Unabhängigkeitstest getrennt für Kranke und Gesunde getestet werden.

Die dritte Möglichkeit ist die bedingte Abhängigkeit:

$$\mathrm{P}(T_A^+ \cap T_B^+|\mathrm{K}^+) = \alpha \cdot \mathrm{P}(T_A^+|\mathrm{K}^+) \cdot \mathrm{P}(T_B^+|\mathrm{K}^+) \tag{2.14}$$

bzw.

$$\mathrm{P}(T_A^+ \cap T_B^+|\mathrm{K}^-) = \alpha' \cdot \mathrm{P}(T_A^+|\mathrm{K}^-) \cdot \mathrm{P}(T_B^+|\mathrm{K}^-) \tag{2.15}$$

mit $\alpha, \alpha' \neq 1$. Für α bzw. $\alpha' > 1$ sind die beiden Tests in der Population der Kranken bzw. Gesunden positiv korreliert.

2.4.3 Sensitivität und Spezifität

Durch die BTP-Regel ist die Zahl der positiven Ergebnisse im Vergleich zu den beiden Einzeltests größer, die Sensitivität der Kombination ist also höher, die Spezifität hingegen niedriger als bei den Einzeltests. Bei Anwendung der BTN-Regel nimmt im Vergleich zu den beiden Einzeltests die Anzahl der negativen Ergebnisse zu, die Sensitivität der Kombination ist somit niedriger, die Spezifität hingegen höher als bei den Einzeltests. Durch die Kombination zweier diagnostischer Tests können Sensitivität und Spezifität somit nicht zugleich verbessert werden. Der Sensitivitätsgewinn bei Verwendung der BTP-Regel hat einen Spezifitätsverlust zur Folge, bei Verwendung der BTN-Regel wird ein Spezifitätsgewinn auf Kosten der Sensitivität erzielt. Dabei gelten folgende Ungleichungen [58]:

$$\text{Sens}_{\text{BTP, parallel}} = \text{Sens}_{\text{BTP, sequentiell}} \geq \max(\text{Sens}_A, \text{Sens}_B) \tag{2.16}$$

$$\text{Spez}_{\text{BTP, parallel}} = \text{Spez}_{\text{BTP, sequentiell}} \leq \min(\text{Spez}_A, \text{Spez}_B) \tag{2.17}$$

$$\text{Sens}_{\text{BTN, parallel}} = \text{Sens}_{\text{BTN, sequentiell}} \leq \min(\text{Sens}_A, \text{Sens}_B) \tag{2.18}$$

$$\text{Spez}_{\text{BTN, parallel}} = \text{Spez}_{\text{BTN, sequentiell}} \geq \max(\text{Spez}_A, \text{Spez}_B) \tag{2.19}$$

Falls die beiden Tests bedingt unabhängig sind, lassen sich die Sensitivität und Spezifität der Testkombination explizit angeben:

$$\text{Sens}_{\text{BTP}} = \text{Sens}_A + (1 - \text{Sens}_A) \cdot \text{Sens}_B$$

$$\begin{aligned} \text{Spez}_{\text{BTP}} &= \text{Spez}_A \cdot \text{Spez}_B \\ \text{Sens}_{\text{BTN}} &= \text{Sens}_A \cdot \text{Sens}_B \\ \text{Spez}_{\text{BTN}} &= \text{Spez}_A + (1 - \text{Spez}_A) \cdot \text{Spez}_B. \end{aligned}$$

Bei der Kombination diagnostischer Test muss man sich entscheiden, ob die Sensitivität oder die Spezifität gegenüber den Einzeltests verbessert werden soll. Wenn eine höhere Sensitivität erwünscht ist, sollte die BTP-Regel verwendet werden und zwei Tests mit hoher Spezifität kombiniert werden, damit der Spezifitätsverlust möglichst gering ausfällt. Analog sollten zwei Tests mit hoher Sensitivität mittels der BTN-Regel kombiniert werden, falls eine Vergrößerung der Spezifität angestrebt wird.

Da der Spezifitäts- bzw. Sensitivitätsverlust umso größer wird, je mehr Tests kombiniert werden (Odysseus-Syndrom), ist es meist nicht sinnvoll, mehr als zwei Tests zu kombinieren. Für eine Testbatterie mit t Tests ist z.B. die Wahrscheinlichkeit eines falsch negativen Ergebnises $P(\text{FN})$ bei Verwenden der BTN-Regel unter Annahme der bedingten Unabhängigkeit gleich $1 - \prod_{i=1}^{t} \text{Sens}_i$. D.h. wenn vier Tests mit einer Sensitivität von jeweils 80% mittels BTN-Regel kombiniert werden, beträgt $P(\text{FN})$ bereits 59%.

2.4.4 Prädiktive Werte

Die prädiktiven Werte der Kombination zweier Tests mittels BTN- oder BTP-Regel lassen sich aus der Bayes-Formel herleiten:

$$\mathrm{P}(\mathrm{K}^+ \mid \mathrm{R}_A, \mathrm{R}_B) = \frac{\mathrm{P}(\mathrm{R}_A, \mathrm{R}_B \mid \mathrm{K}^+) \cdot \mathrm{P}(\mathrm{K}^+)}{\mathrm{P}(\mathrm{R}_A, \mathrm{R}_B \mid \mathrm{K}^+) \cdot \mathrm{P}(\mathrm{K}^+) + \mathrm{P}(\mathrm{R}_A, \mathrm{R}_B \mid \mathrm{K}^-) \cdot (1 - \mathrm{P}(\mathrm{K}^+))},$$

wobei R_A, R_B die Resultate des ersten bzw. zweiten Tests bezeichnen.

So gilt z.B.:

$$\begin{aligned} \mathrm{ppW}_{\mathrm{BTN}} &= \mathrm{P}(\mathrm{K}^+ \mid \mathrm{T}_A^+, \mathrm{T}_B^+) \\ &= \frac{\mathrm{P}(\mathrm{T}_A^+, \mathrm{T}_B^+ \mid \mathrm{K}^+) \cdot \mathrm{P}(\mathrm{K}^+)}{\mathrm{P}(\mathrm{T}_A^+, \mathrm{T}_B^+ \mid \mathrm{K}^+) \cdot \mathrm{P}(\mathrm{K}^+) + \mathrm{P}(\mathrm{T}_A^+, \mathrm{T}_B^+ \mid \mathrm{K}^-) \cdot (1 - \mathrm{P}(\mathrm{K}^+))} \\ &= \frac{\mathrm{Sens}_A \cdot \mathrm{Sens}_B \cdot \mathrm{Präv}}{\mathrm{Sens}_A \cdot \mathrm{Sens}_B \cdot \mathrm{Präv} + (1 - \mathrm{Spez}_A) \cdot (1 - \mathrm{Spez}_B) \cdot (1 - \mathrm{Präv})} \end{aligned}$$

Marshall [59] zeigte, dass für $\mathrm{ppW}_{\text{Test A}} > \mathrm{ppW}_{\text{Test B}}$ und feste Prävalenz P(K+) gilt:

$$\begin{aligned} \mathrm{ppW}_{\mathrm{BTP}} > \mathrm{ppW}_{\text{Test A}} \quad &\Rightarrow \quad \mathrm{ppW}_{\mathrm{BTN}} < \mathrm{ppW}_{\text{Test A}} \\ \mathrm{ppW}_{\mathrm{BTN}} > \mathrm{ppW}_{\text{Test A}} \quad &\Rightarrow \quad \mathrm{ppW}_{\mathrm{BTP}} < \mathrm{ppW}_{\text{Test B}}. \end{aligned}$$

Analog gilt für $\mathrm{npW}_{\text{Test A}} > \mathrm{npW}_{\text{Test B}}$ und feste Prävalenz $\mathrm{P}(\mathrm{K}^+)$:

$$\begin{aligned} \mathrm{npW}_{\mathrm{BTP}} > \mathrm{npW}_{\text{Test A}} \quad &\Rightarrow \quad \mathrm{npW}_{\mathrm{BTN}} < \mathrm{npW}_{\text{Test A}} \\ \mathrm{npW}_{\mathrm{BTN}} > \mathrm{npW}_{\text{Test A}} \quad &\Rightarrow \quad \mathrm{npW}_{\mathrm{BTP}} < \mathrm{npW}_{\text{Test B}}. \end{aligned}$$

Es ist also nicht möglich, sowohl durch Kombination mittels der BTN-Regel als auch durch Kombination mittels der BTP-Regel eine Verbesserung des

positiven bzw. negativen prädiktiven Wertes zu erzielen. Die Umkehrung obiger Aussagen gilt nicht. Es ist möglich, dass sich der positive bzw. negative prädiktive Wert sowohl bei Anwenden der BTP- als auch der BTN-Regel nicht verbessert.

Welcher der drei Fälle:

- BTP führt zu einer Verbesserung des positiven (negativen) prädiktiven Wertes,
- BTN führt zu einer Verbesserung des positiven (negativen) prädiktiven Wertes oder
- weder BTP noch BTN führen zu einer Verbesserung des positiven (negativen) prädiktiven Wertes

bei der Kombination eintritt, hängt von der Abhängigkeitsstruktur der beiden diagnostischen Tests ab.

