Ruprecht-Karls-Universität Heidelberg
Institut für Medizinische Biometrie und Informatik
Abteilung Medizinische Biometrie
Direktor: Prof. Dr. N. Victor

Nr. 27

Methoden zur Konstruktion und Evaluierung klinischer Scores

R. Holle

Habilitationsschrift
Zur Erlangung der Venia Legendi
für das Fach
Medizinische Biometrie
der Hohen Medizinischen Fakultät
der Ruprecht-Karls-Universität Heidelberg

1995

INHALT

1 Einleitung ... 1
1.1 Allgemeine Einführung ... 1
1.2 Ziel und Gliederung der Arbeit ... 2

2 Überblick über Scores ... 4
2.1 Definition klinischer Scores ... 4
2.2 Beispiele klinischer Scores ... 5
2.3 Einteilungsprinzipien ... 6
2.3.1 Einteilung nach Eingangsinformation ... 7
2.3.2 Einteilung nach strukturellen Eigenschaften ... 8
2.3.3 Einteilung nach Funktion ... 9
2.4 Einsatz von Scores ... 10
2.4.1 Scores im Rahmen der Patientenberatung ... 11
2.4.2 Scores als Grundlage für Therapieentscheidungen ... 12
2.4.3 Scores als Zielvariable in Studien ... 14
2.4.4 Scores als Kontrollvariablen in Studien ... 15
2.4.5 Scores in der Qualitätssicherung ... 16
2.5 Kritik an klinischen Scores ... 17

3 Gütekriterien für klinische Scores ... 19
3.1 Theoretische Gütekriterien ... 19
3.2 Überblick über empirische Gütekriterien ... 21
3.2.1 Primäre Gütekriterien ... 21
3.2.2 Sekundäre Gütekriterien ... 23
3.3 Reliabilitätskonzepte ... 24
3.3.1 Varianzquellen ... 24
3.3.2 Zufällige vs. systematische Variation ... 25
3.3.3 Spezielle Reliabilitätskonzepte ... 26
3.3.4 G-Studien und D-Studien ... 28
3.4 Validitätskonzepte ... 29
3.4.1 Konstruktvalidität vs. Kriteriumsvalidität ... 29
3.4.2 Diskrimination ... 31
3.4.3 Änderungssensitivität ... 34
3.4.4 Kalibration ... 36
3.5 Relevanz der Gütekriterien ... 36

4 Statistische Güteparameter für Scores ... 38
4.1 Statistisches Modell des Messens ... 38
4.2 Reliabilitätsmaße ... 40
4.2.1 Reliabilitätsmaße für quantitative Merkmale ... 40
4.2.2 Reliabilitätsmaße für dichotome Merkmale ... 47

4.2.3 Reliabilitätsmaße für ordinale Merkmale 50
4.3 Schätz- und Testverfahren zur Reliabilität 51
4.4 Validitätsmaße 52
4.4.1 Validitätsmaße für quantitative Kriterien 53
4.4.2 Validitätsmaße für dichotome Kriterien 55
4.4.3 Validitätsmaße für zensierte Zeitvariablen als Kriterium 60
4.5 Kalibration 62
4.6 Zufallskorrigierte Validitätsparameter 65
4.7 Änderungssensitivität 70

5 Konstruktion klinischer Scores 74
5.1 Auswahl von Kandidatenvariablen 75
5.2 Variablenselektion 78
5.3 Suche des "besten" multivariaten Modells 81

6 Aspekte der Studienplanung 86
6.1 Planung von Reliabilitätsstudien 86
6.1.1 Ziele von Reliabilitätsstudien 87
6.1.2 Design von Reliabilitätsstudien 88
6.1.3 Ausschaltung systematischer und zufälliger Fehler 91
6.2 Planung von Validierungsstudien 91
6.2.1 Ziele von Validierungsstudien 92
6.2.2 Design von Validierungsstudien 93
6.2.3 Ausschaltung systematischer und zufälliger Fehler 95
6.3 Bewertung des klinischen Nutzens von Scores 98
6.3.1 Bedeutung von Feldstudien 98
6.3.2 Design von Feldstudien 99

7 Zusammenfassung 102

Literatur 103

1 EINLEITUNG

1.1 Allgemeine Einführung

Mit der Zunahme der Informationsflut in der Medizin in den letzten Jahrzehnten ist der Wunsch nach Informationsverdichtung immer stärker geworden. Gleichzeitig führte die empirisch-wissenschaftliche Ausrichtung der Medizin zu einer immer weitergehenden Quantifizierung und Objektivierung medizinischer Informationen. Daraus resultierte eine Tendenz, Sachverhalte meßbar machen zu wollen, die bis dahin nur qualitativ beschreibbar waren, wie beispielsweise das Konzept der Lebensqualität. In den siebziger und achtziger Jahren entstanden so eine praktisch unüberschaubare Zahl von sogenannten klinischen Scores, Indexen oder Skalen, die zum Teil wenig beachtet blieben, zum Teil sich aber auch in der wissenschaftlichen Forschung oder sogar im klinischen Alltag fest etabliert haben.

Kritische Stimmen innerhalb der Medizin fragen allerdings auch nach der Evaluation solcher Scores. Dies ist eine Aufforderung an das Fach Medizinische Biometrie, hierfür Konzepte, Parameter, Studienpläne und statistische Auswertungsverfahren zu entwickeln und zu analysieren. Typisch für die biometrische Herangehensweise ist dabei, daß von den klinischen Anforderungen an Scores ausgegangen wird und dann statistische Parameter und Methoden ausgewählt werden, die diesen Anforderungen bestmöglichst gerecht werden. So ist es nicht notwendig, eine vollkommen neue Methodologie für klinische Scores zu entwickeln, sondern es kann großenteils auf bestehende Methoden zurückgegriffen werden, die gegebenenfalls für die neuen Anforderungen zu adaptieren sind.

Die Methoden zur Evaluierung klinischer Scores sind in unterschiedlichen Bereichen zu suchen, weil der Entwicklung von Scores durchaus unterschiedliche Zielsetzungen zugrunde liegen. Diese Zielsetzungen können klinischer oder aber wissenschaftlicher Natur sein und sie können im Bereich der Diagnose und Prognose, aber auch in der Therapieerfolgsbeurteilung liegen. Je nach Zielsetzung ergeben sich Berührungspunkte mit anderen verwandten Bereichen, etwa mit der Psychometrie als Teilgebiet der Psychologie, mit der Qualitäts-kontrolle im industriellen Bereich und mit der Entscheidungsunterstützung als Teilgebiet der Medizinischen Informatik. Es kann daher hilfreich sein, auch in diese Bereiche außerhalb der Medizinischen Biometrie zu schauen und dortige methodische Entwicklungen aufzugreifen.

Wenn auch die Zahl der Publikationen, in denen Scores vorgestellt, überprüft oder einfach nur verwandt werden, unüberschaubar ist, so sind Publikationen zur Methodologie der Konstruktion und Evaluierung von Scores erst in den letzten fünf bis zehn Jahren in größerer Zahl erschienen und es gibt kaum zusammenfassende Darstellungen in Lehrbüchern. Vor etwa fünf Jahren versuchte Feinstein mit seinem gleichnamigen Buch [1] den Begriff "Clinimetrics" für das Gebiet des Messens in der klinischen Medizin zu etablieren. Stärkere

Aufmerksamkeit als von der klassischen, statistisch ausgerichteten Biometrie erfuhr die Problematik des Messens in der klinischen Medizin von amerikanischen Vertretern der sogenannten "Clinical Epidemiology", die auch Feinstein wesentlich mitgeprägt hat. Da der Begriff Klinische Epidemiologie für die Methodologie der klinischen Forschung in Deutschland jedoch noch nicht etabliert ist [2], befaßt sich hier die Medizinische Biometrie mit diesen Fragestellungen.

1.2 Ziel und Gliederung der Arbeit

Diese Arbeit hat zum Ziel, den aktuellen Stand der Forschung zur Konstruktion und Evaluierung klinischer Scores darzustellen. Dabei sollen die Grundkonzepte der Evaluation von Scores erklärt und die hierbei verwendeten statistischen Parameter und Verfahren verständlich gemacht werden. Besonderes Gewicht findet die Darstellung, Einordnung und kritische Diskussion einiger neuerer Konzepte, die in den letzten Jahren entwickelt und propagiert wurden. Eine wichtige Rolle werden Methoden zur Konstruktion von Scores sowie Prinzipien für die Planung und Durchführung von Studien zur Evaluierung von Scores spielen. Wegen der Breite des Gebietes und im Hinblick auf eine durchgängige Lesbarkeit der Arbeit wird auf mathematische Details weitgehend verzichtet und stattdessen auf die einschlägige Literatur verwiesen. Besondere Berücksichtigung finden dabei aktuellere Arbeiten aus den letzten Jahren.

Das folgende Kapitel 2 gibt einen Überblick über die Vielfalt klinischer Scores und stellt einige Klassifikationsprinzipien vor. Besondere Berücksichtigung findet dabei die Frage, welche Bedeutung Scores im Bereich der klinischen Medizin haben. In den anschließenden Kapiteln erfolgt eine ausführliche Darstellung von Gütekriterien für Scores, wobei bewußt die Parallelität zu der Evaluation klassischer (physikalischer oder chemischer) Meßverfahren betont wird. Zunächst wird zur besseren Orientierung in Kapitel 3 eine Übersicht der in der Literatur existierenden, oft uneinheitlich benannten Konzepte gegeben. Im anschließenden Kapitel 4 werden zu den verschiedenen Konzepten die Gütekriterien als statistische Parameter dargestellt und Methoden zu deren Bewertung genannt. Konkurrierende Güteparameter werden in ihren Vor- und Nachteilen erörtert. Das darauffolgende Kapitel 5 ist den Strategien und Prinzipien zur Konstruktion von Scores gewidmet. Hier wird besonderer Wert darauf gelegt, die Komplexität existierender Konstruktionsverfahren und die dabei auftretenden praktischen Probleme darzustellen, um das hohe Maß an Willkür im Rahmen der Scorekonstruktion sichtbar zu machen. Daraus leitet sich die Notwendigkeit einer sorgfältigen Planung und Durchführung von Studien zur unabhängigen Evaluierung von Scores ab, deren methodische Planungsprinzipien den Inhalt von Kapitel 6 darstellen.

Aus rein methodischem Blickwinkel könnte man sich fragen, ob eine zusammenfassende Arbeit zum Thema der Evaluation von Scores überhaupt sinnvoll ist, da doch Scores mit unterschiedlicher Zielsetzung auch unterschiedliche Güteeigenschaften erfordern. Dies führt ganz zwangsläufig in unterschiedliche Bereiche der statistischen Methodik, die wenig

miteinander zu tun haben. Ein Ziel dieser Arbeit ist es jedoch zu zeigen, daß grundsätzlich die gleichen Konzepte bei allen Scores eine Rolle spielen, wenn auch mit unterschiedlicher Gewichtung und zum Teil in unterschiedlicher Gestalt.

Ein wichtiges Anliegen des Verfassers dieser Arbeit war es, in erster Linie den "state-of-the-art" verständlich darzustellen, da dieses Wissen in der wissenschaftlichen Literatur sehr verstreut ist und nur wenige Lehrbücher, und diese zu Teilgebieten, existieren. Die mathematisch eher anspruchsvolle Darstellung statistischer Einzelaspekte ist in biometrischen und psychometrischen Fachzeitschriften zu finden, während die sachgerechte Auswahl und Anwendung der Konzepte bisher kaum Eingang in die methodische Literatur gefunden hat. Eigene methodische Beiträge sind in diese Arbeit mit aufgenommen und führten dazu, daß einzelne Themen mit größerer Ausführlichkeit dargestellt sind. Zudem enthält die Arbeit Beispiele aus eigenen Kooperationen im Rahmen klinischer Projekte, bei denen es um die Konstruktion und Evaluierung von Scores ging. Aus diesen Kooperationen sind viele Anregungen für die vorliegende Arbeit entstanden.

2 ÜBERBLICK ÜBER SCORES

2.1 Definition klinischer Scores

Wenn man den englischen Begriff "score" auch im Deutschen benutzt und, wie durch den Titel dieser Schrift belegt, bevorzugt, so ist ein kurzer etymologischer Exkurs angebracht. Laut Webster´s Dictionary [3] leitet sich von der ursprünglichen Bedeutung des Wortes ("notch, tally" = Zählholz, Kerbholz) die Bedeutung "an account or reckoning, originally kept by making marks on a tally" ab, diese wurde später erweitert zu "a number that expresses accomplishment (as in a game or test) or excellence (as in quality) either absolutely in points gained or by comparison to a standard". Ähnlich definiert es der Pschyrembel [4]: "Score (ebgl. Punktzahl): zahlenmäßiger Ausdruck einer Leistung; Maßzahl, Wertpunkt".

Wenn wir im folgenden von einem Score sprechen, so ist nicht, wie in den obigen Definitionen, die Maßzahl gemeint, sondern eine Regel, nach der diese Maßzahl ermittelt wird, mithin eine Meßvorschrift im weitesten Sinne. Die Maßzahl selbst werden wir als Scorewert bezeichnen.

Definition: Ein klinischer Score ist eine Meßvorschrift für ein Merkmal aus dem Bereich der klinischen Medizin, die aus einer genau operationalisierten Zusammenfassung von mehreren (i.d.R. mehr als zwei) Komponenten besteht und zu einer mindestens ordinalen Bewertung führt.

Weitgehend synonym werden die Begriffe klinische Skala und klinischer Index verwendet. Eine physikalisch motivierte Zusammenfassung von zwei oder drei Meßgrößen in einer Formel kann als Grenzfall eines Scores betrachtet werden und wird meist als Index bezeichnet, wie zum Beispiel verschiedene Indexe für Übergewicht (Broca, Quetelet). Einen weiteren Grenzfall stellen Stadieneinteilungen oder Gradingsysteme dar, diese basieren in der Regel auch auf der Zusammenfassung mehrerer Komponenten, haben jedoch nur wenige (meist drei bis fünf) ordinale Ausprägungen.

Typisch für die meisten Scores ist, daß das klinische Merkmal komplexer Natur und a-priori nicht eindimensional gegeben ist. Oft spricht man daher von multidimensionalen Konstrukten anstatt von Merkmalen. Die Scorekomponenten können, je nach Zielsetzung des Scores, mehr oder weniger eng verwandte Merkmale sein. Bei der klassischen, an die sprachliche Bedeutung des Wortes angelehnten Definiton eines Scores werden mehrere Merkmale, für deren Ausprägungen bestimmte (i.d.R. ganzzahlige) Punkte vergeben werden, additiv in einem Gesamtscore zusammengefaßt. Oft werden aber auch andere Kombinationen, die zum Teil einfacher, zum Teil mathematisch komplizierter sind, als Score bezeichnet. Der Begriff des klinischen Index umfaßt nach Feinstein [1] auch nicht-numerische Einstufungen, obwohl

diese Abgrenzung wegen der Möglichkeit der nachträglichen zahlenmäßigen Verschlüsselung unwesentlich ist.

2.2 Beispiele klinischer Scores

In diesem Abschnitt wollen wir einige, vor allem bekannte Beispiele klinischer Scores in Erinnerung rufen, um die ganze Spannbreite der Anwendung aufzuzeigen. Die Beispiele werden kurz beschrieben, da zum Teil später auf sie zurückgekommen wird.

Einer der ältesten Scores vom klassischen Typ ist der nach Virginia Apgar benannte *Apgar-Score* [5], den sie 1953 für die Beurteilung des Zustands Neugeborener einführte. Er wird berechnet, indem fünf Merkmale des Neugeborenen, nämlich Herzschlag, Atmung, Muskeltonus, Reflexe und Hautfarbe, nach einem Schema mit 0, 1 oder 2 Punkten bewertet werden und die Summe (entsprechend einem Wert zwischen 0 und 10) gebildet wird. Die einzelnen Scorekomponenten basieren vor allem auf klinischen Eindrücken. Der Score soll eine Minute nach Geburt des Kindes bestimmt werden sowie nochmals nach 5 Minuten. Seiner Einfachheit verdankt es der Score vermutlich, daß er bis heute weltweit in der Neonatologie benutzt wird, obwohl immer wieder Zweifel an seiner Brauchbarkeit geäußert wurden [6] und auch Verbesserungen und Alternativen vorgeschlagen wurden.

Noch älter ist die *Karnofsky-Skala* [7] zur Beschreibung des Allgemeinzustandes von Patienten, die vor allem in der Onkologie häufig eingesetzt wird. Bei diesem Score geschieht die Zuordnung eines Gesamtwertes aus der Information der einzelnen Komponenten nicht arithmetisch, sondern sie wird umgangssprachlich beschrieben. Die Skala umfaßt den Wertebereich von 0 bis 100 in Zehnerschritten, Zwischenwerte werden üblicherweise nicht vergeben. Auch die Karnofsky-Skala basiert auf einer subjektiven Fremdeinschätzung, sie hat sich aber dennoch in der Onkologie als sehr aussagefähig erwiesen und wird sowohl in Prognosestudien als auch in Therapiestudien häufig verwendet.

Das dritte Beispiel ist knapp 20 Jahre alt und stammt aus dem Bereich der Gastroenterologie. Von Best et al. [8] wurde 1976 der *CDAI* (= Crohn´s Disease Activity Index) publiziert, der gelegentlich auch als *Best-Index* bezeichnet wird. Er dient zur Beschreibung der Krankheitsschwere bei Patienten mit Morbus Crohn. Dieser Score ist in seiner Berechnung etwas komplexer als die obigen Beispiele, da die einbezogenen Komponenten unterschiedlicher Natur sind und auch ganz unterschiedlich gewichtet werden. Es werden als Komponenten sowohl völlig subjektive Merkmale (allgemeines Befinden, Schmerzen) als auch objektive Daten (Hämatokritwert, Körpergewicht) einbezogen, einige davon sind zeitlich variabel (Befinden), andere eher konstant (Gewicht). Trotz der inhaltlichen Heterogenität und der etwas umständlichen Erhebung wird der CDAI bis heute viel benutzt, insbesondere in wissenschaftlichen Studien zum Morbus Crohn, wo er zur Operationalisierung sowohl des Behandlungserfolges als auch der Therapieindikation herangezogen wird.

Als viertes Beispiel soll der *APACHE-Score* (=Acute Physiology and Chronic Health Evaluation) vorgestellt werden, der vermutlich der am besten beforschte und am häufigsten eingesetzte Score überhaupt ist. Der APACHE-Score [9] ist auch deshalb interessant, weil er seit seiner ersten Veröffentlichung im Jahre 1981 mehrfach mit großem Aufwand verbessert wurde, was zum APACHE II [10] im Jahre 1985 und schließlich 1991 zu APACHE III [11] führte. Der APACHE-Score beschreibt die Krankheitsschwere von Intensivpatienten und erlaubt eine Abschätzung des Sterberisikos. Die erste Version des Scores basierte auf 34 physiologischen Parametern, deren Ausprägungen mit null bis vier Punkten bewertet und aufaddiert wurden. Die Auswahl der Variablen und die Vergabe der Punktwerte wurden im Rahmen eines Konsensfindungsprozesses von sieben erfahrenen Intensivmedizinern festgelegt. Mit APACHE II ist es gelungen, eine wesentlichen Vereinfachung (Reduktion der Parameterzahl auf 12) ohne Informationsverlust zu erzielen. Die Veränderungen gegenüber der ersten Version basierten auf klinischen Erfahrungen, ihr Effekt wurde jedoch an über 5000 Patienten statistisch überprüft. APACHE II war jahrelang der am weitesten verbreitete Score im Bereich der Intensivmedizin, vermutlich ist er es immer noch. Die Überarbeitung zum APACHE III basierte auf der statistischen Analyse einer Studie an über 17.000 Patienten, sie führte zu einer Verbesserung des Scores zu Lasten seiner Handhabbarkeit. Auf die unterschiedlichen Anwendungen dieses Scores wird noch eingegangen.

Als fünftes und letztes Beispiel eines klinischen Scores dient eine Selbstbeurteilungsskala zur Messung der Lebensqualität bei Krebspatienten. Cella et al. [12] erarbeiteten einen 33 Fragen umfassenden Bogen (FACT = Functional Assessment of Cancer Therapy), der das Befinden der Patienten erfaßt und sowohl durch einen Gesamtscore als auch durch Teilscores für einzelne Bereiche bzw. Dimensionen der Lebensqualität ausdrückt. Die Entwicklung und Evaluierung des Bogens erfolgte in mehreren Phasen an über 800 Patienten über einen Zeitraum von etwa fünf Jahren. Der Score ist in erster Linie für die Anwendung im Rahmen klinischer Studien in der Onkologie konzipiert.

Die genauen Definitionen zur Berechnung der vier erstgenannten Scores sowie eine umfangreiche Übersicht über weitere gebräuchliche Scores werden im Buch von Gerber & Wicki [13] gegeben. Eine Darstellung der bekanntesten Skalen aus dem Bereich der allgemeinen Gesundheitsindikatoren, einschließlich Lebensqualität, ist bei Bowling [14] zu finden.