Der Zusammenhang zwischen der Veränderung der positiven und negativen prädiktiven Werte und α und α' aus (2.14) bzw. (2.15) lässt sich mit Marshalls graphischer Methode [59] untersuchen. Der Wertebereich von α und α' hängt von der Sensitivität und der Spezifität der beiden Einzeltests ab:

$$\begin{aligned} \max\left(0, \frac{\text{Sens}_A + \text{Sens}_B - 1}{\text{Sens}_A \cdot \text{Sens}_B}\right) &\leq \alpha \leq \min\left(\frac{1}{\text{Sens}_A}, \frac{1}{\text{Sens}_B}\right) \\ \max\left(0, \frac{(1-\text{Spez}_A) + (1-\text{Spez}_B) - 1}{(1-\text{Spez}_A) \cdot (1-\text{Spez}_B)}\right) &\leq \alpha' \leq \min\left(\frac{1}{1-\text{Spez}_A}, \frac{1}{1-\text{Spez}_B}\right). \end{aligned}$$

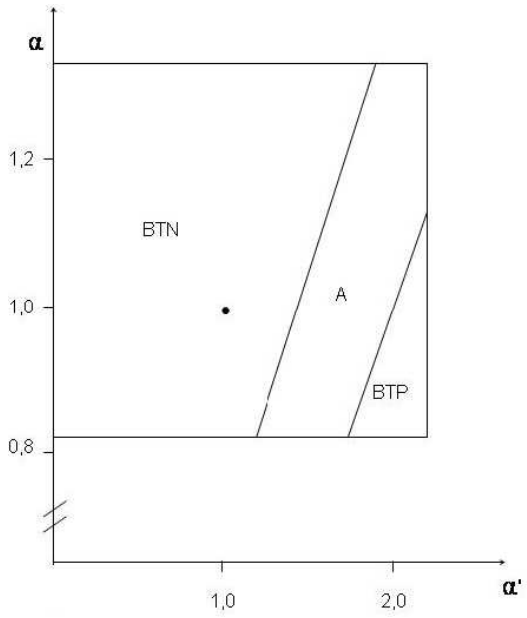

Abbildung 2.3: *Darstellung des Einflusses der Abhängigkeitsstruktur der beiden Tests in der Subpopulation der Kranken bzw. der Gesunden (ausgedrückt duch α und α') auf die positiven prädiktiven Werte der Kombination mittels BTP- bzw. BTN-Regel. Der Kreis markiert den Fall der bedingten Unabhängigkeit beider Tests. Für das Beispiel wurde $Sens_A$=0,75, $Sens_B$=0,65, $Spez_a$=0,60 und $Spez_B$=0,55 gesetzt.*

Trägt man α und α' in einem Koordinatensystem gegeneinander auf, definieren obige Grenzen eine rechteckige Region $\mathrm{R}(\alpha, \alpha')$ (siehe Abbildung 2.3). Diese Region kann in drei Bereiche unterteilt werden, in denen entweder der positive prädiktive Wert der Kombination mittels BTN, der Kombination mittels BTP oder des besseren Einzeltests der größte ist.

Für

$$\alpha > \alpha' \cdot \frac{1 - \mathrm{Spez}_B}{\mathrm{Sens}_B}$$

ist der positive prädiktive Wert der Kombination mit BTN am größten (Re-

gion BTN). Für

$$\alpha < \alpha' \cdot \frac{1 - \text{Spez}_B}{\text{Sens}_B} + \frac{(1 - \text{Spez}_A) \cdot \text{Sens}_B - \text{Sens}_A \cdot (1 - \text{Spez}_B)}{\text{Sens}_A \cdot (1 - \text{Spez}_A) \cdot \text{Sens}_B}$$

ist der positive prädiktive Wert der Kombination mit BTP der größte (Region BTP). In dem Bereich dazwischen hat keine der beiden Kombinationen einen höheren positiven prädiktiven Wert als der bessere der beiden Einzeltests (Region A).

Analog kann $\text{R}(\alpha, \alpha')$ in drei Teilbereiche eingeteilt werden, je nachdem welcher negative prädiktive Wert der höchste ist:
Für

$$e = \frac{(1 - \text{Spez}_A)(1 - \text{Sens}_A)(1 - \text{Spez}_B)}{\text{Sens}_A \cdot \text{Sens}_B \cdot \text{Spez}_A}$$

und

$$\alpha < e\alpha' + \frac{\text{Sens}_A \cdot (1 - \text{Spez}_B) - \text{Sens}_B \cdot (1 - \text{Spez}_A) + \text{Sens}_B - (1 - \text{Spez}_B)}{\text{Sens}_A \cdot \text{Sens}_B \cdot \text{Spez}_A}$$

ist der negative prädiktive Wert der Kombination mit BTP am höchsten. Für

$$\alpha < e\alpha' + \frac{\text{Sens}_A - (1 - \text{Spez}_A)}{\text{Sens}_A \cdot \text{Sens}_B \cdot \text{Spez}_A}$$

ist der negative prädiktive Wert der Kombination mit BTN am höchsten. In dem Bereich dazwischen hat keine der beiden Kombinationen einen höheren negativen prädiktiven Wert als der bessere der beiden Einzeltests.

Durch Übereinanderlegen der drei Regionen des positiven prädiktiven Wertes und der drei Regionen des negativen prädiktiven Wertes erhält man Abbildung 2.4. Der Bereich, in dem BTP zu einer Verbesserung des positiven und

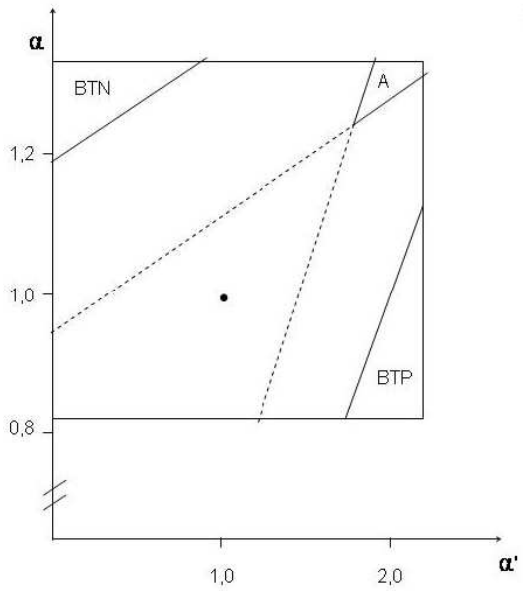

Abbildung 2.4: *Darstellung des Einflusses der Abhängigkeitsstruktur der beiden Tests in der Subpopulation der Kranken bzw. der Gesunden (ausgedrückt duch α und α') auf die positiven und negativen prädiktiven Werte der Kombination mittels BTP- bzw. BTN-Regel. Der Kreis markiert den Fall der bedingten Unabhängigkeit beider Tests. Für das Beispiel wurde $Sens_A$=0,75, $Sens_B$=0,65, $Spez_a$=0,60 und $Spez_B$=0,55 gesetzt.*

des negativen prädiktiven Wertes führt (Region BTP), ist sehr klein und setzt voraus, dass bei den Gesunden eine hohe Abhängigkeit zwischen den Tests besteht (d.h. α' groß ist), bei den Kranken die beiden Tests jedoch (nahezu) unabhängig sind (d.h. α klein). Analog führt die BTN-Regel nur dann sowohl zu einer Verbesserung des positiven wie auch des negativen prädiktiven Wertes (Region BTN), wenn bei den Kranken eine hohe Abhängigkeit der Tests besteht (d.h. α groß), aber bei den Gesunden die beiden Tests (nahezu) unabhängig sind (d.h. α' klein).

Bei starker Abhängigkeit der Tests ist es allerdings möglich, dass durch die Kombination sowohl der positive wie auch der negative prädiktive Wert abnehmen (Region A).

Wenn die Abhängigkeit zwischen den beiden Tests weder zu groß noch zu asymmetrisch ist, führt die BTP-Regel zu mehr (falsch) Positiven und weniger (falsch) Negativen und damit zu einem höheren negativen und einem niedrigeren positiven prädiktiven Wert. Bei der BTN-Regel verhalten sich die prädiktiven Werte genau umgekehrt.

2.4.5 Likelihood Ratios

Ebenso wie beim Vergleich zweier Tests kann man die Likelihood-Ratio-Grafik von Biggerstaff auch dazu nutzen, um Bereiche zu bestimmen, in denen die Kombination zweier Tests den Einzeltests überlegen ist [60].

Bei Anwenden der BTP-Regel erhöht sich die Sensitivität der Kombination im Vergleich zum Ausgangstest A, die Spezifität verringert sich hingegen. D.h. die Koordinaten der Testkombination müssen im schraffierten Bereich liegen (siehe Abbildung 2.5). Dieser Bereich lässt sich in drei Teilbereiche aufteilen. Im Teilbereich K links der Gerade LR^+ liegen alle Kombinationen, bei denen das positive Likelihood Ratio größer ist als beim Ausgangstest A und das negative Likelihood Ratio kleiner ist als beim Ausgangstest A, d.h. Kombinationen, die besser sind als der Ausgangstest A. Im Bereich E unterhalb der Gerade LR^- liegen alle Kombinationen, bei denen das positive Likelihood Ratio kleiner ist als beim Ausgangstest A und das negative Likelihood Ratio größer ist als beim Ausgangstest A, d.h. Kombinationen, die schlechter sind als der Ausgangstest A. Im dazwischenliegenden Bereich NB

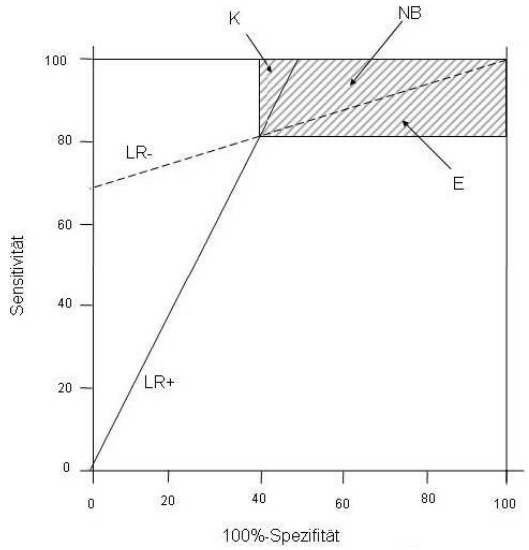

Abbildung 2.5: *Likelihood-Ratio-Grafik nach Biggerstaff: Testkombinationen mittels der BTP-Regel liegen im schraffierten Bereich. Kombinationen in Teilbereich K sind besser, Kombinationen im Teilbereich E schlechter als der Ausgangstest. Bei Kombinationen im Bereich NB ist nicht beurteilbar, ob sich durch die Kombination die diagnostische Leistung verbessert.*

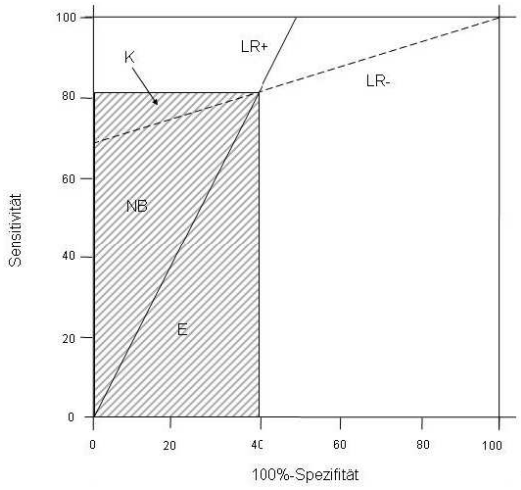

Abbildung 2.6: *Likelihood-Ratio-Grafik nach Biggerstaff: Testkombinationen mittels der BTN-Regel liegen im schraffierten Bereich. Kombinationen in Teilbereich K sind besser, Kombinationen im Teilbereich E schlechter als der Ausgangstest. Bei Kombinationen im Bereich NB ist nicht beurteilbar, ob sich durch die Kombination die diagnostische Leistung verbessert.*

ist zwar das negative Likelihood Ratio bei der Kombination kleiner als beim Ausgangstest A, allerdings ist auch das positive Likelihood Ratio kleiner. In diesem Fall ist nicht beurteilbar, ob die Kombination besser ist als der Ausgangstest A.

Zusammengefasst lässt sich feststellen:

$$\mathrm{LR}^+_{BTP} > \mathrm{LR}^+_A \quad \Rightarrow \quad \mathrm{LR}^-_{BTP} < \mathrm{LR}^-_A \quad \Rightarrow \quad \text{BTP besser}$$
$$\mathrm{LR}^-_{BTP} > \mathrm{LR}^-_A \quad \Rightarrow \quad \mathrm{LR}^+_{BTP} < \mathrm{LR}^+_A \quad \Rightarrow \quad A \text{ besser}$$

Bei Anwenden der BTN-Regel erhöht sich die Spezifität der Kombination im Vergleich zum Ausgangstest A, die Sensitivität hingegen verringert sich,

d.h. die Koordinaten der Kombination liegen im schraffierten Bereich (siehe Abbildung 2.6). Dieser Bereich lässt sich ebenfalls in drei Teilbereiche aufteilen. Im Teilbereich K oberhalb der Gerade LR^- liegen alle Kombinationen, bei denen das negative Likelihood Ratio kleiner ist als beim Ausgangstest A und das positive Likelihood Ratio größer ist als beim Ausgangstest A, d.h. Kombination die besser sind als der Ausgangstest A. Im Bereich E rechts der Gerade LR^+ liegen alle Kombinationen, bei denen das negative Likelihood Ratio größer ist als beim Ausgangstest A und das positive Likelihood Ratio kleiner ist als beim Ausgangstest A, d.h. Kombinationen, die schlechter sind als der Ausgangstest A. Im dazwischen liegenden Bereich NB ist zwar das negative Likelihood Ratio bei der Kombination kleiner als beim Ausgangstest A, allerdings ist auch das positive Likelihood Ratio kleiner. In diesem Fall ist nicht beurteilbar, ob die Kombination besser ist als der Ausgangstest A.

Es gilt somit:

$$\begin{aligned} LR^-_{BTN} < LR^-_A \quad &\Rightarrow \quad LR^+_{BTN} > LR^+_A \quad \Rightarrow \quad \text{BTN besser} \\ LR^+_{BTN} < LR^+_A \quad &\Rightarrow \quad LR^-_{BTN} > LR^-_A \quad \Rightarrow \quad \text{BTN besser} \end{aligned}$$

Um zu überprüfen, ob die Testkombinationen, die im Bereich K liegen, sich wirklich signifikant von dem Ausgangstest A unterscheiden, schlagen Macaskill et al. [60] vor, Konfidenzintervalle für die relative Veränderung des positiven bzw. negativen Likelihood Ratios zu konstruieren.

Die Ergebnisse des Ausgangstests und der Kombination mittels BTP- bzw. BTN-Regel können in einer Kreuzklassifikation gegenübergestellt werden (siehe Tabelle 2.5 und Tabelle 2.6), wobei p^+_{ij} bzw. p^-_{ij} mit $(i, j \in \{+, -\})$, die

Wahrscheinlichkeit angibt, dass ein Kranker bzw. Gesunder im Ausgangstest das Ergebnis i und in der Kombination das Ergebnis j hat. Bei beiden Kombinationsregeln kommt es zu strukturell bedingten Nullen. Bei der BTP-Regel kann die Kombination nie negativ gewertet werden, wenn der Ausgangstest positiv ist. Unabhängig davon wie der zweite Test ausfällt, wird die Kombination positiv gewertet. Bei Kombination mittels BTN-Regel kann die Kombination nie positiv gewertet werden, sobald der Ausgangstest negativ ist.

	Kranke		Gesunde	
	Testkombination BTP		Testkombination BTP	
	+	-	+	-
T_A +	$p^+_{+\cdot}$	0	$p^-_{+\cdot}$	0
T_A -	p^+_{-+}	p^+_{--}	p^-_{-+}	p^-_{--}

Tabelle 2.5: *Kreuzklassifikation der Ergebnisse des Ausgangstests und der Kombination mittels BTP-Regel. Die p^+_{ij} bzw. p^-_{ij} mit $(i, j \in \{+, -\})$ geben die Wahrscheinlichkeit an, dass ein Kranker bzw. Gesunder im Ausgangstest das Ergebnis i und in der Kombination das Ergebnis j hat.*

Unverzerrte Maximum-Likelihood-Schätzer für die Likelihood Ratios bei Kombination mittels BTP- bzw. BTN-Regel sind:

$$\begin{aligned}
\widehat{LR}^+_{BTP} &= \frac{\hat{p}^+_{+\cdot} + \hat{p}^+_{-+}}{\hat{p}^-_{+\cdot} + \hat{p}^-_{-+}} \\
\widehat{LR}^-_{BTP} &= \frac{\hat{p}^+_{--}}{\hat{p}^-_{--}} \\
\widehat{LR}^+_{BTN} &= \frac{\hat{p}^+_{++}}{\hat{p}^-_{++}} \\
\widehat{LR}^-_{BTN} &= \frac{\hat{p}^+_{-\cdot} + \hat{p}^+_{+-}}{\hat{p}^-_{-\cdot} + \hat{p}^-_{+-}}
\end{aligned}$$

Da die Likelihood Ratios schief verteilt sind, basiert das Konfidenzintervall

	Kranke		Gesunde	
	Testkombination BTN		Testkombination BTN	
	+	-	+	-
T_A +	p^+_{++}	p^+_{+-}	p^-_{++}	p^-_{+-}
T_A -	0	$p^+_{-.}$	0	$p^-_{-.}$

Tabelle 2.6: *Kreuzklassifikation der Ergebnisse des Ausgangstest und der Kombination mittels BTN-Regel. Die p^+_{ij} bzw. p^-_{ij} $(i, j \in \{+, -\})$ geben die Wahrscheinlichkeit an, dass ein Kranker bzw. Gesunder im Ausgangstest das Ergebnis i und in der Kombination das Ergebnis j hat.*

auf der Differenz der logarithmierten Likelihood Ratios:

$$\ln(\widehat{\mathrm{LR}}^+_{BTP}) - \ln(\widehat{\mathrm{LR}}^+_A) = \ln(\hat{p}^+_{+.} + \hat{p}^+_{-+}) - \ln(\hat{p}^-_{+.} + \hat{p}^-_{-+}) - \ln(\hat{p}^+_{+.}) + \ln(\hat{p}^-_{+.})$$

Macaskill et al. [60] bestimmten mittels der Delta-Methode asymptotische Schätzer für die Varianz dieser Differenzen:

$$\mathrm{Var}\left(\ln(\widehat{\mathrm{LR}}^+_{BTP}) - \ln(\widehat{\mathrm{LR}}^+_A)\right) = \frac{\hat{p}^+_{-+}}{n_K \hat{p}^+_{+.}(\hat{p}^+_{-+} + \hat{p}^+_{+.})} + \frac{\hat{p}^-_{-+}}{n_G \hat{p}^-_{+.}(\hat{p}^-_{-+} + \hat{p}^-_{+.})}$$

$$\mathrm{Var}\left(\ln(\widehat{\mathrm{LR}}^-_{BTP}) - \ln(\widehat{\mathrm{LR}}^-_A)\right) = \frac{\hat{p}^+_{-+}}{n_K \hat{p}^+_{-.}(\hat{p}^+_{--} + \hat{p}^+_{-+})} + \frac{\hat{p}^-_{-+}}{n_G \hat{p}^-_{--}(\hat{p}^-_{--} + \hat{p}^-_{-+})}$$

$$\mathrm{Var}\left(\ln(\widehat{\mathrm{LR}}^+_{BTN}) - \ln(\widehat{\mathrm{LR}}^+_A)\right) = \frac{\hat{p}^+_{+-}}{n_K \hat{p}^+_{++}(\hat{p}^+_{++} + \hat{p}^+_{+-})} + \frac{\hat{p}^-_{+-}}{n_G \hat{p}^-_{++}(\hat{p}^-_{++} + \hat{p}^-_{+-})}$$

$$\mathrm{Var}\left(\ln(\widehat{\mathrm{LR}}^-_{BTN}) - \ln(\widehat{\mathrm{LR}}^-_A)\right) = \frac{\hat{p}^+_{+-}}{n_K \hat{p}^+_{-.}(\hat{p}^+_{-.} + \hat{p}^+_{+-})} + \frac{\hat{p}^-_{+-}}{n_G \hat{p}^-_{-.}(\hat{p}^-_{-.} + \hat{p}^-_{+-})}$$

Die Grenzen eines asymptotischen Konfidenzintervalls für LR^+_{BTP}/LR^+_A lassen sich somit berechnen durch

$$\exp\left[\left(\ln(\widehat{\mathrm{LR}}^+_{BTP}) - \ln(\widehat{\mathrm{LR}}^+_A)\right) \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}\left(\ln(\widehat{\mathrm{LR}}^+_{BTP}) - \ln(\widehat{\mathrm{LR}}^+_A)\right)}\right],$$

wobei $z_{1-\alpha/2}$ das $1-\alpha/2$-Quantil der Standardnormalverteilung ist. Die Grenzen der anderen Konfidenzintervalle werden analog bestimmt.

Aufgrund von (2.20) ist die Kombination mittels BTP-Regel besser als der Einzeltest, wenn $\mathrm{LR}^+_{BTP}/\mathrm{LR}^+_A > 1$ ist und das Konfidenzintervall die Eins nicht enthält. Wenn hingegen $\mathrm{LR}^-_{BTP}/\mathrm{LR}^-_A > 1$ ist und das Konfidenzintervall die Eins nicht enthält, ist der Ausgangstest besser. Ebenso ist die Kombination auf Grund von (2.20) mittels BTN-Regel besser, wenn $\mathrm{LR}^-_{BTN}/\mathrm{LR}^-_A < 1$ ist und das Konfidenzintervall die Eins nicht enthält. Falls $\mathrm{LR}^+_{BTN}/\mathrm{LR}^+_A < 1$ ist und das Konfidenzintervall die Eins nicht enthält, ist der Ausgangstest besser.