2.3 Einteilungsprinzipien

Aus der obigen Beschreibung einiger weit verbreiteter Scores ist bereits deutlich geworden, daß diese sowohl formal (strukturell) als auch bezüglich ihres Einsatzes (funktional) sehr heterogen sind. Für das folgende erscheint es nützlich, ein möglichst einfaches Beschreibungssystem für Scores und Skalen zu erstellen, um bei der Darstellung von Kriterien und Methoden der Scoreevaluation auch der Vielfalt existierender und prinzipiell möglicher Scores gerecht zu werden. Ein solches System sollte als Beschreibungsdimensionen nur jene

Aspekte enthalten, die für die weiteren Ausführungen relevant sind, indem sie für differenzierte, aber verallgemeinerbare Aussagen benötigt werden. Die von Feinstein [1] vorgeschlagenen Dimensionen für eine "pragmatische Taxonomie" von klinischen Indexen sind hierfür nicht geeignet, sie gehen eher von anatomischen und pathologischen Aspekten aus. Für die methodische Betrachtungsweise ist jedoch vielmehr von Bedeutung, welchem Zweck der Score dienen soll und wie er erhoben wird, zudem spielt die Struktur des Scores eine Rolle. Im folgenden wird daher unterschieden zwischen der Ebene des Input, also der Datenerhebung für den Score, der Ebene der Struktur der Scoredefiniton sowie der Ebene der Anwendung des Scores.

2.3.1 Einteilung nach Eingangsinformation

In Bezug auf die Komponenten ist zu fragen, welche Informationen in den Score eingehen und wie diese erhoben werden. Wieviele getrennt definierte Merkmale werden erfaßt und auf welche Art geschieht dies, d.h. welches sind die Informationsquellen? Der wichtigste Aspekt ist hierbei, ob es sich um sogenannte weiche oder harte Daten handelt, also etwa um Selbst- oder Fremdbeurteilungen des Befindens oder Verhaltens des Patienten oder aber um physikalische oder laborchemische Meßwerte. Dies ist von Bedeutung im Hinblick auf die Beurteilung der Reproduzierbarkeit eines Scores. Weitere wichtige Aspekte der Eingangsinformation sind der zeitliche Bezug der Datenerhebung (z.B. erlaubtes Zeitfenster) sowie die Randbedingungen der Erhebung, da die Reproduzierbarkeit von Scores auch von der zeitlichen und situativen Stabilität der Eingangsinformation abhängt.

Die Frage, ob und wie sich die Reproduzierbarkeit prinzipiell überprüfen läßt, hängt auch davon ab, wieweit die Eingangsinformation "konservierbar" ist, also für eine unabhängige zweite Beurteilung zur Verfügung gestellt werden kann. Dieses gilt zum Beispiel, wenn die Bestimmung eines Merkmals auf der Untersuchung von Material, also etwa einer Blutprobe oder einer Tumorbiopsie, beruht. Ähnlich verhält es sich mit Beurteilungen von bildgebenden Verfahren, z.B. einer CT-Aufnahme. Einige klinische Scores sind so konstruiert, daß zu ihrer Erhebung nur die üblicherweise in der Krankenakte vorhandenen Informationen benötigt werden, dies erleichtert Untersuchungen zur reproduzierbarkeit erheblich. Bezieht sich die Eingangsinformation auf das Verhalten oder das Befinden des Patienten, so ist eine Konservierung nur in manchen Fällen möglich, indem dies genau protokolliert, oder besser noch in Ton- oder Videoaufnahmen festgehalten wird. Eine wiederholte Fremdbeurteilung der so konservierten Eingangsinformation durch den Arzt ist dann durchführbar. Viele Merkmale des subjektiven Befindens sind jedoch nur im Rahmen einer Selbstbeurteilung durch den Patienten erschließbar, in diesem Fall ist die Eingangsinformation nicht konservierbar.

Für die Beurteilung der Praktikabilität eines Scores ist der Aufwand relevant, der mit der Erhebung der Eingangsinformationen verbunden ist. Hier ist zu unterscheiden, ob die Informationen aus der Anamnese und aus diagnostischen Standardverfahren unmittelbar entnommen werden können (z.B. Alter, Geschlecht, Pulsfrequenz, Symptomdauer) oder ob sie

einen besonderen personellen oder apparativen Aufwand erfordern (z.B. Interview durch Psychologe, Magnetresonanz-Tomographie). Die Frage, ob die Eingangsinformationen routinemäßig erhoben werden, ist auch von Bedeutung in Bezug auf die Möglichkeit, einen Score retrospektiv aus den Angaben der Krankenakte zu berechnen, was die Durchführbarkeit von Evaluationsstudien zur Validierung erheblich vereinfacht. Basiert der Score auf einem Patientenfragebogen, so ist eine Evaluierung nur in prospektiven Studien möglich. Auf die Problematik retrospektiver Studien wird am Ende der Arbeit ausführlich eingegangen.

2.3.2 Einteilung nach strukturellen Eigenschaften

Bei der Einteilung nach strukturellen Eigenschaften kann man dreierlei Dinge unterscheiden: erstens die Struktur der Komponenten, also die Skalierung der Eingangsinformation, zweitens die Art der Verknüpfung der Eingangsdaten zum Scorewert, und drittens die Skalierung des Gesamtscores.

Schon bei der Kodierung oder allgemein Bewertung der einzelnen Eingangsvariablen gibt es zahlreiche Möglichkeiten. So kann bei quantitativen Merkmalen die Eingangsinformation erhalten bleiben oder aber dichotom oder mehrstufig klassiert und mit Punkten (z.B. APACHE [10]) versehen werden. In manchen Fällen wird kategorielle, qualitative Information mit Punkten versehen und so auf ein pseudo-quantitatives Niveau angehoben (z.B. Infarktlokalisation bei Norris et al. [15]). Diese Skalierung geschieht auf der Basis von medizinischer Erfahrung oder direkt datengestützt. Ist die Kodierung für die einzelnen Scorekomponenten sehr unterschiedlich, wie etwa beim Best-Index [8], so kann dies die Handhabung des Scores beeinträchtigen und ihn fehleranfällig machen.

Für die Verknüpfung der Eingangsinformation ist der typische Fall sicherlich die Berechnung als gewichtete Summe, indem die Ausprägungen der Eingangsvariablen gemäß ihrer Bedeutung bepunktet werden und anschließend zum Gesamt-Scorewert aufaddiert werden. Jedoch sind auch komplexere arithmetische Formen gebräuchlich, wo zum Beispiel erst eine Berechnung gemäß eines mathematischen Modells erfolgt und dann der Endwert diskretisiert wird. Schließlich können Scores entgegen ihrer etymologischen Herkunft grundsätzlich auch nicht-ganzzahlige Werte annehmen. Neben der arithmetischen ist vor allem die logische Verknüpfung häufig zu finden, die durch und- bzw. oder-Kombinationen der Ausprägungen der Eingangsvariablen charakterisiert ist. Als Beispiel hierfür können viele Stadieneinteilungen in der Onkologie gesehen werden. Ein Spezialfall ist die hierarchische Verknüpfung in Form von Baumdiagrammen, etwa nach der CART-Methode [16], die den Vorteil einer einfachen Handhabung haben. In manchen Fällen stellt sich die Operationalisierung eines Scores noch komplexer dar, weil zunächst aus den Eingangsvariablen Teilscores gebildet werden (auch Achsen des multidimensionalen Scores, oder Subskalen), die dann zum Gesamtscore zusammengefaßt werden.

Es war bereits erwähnt worden, daß der Gesamt-Scorewert grundsätzlich sowohl ordinal als auch ganzzahlig oder sogar reellwertig sein kann. Eine weitere Unterscheidung bei numerischen Scores ergibt sich aus der Frage nach einer Standardisierung. So kann etwa eine Verankerung durch Festlegung des minimalen Scorewertes bei 0 Punkten und des maximalen bei 10 oder 100 Punkten vorgenommen werden, wie etwa bei der Karnofsky-Skala. Dies ist für jeden Score durch die nachträgliche Transformation

$$Y = \frac{100}{X_{max} - X_{min}} \cdot (X - X_{min})$$

möglich. Eine andere Form der Standardisierung orientiert sich an Referenzwerten, die an definierten Referenzpopulationen gewonnen wurden. Dies ist bei vielen psychometrischen Meßinstrumenten üblich, da auf diese Weise alters- und geschlechtsbezogene Scorewerte definiert werden können. Sind der Mittelwert und die Standardabweichung einer Skala in einer Referenzpopulation bekannt, so kann man eine Standardisierung auf sogenannte z-Werte oder T-Werte mittels folgender Transformationen erreichen:

$$z = \frac{X - \bar{X}}{s} \quad \text{bzw.} \quad T = \frac{X - \bar{X}}{s} \cdot 10 + 50 \ .$$

Dieses Vorgehen ist bei solchen Skalen üblich, die auch in Referenzkollektiven von Gesunden anwendbar sind. Für viele klinische Scores ist dies aus ethischen oder praktischen Gründen nicht möglich.

Eine Standardisierung anderer Art betrifft Prognoseskalen, die zur Vorhersage des Auftretens eines klinischen Ereignisses dienen sollen. Hier wird gelegentlich direkt aus dem Scorewert die Auftretenswahrscheinlichkeit des Ereignisses berechnet, wie etwa beim APACHE-Score und anderen Beispielen [10, 17], statt von Standardisierung spricht man in diesem Fall meist von Kalibration.

Die Festlegung der strukturellen Eigenschaften eines Scores liegt meist im Entscheidungsbereich des Autors. Hier kann neben statistischen Argumenten auch die Praktikabilität eine Rolle spielen. Oftmals werden dabei aber die Entscheidungen überwiegend willkürlich gefällt. Insbesondere die Kodierung bzw. Klassierung der Eingangsvariablen und die Auswahl des statistischen Modells hängt oft von den Kenntnissen und Vorlieben des Autors und der Verfügbarkeit von Software ab.

2.3.3 Einteilung nach Funktion

Eine Einteilung von Scores nach funktionellen Eigenschaften erweist sich als besonders relevant und diese Einteilung bestimmt nachhaltig die Güteanforderungen, die an den Score gestellt werden. Diese Sichtweise wurde besonders von Kirshner & Guyatt [18] betont, die von "three purposes of health status measures" sprechen. Sie unterscheiden die

Diskrimination, die Prädiktion und die Evaluation. Als diskriminative Indexe bezeichnen sie solche, die zur Differenzierung von Individuen oder Gruppen bezüglich eines Merkmals dienen, für das es kein klares Referenzkriterium gibt, hierzu zählen sie die Intelligenz oder andere Persönlichkeitseigenschaften. Von einem prädiktiven Index sprechen die Autoren, wenn er zur Vorhersage eines gleichzeitigen oder späteren Außenkriteriums oder Referenzmerkmals dient. Ein evaluativer Index hat nach Kirshner & Guyatt das Ziel, Veränderungen von Individuen oder Gruppen in Längsschnittuntersuchungen zu messen.

Kirshner & Guyatt versuchen in ihrer Arbeit zu zeigen, daß die von ihnen gemachte Unterscheidung einen Einfluß auf fast alle Aspekte der Konstruktion und Evaluation von Skalen hat. Dies ist nicht immer überzeugend, da eine Abgrenzung zwischen diskriminativen und evaluativen Scores schwierig ist. Eine Unterscheidung ließe sich so treffen, daß sich diskriminative Scores eher auf konstante, evaluative Scores hingegen auf veränderliche Merkmale beziehen. Eine Kritik des Konzepts von Kirshner & Guyatt haben Williams & Naylor [19] formuliert. Dabei betonen sie, daß eine derart starre Einteilung weder realistisch noch hilfreich ist, da viele Scores (bzw. bei ihnen Gesundheitsmaße) nicht nur in einer Funktion eingesetzt werden. Ein gutes Beispiel hierfür ist die Karnofsky-Skala, die in der Onkologie sowohl zur Therapieevaluation herangezogen wird als auch ein wichtiges Prognoseinstrument darstellt [20].

Auch andere Autoren haben Einteilungen zur Funktion klinischer Meßinstrumente vorgeschlagen. Feinstein [1] unterscheidet die Zustandsbeschreibung ("status index"), die Verlaufsbeschreibung ("change index"), Verlaufsvorhersage ("prognostic index") und Entscheidungshilfe ("clinical guideline"). Während sich die ersten drei Klassen weitgehend mit der Einteilung von Kirshner & Guyatt decken, fällt die vierte aus dem Schema heraus. Zwar ist der Aspekt, ob ein klinischer Score direkt als Grundlage für eine ärztliche Handlung dient, sei es diagnostischer oder therapeutischer Art, von großer Bedeutung, insbesondere in Bezug auf die Möglichkeit der Evaluation des Nutzens. Jedoch ist dies eine unabhängige Dimension, die zu der anderen Einteilungsdimension hinzukommt. Eine weitere Einteilung, die hier nicht im Detail vorgestellt wird, findet sich bei Kane & Kane [21].

Schaut man sich in der medizinischen Fachliteratur um, in welcher Funktion Skalen oder Scores rein quantitativ am häufigsten verwendet werden, so dominieren ganz klar die Bereiche Prognose ("predictive score") und Evaluation ("outcome scale"). In der vorliegenden Arbeit werden daher vorwiegend die Begriffe prädiktiver bzw, evaluativer Score verwendet, wobei dies keine starren Attribute des Scores sind, sondern Bezeichnungen einer Funktion des Scores.

2.4 Einsatz von Scores

Die fast unüberschaubare Zahl von Scores, die in der wissenschaftlichen Literatur der letzten Dekaden präsentiert wurde, führt zwangsläufig zu der kritischen Frage, ob denn überhaupt ein

entsprechender Bedarf an klinischen Scores vorhanden ist und wo dieser liegt. Der mögliche Nutzen eines neuen Scores liegt sicherlich nicht so klar auf der Hand wie der eines neuen Medikaments. So wie man für jede neue Therapie einen Nachweis der Wirksamkeit bzw. sogar der therapeutischen Überlegenheit gegen den besten verfügbaren Standard nachweisen muß, so sollte auch bei neuen Scores eine Beweislast für einen sinnvollen Einsatz bestehen.

Die möglichen Anwendungen von klinischen Scores sind nun allerdings vielfältiger, sie umfassen insbesondere die Verwendung im klinischen wie im wissenschaftlichen Kontext. Im Bereich der klinischen Routine sind die Bereiche der Information und Beratung des Patienten, der Diagnose- und Prognosestellung sowie der Therapieplanung und -kontrolle zu betrachten, im wissenschaftlichen Kontext die Durchführung von Studien unterschiedlicher Zielsetzungen. Eine ausführliche Darstellung über Einsatzmöglichkeiten, mit Beispielen vor allem aus dem Bereich der Unfallchirurgie und Intensivmedizin, geben Neugebauer & Bouillon [22], sie erwähnen unter anderem auch den gesundheitsökonomischen Bereich, der aktuell an Bedeutung gewinnt.

2.4.1 Scores im Rahmen der Patientenberatung

Die Information und Beratung des Patienten ist eine ganz wesentliche Komponente der ärztlichen Tätigkeit, die gelegentlich vernachlässigt wird. Dabei spielen sicher auch Kommunikationsprobleme eine Rolle, die zwischen dem Patienten, der in der Regel ein medizinischer Laie ist, und seinem Arzt bestehen. Wenn man über den Wert von klinischen Scores in diesem Zusammenhang nachdenkt, so ist hier allerdings nicht vorrangig gemeint, daß der Patient mit Zahlen statt mit Worten informiert werden soll. Vielmehr bieten Scores als zusammenfassende quantitative Merkmale dem Arzt eine einfache Möglichkeit, dem Patienten relevante, ihn interessierende Informationen zu vermitteln, sei es als Zahl oder übersetzt in Worte. Im folgenden wird dies weiter differenziert und an Beispielen erläutert.

Scores können der Zustandsbeschreibung dienen, aus ihren Zahlenwerten lassen sich in erster Linie vergleichende Einordnungen ableiten. Sofern es Normwerte bzw. Referenzbereiche für den Score gibt, ist eine relative Einordnung des Patienten in Bezug auf eine Vergleichspopulation möglich. Nun ist diese Information selbstverständlich noch nicht in praktische Konsequenzen umsetzbar, jedoch sollte man nicht übersehen, daß sie einem anscheinend menschlichen Grundbedürfnis nach Vergleich nachkommt. Ein solcher Nutzen ist noch offensichtlicher, wenn über die Betrachtung der Scorewerte im Behandlungsverlauf die Möglichkeit eröffnet wird, dem Patienten mitzuteilen, ob und in welchem Ausmaß eine Verbesserung seines Zustands eingetreten ist. Dies wird vor allem bei solchen Erkrankungen relevant sein, deren Ausprägungen mehrere Bereiche umfaßt, von denen nicht alle für den Patienten symptomatisch erfaßbar sind. Als Beispiel sei der Morbus Crohn genannt, wo ein häufig angewandter Score (CDAI [8]) sowohl subjektive Parameter als auch Laborwerte beinhaltet. Die häufig geäußerte Kritik, daß ein Score-Summenwert die einzelnen Dimensionen verschleiere [23], ist hier nicht angebracht. Der Arzt hat sehr wohl die

Möglichkeit, dem Patienten Einzeldaten zu vermitteln, sofern diese zusätzliche Information tragen, doch der Score ermöglicht darüberhinaus die quantitative Integration der Einzelbefunde.

Bei prädiktiven Scores ist der Wert im Rahmen der Information und Beratung der Patienten noch wesentlich offensichtlicher, denn sie erlauben nicht nur vergleichende, sondern auch absolute Aussagen im Hinblick auf ein für den Patienten relevantes und vorstellbares Kriterium. Zwar sind diese Aussagen meist stochastischer Natur, also Wahrscheinlichkeitsaussagen über Patienten als Repräsentanten einer bestimmten Population, aber dennoch sind sie leicht verständlich zu machen. Prognostische Aussagen über den möglichen Krankheitsverlauf haben für den Patienten psychologische Bedeutung, sie können Hoffnung vermitteln oder aber Vorsicht gebieten, somit steuern sie wesentlich die Krankheitsverarbeitung. Ganz entscheidend beeinflussen Informationen von Prognosescores aber auch die gesamte Lebensplanung, und damit konkrete Entscheidungen des Patienten. Dieser Aspekt hat selbstverständlich schon aus ethischen Gründen eine unbestrittene Bedeutung. Da er jedoch auch materielle Konsequenzen beinhaltet, kann er sogar juristisch relevant werden. Hierzu gibt es bereits Beispiele, wie etwa der von Annas [24] dargestellte Fall. Darin kam es zu einem Prozeß, weil einem Patienten und den Angehörigen die schlechte Krankheitsprognose nicht mitgeteilt worden war, was unter anderem dazu führte, daß der Patient seine finanziellen Angelegenheiten nicht mehr rechtzeitig vor seinem Tode regeln konnte. Es ist sicher noch zu klären, mit welcher Genauigkeit und auf welche Art einem Patienten seine statistische Lebenserwartung auf Wunsch mitgeteilt werden sollte. Jedoch ist es offensichtlich, daß hierbei prädiktive Scores eine wesentliche Rolle spielen, da sie darauf abzielen, alle prognostisch relevanten Merkmale eines Patienten zu berücksichtigen und damit eine zwar statistische, aber doch möglichst individuelle Aussage zu machen.

2.4.2 Scores als Grundlage für Therapieentscheidungen

Während im vorangegangenen Abschnitt vor allem die Perspektive des Patienten eingenommen wurde, soll in diesem Abschnitt beleuchtet werden, welche Rolle Scores bei der Therapieentscheidung oder Therapieempfehlung des Arztes spielen können. Hier ist zwischen allgemeinen Therapiestrategien und Therapieentscheidungen im individuellen Fall zu unterscheiden.

Allgemeine therapeutische Strategien basieren üblicherweise sowohl auf der Diagnose als auch auf prognostisch relevanten Daten. Die Bedeutung von Scores in diesem Bereich ergibt sich aus der Tatsache, daß sie zum Teil direkt zur Diagnose dienen oder aber diese durch Angabe des Schweregrades der Erkrankung spezifizieren.

> Celani et al. [25] vergleichen in ihrer Untersuchung zwei Scores, die zur Differentialdiagnose zwischen hämorrhagischem und ischämischen Schlaganfall dienen

können. Ein solcher Score ist von Bedeutung für eine schnelle Therapieentscheidung, falls kein Computertomogramm verfügbar ist.

Ein anderes Beispiel stellt der Best-Index [8] für Morbus Crohn dar. Hier gibt es eine weithin akzeptierte Regel, daß bei einem Scorewert über 150 von einem Krankheitsschub gesprochen wird, bei dessen Vorliegen eine Indikation für eine besondere Therapie gesehen wird. Diese Regel wird vor allem im Rahmen von Therapiestudien verwendet, um eine Standardisierung der Behandlung zu erreichen.

Besonders betont werden soll hier der Fall prognoseabhängiger Therapiestrategien. Hierbei erfolgt die Wahl zwischen mehreren Behandlungsalternativen auf der Basis von prognostischer Information, sei es in Form eines Scores oder einer Stadieneinteilung. Die Zielsetzung dabei ist eine differenzierte Therapie unter Abwägung von Nutzen und Risiken der Behandlungsalternativen. An einem Beispiel soll dies veranschaulicht werden:

In der Chemotherapie-Behandlung von Patienten mit hochmalignem Non-Hodgkin-Lymphoms hat man relativ gute Erfolge erzielt. Bei einem überwiegenden Teil der Patienten erreicht man mit den heute angewandten Cytostatikakombinationen über mehrere Jahre anhaltende Remissionen, während ein kleinerer Teil nicht anspricht oder frühzeitig rezidiviert. Es gab und gibt daher zahlreiche Bemühungen, diese Teilpopulation von Patienten mit schlechtem Verlauf durch ein Prognosemodell zu charakterisieren, um ihnen eine intensivere Behandlung, zum Beispiel eine Hochdosis-Chemotherapie unter autologer Knochenmarkstransplantation, anbieten zu können [26-28].

Die am Beispiel beschriebene Situation läßt sich wie folgt verallgemeinern: Bei einer einheitlich nach bestem Standard behandelten Patientengruppe gibt es Patienten, bei denen man aus retrospektiver Sicht mit dem erreichten Krankheitsverlauf nicht zufrieden ist und für die man grundsätzlich eine Therapiealternative zur Hand hat. Hat man nun einen prädiktiven Score zur Verfügung, mit dem eine statistische Vorhersage des Krankheitsverlaufs möglich ist, so wird man eine differenzierte Behandlungsstrategie anwenden, die in Abhängigkeit vom Score einem Teil der Patienten weiterhin die Standardtherapie, einem anderen Teil jedoch die alternative Therapie zukommen läßt.

In ähnlicher Weise können Prognosescores dazu führen, diagnostische Strategien zu verändern. Auch dies kann für die betroffenen Patienten einen spürbaren Nutzen bedeuten, wie etwa im folgenden Beispiel.

Bei Patientinnen mit primärem Mammakarzinom stellt die Lymphknotendissektion eine relativ belastende diagnostische Maßnahme dar, deren Ergebnis allerdings die Wahl der weiteren Therapie maßgeblich beeinflußt. Verschiedene Forschergruppen versuchen derzeit, solche Patientinnen mittels eines prognostischen Scores zu beschreiben, bei denen auf die Lymphknotendissektion verzichtet werden kann, weil diese Maßnahme in

der entsprechenden Subgruppe weder prognostische Aussagekraft noch therapeutische Bedeutung hat [29].

Auch auf der Basis evaluativer Scores lassen sich in ähnlicher Weise Therapiestrategien formulieren, wenn man sich auf therapeutische Entscheidungen im Verlaufe der Behandlung bezieht. Oft wird die Weiterbehandlung eines Patienten vom bisherigen "Ansprechen" abhängig gemacht, und dieses Ansprechen kann mit Hilfe eines evaluativen Scores besonders gut operationalisiert werden.

Die Art und Weise, wie ein Score nach dem bisher Dargestellten eine allgemeine Therapiestrategie definieren kann, läßt sich natürlich auch auf individuelle Therapieentscheidungen übertragen. Oft ist ein Scorewert nur ein Mosaikstein in der gesamten Informationsmenge, die bei einer ärztlichen Entscheidung berücksichtigt wird; sein Beitrag läßt sich in diesem Fall nur schwer abschätzen. Ein anderes Beispiel, bei dem Prognosescores eine Rolle für individuelle Therapieentscheidungen spielen, ist das Triage-Problem, das sich aufgrund der beschränkten Bettenkapazität besonders auf der Intensivstation häufig stellt [30].

2.4.3 Scores als Zielvariable in Studien

Im Rahmen ihrer wissenschaftlichen Verwendung sind Scores weitaus mehr akzeptiert als im klinischen Routineeinsatz. Insbesondere in vergleichenden Therapiestudien wird die Notwendigkeit gesehen, ein quantitatives, objektivierbares und globales Zielkriterium zu haben, an dem die Aussage zum therapeutischen Vergleich letztlich festgemacht werden kann. Ein globales Zielkriterium hat den Vorteil, daß eine prospektiv formulierbare Entscheidungsregel leicht in Form eines einfachen statistischen Signifikanztests festgelegt werden kann, die den zentralen Bezugspunkt fast jeder randomisierten Studie darstellt. Scores sind als globale Kriterien besonders geeignet, da sie mehrere Komponenten zusammenfassen. Natürlich werden Scores häufig auch als Nebenzielgröße in Studien eingesetzt. Eine randomisierte Studie zur medikamentösen Behandlung der rheumatoiden Arthritis [31] stellt ein Beispiel dafür dar, wie Scores in relevanten, aber schwierig meßbaren Krankheitsdimensionen als Zielkriterien dienen können.

Die Objektivierbarkeit ist wichtig, um dem kritischen Leser einer Studie die Bedeutung des gefundenen Therapieunterschieds überhaupt verständlich zu machen. Außerdem ist sie bei multizentrischen Studien aus Gründen der Standardisierung erforderlich. Besonders einleuchtend ist die Verwendung von Scores als Zielgrößen bei solchen Merkmalen, die sich überhaupt nicht klar definieren und damit in unterschiedlichster Weise operationalisieren lassen. Es sei hier etwa das große Gebiet der Lebensqualitätserhebung genannt, das seit etwas über zehn Jahren vor allem in der onkologischen Therapieforschung, aber auch in anderen Gebieten eine große Beachtung gefunden hat. Für die Evaluation sind in den letzten gut zehn Jahren eine Unmenge von Meßinstrumenten entstanden, meist in Form von

Patientenfragebögen, die Aspekte der Gesundheit und Lebensqualität krankheitsspezifisch oder auch allgemein erfassen. Einen guten Überblick hierüber geben die Bücher von McDowell & Newell [32] und von Bowling [14].

Ein quantitatives Kriterium hat den Vorteil einer größeren Differenzierung und kann daher eine bessere statistische Genauigkeit und damit auch kleinere erforderliche Patientenzahlen implizieren. Der letztgenannte Punkt ist allerdings nicht unumstritten und manche Autoren plädieren für möglichst einfache Zielkriterien, was jedoch nicht grundsätzlich gegen die Verwendung von Scores spricht.

Bei allen Vorbehalten, die von manchen Autoren geäußert werden, ist es doch eine Tatsache, daß sich klinische Scores als Zielkriterium für Therapiestudien in manchen Krankheitsbereichen (z.B. Rheumatologie, Psychiatrie) fest etabliert haben. So werden denn klinische Scores in den meisten Lehrbüchern über kontrollierte Therapiestudien im Rahmen der Überlegungen zur Wahl der Zielkriterien explizit erwähnt (z.B. [33-36]).

2.4.4 Scores als Kontrollvariablen in Studien

Wichtig sind Scores im Rahmen von Studien auch unter dem Aspekt der Überprüfung und Gewährleistung der Strukturgleichheit von Vergleichsgruppen. Zum einen erlauben Scores eine einfache Operationalisierung von Ein- und Ausschlußkriterien für die Aufnahme eines Patienten in die Studie. Im psychiatrischen Bereich, wo es viele Borderline-Syndrome gibt, ist dies von besonderer Bedeutung. So wird bei Therapiestudien zur Behandlung der Depression typischerweise ein Mindestwert auf einer Depressionsskala als Aufnahmekriterium definiert.

Wenn in einer empirischen Untersuchung der mögliche kausale Einfluß eines Faktors auf einen bestimmtem Zielparameter untersucht werden soll, so wird man im Rahmen einer experimentellen (d.h. hier randomisierten) oder einer Beobachtungsstudie Patientengruppen bilden, die unterschiedliche Ausprägungen dieses Faktors besitzen und wird nach Unterschieden zwischen den Gruppen in Bezug auf den Zielparameter suchen. Unterscheiden sich die Gruppen jedoch auch systematisch nach anderen, in Bezug auf die Zielvariable relevanten Merkmalen, sogenannten Confoundern, so lassen sich keine zwingenden Schlußfolgerungen zum Einfluß des untersuchten Faktors ziehen. Diese Problematik spielt überall dort eine wichtige Rolle, wo kausale Aussagen das Studienziel darstellen, also insbesondere bei klinischen Therapiestudien und bei epidemiologischen Untersuchungen zur Abklärung von Risikofaktoren.

Bei Therapiestudien kann man mit Hilfe der randomisierten Therapiezuteilung die Strukturgleichheit zumindest mit großer Wahrscheinlichkeit erreichen. In nichtrandomisierten Therapiestudien und in epidemiologischen Beobachtungsstudien ist die Frage der Strukturgleichheit oft der kritischste Punkt bei der Interpretation der Ergebnisse. Deshalb muß eine Strukturgleichheit möglichst auch in diesen Fällen schon durch Maßnahmen des Studiendesigns so weit wie möglich erreicht werden. Wo dies nicht gelingt, hat eine

entsprechende Berücksichtigung des Confounding bei der statistischen Auswertung zu erfolgen.

Existiert nun ein prädiktiver klinischer Score für die in der Studie interessierende Zielvariable, so hat dieser ja die Eigenschaft, daß er die prognostische Information der wichtigsten Merkmale prägnant zusammenfaßt, und er kann deshalb für ein besonders effizientes Matching- oder statistisches Adjustierungsverfahren benutzt werden. Im Zusammenhang mit nicht-randomisierten Therapiestudien haben Abel & Windeler [37] kürzlich ein neues Design vorgeschlagen, das ganz wesentlich von einem solchen Score Gebrauch macht. In einer eigenen Arbeit [38] wurde ein Score benutzt, um in einfacher Weise den differentiellen Effekt der Rekanalisation bei Schlaganfallpatienten unterschiedlicher Krankheitsschwere darzustellen.

Bei randomisierten Studien kann der Einsatz von Scores in dieser Weise ebenfalls sinnvoll sein, obwohl eigentlich von einer Gewährleistung der Strukturgleichheit aufgrund der experimentellen Methodik ausgegangen werden kann. Diese Strukturgleichheit ist jedoch nur im statistischen Erwartungswert gegeben, bei jeder konkreten Studie können Imbalancen wichtiger prognostischer Merkmale nach den Gesetzen der Wahrscheinlichkeitsrechnung auftreten. Insbesondere bei aufwendigen Studien wird man daher diesen Fall möglichst zu verhindern suchen. Hier kommen deshalb häufig Methoden der balancierten Randomisation zum Einsatz, die bei Verwendung eines prädiktiven Scores besonders effektiv sein können [39].

Nur kurz erwähnt werden soll hier das Konzept des sogenannten Propensity-Scores [40], das aus methodischer Sicht einer ähnlichen Zielsetzung dient, nämlich der Elimination von systematischen Unterschieden in den Ausgangsbedingungen zweier Gruppen, die in Beobachtungsstudien miteinander verglichen werden. Allerdings beschreibt der Propensity-Score die Wahrscheinlichkeit der Zugehörigkeit eines Patienten zu einer der zwei zu vergleichenden Gruppen in Abhängigkeit von seinen Basisvariablen, ein solcher Score ist nicht von einer Studie auf eine andere übertragbar und damit nicht von übergreifender Bedeutung.

2.4.5 Scores in der Qualitätssicherung

Mit der zunehmenden Bedeutung von Qualitätssicherungsmaßnahmen in der Medizin ist auch über den Einsatz von Scores im Rahmen der Qualitätskontrolle nachgedacht worden. Dabei können sowohl evaluative als auch prädiktive Scores eine Rolle spielen.

Ein erstes Problem in der Qualitätskontrolle ist die Qualitätsmessung, für die man auf unterschiedliche Qualitätsindikatoren zurückgreift. Eine Möglichkeit besteht, ähnlich wie bei Therapiestudien, in der Messung globaler Zielgrößen, wie etwa der Gesamtsymptomatik oder der Patientenzufriedenheit, durch evaluative Scores.

Das zweite Problem der Qualitätskontrolle, das vor allem bei vergleichenden Studien auftritt, ist der Einfluß eines unterschiedlichen "case mix". Um bei einem Vergleich von verschiedenen Kliniken der Heterogenität der Patienten gerecht zu werden, kann in manchen Fällen auf Prognosescores zurückgegriffen werden.

> Knaus et al. [41] untersuchten die Qualität der intensivmedizinischen Behandlung in 13 amerikanischen Kliniken. Dabei benutzten sie den APACHE II Score, um für jede Klinik die tatsächliche und die aufgrund der Scorewerte erwartete Mortalitätsrate auf der Intensivstation zu vergleichen. Die so gebildeten Mortalitätsquotienten erlaubten eine Rangreihung der Kliniken sowie die Identifikation von zwei Kliniken, die hochsignifikante Abweichungen zeigten. Diese Qualitätsdaten wurden mit strukturellen Merkmalen über die Intensivabteilungen verglichen, um Anhaltspunkte für Determinanten der Qualität zu finden.

Für Maßnahmen der Qualitätskontrolle im klinikinternen Rahmen können Prognosescores ebenfalls eingesetzt werden, indem die Behandlungsergebnisse der einzelnen Patienten mit den aufgrund von Prognosescores zu erwartenden Verläufen verglichen werden.

> Boyd et al. [42] beschreiben, wie man bei schwerverletzten Patienten mittels der Kombination zweier Scores sowohl Einzelfälle mit unerwarteten Verläufen identifizieren als auch Gruppenvergleiche durchführen kann. Guirguis et al. [43] wenden dieses System zum Vergleich der Qualität der Behandlung von Traumapatienten in zwei großen Kliniken an.

2.5 Kritik an klinischen Scores

Ein Hauptkritikpunkt an klinischen Scores bezieht sich auf ihre mangelnde Evaluierung. Schon Feinstein, der Anfang der 80er Jahre im Rahmen eines groß angelegten Projekts einen Überblick über alle in der Literatur vorgestellten klinischen Beurteilungsinstrumente zusammenstellen wollte, beklagt, daß er eine viel größere Zahl von Scores und Indexen vorfand als erwartet (allein 150 im Bereich Magengeschwür), viele davon nur Varianten bereits existierender Indexe, gleichzeitig diese Indexe aber meist entstanden waren "in an ad hoc or laissez-faire manner" [1]. Viele dieser Scores wurden angewandt ohne eine Überprüfung von Qualitätsaspekten wie Reliabilität und Validität.

Kritik an klinischen Scores bezieht sich häufig auf den Versuch, komplexe Phänomene des subjektiven Erlebens meßbar zu machen. Man spricht dann abwertend von "weichen" im Vergleich zu "harten" Daten. Feinstein [1] entgegnet darauf zutreffend: "Because the important clinical and human phenomena are either undescribed, i.e., "unmeasured", or excluded from formal consideration, the data used for evaluating patient care become dehumanized". Besonders deutlich wird dies im Bereich der onkologischen Therapieforschung, wo erst seit Mitte der 80er Jahre der Bereich der Lebensqualität als Forschungsgegenstand in vergleichenden Therapiestudien mit berücksichtigt wird.

Van Gijn & Warlow [23] setzen sich kritisch mit Scores auseinander, die zur Gesamtbeurteilung der Symptomatik von Schlaganfallpatienten dienen. Sie behaupten unter anderem, daß Scores etwas unmögliches versuchen, indem sie einen hochdimensionalen Sachverhalt auf eine einzige Zahl reduzieren. Es ist unbestritten, daß Scores eine erhebliche Informationsreduktion beinhalten, dies ist ja gerade beabsichtigt. Dabei bemüht man sich aber im Rahmen der Scorekonstruktion, daß die für die Anwendung relevante Information erhalten bleibt. Van Gijn & Warlow widersprechen sich letztlich selbst, denn in einem weiteren Argument stellen sie Scores als zu differenziert dar und postulieren, daß die wesentliche Information eines Scores ebensogut in einer vierstufigen Stadieneinteilung repräsentiert werden kann.

Ein weiterer Kritikpunkt an der Verwendung von Scores bezieht sich darauf, daß oft unbegründeterweise mit Scorewerten so gerechnet wird (z.B. bei Mittelwertbildung), als könnte man von einem Intervallskalenniveau ausgehen. Auf diese Thematik wird zu Beginn des nächsten Kapitels noch konkreter eingegangen, jedoch kann schon hier vorweggenommen werden, daß für die meisten Scores überhaupt kein Skalenniveau theoretisch oder empirisch begründet werden kann. Stattdessen muß man Plausibilitätsargumente verwenden, und diese können durchaus ein Intervallskalenniveau als akzeptabel erscheinen lassen. Eine gewisse Zurückhaltung beim Rechnen mit Scorewerten ist allerdings angebracht und man muß mit Skepsis Arbeiten betrachten, die von einer Kommensurabilität von Scorewerten und anderen Merkmalen, wie zum Beispiel der Überlebenszeit, ausgehen, wie Holle [44] kürzlich in einer Kritik des sogenannten QALY-Konzeptes ausgeführt hat.

Gelegentlich wird behauptet, daß Scores in der klinischen Praxis kaum verwandt werden, da ihr Einsatz und Zweck oft unklar ist.

> Eine Umfrage zur Bekanntheit und Nutzung klinischer Scores wurde von Ohmann & Horstmann [45] unter 215 klinisch tätigen Gastroenterologen durchgeführt. Von diesen gaben 37% an, klinische Scores als formale Entscheidungshilfe zu benutzen, dabei handelte es sich allerdings im wesentlichen um drei Scores (Child-Pugh, CDAI, van Hees). Obwohl es etwa im Bereich der Ösophagusvarizenblutung außer dem einfachen, auf einem Expertenurteil basierenden Child-Pugh-Score mehr als 10 Prognosescores gibt [46], sind diese kaum bekannt und werden von den Teilnehmern der Umfrage nicht benutzt. Die Autoren mutmaßen, daß dies an der schwierigeren Verständlichkeit und Handhabbarkeit der konkurrierenden Scores liegt.

Aus einer eigenen Umfrage bei Orthopäden liegen erste Ergebnisse vor, nach denen 30% der Befragten äußerten, daß die Verwendung von Scores ihre klinische Tätigkeit erleichtert. Gleichzeitig gaben aber 25% an, daß ihre klinische Tätigkeit durch Scores erschwert würde, die übrigen waren indifferent. Von 86% der Ärzte wurde angegeben, daß Scores nie (36%) oder selten (50%) ihr ärztliches Handeln beeinflußt.

3 GÜTEKRITERIEN FÜR KLINISCHE SCORES

In diesem Kapitel werden Konzepte vorgestellt, die sich mit der Beurteilung der Güte von Meßmethoden im allgemeinen und von klinischen Scores im speziellen befassen. Besondere Bedeutung haben dabei natürlich jene Gütekriterien, die sich empirisch quantifizieren lassen. Zunächst sollen jedoch einige Gütekriterien genannt werden, deren Beurteilung meist theoretisch oder halb-empirisch, also auf der Basis von Hintergrunderfahrung erfolgt.

3.1 Theoretische Gütekriterien

Zunächst sei die Definition des Messens in Erinnerung gerufen. Unter Messen im weitesten Sinne versteht man die sinnvolle Zuordnung von Zahlen zu Objekten, wobei die arithmetischen Relationen zwischen den Zahlen entsprechende Relationen zwischen den Objekten wiedergeben sollen (nach Stevens [47]). Dabei sind drei Eigenschaften implizit oder explizit gefordert:

1.) Eindeutigkeit: jedem Objekt wird höchstens ein Meßwert zugeordnet.

2.) Vollständigkeit: jedem Objekt wird mindestens ein Meßwert zugeordnet.

Zusammen folgt hieraus, daß jedem Objekt genau ein Meßwert zugeordnet wird.

3.) Relationserhaltung: Die Zuordnung hat die Eigenschaft, relevante Beziehungen (Relationen) zwischen den Objekten durch entsprechende numerische Relationen zwischen den Meßwerten wiederzugeben.

Die Theorie des Messens befaßt sich sehr intensiv mit der Fundierung dieser Beziehung zwischen Zahlen und Objekten und legt großen Wert auf die Unterscheidung von Skalenniveaus [47]. In der klinischen Medizin ist man, sofern es sich nicht um physikalische Messungen handelt, in der Regel weit davon entfernt, entsprechende Begründungen, etwa für das Vorliegen eines Intervallskalenniveaus, empirisch leisten zu können, und bereits die zwei ersten der genannten Eigenschaften bereiten manchmal Probleme.

Eine mangelnde Eindeutigkeit kann aus einer unpräzisen oder inkonsistenten Definition der Zuordnungsregel eines Scores resultieren. Im Falle grenzwertiger Beurteilungen sollte möglichst klar erkennbar sein, wie die Abgrenzung der einzelnen Scorekategorien definiert ist. Wenn Scorekategorien über unterschiedliche Symptome definiert sind, so muß bei gleichzeitigem Vorliegen zweier Symptome klargestellt sein, welches davon den Ausschlag geben soll für die Einstufung.