2.5 Entscheidungsanalyse

Wie in den vorherigen Kapiteln gezeigt wurde, kann die Statistik dazu beitragen, die Leistungsfähigkeit diagnostischer Verfahren zu beurteilen. Da aber fast jeder diagnostische Test falsch Positive und/oder falsch Negative aufweist, müssen viele Entscheidungen im klinischen Alltag unter Unsicherheit getroffen werden. Diese Unsicherheit hat, neben der fehlenden vollständigen Übereinstimmung zwischen klinischer Information und Krankheitszustand, noch weitere Ursachen.

Zunächst kann die Erhebung der Befunde fehlerhaft sein: Aufzeichnungen sind ungenau oder falsch oder Befunde werden fehlinterpretiert. Einige Messverfahren haben nur eine geringe Intra- und Inter-Rater-Reliabilität und es kann vorkommen, dass die Befunde nicht eindeutig sind.

Desweiteren sind die Auswirkungen möglicher Therapien auf den Gesundheitszustand des Patienten nicht mit völliger Sicherheit vorhersagbar. Viele Maßnahmen führen nur zu einer Verbesserung bei der Mehrheit der Patienten, aber es ist nicht sicher, ob auch derjenige Patient, für den die Entscheidung getroffen werden muss, davon profitiert.

Außerdem können bei Patienten Nebenwirkungen auftreten. Zusätzlich bestehen in einigen Fällen auch Meinungsverschiedenheiten über die richtige Vorgehensweise.

Trotz dieser Unsicherheiten muss eine Entscheidung getroffen werden. Somit stellt sich die Aufgabe, unter Berücksichtigung der Unsicherheiten und Abwägung der Risiken und des Nutzens die optimale Vorgehensweise zu fin-

den. Die statistische Entscheidungstheorie, ein Zweig der angewandten Wahrscheinlichkeitstheorie, bietet dazu die theoretische Grundlage.

2.5.1 Statistische Entscheidungstheorie

Die statistische Entscheidungstheorie beschäftigt sich mit dem Problem der Entscheidungsfindung in Situationen, in denen der Eintritt von zukünftigen Zuständen nicht mit Sicherheit vorausgesagt werden kann. Das heißt, dass die Auswirkungen der zur Verfügung stehenden Alternativen nicht vollständig bekannt sind. Je nachdem, ob man Eintrittswahrscheinlichkeiten für die Zustände kennt, wird unterschieden zwischen [61]:

- Entscheidungen unter Risiko: Dem Entscheider sind die von seiner Entscheidung abhängigen Eintrittswahrscheinlichkeiten der Zustände objektiv (z.B. bei einer Lotterie) oder subjektiv (aufgrund von Schätzungen oder von Vergangenheitswerten) bekannt.

- Entscheidungen unter Ungewissheit: Dem Entscheider sind nur die von seiner Entscheidung abhängigen möglichen Zustände bekannt, er kann jedoch keine Aussage über die Wahrscheinlichkeiten treffen, mit denen diese Zustände eintreten werden.

Da im Bereich der Medizin meist Vorwissen über die Eintrittswahrscheinlichkeiten vorliegt bzw. durch Studien ermittelt werden kann, werden hier nur Entscheidungen unter Risiko betrachtet.

In der klassischen Statistik werden zur Schätzung eines unbekannten Parameters θ nur die Annahmen über θ und die Stichprobeninformation benutzt. Darüber hinausgehende Informationen über die Entscheidungssituation und insbesondere die Konsequenzen möglicher Entscheidungen werden

nicht berücksichtigt. Beispielsweise wird anhand der Daten einer Studie die Sensitivität eines bestimmten diagnostischen Verfahrens geschätzt, aber nicht untersucht, was für Folgen die Anwendung dieses Verfahrens in der Praxis hat.

In der Entscheidungsanalyse hingegen werden Stichprobeninformation und andere relevante Aspekte des Problems kombiniert, um so die beste Entscheidung zu finden. Insbesondere werden die Konsequenzen der möglichen Entscheidungen berücksichtigt. Bezogen auf obiges Beispiel würden die Konsequenzen eines (falsch) negativen bzw. positiven Testergebnisses bei den Kranken und den Gesunden in die Analyse mit einbezogen und untersucht, ob es sinnvoll ist, die Patienten zu testen oder ob es besser ist, alle Patienten ungetestet zu behandeln oder alle ungetestet nicht zu behandeln.

Zusätzlich zur Stichprobeninformation wird somit für eine Entscheidungsanalyse eine weitere Art von Information verwendet: die möglichen Konsequenzen der verschiedenen Entscheidungsalternativen, ausgedrückt durch eine Verlustfunktion. Allgemein besteht ein statistisches Entscheidungsproblem aus drei Elementen: Dem Zustandsraum Θ in dem der unbekannte Zustand θ liegt, der Menge aller möglichen Alternativen $\mathcal{A}$, aus der eine Alternative a ausgewählt werden soll und einer Verlustfunktion $L(\theta, a)$ [62, 63].

Der **Zustandsraum** Θ gibt den Bereich aller möglichen Werte von θ an. Im allgemeinen repräsentiert θ den "wahren" aber unbekannten Zustand. Beispielsweise ist θ bei einem Schätzproblem der zu schätzende Parameter (Θ wird dann auch Parameterraum genannt). In obigem Beispiel wäre θ der Gesundheitszustand des Patienten. Die Vorinformation über den unbekannten

Zustand θ wird durch eine Wahrscheinlichkeitsfunktion bzw. Dichtefunktion $\pi(\theta)$ ausgedrückt. D.h. für $A \subset \Theta$ gilt:

$$\begin{aligned} P(\theta \in A) &= \begin{cases} \int_A \pi(\theta) d\theta & \text{stetiger Fall} \\ \sum_A \pi(\theta) & \text{diskreter Fall} \end{cases} \\ &= \int_A dF^{\pi}(\theta), \end{aligned}$$

wobei $F^{\pi}(\theta)$ die Verteilungsfunktion von θ ist [62, 63].

Die **Menge aller möglichen Alternativen** oder Entscheidungen wird mit $\mathcal{A}$ bezeichnet. Die Menge kann diskret sein, z.B. die Entscheidung für eine bestimmte Therapie, oder auch stetig, wie z.B. bei der Schätzung eines Parameters. Bei Schätzproblemen ist die Menge aller möglichen Alternativen $\mathcal{A}$ gleich dem Parameterraum Θ [62].

Die **Verlustfunktion** $L(\theta, a)$ gibt für alle $(\theta, a) \in \Theta \times \mathcal{A}$ den Verlust der Alternative a an, wenn der wahre Zustand gleich θ ist. In dem Beispiel könnte $L(\theta, a)$ der Verlust an Lebensqualität sein, den ein Patient durch seine Krankheit oder die Behandlung erfährt. Der Nutzen einer Alternative a ist der negative Verlust [62, 63].

Oft werden Daten erhoben, bevor eine Entscheidung getroffen wird. Bei Schätzproblemen werden beispielsweise Experimente durchgeführt, in der Qualitätskontrolle Stichproben gezogen und Patienten vor Beginn einer Therapie diagnostischen Untersuchungen unterzogen. Bei manchen Alternativen wird jedoch bewusst auf die Erhebung von Daten verzichtet. So kann z.B. eine Alternative sein, alle Patienten ohne vorher zu testen mit einer Therapie

T zu behandeln.

Falls Daten zu einem Merkmal erhoben werden, repräsentiert X die Stichprobe und die Menge aller möglichen Stichproben, der Stichprobenraum wird mit $\mathcal{X}$ bezeichnet. Meist ist X ein Vektor $X = (X_1, X_2, \ldots, X_n)$ wobei die X_i unabhängige Beobachtungen einer gemeinsamen Verteilung sind. Die Realisationen der Zufallsgröße X werden mit x bezeichnet. Die Wahrscheinlichkeitsfunktion bzw. Dichtefunktion von X hängt vom unbekannten Zustand θ ab. Mit $P_\theta(A)$ bzw. $P_\theta(X \in A)$ wird die Wahrscheinlichkeit des Ereignisses A ($A \subset \mathcal{X}$) bezeichnet, wenn θ der wahre Zustand ist. Dann gilt für :

$$\begin{aligned} P_\theta(A) &= \begin{cases} \int_A f(x|\theta)dx & \text{stetiges } X \\ \sum_A f(x|\theta) & \text{diskretes } X \end{cases} \\ &= \int_A dF^X(x|\theta), \end{aligned}$$

wobei $F^X(x|\theta)$ die Verteilungsfunktion von X ist.

Eine **Entscheidungsregel** $\delta(x)$ ist eine Funktion von $\mathcal{X}$ nach $\mathcal{A}$, die die Daten, falls vorhanden, mit der Menge der möglichen Alternativen verbindet. Falls $X = x$ der beobachtete Wert ist, wird die Alternative $\delta(x)$ gewählt. Der Raum aller möglichen Entscheidungsregeln wird mit $\mathcal{D}$ bezeichnet. Bei Schätzproblemen entspricht die Entscheidungsregel dem Schätzer. Falls keine Daten erhoben werden, ist die Entscheidungsregel einfach eine Alternative. Für die Lösung des Entscheidungsproblems soll die Entscheidungsregel so gewählt werden, dass der erwartete Verlust minimiert bzw. der erwartete Nutzen maximiert wird [63].

Da der wahre Verlust $L(\theta, a)$ nie mit Sicherheit bekannt ist, wird bei der Auswahl der besten Alternative der erwartete Verlust betrachtet und die Alternative gesucht, die den geringsten erwarteten Verlust hat. Der erwartete Verlust einer Entscheidungsregel wird in der Entscheidungstheorie Risiko der Entscheidungsregel (bzw. der Alternative) genannt. Je nachdem, auf was der Verlust bezogen wird, gibt es verschiedene Möglichkeiten den erwarteten Verlust einer Entscheidungsregel zu berechnen.

Das **(frequentistische) Risiko** einer Entscheidungsregel wird mit $R(\theta, \delta(x))$ bezeichnet und ist definiert als:

$$R(\theta, \delta(x)) = E_\theta^X \left[L(\theta, \delta(X))\right] = \int_{\mathcal{X}} L(\theta, \delta(x)) f(x|\theta) \mathrm{d}x.$$

Falls keine Daten erhoben werden, ist $R(\theta, \delta(x)) \equiv L(\theta, \delta(x))$. Das frequentistische Risiko ist der durchschnittliche Verlust bei Verwendung einer Entscheidungsregel $\delta(x)$, wenn der wahre Wert des unbekannten Parameters θ ist [62].

Das **bedingte Bayesrisiko** einer Alternative oder Entscheidungsregel ist der erwartete Verlust der Alternative oder Entscheidungsregel unter der Annahme, dass die Daten bekannt sind. D.h. für die Bestimmung des bedingten Bayesrisikos wird angenommen, dass die Daten x fest sind, so dass auch die Alternative $a_x = \delta(x)$ fest ist, falls die Entscheidungsregel nicht randomisiert ist. Das bedingte Bayesrisiko wird mit $\rho(\pi^*(\theta), a_x)$ bezeichnet und ist definiert als [62]:

$$\rho(\pi^*(\theta), \delta(x)) = \rho(\pi^*(\theta), a_x) = E_x^{\pi^*}[L(\theta, a_x)] = \int_\Theta L(\theta, a_x) \pi^*(\theta|x) \mathrm{d}\theta.$$

Das bedingte Bayesrisiko gibt den erwarteten Verlust einer bestimmten Aktion gegeben die Daten an. [6] Es wird auch *posteriori erwarteter Verlust* genannt [62].

Das **Bayesrisiko** einer Entscheidungsregel ist der erwartete Verlust einer Entscheidungsregel, bevor die Daten bekannt sind. Vorinformation über θ wird durch $\pi^*(\theta)$ ausgedrückt. Das Bayesrisiko wird mit $r(\pi^*, \delta(x))$ bezeichnet und ist definiert als [62]:

$$\begin{aligned} r(\pi^*(\theta), \delta(x)) &= E^{\pi^*}[R(\theta, \delta(x))] = E^{\pi^*}[E_\theta^X L(\theta, \delta(X))] \\ &= \int_\Theta \int_{\mathcal{X}} L(\theta, \delta(X)) f(x|\theta) \pi^*(\theta) \mathrm{d}x \mathrm{d}\theta. \end{aligned} \tag{2.20}$$

Es gibt somit den *a priori erwarteten Verlust* einer bestimmten Aktion an.

Unterschiedliche Kriterien können zur Beurteilung welche Entscheidungsregeln besser als andere oder gar optimal sind, herangezogen werden. Eines der Kriterien ist die **Zulässigkeit**: Eine Entscheidungsregel $\delta_1(x)$ ist zulässig bezüglich einer anderen Entscheidungsregel $\delta_2(x)$, falls es mindestens einen Wert θ gibt, für den das frequentistische Risiko von $\delta_1(x)$ kleiner ist als das von $\delta_2(x)$ [62].

Eine Entscheidungsregel $\delta(x)$ ist eine **Bayesregel** falls sie das Bayesrisiko minimiert. Bayesrgeln werden mit δ^* bezeichnet. Im Allgemeinen hängt die Bayesregel von $\pi^*(\theta)$ ab. Das kleinste Risiko einer Bayesregel für ein $\pi^*(\theta)$

[6] Das bedingte Bayesrisiko entspricht der (overall) Fehlerrate in einem Klassifikations- bzw. Diskriminationsproblem, wenn $L(\theta, a_x) = 0$ für eine richtige bzw. $L(\theta, a_x) = 1$ für eine falsche Zuordnung gesetzt wird [64].

wird mit r^* bezeichnet. Jede andere Entscheidungsregel, die dasselbe Bayesrisiko hat, ist auch eine Bayesregel [62].

Wenn eine Entscheidungsregel $\delta_1(x)$ zulässig bezüglich aller anderen Entscheidungsregeln auf $\mathcal{D}$ ist, d.h. es mindestens ein θ gibt, für das das frequentistische Risiko für δ_1 kleiner ist als für alle anderen Entscheidungsregeln $\delta_i \in \mathcal{D}$, $i \neq 1$, dann ist δ_i Bayesregel für mindestens ein $\pi^*(\theta)$ [62].

2.5.2 Klinische Entscheidungsanalyse

Die klinische Entscheidungsanalyse ist ein mathematisch-formaler Ansatz zur medizinischen Entscheidungsfindung unter Unsicherheit, bei dem Nutzen, Risiken und gegebenenfalls Kosten gewichtet und Unterschiede zwischen den verschiedenen Handlungsalternativen verdeutlicht werden. Sie ist definiert als systematischer, expliziter und quantitativer Ansatz zur Entscheidungsfindung unter Unsicherheit [65, 66]. Systematisch heißt, dass das Entscheidungsproblem in seine Komponenten zerlegt, diese dann individuell analysiert und anschließend systematisch wieder zusammengesetzt werden. Explizit bedeutet, dass die Entscheidungsanalyse den Entscheidungsträger dazu verpflichtet alle Handlungsalternativen und Annahmen detailliert darzustellen. Mit quantitativ ist gemeint, dass Unsicherheiten in Wahrscheinlichkeiten ausgedrückt und den Endpunkten Nutzwerte (sogenannte Utilities) zugeordnet werden, die die Präferenzen des Entscheidungsträgers ausdrücken [66].

Die Entscheidungsanalyse verfolgt das Prinzip der Nutzenmaximierung (Utilitarismus). Für jede Handlungsalternative werden die Nutzwerte mit den zugehörigen Wahrscheinlichkeiten gewichtet und daraus der erwartete Nutzen berechnet. Die Handlungsstrategie mit dem größten erwarteten Nutzen,

also diejenige, die im Mittel den größten Nutzen hat, wird als beste gewählt. Der zu maximierende Nutzen ist vor der Analyse festzulegen bzw. zu operationalisieren. Dies kann ein einzelner Parameter wie beispielsweise höhere Überlebenswahrscheinlichkeit, bessere Lebensqualität oder geringere Kosten sein, oder eine gewichtete Kombination oder wie bei Kosten-Effektivitäts-Analysen ein Verhältnis verschiedener Parameter [66].

Da verschiedene Parameter der Krankheit, ihrer Behandlung und der daraus resultierenden Konsequenzen innerhalb eines mathematischen Modells zusammengeführt werden, spricht man auch von entscheidungsanalytischer Modellierung [66]. Besonders sinnvoll ist eine klinische Entscheidungsanalyse bei komplexen Fragestellungen, bei denen es schwierig ist, ohne entscheidungsanalytische Modellierung alle Aspekte zu erfassen.

Bevor ein Entscheidungsproblem quantitativ analysiert werden kann, ist ein Entscheidungsmodell aufzustellen. Mit einem solchen Modell wird versucht, die relevanten Aspekte der Wirklichkeit möglichst gut abzubilden und die weniger relevanten Aspekte zu vereinfachen. Dabei kann man zwei entscheidungsanalytische Modelltypen unterscheiden: Entscheidungsbäume und Markov-Modelle. Entscheidungsbäume kommen bei einfacheren Problemen zum Einsatz, insbesondere wenn alle relevanten Ereignisse innerhalb eines kurzen Zeithorizontes eintreten. Demgegenüber werden Markov-Modelle vorwiegend bei Problemen mit komplexen Strukturen und längeren Zeithorizonten verwendet. Kombinationen der beiden Verfahren sind möglich [66, 67]

Entscheidungsbäume

Mit Hilfe eines Entscheidungsbaumes können mögliche Entscheidungen, Ereignisse und Endpunkte strukturiert werden. Ein Entscheidungsbaum stellt die zeitliche und logische Struktur des Entscheidungsproblems, alle relevanten Alternativen, unsicheren Ereignisse, deren Eintrittswahrscheinlichkeiten sowie die zu erwartenden Konsequenzen dar [66, 67].

Dabei werden folgende Strukturelemente verwendet:

- Entscheidungsknoten, dargestellt durch ein Rechteck
- Ereignis- oder Zufallsknoten, dargestellt durch einen Kreis
- Endknoten, dargestellt durch ein Dreieck

In Abbildung (2.7) ist als Beispiel ein Entscheidungsbaum für das Problem, ob bei Patienten mit Verdacht auf eine Krankheit K alle behandelt, keiner behandelt oder alle mit einem Test T getestet und nur die Positiven behandelt werden sollen, abgebildet. Ein Entscheidungsbaum wird von links nach rechts betrachtet. An der Wurzel des Entscheidungsbaums wird die Entscheidungssituation bzw. die Zielpopulation eingetragen, hier also Patienten mit Verdacht auf Krankheit K.

Der nachfolgende rechteckige Entscheidungsknoten symbolisiert, dass es drei verschiedene Handlungsalternativen gibt: „Patienten nicht testen und alle behandeln", oder „Patienten nicht testen und keinen behandeln" oder „Patienten testen und nur die Positiven behandeln".

Die zweite Verzweigung ist bei allen drei Alternativen gleich und gibt an, ob der Patient die Krankheit K hat ($K+$) oder nicht ($K-$). Die Unsicherheit des

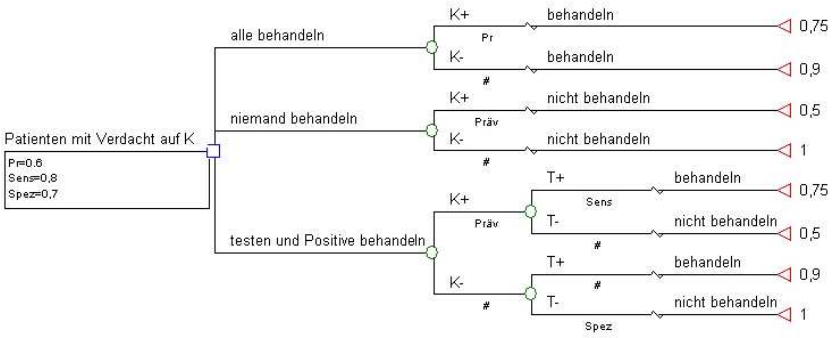

Abbildung 2.7: *Entscheidungsbaum für das Problem, ob bei Patienten mit Verdacht auf eine Krankheit K alle behandelt, keiner behandelt oder alle mit einem Test T getestet und nur die Positiven behandelt werden sollen*

Entscheiders über den Zustand des Patienten wird dabei durch einen runden Ereignis- oder Zufallsknoten symbolisiert. Das Vorwissen, mit welcher Wahrscheinlichkeit Patienten der betrachteten Population die Krankheit K haben, wird durch die Prävalenz ausgedrückt. In dem Beispiel beträgt die Prävalenz 60%.