> MacKenzie & Charlson [48] haben klinische Scores, die in Publikationen als Zielvariable verwandt wurden, auf ihre Eindeutigkeit untersucht und fanden in 3 von 27 Skalen eine Verletzung dieser Voraussetzung.

> Hutchinson et al. [49] berichten über Schwierigkeiten bei der Erhebung des Karnofsky-Index mittels eines Selbstbeurteilungsbogens. Neun von 26 (35%) Patienten waren nicht eindeutig einstufbar.

Eine unvollständige Meßbarkeit ergibt sich bei eingeschränkter Anwendbarkeit des Scores für manche Patienten. Starmark et al. [50] nennen das diesbezügliche Gütekriterium die "coverage" eines Scores.

> Die Glasgow Coma Scale (GCS) wurde von Starmark et al. [50] auf neurochirurgische Patienten angewendet, dazu ist eine Beurteilung der verbalen, der motorischen und der Augenreaktion erforderlich. Die Untersucher stellten fest, daß der GCS Summenscore nur bei 40% der Patienten direkt erhoben werden konnte, weil entweder die Augenreaktion wegen geschwollener Augen oder die verbale Reaktion aufgrund Intubation nicht festgestellt werden konnten.

Eine weitere Ursache der Unvollständigkeit kann darin liegen, daß bei fehlenden Informationen zu einzelnen Scorekomponenten keine Regel angegeben ist, nach der diese Teilinformationen bei der Berechnung des Gesamtscores zu ersetzen sind.

> Beim APACHE II, der aus den Krankenakten erhoben wird, werden fehlende Angaben grundsätzlich als Normalbefunde gewertet [51].

In Bezug auf die Relationserhaltung ist zunächst zu klären, welche numerischen Relationen bei der Verwendung von Scores eine Rolle spielen. Dies ist zunächst vor allem die Ordnungsrelation, die voraussetzt, daß zu je zwei Patienten angegeben werden kann, ob bei dem einen das Merkmal stärker, gleich oder weniger stark ausgeprägt ist als bei dem anderen und daß dieser Vergleich auch durch die Scorewerte wiedergegeben wird. Bei einfach konstruierten Scores, die nur wenige Komponenten berücksichtigen, läßt sich die Plausibilität dieser Relation auf der Basis von klinischer Erfahrung beurteilen. MacKenzie & Charlson fanden in ihrer oben zitierten Untersuchung [48], daß bei 3 von 25 beurteilbaren Scores die Rangfolge der Kategorien nicht einer klaren Hierarchie entsprach.

Gelegentlich läßt sich die Gültigkeit der Ordinalskala auch empirisch überprüfen, indem man die verschiedenen Kategorien als unabhängige Merkmale erhebt und die Antwortmuster auf Vorliegen einer hierarchischen Anordnung überprüft (Guttman-Skala, vgl. [52]).

> In einer eigenen Untersuchung [53] wurde die Lebensqualität von Patienten mit kleinzelligem Bronchialkarzinom im Rahmen einer Chemotherapiestudie mit Hilfe eines Fragebogens erhoben. Darunter waren sechs Fragen zur körperlichen Verfassung und Selbstversorgung, mit denen eine der Karnofsky-Skala entsprechende Selbstbeurteilung erfaßt werden sollte. Über 90% der Patienten zeigten bei diesen mit "ja" und "nein" zu beantwortenden Fragen ein Antwortmuster, das mit einer siebenstufigen Rangskala konsistent war.

Das gerade dargestellte Vorgehen ist allerdings bei den meisten Scores nicht anwendbar, insbesondere wenn sie als gewichtete Summe unterschiedlicher Komponenten definiert sind. Man muß davon ausgehen, daß für diese Form einer abgeleiteten Messung das Vorliegen einer Rangskala nicht begründbar ist, erst recht gilt dies für das Intervallskalenniveau. Bei genauerem Hinsehen zeigt es sich, daß es nicht einmal eine Begründung für die Gültigkeit der fundamentalsten aller numerischen Relationen, der Gleichheitsrelation, gibt. Hierfür müßte man beispielsweise rechtfertigen, daß zwei Patienten mit gleichem Summenwert in einem Lebensqualitätsscore tatsächlich auch das gleiche Ausmaß an Lebensqualität hätten, auch wenn bei einem von ihnen in drei Bereichen die Lebensqualität leicht eingeschränkt und beim zweiten Patienten nur in einem Bereich, dort aber stark eingeschränkt ist. Es ist völlig klar, daß diese Gleichsetzung weder theoretisch noch empirisch zu leisten ist und daß man in diesem Bereich auf Plausibilität und klinischem "common sense" bauen muß.

Als theoretische Gütekriterien lassen sich dementsprechend auch die folgenden zwei häufig zu findenden Begriffe einordnen, nämlich die "face validity" (=Augenschein-Validität) und "content validity" (=Inhaltsvalidität). Sie sind nichts anderes als Begriffe dafür, ob ein Score oder eine Skala inhaltlich plausibel erscheint und ob alle relevanten inhaltlichen Dimensionen des zu erfassenden Merkmals abgedeckt werden. Feinstein [1] ordnet diese beiden Begriffe einem von ihm als "sensibility" bezeichneten Gütekriterium eines klinischen Index unter, zu dem er eine Checkliste von 21 Aspekten zusammengestellt hat. Ein empirischer Zugang zu diesen Kriterien ergibt sich jedoch nur durch die Erhebung von Einschätzungen von Experten, durch die vor allem vergleichende Bewertungen verschiedener Skalen bezüglich dieser Validitätskriterien möglich sind.

3.2 Überblick über empirische Gütekriterien

Bei der Evaluation von klinischen Scores spielen vor allem die empirischen Gütekriterien eine Rolle, also jene Qualitätseigenschaften, die sich erst in der praktischen Anwendung erkennen und quantitativ beschreiben lassen. Man kann hier unterscheiden zwischen primären Gütekriterien, die im engeren Sinne die Meßqualität beschreiben, und sekundären Gütekriterien, die den Aufwand und den Nutzen bei der praktischen Anwendung des Scores betreffen.

Letztlich hat eine rationale Entscheidung für oder gegen die routinemäßige Verwendung eines Scores in einem Bereich sowohl die primären als auch die sekundären Gütekriterien zu berücksichtigen. Eine solche Kosten-Nutzen-Abwägung läßt sich allerdings nur schwer quantitiv durchführen, sondern ist eher auf qualitativer Ebene denkbar.

3.2.1 Primäre Gütekriterien

In der Literatur zur Meßproblematik in den verschiedensten Bereichen findet man stets zwei grundlegende Konzepte, die allerdings nicht immer einheitlich benannt werden. Das eine

Konzept betrifft die Frage, ob das Meßergebnis reproduzierbar ist, wenn man die Messung unter praktisch gleichen Bedingungen wiederholt. Das zweite Konzept befaßt sich mit dem Aspekt, ob das Meßergebnis qualitativ und quantitativ das wiedergibt, was man eigentlich zu messen beabsichtigt. Im erstgenannten Fall, der durch Begriffe wie Reproduzierbarkeit oder Reliabilität beschrieben wird, bleibt unberücksichtigt, ob man überhaupt das richtige mißt. Der zweite Fall hingegen beinhaltet die zusätzliche Forderung, daß das zu messende Merkmal auch wirklich quantitiv korrekt erfaßt wird, hier spricht man meist von Richtigkeit oder Validität der Meßmethode.

Tatsächlich stellen die Begriffe Reliabilität und Validität nur zwei allgemeine Konzepte dar, die bei genauerer Betrachtung in unterschiedliche Facetten differenziert werden können, die in verschiedenen Anwendungszusammenhängen von unterschiedlicher Bedeutung sind. Schaut man in den Bereich physikalischer oder laborchemischer Messungen, so sind die am häufigsten gebrauchten Begriffe Genauigkeit, Präzision und Richtigkeit, allerdings werden sie nicht immer übereinstimmend verwandt. Die dadurch entstehende Verwirrung wird durch Unklarheiten in der Zuordnung der englischen Begriffe "accuracy" und "precision" noch größer. Die Inkonsistenzen gehen von den Begriffen Richtigkeit und Genauigkeit aus, da diese bei manchen Autoren zwei unabhängige Eigenschaften bezeichnen, bei anderen jedoch teilweise voneinander abhängen. Um dies im folgenden genauer darlegen zu können, gehen wir von zwei anderen, in eindeutiger Weise gebrauchten Begriffen aus.

Unter Präzision ("precision") versteht man, daß zufällige Schwankungen der Meßwerte bei Meßwiederholung unter praktisch gleichen Bedingungen möglichst gering sind. Mit Unverzerrtheit ("unbiasedness") wollen wir die Eigenschaft einer Meßmethode bezeichnen, daß die sytematische (d.h. zum Beispiel mittlere) Abweichung der Werte aus Meßwiederholungen vom durch ein Referenzverfahren bestimmten wahren Wert minimal ist. Nach dieser Definition sind Präzision und Unverzerrtheit unabhängige Eigenschaften einer Meßmethode. Für die Unverzerrtheit finden sich in der Literatur auch die Bezeichnungen Richtigkeit [54, 55] bzw. "accuracy" [56] und für die Präzision die Bezeichnung "repeatability" [57]. Von anderen Autoren [57-59] wird Richtigkeit bzw. "accuracy" als zusammenfassende Eigenschaft angesehen, die sowohl Unverzerrtheit als auch Präzision voraussetzt, wieder andere verwenden hierfür den Begriff Genauigkeit [55].

Wendet man sich nun den klinischen oder psychometrischen Meßmethoden zu, so werden ganz überwiegend die Begriffe Reliabilität und Validität benutzt. Andere deutschsprachige Begriffe für diese zwei Konzepte, wie Zuverlässigkeit und Gültigkeit, werden wir hier nicht verwenden. Die Reliabilität ist dabei als weitgehend identisch mit der Präzision im obigen Sinne anzusehen, synonym werden wir auch von Reproduzierbarkeit sprechen. Der von Feinstein [1] gemachte Vorschlag, den Begriff Konsistenz anstelle von Reliabilität zu benutzen, da das Wort Reliabilität auch die Richtigkeit suggeriert, hat bisher wenig Verbreitung gefunden. Die Objektivität [60], die sich auf die Unabhängigkeit vom Anwender

eines Scores bezieht, ist kein eigenständiges Konzept, sondern kann der Reliabilität zugeordnet werden.

Der Validitätsbegriff hat bei der Verwendung im Zusammenhang mit klinischen Skalen eine wesentliche Erweiterung gegenüber dem Begriff der Unverzerrtheit erfahren, da im Falle des Fehlens einer Standardmeßmethode der Aspekt der quantitativen Richtigkeit sinnlos wird und hinter den der qualitativen oder inhaltlichen Richtigkeit zurücktritt. Da es in diesem Bereich keine natürliche Skalierung der Merkmale gibt, wird als Referenzverfahren meist ein anders skaliertes Meßinstrument oder sogar ein ganz anderes Merkmal herangezogen. Dies hat zur Konsequenz, daß es einen vom Meßinstrument unabhängigen "wahren" Wert nicht gibt und die Validität daher skalenunabhängig definiert werden muß. Hierzu greift man auf die Eigenschaft des Zusammenhangs, also der korrelativen Beziehung, zwischen Meßinstrument und Referenzkriterium zurück. Dies hat allerdings zwangsläufig den Verlust der Unabhängigkeit von Validität und Reliabilität zur Folge, in dem Sinne daß Validität automatisch Reliabilität mit voraussetzt, ähnlich wie bei dem oben dargestellten erweiterten Konzept der Richtigkeit eines Labormeßverfahrens.

3.2.2 Sekundäre Gütekriterien

Bei den sekundären Gütekriterien eines klinischen Scores, oder ganz allgemein eines Meßverfahrens, geht es stets um Aspekte im Zusammenhang mit Aufwand und Kosten der praktischen Anwendung, aber auch um den Nutzen.

Der Nutzen des Einsatzes eines Scores ist, wenn man die verschiedenen Einsatzmöglichkeiten bedenkt, in vielen Fällen empirisch kaum zu quantifizieren. Eine wesentliche Ausnahme bilden die Situationen, in denen ein Score klinisches Handeln beeinflußt, also vor allem in der Diagnostik, Prognostik und Therapieüberwachung. Der Bereich der therapeutischen Entscheidungsunterstützung ist derjenige, wo sich der tatsächliche Nutzen eines Scores am besten empirisch untersuchen und bewerten läßt. Dies ist am ehesten der Fall, wenn der Score fest in eine Therapiestrategie eingebunden ist.

Es wird später ausgeführt, daß der Nutzen eines Scores sehr eng mit der Validität in Verbindung steht und daß daher solche Validitätsparameter eine große Rolle spielen, die einen quantitativen Zugang zur Nutzenmessung erlauben. Wenn die Frage nach dem Nutzen eines Scores auch von vorrangiger klinischer Relevanz ist, so muß aber unbedingt erwähnt werden, daß der klinische Nutzen nicht allein ein Gütekriterium des Scores ist, sondern von seiner konkreten klinischen Verwendung abhängt. Auch ein für eine Therapieentscheidung herangezogener Score kann also, abhängig von der verwandten Therapie, nützlich oder aber ohne Nutzen sein.

Lienert [60] nennt die Ökonomie eines der Nebengütekriterien für psychodiagnostische Tests und dies gilt auch für alle klinischen Scores. Wichtige Eigenschaften hierbei sind die Einfachheit, mit der der Score erlernt und angewandt werden kann, die hierzu erforderliche

Qualifikation des Personals und der nötige Zeitaufwand. Alle diese Aspekte schlagen sich letztlich auch in den Kosten nieder, die die Anwendung des Scores verursacht. Dazu kommt die Tatsache, daß Praktikabilitätsaspekte oft die Qualität der Scoreerhebung beeinflussen und damit direkt auf die primären Gütekriterien rückwirken können. Dies kann soweit gehen, daß bei mangelnder Akzeptanz eines Scores sein Einsatz nicht sinnvoll ist.

Der finanzielle Aufwand ist zu berücksichtigen, wenn die Bestimmung des Scores zusätzliche, über die Kliniksroutine hinausgehende diagnostische Maßnahmen erfordert. Schließlich gibt es einige wenige Scores, für die sogar Anschaffungskosten zu veranschlagen sind. Im Bereich der psychodiagnostischen Meßinstrumente ist dies bekannt, hier werden schon seit Jahren Manuale, Testbögen, Auswertungsanleitungen und auch Auswertungssoftware kommerziell vertrieben. Einen besonderen Fall stellt aber ein prädiktiver Score für den Bereich der Intensivmedizin dar, das vor kurzem vorgestellte APACHE III System [11]. Dieser Score wurde mit großem Aufwand an Wissenschaftlern und an Patienten (über 17.000 Patienten aus 40 amerikanischen Kliniken) als Weiterentwicklung des APACHE II konstruiert. Während der Score als solcher publiziert und damit frei verfügbar ist, wird ein Computerprogramm, das aus dem Score unter Zugriff auf eine Patientendatenbank Vorhersagewerte, d.h geschätzte Mortalitätswahrscheinlichkeiten, berechnet, teuer verkauft [61].

3.3 Reliabilitätskonzepte

3.3.1 Varianzquellen

Als Reliabilität haben wir die Eigenschaft einer Meßmethode bezeichnet, reproduzierbare, d.h. identische oder zumindest ähnliche Meßergebnisse zu liefern, wenn man die Messung unter praktisch gleichen Bedingungen wiederholt. Eine empirische Bestimmung der Reliabilität erfordert also zunächst die Überlegung, was man unter praktisch gleichen Bedingungen verstehen möchte. Hierzu ist als erstes zu überlegen, welche Bedingungen den gesamten Meßprozeß bestimmen und welche davon in ihrem Einfluß auf die Varianz der Meßergebnisse untersucht werden sollen. Da dies im Bereich der Labormessung besonders deutlich darstellbar ist, seien zunächst die dort anzutreffenden typischen Varianzquellen beispielhaft beschrieben [62]. Dies sind unter anderen:

- das Gerät: technische Funktion, Eichung, Einstellungen
- das Material/die Reagenzien: Menge, Reinheit, Temperatur
- das Personal: Ablesung, Handhabung
- das Labor: Umgebungsbedingungen, Arbeitsablauf
- die Probe: Lagerung (Dauer, Temperatur, Material)
- die Probennahme: Durchführungsbedingungen, Stichprobe
- der Patient/ die Patientin: Tagesrhythmik, Tagesform

Auch bei klinischen Scores spielen diese oder ähnliche Varianzquellen eine Rolle, abhängig davon, auf welche Art von Eingangsinformation sich der Score bezieht. Viele Scores basieren auf der Fremdbeurteilung, z.B. durch den Arzt. Hier ist in gewissem Sinne der Beurteiler das Meßgerät und stellt oft die wichtigste Varianzquelle dar. Hat man es mit einem Score zu tun, der mittels Selbstbeurteilung durch einen Fragebogen erhoben wird, so kann man den Fragebogen als das Instrument ansehen. Der Patient kann aufgrund seiner intraindividuellen Schwankungen als Varianzquelle angesehen werden. Es hängt im wesentlichen vom Zeitbezug eines Scores ab, welche Bedeutung diese Schwankungen für die Messung haben. Eine häufig zu findende Einteilung der möglichen Variationsquellen unterscheidet den Beobachter (observer variation, user variability), das Meßinstrument (instrument variation, procedure variability) und den zu messenden Merkmalsträger (subject variation, input variability), wobei die englischen Bezeichnungen nach Daly et al. [59] und Feinstein [1] zitiert sind.

Für die Untersuchung einer einzelnen Varianzquelle ist es wichtig, daß die anderen Varianzquellen gleichzeitig möglichst ausgeschaltet sind. Will man etwa die Varianz durch den Beobachter untersuchen, so sollte dabei das Merkmal unverändert bleiben. Eine wichtige Rolle bei der Untersuchung der Reliabilität einer Meßmethode spielt daher die Frage, ob es möglich ist, eine "Probe" des zu messenden Merkmals zu konservieren, um für eine Wiederholung der Messung die Konstanz des Merkmals zu garantieren (vgl. Abschnitt 2.3.1).

3.3.2 Zufällige vs. systematische Variation

Als Voraussetzung für das Verständnis der statistischen Herangehensweise an die Reliabilitätsbestimmung ist es wichtig, zwischen systematischen und zufälligen Einflüssen der Varianzquellen auf das Meßergebnis zu unterscheiden. Dieser Unterschied liegt allerdings weniger in der Natur des Einflusses, sondern in unserer Kenntnis darüber. Erst wenn man verschiedene Meßwiederholungen im Hinblick auf einen bestimmten Faktor unterscheiden kann, läßt sich der systematische Einfluß dieses Faktors untersuchen. Es muß hier aber betont werden, daß man den systematischen Effekt eines Einflußfaktors auf das Meßergebnis nicht mit dem oben beschrieben systematischen Meßfehler im Sinne der Unverzerrtheit oder Richtigkeit gleichsetzen darf, da letzterer ja die Kenntnis des "wahren" Wertes voraussetzt.

Meßwiederholungen, die keine weiteren differenzierenden Merkmale aufweisen, bezeichnet man meist als Replikationen, im Gegensatz zu Meßwiederholungen mit einer gewissen systematischen Struktur. Dies soll an einem Beispiel veranschaulicht werden: Als reine Replikationen können im Labor solche Meßwiederholungen angesehen werden, die zum Beispiel dadurch entstehen, daß jede Serumprobe in zwei oder mehr Teilproben aufgeteilt wird, die dann im gleichen Meßvorgang parallel mitbestimmt werden (Intra-Assay-Variation). Werden die zwei Teilproben jedoch grundsätzlich unterschiedlich behandelt, indem sie unterschiedlich lange gelagert werden oder indem sie in verschiedenen Labors analysiert

werden, so liegt eine Zuordnung der Meßwiederholungen zu einem zwei- oder mehrstufigen Faktor vor (Inter-Assay-Variation, Inter-Labor-Variation).

Bei der Reliabilitätsuntersuchung klinischer Skalen ist der Fall einer reinen Replikation selten, da Verhaltensstichproben in Form von Fragebögen, Interviews oder Untersuchungen nicht beliebig geteilt und Messungen daran wiederholt werden können. Als Beispiel für eine Ausnahme sei das Five-Minute-Speech-Sample aus dem Bereich der Schizophrenieforschung [63] erwähnt, bei dem aus einem längeren Gesprächsmitschnitt eine fünfminütige Stichprobe ausgewählt und nach bestimmten Kriterien beurteilt wird. Stattdessen kommt der Fall des Meßwiederholungsfaktors häufig vor, wenn etwa Beurteilungen in einem vorgegebenen Zeitabstand wiederholt werden oder aber die Untersuchungen durch zwei bestimmte, für alle Patienten gleiche Beurteiler durchgeführt werden. Wird hingegen jeder Patient von jeweils zwei anderen Ärzten untersucht, was in multizentrischen Reliabilitätsstudien vorkommen kann, so ist dies wiederum als Replikation anzusehen.

Im Falle von Meßwiederholungsfaktoren ist eine weitere Unterscheidung erforderlich, die sich auf die Auswahl der Faktorstufen bezieht. Die Frage ist hierbei, ob man alle interessierenden Ausprägungen des Einflußfaktors im Rahmen einer Reliabilitätsuntersuchung erfassen kann oder nur eine Stichprobe derselben. Man spricht im ersten Fall von einem festen ("fixed") Faktor und im zweiten Fall von einem zufälligen ("random") Faktor. Bei der Untersuchung des Einflusses des Beurteilers auf einen Score wird man in der Regel von einem zufälligen Faktor ausgehen, da man aus allen potentiellen Beurteilern nur eine kleine Stichprobe in die Untersuchung aufnehmen kann. Ein fester Einflußfaktor liegt hingegen vor, wenn man bei einem auf einem kurzen Fragebogen basierenden Score untersucht, welchen Einfluß es hat, ob die Erhebung schriftlich, telefonisch oder im direkten Gespräch erfolgt.

3.3.3 Spezielle Reliabilitätskonzepte

Untersucht man einmal, welche der verschiedenen möglichen Varianzkomponenten in Studien am häufigsten untersucht werden, so stößt man vor allem auf drei spezielle Reliabilitätskonzepte: die Inter-Rater-Reliabilität, die Test-Retest-Reliabilität und die interne Konsistenz.

Bei Scores, die auf subjektiven Einschätzungen, sei es von Verhalten oder von Befunden, basieren, wird vorrangig die sogenannte Inter-Rater-Reliabilität untersucht, die manchmal auch als Objektivität [60] bezeichnet wird. Elmore & Feinstein [64] geben eine ausführliche Bibliographie von Studien aus den verschiedensten medizinischen Bereichen, in denen die Inter-Rater-Reliabilität von klinischen Beurteilungen untersucht wurde. In vielen Fällen war die Übereinstimmung zwischen verschiedenen Beurteilern schlechter als von den Beteiligten erwartet worden war. Der Einfluß der subjektiven Beurteilung kann sich auch bei anscheinend harten Daten ergeben, wie das folgende Beispiel zeigt:

In einer multizentrischen Reliabilitätsstudie zum APACHE II [51] wurde der Gesamtscore sowie alle Einzelkomponenten, also insbesondere auch die physiologischen Variablen, auf ihre Inter-Rater-Reliabilität hin untersucht. Dabei ist anzumerken, daß der APACHE II aus der Krankenakte des Patienten erhoben wird. Zwar hatte der Gesamtscore eine recht hohe Reliabilität, jedoch galt dies nicht für einige Komponenten (z.B. Körpertemperatur). Der Grund hierfür liegt darin, daß bei jeder Komponente der ungünstigste Wert in den ersten 24 Stunden nach Aufnahme des Patienten gewertet wird, hier können bei häufig gemessenen Merkmalen leicht Bewertungsfehler auftreten.

Eine Voraussetzung für eine akzeptable Übereinstimmung zwischen verschiedenen Beurteilern ist die Konsistenz der Urteile eines jeden Raters, also die Intrarater-Reliabilität.

Basiert der Score auf der Selbstbeurteilung der Patienten, etwa im Falle eines Lebensqualitätsfragebogens, so findet man meist Untersuchungen zur Test-Retest-Reliabilität, bei denen eine Messung in einem geeigneten zeitlichen Abstand wiederholt wird. Dies macht natürlich nur Sinn, wenn man dabei die grundsätzliche Konstanz des zu messenden Merkmals unterstellen kann. In der Medizin hat man es oft mit Merkmalen zu tun, die sich in dem Zeitrahmen üblicher Behandlungsverläufe, also in Tagen, Wochen oder Monaten, in relevanter Weise ändern, während sie in kurzen Zeiträumen, etwa Stunden oder Tagen, nur geringfügig schwanken. Hier würde man Test-Retest-Untersuchungen in solchen zeitlichen Abständen durchführen, bei denen man von nur unerheblichen Veränderungen des Merkmals ausgeht. Es ist aber klar, daß mit einer solchen Studie gleichzeitig die Stabilität des Merkmals untersucht wird. Die Aussagekraft der Untersuchung hängt somit sehr von der geeigneten Wahl des zeitlichen Abstands der Messungen ab, hierauf kommen wir in Abschnitt 6.1.2 nochmal zurück.

Eine Reliabilitätsuntersuchung, bei der sowohl der Beobachter als auch der Meßzeitpunkt konstant gehalten werden, ist bei vielen Meßmethoden kaum machbar, weil unter diesen Einschränkungen die Messung nicht wiederholbar ist. Im Fall von klinischen Scores, die aus vielen einzelnen Komponenten zusammengesetzt sind, besteht jedoch die Möglichkeit, eine gewisse Redundanz im Score zur Untersuchung der sogenannten internen Konsistenz zu nutzen. Dieses Vorgehen ist insbesondere bei Scores, die sich auf Fragebögen stützen, üblich. Man versucht dabei, den Score anhand seiner Komponenten in zwei etwa gleichwertige (parallele) Hälften aufzuteilen und faßt die beiden als alternative Realisationen des Meßinstruments auf, ähnlich wie zwei Labormeßgeräte des gleichen Fabrikats. Dies ist leicht möglich, wenn man entweder bereits eine gewisse Redundanz in den Fragen hat oder aber sogar zwei parallele Fragebogenversionen existieren, die sich mischen lassen. Diese Testhalbierungs-Reliabilität läßt sich noch fortsetzen auf kleinere Teile des Meßinstruments bis hin zu den einzelnen Kompoenenten oder Items.

3.3.4 G-Studien und D-Studien

Im vorangegangenen Abschnitt wurden die Reliabilitätsvarianten dargestellt, die am häufigsten empirisch untersucht werden. Dabei wird fast immer nur eine einzige Varianzkomponente isoliert betrachtet. Die unterschiedlichen Möglichkeiten der Bestimmung der Reliabilität eines Scores lassen sich jedoch als Spezialfälle eines gemeinsamen Ansatzes auffassen. Diese Sichtweise hat für den Bereich der Messung psychischer Merkmale Cronbach [65] im Rahmen seiner "generalizability theory" formuliert (s.a Shavelson et al. [66]). Der gemeinsame Studienansatz für die Untersuchung der verschiedenen Reliabilitäten kann als umfassende Varianzkomponentenanalyse beschrieben und mittels des gleichnamigen statistischen Verfahrens realisiert werden. Ein solches Vorgehen ist schon lange in Anwendungen bei technischen Meßverfahren üblich. Im Zusammenhang mit Reliabilitätsuntersuchungen von Skalen in der Medizin oder Psychologie gibt es aber nur relativ wenige Anwendungen dieses als "G-Studien" [67] bezeichneten systematischen Zugangs ("G" steht für "generalizability"), z.B. von Chambers et al. [68] und von Evans et al. [69]. Der Grund liegt vermutlich darin, daß solche G-Studien vom Versuchsdesign und von der Auswertungsmethodik wesentlich anspruchsvoller sind als einfaktorielle Untersuchungen, allerdings lassen sie auch weitergehende Erkenntnisse zu. Die Zielsetzung von G-Studien ist es letztlich, alle wesentlichen Variationsquellen und ihre Interaktion zu untersuchen sowie diese Einflüße zu quantifizieren.

Die oben geschilderte Unterscheidung von systematischen und zufälligen Einflüssen der verschiedenen Variationsquellen hat für die Reliabilitätsuntersuchung zunächst nur eine geringe Bedeutung, da sie sich auf die quantitative Reliabilitätsschätzung, also den Vergleich des Einflusses verschiedener Varianzkomponenten, in der Praxis meist nur wenig auswirkt. Für die Umsetzung des Ergebnisses einer Reliabilitätsstudie ist sie aber essentiell, denn der Einfluß systematischer und zufälliger Variationsquellen kann in unterschiedlicher Weise in der Praxis reduziert werden.

Hat man eine wesentliche Varianzkomponente eines Scores identifiziert, so gibt es grundsätzlich zwei unterschiedliche Herangehensweisen, wie man die Reliabilität verbessern kann. Bei zufälligen Einflußfaktoren geschieht dies durch das (in der Regel kostenverursachende) Prinzip der Meßwiederholung. Kann man jedoch diesen Einfluß als systematischen Effekt eines Faktors beschreiben, so ist durch Standardisierungsmaßnahmen eine einfache Möglichkeit der Reliabilitätsverbesserung zu erreichen. Ein Beispiel soll dies verdeutlichen: Ist bei der Beurteilung eines Merkmals anhand einer Rating-Skala eine erhebliche Zufallsstreuung zwischen den Beurteilern festzustellen, so läßt sich eine Verbesserung der Inter-Rater-Reliabilität erreichen, wenn man das Merkmal grundsätzlich von mehreren Ratern unabhängig beurteilen läßt und deren Mittelwert oder Konsensusurteil als Meßwert nimmt. Dieser Aufwand wäre aber nicht nötig, wenn man den Ratereffekt als einen systematischen Einfluß der unterschiedlichen Qualifikation zurückführen kann, der sich

dann entweder durch die Beschränkung auf erfahrene Beurteiler oder durch ein vorangehendes Ratertraining weitgehend beheben ließe.

Die Untersuchung von Möglichkeiten der Reliabilitätsverbesserung wird in ihrer Bedeutung oft noch vernachlässigt, u.a. weil die Konsequenzen von mangelnder Reliabilität von vielen Forschern unterschätzt werden (vgl. Abschnitt 3.5). Hat man im Rahmen einer G-Studie die wesentlichen Varianzkomponenten in ihrem Einfluß auf das Scoreergebnis quantifiziert, so läßt sich im Rahmen von Modellrechnungen untersuchen, welchen Effekt verschiedene Entscheidungen in Bezug auf mögliche Maßnahmen zur Reliabilitätsverbesserung hätten. Solche Berechnungen werden auch als D-Studien bezeichnet, das "D" steht dabei für "decision" (vgl. hierzu Streiner & Norman [67], wo auch ein Rechenbeispiel vorgeführt wird).

3.4 Validitätskonzepte

3.4.1 Konstruktvalidität vs. Kriteriumsvalidität

Der Begriff der Validität gibt leicht zu Mißverständnissen Anlaß, weil er teils in einem sehr weiten Sinne, teils spezifisch gebraucht wird. Wir haben ihn hier zunächst etwa als Synonym für die Richtigkeit einer Meßmethode eingeführt, aber bereits erwähnt, daß bei klinischen Merkmalen in Abwesenheit eines "wahren" Wertes diese Übertragung schwierig ist. Eine häufig zitierte Definition aus der Testpsychologie [60] besagt: "Die Validität eines Tests gibt den Grad der Genauigkeit an, mit dem dieser Test dasjenige Persönlichkeitsmerkmal oder diejenige Verhaltensweise, das (die) er messen soll oder zu messen vorgibt, tatsächlich mißt." Feinstein [1] betont, daß das Wort Validität in drei verschiedenen Bedeutungen gebraucht wird, nämlich für Konsistenz (im Sinne von Reliabilität), für Richtigkeit und "suitability". Validität im Sinne von Feinstein´s "suitability" haben wir oben als Plausibilität bezeichnet und im Rahmen der theoretischen, nicht empirisch überprüfbaren Gütekriterien von Scores eingeordnet. Hier werden wir es nur auf die mittlere der drei Bedeutungen beziehen.

Die Definition bzw. Untersuchung der Richtigkeit einer Messung setzt voraus, daß ein richtiger (wahrer) Wert überhaupt existiert und möglichst bestimmbar ist. Bei physikalischen oder chemischen Meßverfahren ist die Existenz eines wahren Wertes, der unabhängig vom Meßverfahren existiert, in der Regel kein Problem, obwohl seine praktische Bestimmung auch bei noch so aufwendiger Messung gewissen Einschränkungen unterworfen ist. Anders ist dies bei Merkmalen oder Konstrukten, die nicht physikalisch oder materiell erfaßbar sind, wie etwa bei der Messung von Intelligenz, Schmerz oder Lebensqualität. Hier gibt es erst recht kein optimales Meßinstrument, keinen sogenannten gold standard, und es fällt sogar schwer, sich die Existenz einer wahren Merkmalsausprägung unabhängig von einem Meßverfahren vorzustellen, da das Konstrukt vielschichtig ist und viele Varianten zuläßt. Oft gelingt es nicht einmal, eine faßbare Definition des Konstrukts zu geben und man landet schließlich bei der bekannten Tautologie "Intelligenz ist das, was der Intelligenztest mißt". In solchen Fällen

versteht man unter dem Begriff der Konstruktvalidität einen empirischen Ansatz, aufgrund der Untersuchung des Zusammenhangs des Meßinstruments mit anderen Merkmalen oder Meßinstrumenten annähernd zu bestimmen, was denn nun eigentlich gemessen wird.

Der Hauptansatzpunkt für eine empirisch faßbare Validitätsbestimmung muß eine Referenzmethode oder ein Referenzkriterium sein, mit dem man definiert, was der Score eigentlich messen soll. Die Validität in diesem Sinne bezeichnet man als Kriteriumsvalidität, mit ihr werden wir uns hauptsächlich beschäftigen. Die Art des Kriteriums wird sich aber bei evaluativer oder prädiktiver Anwendung des Scores unterscheiden, zum Beispiel kann es ein bereits etabliertes Meßinstrument, ein Expertenrating oder ein späteres klinisches Ereignis sein.

Der Begriff Konstruktvalidität ist weniger leicht zu konkretisieren als die Kriteriumsvalidität. Dies liegt wohl unter anderem daran, daß die beiden Begriffe sich konzeptuell kaum unterscheiden, da sich auch die Konstruktvalidität auf Kriterien bezieht. Charakteristisch für die Konstruktvalidität ist allerdings, daß man erstens meist mehrere Kriterien heranzieht und daß zweitens deren enger Zusammenhang mit dem Merkmal selbst nur theoretisch oder unzureichend empirisch belegt ist. Feinstein [1] bezeichnet die Konstruktvalidität als einen Ersatz für die Kriteriumsvalidität, wenn kein befriedigendes Kriterium existiert, man also weder eine a-priori bessere (richtigere) Meßmethode des Merkmals noch ein für die Anwendung des Scores relevantes Außenkriterium gefunden hat. Das daraus resultierende Problem der Konstruktvalidierung besteht darin, daß im Falle eines negativen Ergebnisses eines Validierungsversuches stets die unentscheidbare Frage bleibt, ob nun der untersuchte Score nicht valide ist oder ob die Kriterien schlecht gewählt waren, also die theoretischen Annahmen nicht gerechtfertigt.

Konstruktvalidität ist nicht durch einen Parameter ausdrückbar, stattdessen erhält man ein ganzes Muster von Beziehung zwischen dem Score und den verschiedenen verwendeten Kriterien. Dabei verwendet man nicht nur Kriterien, bei denen man einen Zusammenhang mit dem Score erwartet, sondern bezieht ausdrücklich auch Kriterien mit ein, von denen sich der Score aus theoretischen Gründen unterscheiden soll. Im ersten Fall spricht man von konvergenter, im zweiten von diskriminanter Validität. Eine Strategie zur Untersuchung der Konstruktvalidität stellt die "multi-trait-multi-method"- Methode dar. Für eine anschauliche Darstellung dieses Vorgehens wird auf die Literatur verwiesen [67].

In den folgenden Abschnitten werden drei spezielle Fälle der Kriteriumsvalidität dargestellt und in ihrer Bedeutung diskutiert. Der erste Fall bezieht sich auf dichotome Außenkriterien, deren Relevanz ausführlich begründet wird. Als spezielles Konzept der Kriteriumsvalidität wird hier die Diskrimination eiingeführt.Als Gütekriterium für evaluative Scores wurde von Guyatt und Mitarbeitern [70] der Begriff "responsiveness" eingeführt, den wir hier mit dem im Deutschen bekannten Ausdruck Änderungssensitivität gleichsetzen. Während die genannten Autoren die Änderungssensitivität als eigenständiges Gütekriterium ansehen, wird in Abschnitt 3.4.3 die Ansicht begründet, daß es nichts anderes als eine Facette der

Kriteriumsvalidität ist. Ein weiterer spezieller Validitätsbegriff taucht im Zusammenhang mit prognostischen Scores auf. Diese lassen sich als Meßverfahren für ein Merkmal wie Heilungswahrscheinlichkeit oder Sterberisiko ansehen und die Scorewerte lassen sich in Vorhersagewahrscheinlichkeiten übersetzen. Die Richtigkeit dieser geschätzten Wahrscheinlichkeiten bezeichnet man als Kalibration. Die Problematik dieses Gütekriteriums wird in Abschnitt 3.4.4 behandelt.

Es bleibt zu fragen, ob ein Score global als valide oder unvalide bezeichnet werden kann, oder ob es so viele Validitätsaussagen wie sinnvolle Kriterien gibt. Letzteres ist der Fall, da ja ein Score in der Regel keine direkte Messung eines Merkmals darstellt, sondern eine indirekte Messung. Somit kann ein Score durchaus in Bezug auf ein Kriterium eine hohe Validität aufweisen, in Bezug auf ein anderes jedoch unvalide sein.

3.4.2 Diskrimination

Bei der Kriteriumsvalidität sind zweierlei Unterscheidungen zu treffen. Zum einen betrifft dies die Art des Kriteriums, das sowohl quantitativ als auch kategoriell sein kein. In diesem Abschnitt wird es um dichotome Kriterien gehen, die in der Praxis eine ganz wichtige Rolle spielen. Dies soll im folgenden etwas ausführlicher erläutert werden. Des weiteren wird in der Literatur danach unterschieden, ob es sich um ein gleichzeitiges ("concurrent validity") oder ein zeitlich späteres ("predictive validity") Kriterium handelt. Die diesbezügliche Entscheidung hängt teilweise vom Verwendungszweck des Scores ab, jedoch ist ein gleichzeitiges Kriterium, z.B. ein als Standard angesehener Score, grundsätzlich möglich.

Die Kriteriumsvalidität hat eine vorrangige Bedeutung bei Scores, die im Rahmen der individuellen Diagnostik bzw. Prognostik eingesetzt werden sollen. Nun ist es bei diagnostischen Fragestellungen naheliegend, von einem dichotomen Kriterium auszugehen, da man primär an der Unterscheidung von Erkrankten (genauer: an der Krankheit K Erkrankten, im folgenden K+) und Nicht-Erkrankten (im folgenden K-) interessiert ist. So wird man aus dem Score, genauso wie aus einem quantitativen Labortest, eine dichotome diagnostische Aussage ableiten, die wiederum mit einer dichotomen therapeutischen Handlungskonsequenz verknüpft ist. Die doppelte Dichotomie wird also bereits durch die Problemstellung bedingt und führt zu der bekannten Darstellung in Form der folgenden Vierfeldertafel.

	Krankheit ja K+	Krankheit nein K-
Test positiv T+	richtig positiv RP	falsch positiv FP
negativ T-	falsch negativ FN	richtig negativ RN

Abb. 1: Modell eines einfachen diagnostischen Tests

Dabei bedeuten T+ und T- das auf der Dichotomisierung des Scores basierende positive bzw. negative Testergebnis und FP und FN die falsch-positive bzw. falsch-negative Testentscheidung (analog RP und RN). Diese naheliegende Modellierung der Kriteriumsvalidität bei Scores für diagnostische Zwecke bietet auch eine einfache Möglichkeit, die Quantifizierung der Validität mit Aspekten des klinischen Nutzens zu verbinden, da sich dieser in den Konsequenzen richtiger und falscher Entscheidungen ausdrückt.

Anders ist es im Falle der prognostischen Anwendungen von Scores, wo als eigentliches Zielmerkmal der Krankheitsverlauf betrachtet wird, der sehr vielgestaltig sein kann und a-priori keinesfalls eine Dichotomie nahelegt. Selbst in Fällen, wo eine Einteilung in "Heilung" vs. "keine Heilung" scheinbar auf ein dichotomes Zielkriterium hinweist, liegen die Dinge komplizierter, da man stets auch die zeitliche Dimension, also etwa die Dauer bis zum Eintreten der Heilung, zu berücksichtigen hat. Bei vielen chronischen Krankheiten ist eine echte Heilung kaum zu erzielen und prognostische Aussagen beziehen sich eher auf die Zeitdauer bis zum Auftreten einer Verschlechterung des Zustands, in der Onkologie etwa die Zeit bis zum erneuten Tumorprogreß oder bis zum Tode. Welche Gründe sprechen nun dafür, auch in Fällen, wo das interessierende Merkmal eine solche Zeitdauer ist, diese zu einem dichotomen Kriterium für die Bestimmung der prädiktiven Validität zu vereinfachen? Ein Grund ist die einfache statistische Handhabbarkeit, vor allem aber die bessere Interpretierbarkeit der statistischen Parameter in diesem vereinfachten Modell. Um dies genauer zu erläutern, gehen wir im folgenden auf den Zusammenhang zwischen Validität und klinischem Nutzen ein.