Bei den ersten beiden Handlungsalternativen gibt es keine weiteren Verzweigungen, die Patienten werden entweder alle behandelt (Alternative 1) oder alle nicht behandelt (Alternative 2). Bei der dritten Alternative kommt noch eine weitere Unsicherheit ins Spiel, der Ausgang des diagnostischen Tests. Dies wird durch einen zweiten Ereignis- oder Zufallsknoten dargestellt. Sensitivität und Spezifität geben hier das Vorwissen über den Test an.

Der Baum endet an der rechten Seite mit den Endknoten, an denen die Konsequenzen, d.h. die Ausprägungen der betrachteten Zielgröße eingetragen werden. In dem Beispiel soll die 5-Jahres-Überlebenswahrscheinlichkeit betrachtet werden. Patienten ohne Krankheit K, die nicht behandelt werden,

haben eine 5-Jahres-Überlebenswahrscheinlichkeit von 1. Gesunde Patienten, die fälschlicherweise behandelt werden, haben durch Nebenwirkungen und Komplikationen eine etwas geringere 5-Jahres-Überlebenswahrscheinlichkeit von 0,9. Patienten, die an der Krankheit K erkrankt sind und behandelt werden, haben eine 5-Jahres-Überlebenswahrscheinlichkeit von 0,75, unbehandelte Patienten hingegen nur von 0,5. Nach Konstruktion des Baumes wird für jede Entscheidungsalternative der Erwartungswert für die untersuchte Zielgröße berechnet.

In der Notation der statistischen Entscheidungstheorie kann das Beispiel folgendermaßen beschrieben werden: Der unbekannte Zustand θ ist entweder „gesund" oder „krank", d.h. Θ ={gesund, krank}. Die Vorinformation über θ wird durch π(gesund)=0.4 und π(krank)=0.6 ausgedrückt. Die Menge aller Alternativen $\mathcal{A}$ besteht aus den drei Elementen a_1=„alle Patienten behandeln", a_2=„keinen Patienten behandeln" und a_3=„alle Patienten testen und die Positiven behandeln".

Bei den beiden ersten Alternativen werden keine Daten erhoben, die Entscheidungsregel lautet immer „alle Patienten behandeln" bzw. „keinen Patienten behandeln". Bei der dritten Alternative werden Daten erhoben. Dabei ist X Null-Eins-verteilt mit der Wahrscheinlichkeit P(Test positiv|Patient erkrankt) = Sensitivität für kranke Patienten und mit der Wahrscheinlichkeit P(Test positiv|Patient gesund) = 1 - Spezifität für gesunde Patienten. In dem Beispiel beträgt die Sensitivität 0,8 (80%) und die Spezifität 0,7 (70%). Die Entscheidungsregel lautet im Fall der dritten Alternative δ(Test positiv)=„Patient behandeln" bzw. δ(Test negativ)=„Patient nicht behandeln".

Da der Verlust der negative Nutzen ist, gibt die Verlustfunktion $L(\theta, a(x))$ an, um wieviel sich die 5-Jahres-Überlebenswahrscheinlichkeit des Patienten verringert. D.h. L(gesund, unbehandelt) =0 , L(gesund, behandelt) = 0,1, L(krank, unbehandelt) = 0,5 und L(krank, behandelt) = 0,25. Da die Entscheidung, welche der drei Alternativen die beste ist, getroffen werden muss, bevor die Daten erhoben werden, d.h. bevor die Patienten getestet werden, wird das Bayesrisiko (2.20) berechnet, um den erwarteten Verlust aller drei Alternativen zu ermitteln:

$$\begin{aligned}
r(\pi^*(\theta), a_1) &= \textstyle\sum_{\theta\in\Theta,} L(\theta, a_1)\pi^*(\theta) \\
&= L(\text{gesund, behandeln})\pi(\text{gesund}) + \\
&\quad L(\text{krank, behandeln})\pi(\text{krank}) \\
&= 0,1 \cdot 0,4 + 0,25 \cdot 0,6 \\
&= 0{,}190 \\
r(\pi^*(\theta), a_2) &= \textstyle\sum_{\theta\in\Theta} L(\theta, a_1)\pi^*(\theta) \\
&= L(\text{gesund, nicht behandeln})\pi(\text{gesund}) + \\
&\quad L(\text{krank, nicht behandeln})\pi(\text{krank}) \\
&= 0 \cdot 0,4 + 0,5 \cdot 0,6 \\
&= 0{,}300
\end{aligned}$$

$$\begin{aligned}
r(\pi^*(\theta), a_3) &= \sum_{\theta\in\Theta, x\in\mathcal{X}} L(\theta, \delta(x)) f(x|\theta)\pi^*(\theta) \\
&= L(\text{gesund, nicht behandeln})P(\text{Test negativ}|\text{gesund})\pi(\text{gesund}) + \\
&\quad L(\text{gesund, behandeln})P(\text{Test positiv}|\text{gesund})\pi(\text{gesund}) + \\
&\quad L(\text{krank, nicht behandeln})P(\text{Test negativ}|\text{krank})\pi(\text{krank}) + \\
&\quad L(\text{krank, behandeln})P(\text{Test positiv}|\text{krank})\pi(\text{krank}) \\
&= L(\text{gesund, nicht behandeln}) \cdot \text{Spezifität} \cdot (1 - \text{Prävalenz}) + \\
&\quad L(\text{gesund, behandeln}) \cdot (1 - \text{Spezifität}) \cdot (1 - \text{Prävalenz}) +
\end{aligned}$$

$$
\begin{aligned}
& L(\text{krank, nicht behandeln}) \cdot (1 - \text{Sensitivität}) \cdot \text{Prävalenz}) + \\
& L(\text{krank, behandeln}) \cdot \text{Sensitivität} \cdot \text{Prävalenz}) \\
= \; & 0 \cdot 0,7 \cdot 0,4 + 0,9 \cdot (1 - 0,7) \cdot 0,4 + 0,5 \cdot (1 - 0,8) \cdot 0,6 + \\
& 0,25 \cdot 0,8 \cdot 0,6 \\
= \; & 0,192
\end{aligned}
$$

Die Alternative a_1, d.h. alle Patienten behandeln, ohne vorher zu testen, hat mit einer erwarteten Verringerung der 5-Jahres-Überlebenswahrscheinlichkeit von 0,190 das kleinste Bayesrisiko aller drei Alternativen. Sie ist also unter $\pi^*(\theta)$ Bayes-Regel, d.h. beste Alternative.

Wird statt des erwarteten Verlustes, wie in der klinischen Entscheidungsanalyse üblich, der erwartete Nutzen, d.h. 5-Jahres-Überlebenswahrscheinlichkeit verwendet, wird das Maximum gesucht. Analog ist die Alternative a_1 mit einer erwarteten 5-Jahres-Überlebenswahrscheinlichkeit von 0,810 die beste.

Zusammengefasst lässt sich feststellen, dass die klinische Entscheidungsanalyse ein Spezialfall der statistischen Entscheidungstheorie ist, bei dem sowohl der Zustandsraum $\Theta =$ als auch die Menge aller Alternativen $\mathcal{A}$ diskret sind. Statt den Verlust $L(\theta, a)$ zu minimieren, wird versucht den Nutzen zu maximieren. Dazu wird das Bayesrisiko betrachtet.

Der Vorteil des Entscheidungsbaums ist, dass er sehr anschaulich die Struktur des Entscheidungsproblems darstellt und deswegen, insbesondere auch von nicht mit den Methoden der Entscheidungsanalyse vertrauten Leuten, gut akzeptiert wird. Wenn das Entscheidungsproblem jedoch zeitveränderliche

Parameter beinhaltet, der Zeitpunkt des Eintretens eines bestimmten Ereignisses eine Rolle spielt oder relevante Ereignisse mehrmals auftreten können, ist an Stelle eines Entscheidungsbaumes ein Markov-Ansatz das adäquate Modell. [66].

Markov-Modelle werden nachfolgend nur kurz beschrieben, weil beim Pankreaskarzinom der Zeithorizont sehr kurz ist (5-Jahres-Überlebensrate $< 5\%$) und somit zur Modellierung ein Entscheidungsbaum verwendet werden kann. Für eine weiterführende Einführung siehe [68].

Markov-Modelle

Bei Verwendung eines Markov-Modells wird die Entscheidungssituation als stochastischer Prozess, genauer als Markov-Kette, formuliert. Dadurch kann die Veränderung einer Zufallsgröße X, z.B. des Gesundheitszustands, über die Zeit beschrieben werden.

Allgemein ist ein stochastischer Prozess definiert als Familie $X = \{X_T, t \in T\}$ von Zufallsgrößen auf einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathcal{P})$ mit dem gemeinsamen Wertebereich E. T wird Parameterraum, E Zustandsraum des stochastischen Prozesses genannt [69]. Wenn X der Gesundheitszustand ist, würde der Wertebereich E die möglichen Gesundheitszustände (z.B. „gesund" und „krank"), die zu den verschiedenen Zeitpunkten t beobachtet werden, repräsentieren.

Wenn die Übergangswahrscheinlichkeiten zu jedem Beobachtungszeitpunkt nur vom aktuellen Zustand und nicht von den davor durchlaufenen Zuständen

abhängen, d.h. wenn

$$P(X_{t+1} = j|X_t, X_{t-1}, \ldots, X_0) = P(X_{t+1} = j|X_t) \tag{2.21}$$

für alle $t \in N_0$ gilt, nennt man den Prozess „gedächtnislos". Bedingung (2.21) wird auch Markov-Eigenschaft genannt.

Wenn die Markov-Eigenschaft gilt und der Zustandsraum E eines stochastischen Prozesses $X = \{X_t, t \in N_0\}$ abzählbar ist, heißt der stochastische Prozess Markov-Kette [69].

Sind die (einschrittigen) Übergangswahrscheinlichkeiten

$$p_{ij}(t+1, t) := P(X_{t+1} = i|X_t = j)$$

von t unabhängig, wird der stochastische Prozess homogen bzw. Markov-Kette mit stationären Übergangswahrscheinlichkeiten genannt.