Will man einen in einer Studie ermittelten Wert eines Validitätsparameters ähnlich sachgerecht beurteilen wie einen Erfolgsparameter einer Therapie, so muß die klinische Bedeutung des Parameterwertes klar erkennbar sein. Dies geht nur, wenn der Validitätsparameter in einen definierten klinischen Zusammenhang gestellt wird. Im Gegensatz zur Therapie ist der klinische Nutzen der Prognose jedoch nicht unmittelbar erkennbar und nur in manchen Fällen in einen therapeutischen Nutzen übersetzbar. Am ehesten ist dies möglich, wenn abhängig vom Prognosescore eine therapeutische

Entscheidung gefällt wird. Dieser Fall einer "individualisierten", d.h. vom Scorewert abhängigen Therapie war bereits als eine der wichtigen Anwendungen von Scores dargestellt worden. Für die weiteren Überlegungen wird davon ausgegangen, daß bei einer geeignet definierten Patientengruppe für die Therapiewahl nur eine auf dem Scorewert basierende Entscheidungsregel ausschlaggebend ist.

Therapieentscheidungen kommen oft als dichotome Alternativen vor (Behandlung ja oder nein, palliativ oder kurativ, Therapiefortsetzung oder Beendigung). Sind sie es nicht, so lassen sie sich in eine Abfolge dichotomer Teilentscheidungen zerlegen. Die entscheidende Überlegung im Hinblick auf eine einfache Modellierung der klinischen Situation ist nun, daß die Dichotomie der therapeutischen Entscheidung auch eine Dichotomisierung des Zielkriteriums, also des Krankheitsverlaufs, nahelegt. Wenn sich das prognostische Anwendungsproblem so einfach modellieren läßt wie gerade beschrieben, dann kann man es in analoger Weise zum Diagnoseproblem als Vierfeldertafel darstellen.

		Verlauf	
		schlecht	gut
Prognose	schlecht	RP	FP
	gut	FN	RN

Abb. 2: Modell einer einfachen Prognoseregel

Dieses einfache, für diagnostische und prognostische Fragestellungen gleichermaßen anwendbare Modell hat zum einen den Vorteil, daß zwei Parameter ausreichen, um die gesamte Information zur Validität der Entscheidungsregel auszudrücken. Darüberhinaus ist aber von entscheidender Bedeutung, daß diese Parameter in direkten Zusammenhang mit dem klinischen Nutzen einer Diagnose oder Prognose gebracht werden können, indem eine Bewertung der zwei möglichen Fehlentscheidungen FP und FN vorgenommen wird [71].

Im Falle der Anwendung des Modells auf die prognostische Fragestellung geht eine gewisse Differenzierung verloren, wenn in dem Spektrum "guter" Verläufe noch erhebliche interindividuelle Unterschiede bestehen. Die Dichotomisierung quantitativer Zielgrößen ist jedoch auch im Rahmen von Therapiestudien durchaus üblich und der entstehende Informationsverlust wird unter bestimmten Bedingungen in Kauf genommen. Allerdings besteht bei Prognosestudien die zusätzliche Problematik, daß eine artifizielle Differenzierung erzeugt wird, indem zwei fast identische Verläufe als "gut" bzw. "schlecht" unterschieden werden, weil im einen Fall das Zielereignis gerade noch im, im anderen kurz nach dem Ende des definierten Zeitintervalls auftrat. So wird die Möglichkeit, eine trennscharfe Prognoseregel zu finden, grundsätzlich eingeschränkt, wenn die künstliche Abgrenzung guter und schlechter Verläufe nicht einem natürlichen Wert entspricht, sondern in einem Bereich hoher Dichte liegt.

3.4.3 Änderungssensitivität

Das Gütekriterium der Änderungssensitivität ("responsiveness" oder "sensitivity to change") wurde in der amerikanischen Literatur vor allem von Guyatt und Mitarbeiten stark propagiert [18, 70, 72-74]. Es soll ein Maß sein für die Eigenschaft eines Scores, klinisch relevante Änderungen des Gesundheitszustandes zu erfassen, womit es bereits konzeptuell der Validität zuzuordnen ist. Ein solches Gütekriterium ist für evaluative Scores relevant. Es unterscheidet sich von den bisher erörterten Validitätskonzepten insofern, als es keine Aussage über einmalig erhobene Scorewerte macht, sondern über die Änderung der Scorewerte bei mehrmaliger Erhebung im Behandlungsverlauf. Der Grund für die Betonung der Bedeutung dieses Kriteriums war die Feststellung, daß sich ein für diagnostische oder prognostische Zwecke valider Score nicht automatisch als Zielkriterium in Studien eignet.

Eine Vorgehensweise zur Quantifizierung der Änderungssensitivität zeigt die Analogie zur Kriteriumsvalidität. Dabei werden die Änderungen der Scorewerte in Beziehung gesetzt zu den tatsächlichen, mit einer als Standard akzeptierten Referenzmethode gemessenen Änderungen des Gesundheitszustandes eines Patienten [75]. Leider wurde dieser naheliegende Vorschlag in der Literatur relativ wenig beachtet, da von Guyatt und Koautoren frühzeitig [18] und wiederholt eine andere mögliche Strategie zur Erfassung der Änderungssensitivität vorgeschlagen wurde. Diese Sichtweise konzentriert sich auf die Veränderung des Scores bei Vorliegen einer realen Veränderung. Hierzu wird die Untersuchung des Scores bei Patienten unter einer Therapie mit nachgewiesener Effektivität vorgeschlagen. Ein solches Vorgehen ist aus mehreren Gründen ausgesprochen problematisch. Erstens entsteht dabei eine Abhängigkeit des Gütekriteriums von der eingesetzten Therapie, wobei unklar bleibt, wieweit die unterstellte Wirksamkeit dieser Therapie wirklich gesichert ist. Zweitens wird dabei nicht unterschieden zwischen Patienten, bei denen die Therapie gewirkt hat und solchen, bei denen dies nicht der Fall war. Damit wird zwar das Problem vermieden, für die Änderung des interessierenden Merkmals (z.B. Lebensqualität) ein als Referenz dienendes anderes Meßverfahren zur Verfügung haben zu müssen. Stattdessen muß man aber unterstellen, daß eine systematische Änderung der Scorewerte tatsächlich Ausdruck der Änderung des Merkmals ist, was nicht immer plausibel ist.

Es ist zu bezweifeln, daß es sich bei der Änderungssensitivität um ein eigenständiges Gütekriterium für evaluative Skalen handelt, wie Guyatt und Mitarbeiter behaupten. In einer späteren Arbeit [73] versuchten diese Autoren dem Einwand, daß "...responsiveness is simply another aspect of validity, and ... its introduction as a separate concept merely confuses the issue" mit einigen Beispielen zu begegnen.

> In einer randomisierten Therapiestudie bei Patientinnen mit Brustkrebs wurden eine kurze (12 Wochen) vs. eine lange (36 Wochen) Chemotherapiebehandlung miteinander verglichen. Mit vier unterschiedlichen Meßinstrumenten wurden Aspekte der Lebensqualität erfaßt und die Scoredifferenzen zwischen zwei relevanten Zeitpunkten wurden zwischen den zwei Behandlungsgruppen verglichen. Als Argument für die

> Notwendigkeit der Unterscheidung zwischen Validität und Änderungssensitivität wird vorgebracht, daß zum einen die ECOG Toxizitätsskala zwar hochsignifikante Unterschiede zwischen den Behandlungen anzeigt, also responsiv sei, jedoch nicht valide, weil sie keine Lebensqualitätsskala sei. Zum anderen sei die auf fünf Items gekürzte Skala des Rand-Fragebogens zur emotionalen Befindlichkeit bekanntlich valide, obwohl sie in dieser Untersuchung keine Therapieunterschiede abbildet [73] .

Dieses Beispiel zeigt aber nur, wie unsinnig eine Definition von Änderungssensitivität ist, wenn sie nicht beinhaltet, welches Merkmal eigentlich in seiner Änderung erfaßt werden soll. Daraus resultiert, daß die Änderungssensitivität ein Spezialfall der Kriteriumsvalidität ist, mit der einzigen Besonderheit, daß sie sich auf Änderungen von Skalenwerten im zeitlichen Verlauf anstatt auf einmal gemessene Werte bezieht. Es ist klar, daß die Änderungssensitivität eines Scores nicht mit seiner Validität gleichzusetzen ist, denn es gibt grundsätzlich mehrere Validitätsaussagen zu einem Score und die Änderungssensitivität ist nur eine davon. In einer weiteren Arbeit versuchen Guyatt et al. [70] sogar, die Änderungssensitivität als eine Art der Reliabilität für evaluative Skalen darzustellen, ohne dies jedoch überzeugend leisten zu können.

Tatsache ist, daß die Änderungssensitivität als Gütekriterium für evaluative Skalen von den meisten Autoren vernachlässigt oder sogar völlig übersehen wurde, sowohl in theoretischen Publikationen [1, 76, 77] als auch in vergleichenden empirischen Evaluationen. Erst in jüngster Zeit ist die Beurteilung der Änderungssensitivität häufiger in Studien zum Vergleich evaluativer Scores zu finden [78]. Auf die dabei verwendeten Operationalisierungen der Änderungssensitivität wird später eingegangen .

Bei der vorangegangenen Diskussion wurde die Frage, wieweit die Verwendung von Scoredifferenzen zur Quantifizierung von Veränderungen überhaupt sinnvoll ist, nur kurz angesprochen. Speziell für den Fall randomisierter Therapiestudien, wo eine Gleichheit der Ausgangswerte unterstellt werden kann, ist dies sehr umstritten. Die wichtigsten Argumente gegen die Verwendung von Scoredifferenzen (und stattdessen für eine einfache Auswertung der Scorewerte nach Therapie) beziehen sich auf die verminderte Reliabilität von Differenzen sowie auf mögliche Boden- oder Deckeneffekte. Eine ausführlichere Darstellung ist bei Streiner & Norman [67] sowie bei Stelzl [79] zu finden. Entscheidet man sich aus diesen Gründen gegen die Auswertung von Scoreänderungen, so wird das Konzept der Änderungssensitivität überflüssig und stattdessen spielt die diskriminative Validität die entscheidende Rolle.

Es bleibt an dieser Stelle noch zu erwähnen, daß die begriffliche Verwirrung im Zusammenhang mit der Änderungssensitivität gar nicht erst auftaucht, wenn man Scores, die allein dem Zweck des Messens von Veränderungen dienen sollen, von vornherein so konstruiert, daß in ihren Komponenten eine direkte Beurteilung der Änderung erfolgt. Wright & Feinstein [52] weisen allerdings darauf hin, daß dies auch Nachteile hat. Hierauf wird im Zusammenhang mit der Konstruktion von Scores noch eingegangen.

3.4.4 Kalibration

Im Zusammenhang mit prädiktiven Scores hat sich der Begriff der Kalibration als Gütekriterium etabliert. Die Kalibration spielt eine Rolle, wenn von Scorewerten Wahrscheinlichkeitsaussagen über ein klinisches Ereignis abgeleitet werden. Unter der Kalibration versteht man dann die Unverzerrtheit (Richtigkeit) dieser Wahrscheinlichkeitsaussagen, also die Übereinstimmung von vorhergesagten und "wahren" Wahrscheinlichkeiten. Damit ist die Kalibration eindeutig ein Aspekt der Validität, obwohl sie in manchen Publikationen auch als Zuverlässigkeit [80] bzw. "reliability" [81] bezeichnet wird.

Der Begriff der Kalibration ist jedoch nicht ganz unproblematisch. Zunächst ist zu beachten, daß für den einzelnen Patienten zwar das Auftreten des klinischen Ereignisses (in einem definierten Zeitraum) ein meist ohne Probleme zu beobachtendes Merkmal ist, jedoch nicht die "wahre" Wahrscheinlichkeit. Diese definiert sich ja erst durch die Zugehörigkeit des einzelnen Patienten zu einer bestimmten Population. Im Gegensatz zur Labormessung mit vorliegendem "gold standard" wird hier also ein Scorewert mit einem aus der Stichprobe geschätzten "wahren" Wert verglichen, und dieser Wert ist auch noch abhängig von der willkürlichen Definition einer Population. Die Möglichkeit, eine gute Kalibration überhaupt erreichen zu können, steht in gewissem Widerspruch zu der aus der Diagnostik bekannten Tatsache, daß Vorhersagewahrscheinlichkeiten (prädiktive Werte) prävalenzabhängig sind, d.h sich in Populationen mit unterschiedlicher Prävalenz zum Teil erheblich unterscheiden. Diese Tatsache ist bei prognostischen Vorhersagewahrscheinlichkeiten grundsätzlich genauso gültig, so daß die Bestätigung einer guten Kalibration eigentlich nur besagt, daß die Population, auf die der Score anzuwenden ist, hinreichend exakt beschrieben und eingehalten wurde.

Zur Populationsabhängigkeit kommt bei prognostischen Scores noch die Therapieabhängigkeit hinzu, da auch die Behandlung die Prävalenz des interessierenden Ereignisses modifiziert. Findet man in einer Evaluationsstudie zu einem Score eine schlechte Kalibration, so kann dies also auch an Charakteristika des Patientenkollektivs in der Studie liegen oder an der Qualität der Behandlung, es muß nicht zwangsläufig ein Mangel des Scores sein.

3.5 Relevanz der Gütekriterien

Nachdem die den empirischen Gütekriterien zugrundeliegenden Konzepte diskutiert wurden, soll eine vergleichende Betrachtung der Relevanz der einzelnen Gütekriterien für verschiedene Anwendungen von Scores abgeleitet werden.

Die wichtigste Erkenntnis in diesem Zusammenhang beinhaltet, daß die Reliabilität eine notwendige, aber keinesfalls eine hinreichende Voraussetzung für die Validität darstellt. Ein unreliabler Score kann somit nicht valide sein, unabhängig davon, auf welches Kriterium sich die Validität bezieht. Diese Aussage gilt nicht, wenn man Validität nur im Sinne von

Unverzerrtheit versteht. In Abschnitt 3.2.1 war jedoch bereits darauf hingewiesen worden, daß für klinische Scores die Unverzerrtheit nur selten eine Rolle spielt, da dieser Begriff nur dann Sinn macht, wenn für den Score ein Referenzkriterium mit gleicher Skalierung existiert. Als Ausnahme muß der im vorigen Abschnitt besprochene Fall der kalibrierten Prognosescores angesehen werden, für den der Zusammenhang zwischen verschiedenen Gütekriterien später noch diskutiert wird.

Die oben gemachte Aussage bedeutet aber auch, daß aus dem Nachweis einer hohen Validität, egal für welches Kriterium, automatisch ein Nachweis der Reliabilität des Scores folgt. Dieses Argument hat wohl dazu geführt, daß von vielen Forschern die Reliabilität als Gütekriterium vernachlässigt oder ganz übersehen wurde, vor allem im Bereich prädiktiver Scores, wo eine Priorität der Validität auf der Hand liegt. Diese Sichtweise ist kurzsichtig, denn sie verwechselt die Validität in der Stichprobe mit der tatsächlichen Validität. Gerade im Fall einer geringen Reliabilität kann aufgrund des Zufallseinflusses bei der Scoreerhebung besonders leicht in einer Studie fälschlich eine zu hohe Validität beobachtet werden.

Die Reliabilität spielt eine unbestrittene Rolle bei evaluativen Scores. Im Zusammenhang mit der Verwendung von Scores als Zielkriterien in Therapiestudien wurde der Einfluß mangelnder Reliabilität auf die erforderliche Patientenzahl unter anderem von Fleiss [82] beschrieben. Viel zu selten wird in Studien von der Möglichkeit Gebrauch gemacht, durch eine Erhöhung der Reliabilität der Zielvariablen mittels geeigneter Maßnahmen eine höhere statistische Power zu erreichen. Die Änderungssensitivität wird in ihrer Bedeutung für evaluative Scores überschätzt, da Scoredifferenzen wegen verschiedener statistischer Probleme nicht allzu oft als Zielgrößen verwendet werden.

Schließlich muß noch betont werden, daß die Validität nicht in jedem Fall ausreichend als Gütekriterium ist. Bei der Umsetzung von klinischen Scores in Entscheidungshilfen für therapeutisches Handeln ist eine hohe Validität zwar notwendig, aber nicht hinreichend für einen daraus ableitbaren klinischen Nutzen. Hierauf wird am Ende der Arbeit noch eingegangen.

4 STATISTISCHE GÜTEPARAMETER FÜR SCORES

In diesem Kapitel geht es um die empirische Erfassung der in Kapitel 3 dargestellten Konzepte. Hierzu sollen statistische Parameter zur Quantifizierung der Gütekriterien vorgestellt und diskutiert werden. Die Darstellung und der Vergleich dieser statistischen Parameter wird relativ viel Raum einnehmen, so daß technische Details im Zusammenhang mit der Schätzung und Testung dieser Parameter nur unter Hinweis auf die Literatur genannt werden können.

Die Darstellung beginnt mit einer kurzen Einführung des klassischen statistischen Meßmodells, das aus dem Bereich der physikalischen Messung stammt und an dem die Umsetzung der Konzepte in statistische Parameter besonders einfach klargemacht werden kann. Danach werden die Übertragbarkeit des Modells auf den Bereich des klinischen Messens diskutiert und anschließend Parameter für die Reliabilität und für die Validität vorgestellt.

4.1 Statistisches Modell des Messens

Bei der statistischen Modellierung des Messens, die einen numerischen Zugang zu den bisher besprochenen Konzepten liefern soll, beschränken wir uns auf relativ einfache Modelle, die für ein Grundverständnis der danach beschriebenen statistischen Parameter und Methoden erforderlich sind. Tatsächlich betritt man hier ein Gebiet, das für den Laien fast unüberschaubar groß und mathematisch komplex wird. Dies hat zur Folge, daß in manchen Abhandlungen zu diesem Thema (z.B. [83]) die Beziehung zwischen dem mathematischen Modell und der Praxis des Messens, also die praktische Bedeutung der Modellannahmen und die Interpretation der Modellparameter, weitgehend verlorengeht.

Das klassische statistische Modell des Messens (in der Psychologie spricht man von der klassischen Testtheorie) geht davon aus, daß es eine wahre Merkmalsausprägung gibt, den sogenannten wahren Wert T (für "true value"), der bei einer Messung bis auf einen Meßfehler E (für "error") genau bestimmt wird. Weiterhin wird unterstellt, daß theoretisch eine beliebig häufige, unabhängige Wiederholung der Messung unter identischen Bedingungen, also ohne daß die Meßergebnisse sich gegenseitig beeinflussen und ohne daß der wahre Wert T sich ändert, denkbar ist. In der Praxis ist man zufrieden, wenn man eine einigermaßen gute Annäherung an die idealen Modellvoraussetzungen plausibel machen kann, also etwa eine mehrmalige Messung unter fast gleichen Bedingungen erreicht. Selbst wenn dabei eine Konstanz von T nicht realistisch erscheint, so macht dies nichts: Solange die Änderungen des Merkmals zwischen den Messungen a priori als geringfügig angenommen werden können, läßt sich der wahre Wert T als mittlere Ausprägung des Merkmals definieren und die kleinen

Schwankungen um diesen Wert herum als Teil des zufälligen Meßfehlers, hierauf wurde bei den Ausführungen zur Test-Retest-Reliabilität schon hingewiesen.

Das einfachste statistische Modell sieht also Meßwiederholungen unter gleichen Bedingungen (Replikationen im Sinne von Abschnitt 3.3.2) an einem Objekt mit wahrem Wert T vor und lautet:

$$X_j = T + E_j$$

d.h. daß der Meßwert X_j bei der j-ten Messung sich additiv aus dem wahren Wert T und dem zufälligen Meßfehler E_j zusammensetzt. In diesem Modell ist der wahre Wert T eine Konstante und X_j und E_j sind Zufallsvariablen.

Die Annahme der Additivität ist hier im einfachsten Fall noch nicht von Bedeutung und macht erst einen Sinn, wenn weitere Spezifikationen über die Art des Meßfehlers getroffen sind. Bei den folgenden Modellen erweist sich die Additivität, abgesehen von einer gewissen empirischen Plausibilität, als vorteilhaft, weil sie eine einfache mathematische Handhabbarkeit der Modelle erlaubt. Die Spezifikation des zufälligen Meßfehlers ist von Bedeutung sowohl für die Auswahl geeigneter statistischer Parameter als auch besonders für die Formulierung und Begründung statistischer Schätz- und Testverfahren. Dabei geht man in der Regel von der Annahme der Normalverteilung zufälliger Meßfehler aus, die in vielen Anwendungsfällen empirisch überzeugend nachgewiesen werden kann. Die Parameter der Normalverteilung sind bekanntlich der Erwartungswert μ_E und die Standardabweichung σ_E (vgl. Abbildung 3).

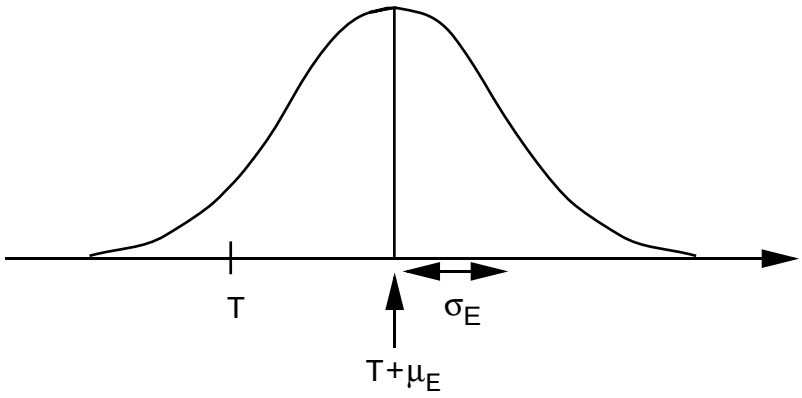

Abb. 3: Veranschaulichung des einfachsten Meßfehlermodells (mit Normalverteilung)

An diesem einfachsten Modell lassen sich die zwei Konzepte der Meßqualität besonders klar darstellen. Der Erwartungswert μ_E beschreibt hier den systematischen Meßfehler und ist somit ein Maß für die Unverzerrtheit der Meßmethode. Er spielt bei den Betrachtungen zur Reliabilität praktisch keine Rolle und wird dort als 0 angenommen. Die Standardabweichung σ_E ist der statistische Parameter für die Präzision, sie charakterisiert die Größe des zufälligen Meßfehlers und wird häufig auch als Standardmeßfehler bezeichnet. Es war oben erwähnt

worden, daß von manchen Autoren der Begriff Richtigkeit für das Zusammentreffen von Präzision und Unverzerrtheit gebraucht wird. Dies läßt sich in einem statistischen Parameter ausdrücken, den man als mittlere quadratische Abweichung vom wahren Wert definieren kann, der aber relativ ungebräuchlich ist.

Der nächste Schritt sieht eine Erweiterung des Modells auf mehrere Objekte mit wahren Werten T_i vor:

$$X_{ij} = T_i + E_{ij}$$

wobei X_{ij} der j-te Meßwert am i-ten Objekt ist, T_i der wahre Wert des i-ten Objekts und E_{ij} der Meßfehler bei der j-ten Messung am i-ten Objekt.

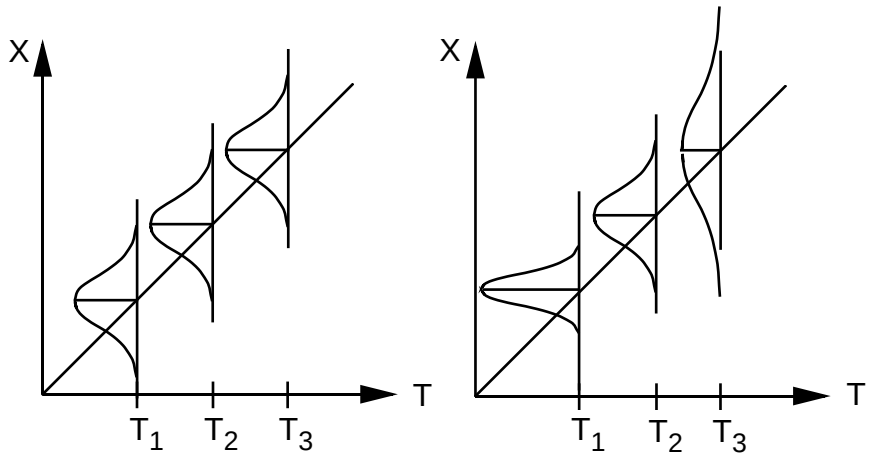

Abb. 4: Meßfehlermodells mit gleicher (links) bzw. unterschiedlicher (rechts) Meßfehlerverteilung

In diesem erweiterten Modell kann auch T_i eine Zufallsvariable sein, wenn die Objekte zufällig aus einer Grundgesamtheit ausgewählt wurden. Wenn man in diesem Modell die Begriffe Unverzerrtheit und Präzision gebraucht, so nimmt man in der Regel an, daß diese für alle Objekte gleich sind. Dies impliziert die Annahme der Gleichheit der Meßfehlerverteilungen, also speziell ihrer Varianzen, über die verschiedenen Objekte (vgl. Abb. 4 links). Als zusätzliche Annahme muß die Unabhängigkeit der Meßfehler, hier nicht nur innerhalb der Replikationen, sondern auch zwischen den Objekten, gegeben sein. Der Fall unterschiedlicher Meßfehlerverteilungen (vgl. Abb. 4 rechts) ist statistisch schwieriger zu handhaben, hierauf wird am Ende von Abschnitt 4.2.1 kurz eingegangen.

4.2 Reliabilitätsmaße

4.2.1 Reliabilitätsmaße für quantitative Merkmale

Im folgenden werden Reliabilitätsmaße vorgestellt, die quantitative Merkmale mit Intervallskalenniveau voraussetzen. In Abschnitt 3.1 war darauf hingewiesen worden, daß bei klinischen Scores diese Annahme theoretisch nicht zu begründen ist, man in der Praxis jedoch

bei gegebener Plausibilität gut damit arbeiten kann. Grundsätzlich müssen zur Bestimmung der Reliabilität (ein oder) mehrere Objekte verfügbar sein, deren wahrer Wert nicht bekannt zu sein braucht und an denen jeweils Mehrfachmessungen durchgeführt wurden. Zunächst sei von der Gültigkeit des oben beschriebenen Modells mit gleicher Meßfehlervarianz für jedes Objekt ausgegangen, man spricht in diesem Fall von Homoskedastizität. Möglichkeiten der Überprüfung dieser Annahme sowie Auswertungsstrategien im Falle des Vorliegens von Heteroskedastizität werden weiter unten beschrieben.

Man unterscheidet grundsätzlich zwei Typen von Reliabilitätsparametern. Beide beziehen sich auf die zufällige Streuung, die im einen Fall skalenabhängig und im anderen skalenunabhängig quantifiziert wird. Zunächst betrachten wir die einfachste Studiensituation, in der an n Objekten jeweils k_i unabhängige Meßwiederholungen (im Sinne von Replikationen) durchgeführt werden. Hier gilt das oben bereits erwähnte einfache Modell

$$X_{ij} = T_i + E_{ij} ,$$

in dem die Objekte als zufällig aus einer Population ausgewählt betrachtet werden. Aufgrund der über die Unabhängigkeit der Meßfehler gemachten Annahmen gilt nun $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$, d.h. daß die Varianz der Meßwerte sich additiv aus der Varianz der wahren Werte der Objekte und der Varianz des Meßfehlers zusammensetzt. Der gesuchte skalenabhängige Reliabilitäts-parameter wird durch die Standardabweichung σ_E gegeben und auch als Standardmeßfehler bezeichnet. Der Standardmeßfehler entspricht der Standardabweichung der Meßfehlerverteilungen in der linken Graphik von Abbildung 4.

Für jedes einzelne Objekt erhält man eine Schätzung für den Standardmeßfehler aus der empirischen Standardabweichung s_i der Replikationen an diesem Objekt, wobei

$$S_i = \sqrt{\frac{1}{k_i - 1} \cdot \sum_{j=1}^{k_i} (x_{ij} - \bar{x}_i)^2} .$$

Die geschätzten s_i an n einzelnen Objekten lassen sich zu einer gepoolten Schätzung s_E zusammenfassen. Dabei ist

$$S_E = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} S_i^2}$$

der Schätzer für den Standardmeßfehler, wenn die Anzahl der Replikationen pro Objekt gleich ist ($k_i = k$ für alle i), andernfalls muß ein gewichteter Schätzer berechnet werden.

Der Vorteil des Standardmeßfehlers besteht in einer recht anschaulichen Interpretationsmöglichkeit. Geht man trotz der diskreten Natur der Scorewerte von der annähernden Gültigkeit eines Normalverteilungsmodells für den zufälligen Meßfehler aus, so weiß man, daß bei wiederholter Messung eines Patienten mit wahrem Wert μ etwa 68% der Meßwerte im Bereich $\mu \pm \sigma_E$ und etwa 95% der Meßwerte im Bereich $\mu \pm 2 \cdot \sigma_E$ liegen.

Umgekehrt läßt sich auf diese Weise ein Konfidenzintervall für den wahren Wert eines Patienten bei gegebenem Meßwert x berechnen Das übliche Konfidenzintervall mit einer Vertrauenswahrscheinlichkeit von 95% ist annähernd das Intervall

$$(x - 2 \cdot s_E, x + 2 \cdot s_E),$$

es beschreibt die Unsicherheit der Schätzung des wahren Wertes aufgrund des beobachteten Scorewerts. Ganz ähnlich kann man den Standardmeßfehler anwenden, wenn von einem Patienten zwei zeitlich versetzte Scorewerte x_1 und x_2 vorliegen und man sich dafür interessiert, ob die Differenz der Werte im Rahmen des Meßfehlers (z.B. Intra-Rater-Variation) liegt oder ob man von einer realen Veränderung des Merkmals ausgehen kann. Das entsprechende 95%-Konfidenzintervall für die Differenz der wahren Werte zweier Messungen ist annähernd

$$(x_1 - x_2 - 2 \cdot \sqrt{2} \cdot s_E, x_1 - x_2 + 2 \cdot \sqrt{2} \cdot s_E).$$

Die halbe Breite dieses Konfidenzintervalls (also ca. 2.83 s_E) wird manchmal auch als kritische Differenz bezeichnet.

Der Standardmeßfehler ist als Reliabilitätsmaß deshalb gut geeignet, weil er auf die Skaleneinheit der Messung bezogen ist und, wie gerade gezeigt, leicht interpretierbar ist. Der Hauptnachteil des Standardmeßfehlers, vor allem bei Anwendungen im Zusammenhang mit klinischen Skalen, ist eng mit diesem Vorteil verknüpft. Bei den dort untersuchten Merkmalen gibt es nämlich in der Regel keine "natürliche" Meßeinheit, so daß jedes Meßinstrument typischerweise seine eigene Skalierung besitzt. Damit sind aber verschiedene Skalen nicht mehr anhand des Standardmeßfehlers bezüglich ihrer Reliabilität direkt vergleichbar. Hierzu ist eine Standardisierung erforderlich, die diese Skalenabhängigkeit beseitigt, indem man die zufällige Streuung relativ ausdrückt. Dies erreicht man, indem man die Fehlervarianz σ_E^2 auf die beobachtete Gesamtvarianz σ_X^2 der Meßwerte bezieht. Der Ausdruck

$$ICC = 1 - \frac{\sigma_E^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_X^2}$$

wird als Intraklass-Korrelationskoeffizient (ICC) bezeichnet. Die Schätzung des ICC ergibt sich in diesem einfachen Fall direkt aus der Schätzung des Standardmeßfehlers und der Standardabweichung des Merkmals in der Stichprobe. Da sich die Varianz der Meßwerte additiv aus Fehlervarianz und Varianz der wahren Werte zusammensetzt, entspricht der ICC auch dem relativen Anteil der Varianz der wahren Werte an der Varianz der Meßwerte.

Der Vorteil des Intraklass-Korrelationskoeffizienten, als relatives, skalenunabhängiges Maß einen Vergleich der Reliabilität verschiedener Scores zu ermöglichen, ist aber wiederum mit einem Nachteil eng verknüpft. Die Standardisierung durch Bezug auf die Merkmalsvarianz in der Stichprobe bzw. Population macht dieses Reliabilitätsmaß abhängig von der Zusammensetzung derselben. So ist zum Beispiel sofort ersichtlich, daß ein und derselbe

Score mit vorgegebenem Standardmeßfehler bei Einschränkung auf eine homogenere Population (kleineres σ_T^2 und somit kleineres σ_X^2) automatisch einen niedrigeren ICC zeigt. Umgekehrt läßt sich auf diese Weise auch ein eindrucksvoller ICC gewinnen, wenn man eine extrem heterogene Stichprobe für die Untersuchung auswählt. Intraklass-Korrelationskoeffizienten, die aus Studien an unterschiedlichen Populationen gewonnen wurden, lassen sich daher nicht miteinander vergleichen.

Falls in einer Publikation zu einem Score nur der Intraklass-Korrelationskoeffizient angegeben ist, so kann man bei gegebener Standardabweichung des Scores aus den obigen Formeln leicht eine Umrechnung auf den Standardmeßfehler ableiten, diese lautet:

$$\sigma_E = \sigma_X \cdot \sqrt{1 - ICC}$$

Die Berechnung der Reliabilitätsparameter war soweit elementar und es wurde angenommen, daß sich die Varianzen direkt aus der Stichprobe schätzen lassen. Für komplexere Modelle lassen sich hierfür keine einfachen Formeln mehr angeben und man muß die Berechnungen mit Hilfe einer statistischen Verfahrensklasse, der Varianzkomponentenanalyse, durchführen. Dabei wird die gesamte Streuung der Meßwerte entsprechend der Modellspezifikation in Varianzkomponenten zerlegt und der Einfluß jeder Komponente quantifiziert. Auf eine detaillierte Darstellung dieser Methode muß hier verzichtet werden, stattdessen sei auf die einschlägige Fachliteratur [84, 85] hingewiesen. In der Regel wird man die Berechnungen ohnehin mit Hilfe eines statistischen Auswertungsprogrammes durchführen, beispielsweise mit der Prozedur VARCOMP des statistischen Auswertungssystems SAS.

Das eben vorausgesetzte Modell ist den meisten Reliabilitätsuntersuchungen an Scores nicht angemessen, da die Meßwiederholungen nur selten wirkliche Replikationen sind. Entstehen die Meßwiederholungen durch eine erneute Messung im zeitlichen Abstand, so hat man es bereits mit einem Faktor zu tun, dem möglicherweise auch systematische Effekte zugeschrieben werden können. Diese können entweder durch die Änderung des Merkmals selbst oder durch eine Änderung in der Beurteilung (Lerneffekte) bedingt sein. Werden die Meßwiederholungen durch die Einbeziehung verschiedener Beurteiler realisiert, so ist erst recht von systematischen Effekten auszugehen. Echte Replikationen liegen allerdings dann vor, wenn jeder Patient von unterschiedlichen Personen beurteilt wurde.

Im folgenden wollen wir die Behandlung der im Zusammenhang mit klinischen Scores wichtigsten Reliabilitätsarten getrennt diskutieren.

1.) Inter-Rater-Reliabilität: In diesem Fall wird die dem Score zugrundeliegende Beurteilung pro Patient durch mehrere Beurteiler durchgeführt. Sind diese für alle Patienten unterschiedlich, so haben wir es, wie gerade angemerkt, mit Replikationen zu tun und das obige Modell und die genannten Formeln gelten. Häufig findet man jedoch den Fall, daß alle Patienten von den gleichen zwei oder mehr Personen beurteilt wurden. Ist dies der Fall, dann besteht die Streuung zwischen den Meßwiederholungen aus einem zufälligen und einem systematischen Anteil. Hier gilt das um einen Faktor erweiterte Modell, das den Beurteiler

(Rater) mit einbezieht. Dies lautet im üblichen Fall, d.h. wenn nur ein Meßwert pro Rater und Objekt vorliegt,

$$X_{ij} = T_i + R_j + E_{ij}$$

In den meisten Fällen ist es sinnvoll anzunehmen, daß die Beurteiler zufällig aus einer Population geeigneter Personen ausgewählt wurden, also der Rater-Faktor einen zufälligen Effekt (vgl. Abschnitt 3.3.2) darstellt. Die Annahmen über Unabhängigkeit und Homoskedastizität der Meßfehler müssen mit zunehmender Komplexität des Modells verschärft werden. Wenn diese erfüllt sind, so kann für dieses Modell eine Varianzkomponentenzerlegung wie folgt vorgenommen werden:

$$\sigma_X^2 = \sigma_T^2 + \sigma_R^2 + \sigma_E^2$$

Der Standardmeßfehler setzt sich nun aus zwei Komponenten zusammen, nämlich der Variation zwischen den Beurteilern und der innerhalb der Beurteiler, beide Komponenten können getrennt geschätzt werden. Die oben dargestellte, sehr einfache Berechnung der beiden vorgestellten Reliabilitätsparameter gilt auch weiterhin, wenn man den Ratereffekt als Teil des zufälligen Meßfehlers auffaßt. Das absolute Reliabilitätsmaß ist dann der erweiterte Standardmeßfehler $\sqrt{\sigma_R^2 + \sigma_E^2}$, der geschätzt werden kann als $\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} S_i^2}$,und das relative Maß ist $ICC = 1 - \frac{\sigma_R^2 + \sigma_E^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_X^2}$.

Einen geringfügig anderen Reliabilitätsindex schlägt Lin [86] vor, der jedoch nur minimal vom ICC abweicht [87]. Einen Überblick über diese und weitere, auch nichtparametrische Alternativen gibt Müller [88].

Es sei noch einmal betont, daß die zufällige Auswahl der Rater hier durchaus eine wichtige Rolle spielt, da man nur in diesem Falle den Einfluß der Beurteiler als Teil des zufälligen Meßfehlers auffassen kann. Der Standardmeßfehler ist damit auch auf andere, zufällig aus der gleichen Population ausgewählte Beurteiler übertragbar. Hierauf wird im Abschnitt zur Versuchsplanung von Reliabilitätsstudien nochmals eingegangen.

2.) Test-Retest-Reliabilität: In einer Testwiederholungsstudie wird der Score an allen Patienten zwei- oder mehrmals erhoben. In der Regel ist die Zahl der Wiederholungen für alle Patienten gleich und die Zeitpunkte der Erhebung, bzw. ihr Abstand, sind relativ klar definiert. Dabei wählt man den Zeitraum so, daß größere, insbesondere systematische Änderungen des Scores nicht anzunehmen sind. Im Gegensatz zum Fall mehrerer Beurteiler ist hier der systematische Unterschied zwischen den Meßwiederholungen nicht unbedingt als Teil des zufälligen Fehlers zu sehen, die Meßwiederholung ist fast immer ein fester Faktor des Studienplans. Die systematischen Effekte dieses Faktors lassen sich bei Anwendung des Scores durch Adjustierung berücksichtigen, sie tragen also nicht zur Meßungenauigkeit bei.

Dieser Fall liegt zum Beispiel vor, wenn man einen Score kurz nach Aufnahme in die Klinik erhebt sowie 24 Stunden später noch einmal. Ergibt sich dabei eine systematische Abnahme um einen bestimmten Betrag, so läßt sich dies bei der praktischen Anwendung des Scores leicht in Rechnung stellen.

Die eben beschriebene Situation war durch eine Vermischung von Reliabilität des Instruments und Stabilität des Merkmals charakterisiert. Dient der Score zur Erfassung der mittleren Ausprägung eines Merkmals in einem Zeitintervall, so wird man die zufälligen Schwankungen um diesen Mittelwert in den zufälligen Meßfehler einbeziehen. Auch die systematischen zeitlichen Einflüsse sind in der Anwendung des Scores relevant, können jedoch direkt berücksichtigt werden. Im folgenden Fall gibt es ebenfalls systematische Zeiteffekte, die aber für die Praxis gar nicht von Relevanz sind, da sie nur aufgrund des Studiendesigns zustandekommen.

Man denke dazu an den Fall, in dem nicht die Fluktuaktion des Merkmals interessiert, sondern nur die alleinige Reliabilität des Meßinstruments, man aber dennoch eine Studie mit Meßwiederholung, diesmal mit besonders engem zeitlichen Abstand, durchführen muß, weil eine andere Art der Reliabilitätsmessung nicht möglich ist (z.B. bei Selbsteinschätzungsfragebögen). Auch hier muß man mit systematischen Effekten rechnen, die aber nicht die Veränderung des Merkmals, sondern eher eine gewisse Adaptation an den Meßvorgang, z.B. Lern- oder Motivationseffekte, widerspiegeln. Diese Effekte treten nur in der Studiensituation auf, in der Praxis der Anwendung des Scores jedoch nicht, sie sollen daher keinen (negativen) Beitrag zur Reliabilität liefern.

Die beiden dargestellten Anwendungssituationen lassen sich durch ein ähnliches Modell wie im Falle der Inter-Rater-Reliabilität beschreiben, wobei der Faktor R nun nicht für den Beurteiler, sondern für die Meßwiederholung steht. Will man den systematischen Effekt des Meßwiederholungsfaktors aus den genannten Gründen ignorieren, so ist der Intraklass-Korrelationskoeffizient hier

$$ICC = 1 - \frac{\sigma_E^2}{\sigma_E^2 + \sigma_T^2} = \frac{\sigma_T^2}{\sigma_E^2 + \sigma_T^2}.$$

Trägt die Meßwiederholung jedoch in rein zufälliger Weise zur Messung bei, indem beispielsweise pro Patient unterschiedliche Zeitpunkte gewählt wurden, so gilt die gleiche Darstellung für den ICC wie im Falle der Inter-Rater-Reliabilität mit zufällig ausgewählten Beurteilern. Ganz ähnliche Überlegungen gelten für die Beurteilung der Intra-Rater-Reliabilität, diese Situation entspricht logisch einer Testwiederholungsstudie.