Bei einem entscheidungsanalytischen Markov-Modell besteht der Zustandsraum E aus einer endlichen Zahl von disjunkten[7] und erschöpfenden[8] Gesundheitszuständen, die von Patienten durchlaufen werden können. Die Zeit wird in äquidistante Intervalle, die Zyklen, eingeteilt, die den Parameterraum T bilden. In jedem Zyklus geben die Übergangswahrscheinlichkeiten an, welche Übergänge zwischen den einzelnen Gesundheitszuständen möglich sind. Meist wird dabei angenommen, dass die Übergangswahrscheinlichkeiten die Markov-Eigenschaft (2.21) erfüllen. Auf diese recht strenge Voraussetzung

[7] d.h., ein Patient kann nicht gleichzeitig zwei verschiedene Gesundheitszustände haben

[8] d.h., alle möglichen Gesundheitszustände werden abgedeckt

kann aber gegebenenfalls verzichtet und flexiblere Modelle formuliert werden.

Sensitivitätsanalysen

Da die Bayes-Regel, d.h. die beste Alternative von $\pi^*(\theta)$ abhängt, werden zur Prüfung der Stabiliät der Ergebnisse Sensitivitätsanalysen durchgeführt. Dabei wird durch systematische Veränderung der Parameter (z.B. Prävalenz, Sensitivität) untersucht, inwieweit die Entscheidung für eine Alternative von den Werten der Parameter abhängt.

Je nach Anzahl der gleichzeitig variierenden Parameter wird zwischen univariaten und multivariaten Sensitivitätsanalysen unterschieden. Außerden kann zwischen deterministischen und probabilistischen Sensitivitätsanalysen differenziert werden. Bei deterministischen Sensitivitätsanalysen werden ein oder auch mehrere Parameter in einem vorgegebenen Bereich variiert und die Zielgrößen in Abhängigkeit der Parameter dargestellt.

In obigem Beispiel könnte z.B. eine Einweg-Sensitivitätsanalyse bezüglich der Prävalenz durchgeführt werden, um zu untersuchen, inwieweit die Wahl der besten Strategie von der Prävalenz der Krankheit K in der betrachteten Population abhängt.

Wie in Abbildung 2.8 dargestellt, ist bis zu einer Prävalenz von 0,13 die Strategie „keinen Patienten behandeln" die beste. Bei einer Prävalenz zwischen 0,13 und 0,58 ist die Strategie „alle Patienten testen und Positive behandeln" am besten. Erst bei einer Sensitivität größer als 0,58 ist die Strategie „alle Patienten behandeln" optimal.

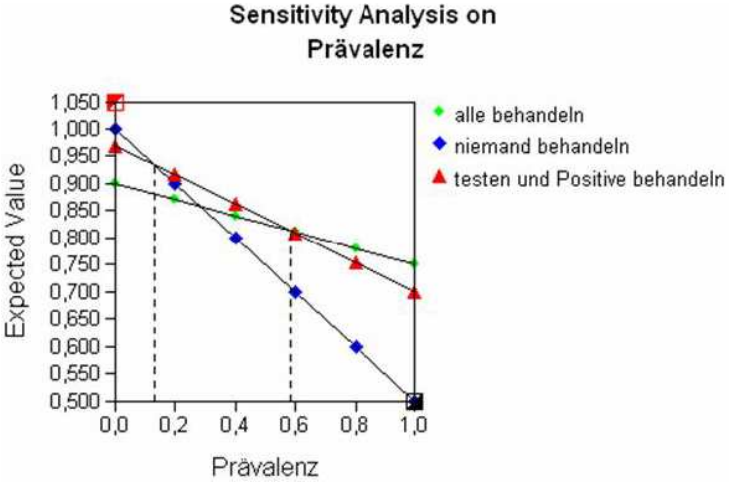

Abbildung 2.8: *Ergebnis der Einweg-Sensitivitätsanalyse bezüglich der Prävalenz. Dargestellt ist die erwartete 5-Jahres-Überlebenswahrscheinlichkeit (expected value) der drei Strategien in Abhängigkeit von der Prävalenz.*

Bei der probabilistischen Sensitivitätsanalyse werden die Parameter mehrfach mittels ihrer Verteilungen simuliert und jeweils die Bayesrisiken bzw. Bayesnutzen für alle Alternativen berechnet. Aus der Häufigkeit mit der die verschiedenen Alternativen bei den einzelnen Simulationen die beste Alternative sind, kann abgeschätzt werden, wie stabil die Ergebnisse bei gleichzeitiger Variation der Parameter sind.

Beispielsweise könnte eine probabilistische Sensitivitätsanalyse gleichzeitig für die drei Parameter Sensitivität, Spezifität und Prävalenz durchgeführt werden. Da diese Größen binomialverteilt sind, wird die Beta-Verteilung als a-priori-Verteilung gewählt. In Abbildung (2.9) ist das Ergebnis einer Monte-Carlo-Simulation mit 1000 Wiederholungen dargestellt. Die Strategie „alle Patienten behandeln" war in 54% der Simulationen die beste, die Strategie

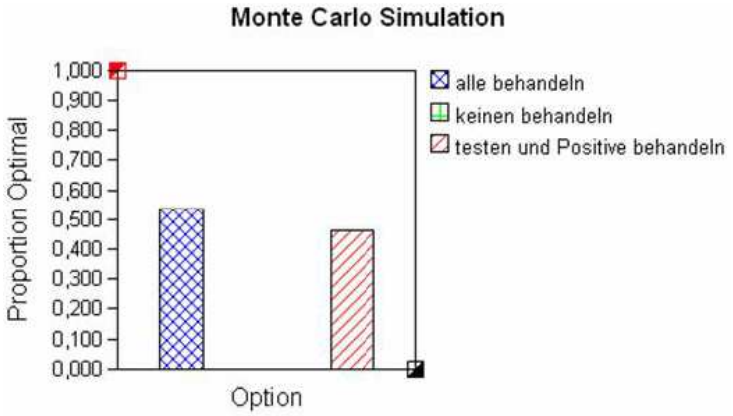

Abbildung 2.9: *Ergebnis der probabilistischen Sensitivitätsanalyse. Dargestellt sind die relativen Häufigkeiten, mit der die Strategien in der probabilistischen Sensitivitätsanalyse die beste Strategie waren.*

„alle Patienten testen und Positive behandeln" hingegen in 46% der Simulationen. Die Strategie „keinen Patienten behandeln" war nie die optimal.

Der Vorteil der probabilistischen Sensitivitätsanalyse ist, dass sie simultan für eine große Zahl von Parametern durchführbar ist. Zudem liefert sie ein Maß für die Unsicherheit, z.B. ein Konfidenzintervall für die erwartete 5-Jahres-Überlebenswahrscheinlichkeit.

Der Median der erwarteten Überlebenszeit beträgt für Strategie „Alle Patienten behandeln" 0,810 (95%-Konfidenzintervall $[0,796; 0,825]$). Für die Strategie „Alle Patienten testen und Positive behandeln" ist der Median der erwarteten 5-Jahres-Überlebenswahrscheinlichkeit 0,809 (95%-Konfidenzintervall $[0,777; 0,836]$).

Kosten-Effektivitäts-Analyse

Die Kosten-Effektivitäts-Analyse ist eine Methode zur Evaluation der Kosten und Folgen von medizinischen Interventionen. Dabei werden die verschiedenen Interventionen im Hinblick auf ihre Kosten im Verhältnis zu ihrer Effektivität verglichen.

Je nachdem welche Perspektive bei der Analyse eingenommen wird, unterscheiden sich die betrachteten Kosten. So wird beispielsweise eine Versicherung andere Kosten für einen stationär aufgenommenen Patienten ansetzen als das Krankenhaus, das die tatsächlich anfallenden Kosten berücksichtigt. Bei der gesellschaftlichen Perspektive kommen neben diesen direkten Kosten noch indirekte Kosten, z.B. durch Arbeitsausfall, hinzu.

Da in das Entscheidungsmodell nur die durch die Diagnostik direkt verursachten Kosten mit einbezogen werden sollen, wird auf eine ausführliche Darstellung der Kosten-Effektivitäts-Analyse verzichtet. Ausführlichere Einführungen geben z.B. [70, 71].

Zentrale Größe bei Kosten-Effektivitäts-Analysen ist das Kosten-Effektivitäts-Verhältnis, welches die inkrementellen Kosten einer Alternative A im Vergleich zu einem „Standard" S zu den korrespondierenden inkrementellen medizinischen Effekten angibt:

$$\text{ICER} = \frac{\text{Kosten}_A - \text{Kosten}_S}{\text{Effektivität}_A - \text{Effektivität}_S} = \frac{\Delta\text{Kosten}}{\Delta\text{Effektivität}}$$

ICER steht dabei für die auch in Deutschland verwendete Abkürzung des englischen Begriffs *incremental cost-effectiveness ratio*.

Wenn A sowohl kostengünstiger als auch effektiver ist als S, d.h. ΔKosten < 0 und ΔEffektivität > 0 ist, ist A besser als S und man sagt A dominiert die Standardstrategie S. Im umgekehrten Fall, d.h. die Alternative A ist teurer und weniger effektiv, wird diese von S dominiert. In den beiden übrigen Fällen - A ist effektiver, aber teurer bzw. A ist günstiger aber weniger effektiv - muss abgewogen werden, ob der Effektivitätsgewinn die Kostenzunahme bzw. die Kosteneinsparung den Effektivitätsverlust rechtfertigt.

Werden mehr als zwei Strategien miteinander verglichen, kann es auch zur so genannten erweiterten Dominanz kommen. Diese tritt dann auf, wenn eine Strategie sowohl weniger effektiv, als auch teurer ist als eine Linearkombination zweier anderer Strategien.

2.6 Auswertungsmethoden

2.6.1 Vergleich und Kombination der Verfahren

Die Konfidenzintervalle der unterschiedlichen Wahrscheinlichkeiten (Sensitivität, Spezifität und prädiktive Werte) wurden nach der Methode von Wilson (siehe (2.4)) berechnet, die Konfidenzintervalle der Likelihood Ratios mittels der Log-Methode (2.12). Der Vergleich der Sensitivitäten und der Spezifitäten erfolgte mit dem McNemar-Test.

Zur Bestimmung der Korrelation und zur Überprüfung der bedingten Unabhängigkeit der Tests innerhalb der Subpopulation der Kranken und der Subpopulation der Gesunden wurden ϕ-Koeffizienten bestimmt und exakte Tests von Fisher gerechnet.

2.6.2 Entscheidungsanalyse

Bei sechs verschiedenen bildgebenden Verfahren ist die Menge möglicher Kombinationen sehr groß. Allerdings sind nur einige der Kombinationen klinisch relevant und durchführbar. Deswegen wurde eine Vorauswahl der Strategien mittels folgender Kriterien getroffen:

- Aufgrund des Odysseus-Syndroms wurden nur Strategien mit einem oder zwei Verfahren berücksichtigt (vgl. Seite 38).
- Bei Einzeltest-Strategien wurden nur Modalitäten berücksichtigt, die sowohl die Dignität als auch die Resektabilität beurteilen können (d.h. US, EUS, CT und MR).
- Bei Zweitest-Strategien wurden nur Kombinationen, in denen mindestens eine Modalität die Resektabilität beurteilen kann, berücksichtigt.

- Da die Reihenfolge der Untersuchungen keinen Einfluss auf die Sensitivität und Spezifität der Kombination hat, wurde die weniger invasive Modalität an die erste Stelle gesetzt, um die Zahl der invasiven Untersuchungen möglichst klein zu halten. Unabhängig voneinander ordneten vier erfahrene Gastroenterologen in völliger Übereinstimmung die sechs Modalitäten nach zunehmender Invasivität in folgender Reihenfolge: US < MR < CT < PET < EUS < ERCP.

Dies führte zu 18 möglichen Testkombinationen (vgl. Tabelle 2.7). Anwendung der beiden Kombinationsregeln BTP und BTN auf die Beurteilung der Dignität und die Beurteilung der Resektabilität bietet bei Kombinationen, in denen beide Tests die Resektabilität beurteilen können, vier Möglichkeiten:

1. BTP für Diagnose und für die Beurteilung der Resektabilität,
2. BTP für Diagnose und BTN für die Beurteilung der Resektabilität,
3. BTN für Diagnose und für die Beurteilung der Resektabilität und
4. BTN für Diagnose und BTP für die Beurteilung der Resktabilität.

Bei Kombinationen, in denen nur eine Modalität die Resektabilität beurteilen kann, gibt es nur zwei Möglichkeiten:

1. BTP für die Diagnose oder
2. BTN für die Diagnose. Dies führt zu insgesamt 44 verschiedenen diagnostischen Strategien (siehe Tabelle 2.7)

In Abbildung 2.10 ist die Struktur eines Astes des Entscheidungsbaums abgebildet. Die ersten beiden Zufallsknoten innerhalb einer Strategie geben den wahren Gesundheitszustand (D+: maligne Läsion, D-: benigne Läsion) und

Kombination	Strategien	Kombination	Strategien
nur US	1	nur CT	1
US gefolgt von MR	4	CT gefolgt von EUS	4
US gefolgt von CT	4	CT gefolgt von PET	2
US gefolgt von EUS	4	CT gefolgt von ERCP	2
US gefolgt von PET	2	PET gefolgt von EUS	2
US gefolgt von ERCP	2	nur EUS	1
nur MR	1	EUS gefolgt von ERCP	2
MR gefolgt von CT	4		
MR gefolgt von EUS	4		
MR gefolgt von PET	2		
MR gefolgt von ERCP	2		

Tabelle 2.7: *Übersicht der 18 untersuchten Kombinationen mit den 44 daraus resultierenden Strategien*

für Patienten mit einem Malignom die Resektabilität (R+: resektabel, R-: irresektabel) wieder. Beim ersten Zufallsknoten werden die Patienten abhängig von der Prävalenz in maligne und benigne getrennt. Die Patienten mit maligner Läsion werden anschließend abhängig von der Resektabilitätsrate in resektabel und irresektabel getrennt. An der nächsten Gabelung werden die Patienten abhängig vom wahren Gesundheitszustand und der Sensitivität bzw. Spezifität des ersten Verfahrens diagnostiziert. Bei der Diagnose einer malignen Läsion wird anschließend die Resektabilität beurteilt. Die letzten beiden Gabelungen geben die Evaluation des Patienten durch das zweite Verfahren wieder. Da in der dargestellten Strategie mittels BTN-Regel kombiniert wird, wird bei einem negativen US Befund kein MR durchgeführt.

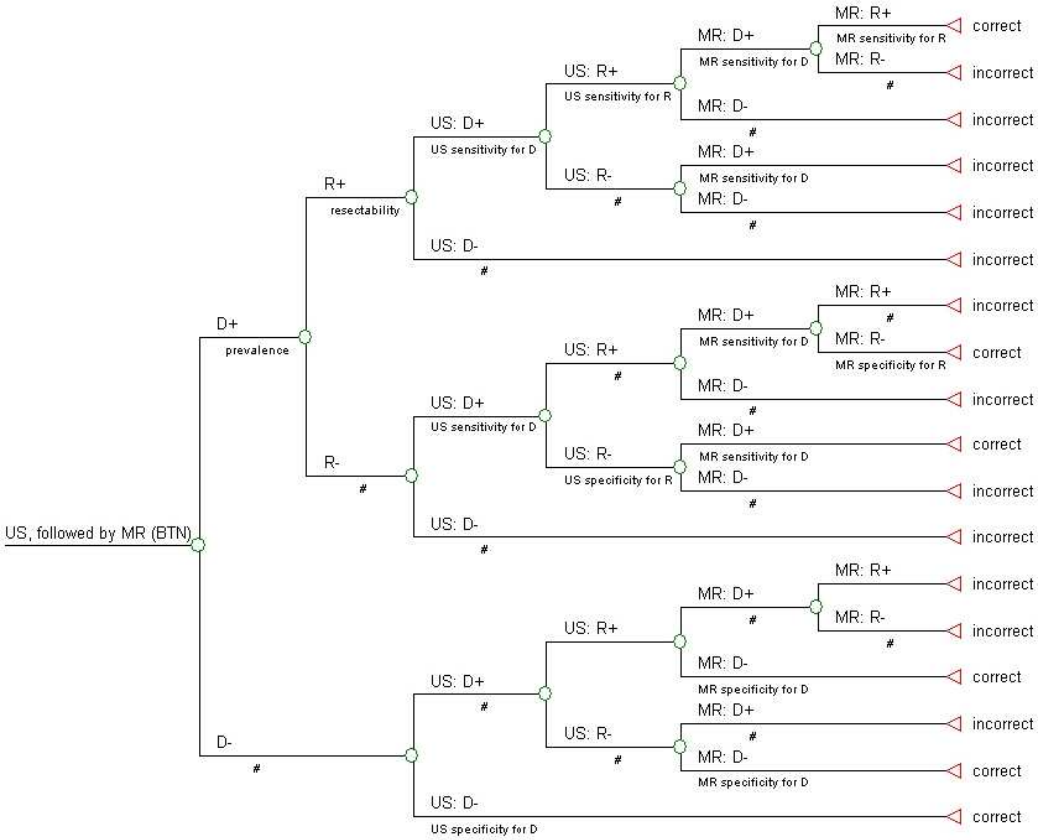

Abbildung 2.10: Struktur eines Astes des Entscheidungsbaums

Da die korrekte Diagnose sowohl der Dignität (maligne oder benigne Läsion) als auch der Resektabilität wichtig für eine adäquate Therapie sind, wurde als Endpunkt die Zahl der richtig klassifizierten Patienten gewählt. Somit ist die Verlustfunktion $L(\theta, a)$ dichotom: $L(\theta, a) = 1$ für eine richtige Zuordnung bzw. $L(\theta, a) = 1$ für eine falsche Zuordnung.

In der Basisfallanalyse wurden die empirischen Daten der Studie benutzt, um Punktschätzer der (bedingten) Wahrscheinlichkeiten zu bestimmen. Anschließend wurde eine Kohortensimulation durchgeführt (deterministische

Analyse), in der eine hypothetische Kohorte Patienten durch den Entscheidungsbaum läuft.

Anschließend wurde eine Einweg-Sensitivitätsanalyse durchgeführt um zu untersuchen, inwieweit die Ergebnisse von der Prävalenz des Pankraskarzinoms in der Population abhängen.

Da die Fallzahlen in einigen Unterästen recht klein waren, wurde zusätzlich eine probabilistische Sensitivitätsanalyse mittels Monte-Carlo-Simulation durchgeführt um zu untersuchen, inwieweit die Ergebnisse durch die Unsicherheiten bezüglich der Modellparameter beeinflusst werden. Für die Monte-Carlo-Simulation wurden für jeden unbekannten Parameter a-priori-Verteilungen festgelegt, aus denen dann wiederholt simultan Stichproben gezogen wurden. Da die (bedingten) Wahrscheinlichkeiten binomialverteilt sind, wurde die Beta-Verteilung als a-priori-Verteilung ausgewählt. Bei denjenigen Parametern, bei denen die empirischen Daten eine Punktschätzung von 0 oder 1 ergaben, wurden exakte Konfidenzintervalle berechnet und deren Grenzen als Parameter einer Dreiecksverteilung verwendet.

Für die Ermittlung der Kosten der Diagnostik wurde der Nebenkostentarif der Deutschen Krankenhausgesellschaft (DKG-NT) verwendet.

2.6.3 Verwendete Software

Sensitivitäten, Spezifitäten und die prädiktiven Werte wurden ebenso wie die bedingten Wahrscheinlichkeiten für den Entscheidungsbaum mit SPSS 11.0 (SPSS INC., Chicago, USA) ermittelt. Für die Berechnung der ϕ-Koeffizienten und der p-Werte des exakten Tests nach Fisher wurde ebenfalls

SPSS verwendet.

Die Berechnung der Likelihood Ratios erfolgte ebenso wie die Berechnung der Konfidenzintervalle der Wahrscheinlichkeiten mit Hilfe eines selbstgeschriebenen Programms in Excel 2002 (Microsoft Corporation, Redmond, USA).

Für die Berechnung der exakten Konfidenzintervalle, die für Dreiecksverteilungen in den Monte-Carlo-Simulationen benötigt wurden, wurden StatXact 5 (Cytel Software Corporation, Cambridge, USA) verwendet. Der Entscheidungsbaum wurde in TreeAge Pro 2004 (TreeAge Software, Williamstown, USA) programmiert und ausgewertet.