Eine noch größere Komplexität von Reliabilitätsuntersuchungen, etwa durch die Kombination von Meßwiederholungen und Beurteilern, läßt sich durch eine entsprechende Erweiterung des Modells beschreiben. Zwar werden die Voraussetzungen an die Durchführung der Studie damit immer strenger, aber gleichzeitig gewinnt man die Möglichkeit, die verschiedenen

Einflußquellen von Meßfehlern (im weitesten Sinne) in ihrem Zusammenwirken zu erfassen und zu quantifizieren. Dies spielt in anderen Anwendungsbereichen, etwa in der industriellen Qualitätskontrolle oder bei der Durchführung von Ringversuchen in verschiedenen klinisch-chemischen Labors, eine große Rolle. Im Zusammenhang mit klinischen Skalen gibt es hierfür bisher aber nur wenige Anwendungsbeispiele [68, 69]. Eine praxisnahe Beschreibung verschiedener komplexerer Situationen und der dazugehörigen Modelle findet sich bei Healy [57].

In Abschnitt 3.3.3 war ein weiteres Reliabilitätskonzept kurz erwähnt worden, die interne Konsistenz bzw. Testhalbierungs-Reliabilität, das in der psychometrischen Literatur eine wichtige Rolle spielt, jedoch allgemein bei Scores wenig verwendet wird. Das Konzept der internen Konsistenz basiert auf der inhaltlichen Homogenität der Komponenten eines Scores und ist deshalb bei prädiktiven Scores, die meist absichtlich sehr unterschiedliche Prognosevariablen einbeziehen, völlig inadäquat. Will man jedoch eine inhaltlich homogene Skala bilden, um etwa einen bestimmten Aspekt der Lebensqualität zu beschreiben, so kann das Ausmaß der Gemeinsamkeit der Items durch einen statistischen Parameter beschrieben werden. Dieser Koeffizient soll hier nur kurz genannt werden, es handelt sich um das Cronbach´sche α. Der Parameter läßt sich anschaulich wie folgt interpretieren: Wenn man die Skala auf alle möglichen Arten anhand der Items in zwei gleich große Hälften aufteilt und jeweils die Testhalbierungs-Reliabilität als Korrelation dieser zwei Hälften berechnet, so entspricht der Parameter α dem Mittelwert aller dieser Korrelationskoeffizienten [67].

In der Literatur findet man gelegentlich Beispiele von Studien [20], bei denen als Parameter für die Reliabilität der einfache Pearson´sche Korrelationskoeffizient r bzw. sein p-Wert berechnet werden. Dies ist eine ungeeignete Art und Weise, die Reliabilität eines Scores zu belegen. Der Korrelationskoeffizient beschreibt nicht die tatsächliche Übereinstimmung, sondern nur den linearen Zusammenhang zwischen zwei Meßwiederholungen oder Beurteilern. Er ignoriert also systematische Effekte dieses Faktors, beziehungsweise geht davon aus, daß diese gleich Null sind und dies nicht nur im Mittel. Der p-Wert schließlich hängt außerdem noch von der zugrundeliegenden Fallzahl ab und ist somit als Reliabilitätsmaß völlig ungeeignet.

Bevor man auf die beschriebene Weise die Reliabilität eines Scores ermittelt, sollte man prüfen, ob das Modell mit gleicher Meßfehlerverteilung überhaupt passend ist. Dies geschieht am einfachsten mittels einer Graphik, bei der man für jedes Objekt den Mittelwert der Meßwiederholungen als Abszisse und die Standardabweichung (bei Doppelmessungen einfach die Differenz) als Ordinate abträgt. Falls das Modell gilt, so sollten die Punkte um eine Gerade herum verteilt sein, die in etwa parallel zur Abszisse liegt. Die in der Praxis am häufigsten auftretende Abweichung zeigt eine Streuung der Punkte um eine ansteigende Gerade herum, man nennt dies den Fall des proportionalen Meßfehlers.

Bei proportionalem Meßfehler, wenn also σ_E im gleichen Verhältnis wie der wahre Wert T zunimmt, ist der Variationskoeffizient ein möglicher skalenunabhängiger Parameter für die Reliabilität:

$$VK = \frac{\sigma_E}{T}$$

Er ist in manchen Bereichen, z.B. in der klinischen Chemie, gebräuchlicher ist als der Standardmeßfehler. Die obige Definition bezieht sich nur auf Meßwiederholungen an einem Objekt. Für eine zusammengefaßte Schätzung aus mehreren Objekten wird oft ein pragmatischer, aber theoretisch nicht optimaler Weg beschritten, indem die Schätzungen an den einzelnen Objekten in Form eines Mittelwerts oder Medians zusammengefaßt werden. Besser ist es, den Umweg über die logarithmische Transformation der Scorewerte zu gehen und für diese die gepoolte Schätzung des Standardmeßfehlers berechnen [89].

4.2.2 Reliabilitätsmaße für dichotome Merkmale

Der im folgenden behandelte Fall der Reliabilitätsmessung für dichotome Merkmale spielt für Scores nur eine untergeordnete Rolle, etwa wenn sie dichotome Komponenten haben, deren Reliabilität untersucht und verglichen werden soll. Die entsprechenden statistischen Reliabilitätsparameter werden hier aber auch deshalb besprochen, weil sich am dichotomen Fall der Grundgedanke der zufallskorrigierten Konkordanzmaße besonders gut einführen läßt, der dann im folgenden Abschnitt auf ordinale Merkmale übertragen wird.

Zunächst wird der einfachste Fall betrachtet, in dem mehrere Objekte vorliegen, deren wahrer Wert nicht bekannt sein muß, und für die je zwei Messungen vorgenommen wurden, etwa im Rahmen eines Inter-Rater-Vergleiches. Die erhobenen Daten lassen sich als Vierfeldertafel darstellen, wobei die Zeilen durch Messung A und die Spalten durch Messung B definiert werden.

	B=0	B=1
A=0	a	b
A=1	c	d

Das intuitiv einleuchtendste Maß für die Reliabilität, hier auch als Konkordanz bezeichnet, ist die Übereinstimmungsrate

$$\ddot{U}R = \frac{a+d}{a+b+c+d}\ .$$

Diese hat allerdings den Nachteil, daß sie auch bei rein zufälligem Urteilsverhalten der Rater von Null verschiedene, fast beliebig nahe an Eins liegende Werte annehmen kann. Die folgenden zwei Beispiele zeigen Vierfeldertafeln, die man erhält, wenn man sich jeweils zwei unabhängig voneinander urteilende Rater vorstellt, die im ersten Fall eine Münze mit p=0.5 und im zweiten mit p=0.9 werfen

25	25
25	25

81	9
9	1

Die Übereinstimmungsraten betragen im ersten Beispiel 50% und im zweiten sogar 82%, obwohl beiden Fällen nur Zufallsmechanismen zugrundeliegen. Es wurde daher ein zufallskorrigiertes Maß für die Konkordanz gesucht, das bei optimaler Übereinstimmung den Wert 1 annimmt und bei rein zufälliger Übereinstimmung den Wert 0. Ein solches Maß ist der Kappa-Koeffizient [90], der berechnet wird als

$$\kappa = \frac{\ddot{U}R_b - \ddot{U}R_e}{1 - \ddot{U}R_e}$$

wobei $\ddot{U}R_b$ die beobachtete Übereinstimmungsrate und $\ddot{U}R_e$ die bei unabhängiger Beurteilung erwartete Übereinstimmungsrate ist. Die Übertragung der Definitionen auf den Fall von mehr als zwei Beurteilern, deren Zahl bzw. sogar deren Zusammensetzung pro Objekt noch unterschiedlich sein dürfen, ist möglich und wird zum Beispiel bei Bortz [91] beschrieben.

Eine Interpretation des Kappa-Wertes ist schwierig, da es sich nicht mehr um den Schätzwert für eine Wahrscheinlichkeit handelt. Häufig stützt man sich auf die folgende willkürliche Einteilung von Landis & Koch [92]: 0.21-0.40: "fair" (schwach), 0.41-0.60: "moderate" (mittel), 0.61-0.80: "substantial" (gut), 0.81-1.00: "almost perfect" (sehr gut).

Der Kappa-Koeffizient ist der am häufigsten benutzte Parameter zur Beschreibung der Übereinstimmung zweier Beurteiler bezüglich eines dichotomen Merkmals. Daß er hier so ausführlich dargestellt wird, ist zum einen darin begründet, daß er in vielen Lehrbüchern der Medizinischen Statistik nicht erwähnt wird und zum anderen, weil er durchaus nicht unumstritten ist. Hier besteht, ähnlich wie in anderen Bereichen der Biometrie, die Situation, daß für ein bestimmtes Konzept mehrere Parameter existieren, von denen keiner allen Anforderungen gerecht wird. Die Kritik am Kappa-Koeffizienten haben Feinstein & Cicchetti [93, 94] in sehr einleuchtender Weise in Form zweier Paradoxa beschrieben. Diese lassen sich an den folgenden Beispielen illustrieren.

Angenommen, zwei Beurteiler würden eine Stichprobe von 20 Objekten bezüglich des Vorliegens eines dichotomen Merkmals bewerten. Die beiden folgenden Vierfeldertafeln geben zwei mögliche Ergebnisse wieder:

10	1
0	9

18	1
0	1

Während nun in beiden Fällen eine Übereinstimmung in 19 von 20 Fällen (entspricht ÜR = 95%) erreicht wurde, beträgt der Kappa-Koeffizient im ersten Fall 0.9 und im zweiten nur 0.64. Manche Autoren sehen dies als Effekt der Prävalenz auf Kappa an, obwohl hier über die wahre Prävalenz gar keine Aussage gemacht wird.

Ein zweites Paradoxon wird durch das folgende Beispiel verdeutlicht:

9	3
5	3

5	7
1	7

Wieder sind zwei Situationen mit gleicher Übereinstimmungsrate (12 von 20 = 60%) gegeben. Hier hat der Kappa-Koeffizient im ersten Fall (links) den Wert 0.13 und im zweiten (rechts) 0.26, obwohl im letzteren Fall die Beurteiler in den Randverteilungen wesentlich deutlicher divergieren.

Das Zustandekommen dieses und ähnlicher Paradoxa kann darauf zurückgeführt werden, daß bei der Berechnung des Kappa-Koeffizienten auf die Randverteilung der beiden Beobachter bedingt wird, d.h. daß in dieser Randverteilung keine Information zur Übereinstimmung gesehen wird. Dies läßt sich motivieren durch die Unterstellung, daß schlimmstenfalls die beiden Beurteiler ihre Urteile jeweils durch einen Zufallsprozeß mit einer bestimmten Wahrscheinlichkeitsverteilung (Wahrscheinlichkeit p für Urteil 1, bzw. 1-p für Urteil 0) fällen. Bei einer weniger skeptischen Sichtweise kann man jedoch auch die Ähnlichkeit der Randverteilungen schon als Indiz für eine Konkordanz der Beurteiler werten, was jedoch durch den Kappa-Koeffizienten nicht berücksichtigt wird.

Feinstein & Cicchetti [94] kritisieren das hier unterstellte Modell, daß jeder Beurteiler mit einer a-priori festgelegten Wahrscheinlichkeitsverteilung an die Beurteilung geht. Auch Byrt et al. [95] greifen diese Paradoxa auf, die sie als Prävalenzabhängigkeit und Biaseffekt bezeichnen und schlagen zwei korrigierte Versionen von Kappa vor. Dabei ergibt sich, daß der resultierende prävalenz- und biasadjustierte Kappa-Koeffizient (PABAK) fast identisch mit der einfachen Übereinstimmungsrate ist. Cicchetti & Feinstein [93] zeigen, daß die beschriebenen Paradoxa durch kein anderes globales Konkordanzmaß behoben werden. Sie schlagen daher vor, daß zusätzlich zum Kappa-Koeffizienten auch stets zwei weitere

Parameter berechnet werden, die als positive bzw. negative Übereinstimmungsrate bezeichnet werden könnten.

Dieser Vorschlag ist bisher aber kaum auf Akzeptanz gestoßen, so daß man weiterhin die Berechnung des Parameters Kappa als "state-of-the-art" ansehen kann. Es sollten in Publikationen jedoch stets entweder die Übereinstimmungsrate oder die Originalwerte in Form einer Vierfeldertafel mitpräsentiert werden. Festzuhalten bleibt aber auch, daß der Kappa-Koeffizient von der Zusammensetzung der zu beurteilenden Stichprobe mit beeinflußt wird, er also tatsächlich auch abhängig von der wahren Prävalenz ist. Dies macht es schwierig, Kappa-Koeffizienten zu vergleichen, die an unterschiedlichen Stichproben ermittelt wurden.

Die Abhängigkeit des Kappa-Koeffizienten von der wahren Prävalenz läßt sich in einem Modell untersuchen, das die Kenntnis des wahren Wertes sowie der Validitätsparameter der Beurteiler beinhaltet. Guggenmoos-Holzmann [96] zeigt in diesem Rahmen ein Analogon des Simpson´schen Paradoxons auf. Sie konstruiert dazu ein Beispiel, bei dem in den Teilpopulationen der Patienten, die die zu beurteilende Eigenschaft tatsächlich haben bzw. nicht haben, beide Beurteiler (oder Tests) völlig unabhängig voneinander urteilen (Kappa = 0), und in der Gesamtpopulation dennoch ein Kappa-Wert von 0.48 zustande kommt. Als möglicher statistischer Ansatz zur Lösung dieser Probleme wird die "Latent-Class-Analyse" genannt, die allerdings höhere Anforderungen an das Studiendesign stellt (vgl. hierzu auch Walter & Irwig [97]).

4.2.3 Reliabilitätsmaße für ordinale Merkmale

Die Übertragung der eingeführten Begriffe (Übereinstimmungsrate und Kappa-Koeffizient) auf den Fall mehrkategorieller Merkmale ist ohne weiteres möglich. Bei der Konkordanzbeurteilung wird dann allerdings unterstellt, daß eine Nichtübereinstimmung immer gleich zu bewerten ist, egal welche Kategorien davon betroffen sind. Ist diese Annahme nicht adäquat, z.B. im Falle von ordinalen Merkmalen, so läßt sich als Verallgemeinerung ein gewichteter Kappa-Koeffizient [98] definieren. Dabei wird jedem Feld der Kontingenztafel ein Gewicht w_{ij} zugewiesen, wobei die Gewichte maximal als 1 (in der Diagonalen) und minimal als 0 gewählt werden.

Für den Fall ordinaler Merkmale gibt es unterschiedliche Vorschläge für die Gewichtswahl, z.B.

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2} \qquad \text{oder} \qquad w_{ij} = 1 - \frac{|i-j|}{k-1}.$$

Die entsprechenden Werte sind beispielsweise für ein ordinales vierstufiges Merkmal :

	j=1	2	3	4
i=1	1	0.89	0.56	0
2	0.89	1	0.89	0.56
3	0.56	0.89	1	0.89
4	0	0.56	0.89	1

	j=1	2	3	4
i=1	1	0.67	0.33	0
2	0.67	1	0.67	0.33
3	0.33	0.67	1	0.67
4	0	0.33	0. 67	1

Zur Berechnung des gewichteten κ-Koeffizienten setzt man nun

$$\ddot{U}R_b = \sum_{i,j} w_{ij} \cdot r_{ij} \quad \text{und} \quad \ddot{U}R_e = \sum_{i,j} w_{ij} \cdot r_{i.} \cdot r_{.j} \ ,$$

wobei r_{ij} die relativen Häufigkeiten in der Konkordanzmatrix und die $r_{i.}$ und $r_{.j}$ die relativen Häufigkeiten der Urteilskategorien für Rater A und B.

Der gewichtete Kappa-Koeffizient wird in Reliabilitätsstudien zu klinischen Scores verwendet, wenn die Zahl der möglichen Werte des Scores nicht zu hoch ist. Ein gleichzeitiger Vor- und Nachteil des gewichteten Kappa ergibt sich aus der willkürlichen Wahl der Gewichte. Der Vorteil besteht in der Möglichkeit der Berücksichtigung einer spezifischen Einschätzung der verschiedenen Fehlklassifikationen durch den Untersucher, der Nachteil ist die eben dadurch bedingte Subjektivität und Manipulierbarkeit. Von Cohen [98] wurde gezeigt, daß der gewichtete Kappa-Koeffizient mit quadratischen Gewichten genau einem Spezialfall des Intraklass-Korrelationskoeffizienten entspricht. Dies wird von vielen Autoren als Argument für die Verwendung dieser Gewichte genannt, während MacLure & Willett [99] hingegen aus dem gleichen Grund das gewichtete Kappa für generell überflüssig halten.

Es sei zum Schluß dieses Abschnitts betont, daß der ungewichtete Kappa-Koeffizient für ordinale Merkmale, also mithin für klinische Scores, völlig ungeeignet ist, weil er das Ausmaß der Nichtübereinstimmung nicht adäquat berücksichtigt. Dennoch ist eine solche inadäquate Auswertung in manchen Publikationen zu finden (z.B. [20, 49]).

4.3 Schätz- und Testverfahren zur Reliabilität

Statistische Testverfahren sowie Konfidenzintervalle für Reliabilitätsmaße werden in Studien selten berichtet. Die entsprechenden Verfahren sind weder in den gängigen Statistik-Auswertungssystemen enthalten noch lassen sie sich leicht in der einschlägigen Literatur finden.

Eine Ausnahme zum letzteren stellt der Kappa-Koeffizient dar, für den Fleiss [100] asymptotisch gültige Standardfehler angibt, mit deren Hilfe Konfidenzintervalle berechnet

werden können. Diese werden auch in einigen Studien angegeben und manche Autoren gehen sogar so weit, p-Werte eines entsprechenden statistischen Tests mitzuteilen. Dies ist zwar technisch richtig, wenn aufgrund der Fallzahlen die Gültigkeit der approximativen Schätzung des Standardfehlers gewährleistet ist, jedoch inhaltlich meist nicht relevant, da als Nullhypothese die Annahme völlig unabhängiger Beurteilungen gewählt wird. Eindrucksvolle p-Werte werden selbst bei mäßigem Kappa und mittleren Fallzahlen erreicht (z.B. $p < 0.00001$ bei $\kappa = 0.35$ und $n = 46$ bei Tesseris et al. [101]), jedoch besagt ein solches Ergebnis nur, daß die Übereinstimmung besser als bei zufälligem Raten ist.

Stattdessen sollte ein Test gegen einen vorgegebenen Wert durchgeführt werden, mit dem man prüft, ob die Reliabilität eine bestimmte Schranke nicht unterschreitet. Hierzu muß man allerdings eine Vorstellung haben, welche Werte des Reliabilitätsmaßes noch als akzeptabel anzusehen sind. In diesem Zusammenhang wird häufig auf Empfehlungen aus der Literatur, z.B. die oben erwähnte Einteilung von Landis & Koch [92] zurückgegriffen. Es fragt sich, ob aufgrund der Populationsabhängigkeit solche Vorgaben überhaupt sinnvoll sein können. Zumindest muß man eine zufällige oder repräsentative Auswahl aus einer wohldefinierten, relevanten Grundgesamtheit voraussetzen. Auch hängt die Strenge der Anforderung von der Verwendung des Scores ab, insbesondere ob er auf den einzelnen Patienten im Rahmen der Diagnose oder Prognose eingesetzt werden soll oder ob nur Aussagen über Patientengruppen im Rahmen evaluativer Studien erfolgen sollen.

Die statistischen Verfahren zur Konstruktion von Konfidenzintervallen und Tests für Reliabilitätsmaße in Form von Varianzkomponenten sind anspruchsvoll und nicht sehr verbreitet. Nur für den Fall von k Meßwiederholungen an einem Objekt lassen sich einfache Tests angeben für den Vergleich einer Varianz mit einem vorgegebenen Wert oder für den Vergleich zweier Varianzen. Unter den üblichen Annahmen stehen hierzu ein Chi-Quadrat-Test bzw. ein F-Test zur Verfügung [57, 102].

Einen Test zum Vergleich eines Intraklass-Korrelationskoeffizienten mit einem vorgegebenen Wert für den Fall einer Studie mit n Replikationen an k Objekten geben Donner & Eliasziw [103] an und beschreiben die Power dieses Tests für verschiedene Werte von n und k. Für komplexere Designs sei auf das Buch von Burdick & Graybill [84] verwiesen, das sich ganz dieser statistischen Problematik widmet. Dunn [85] beschreibt die Möglichkeit, mit Hilfe von Bootstrap-Methoden die statistische Genauigkeit von Reliabilitätsparametern zu bestimmen.

4.4 Validitätsmaße

Bei der Darstellung der verschiedenen Validitätskonzepte war bereits angedeutet worden, daß nur die Kriteriumsvalidität hinreichend scharf definierbar ist, um eine quantitative Beschreibung mittels eines statistischen Güteparameters zu erlauben. Für die Bestimmung der Validität eines Scores wird also im folgenden stets angenommen, daß es entweder eine quantitative Referenzmethode für das zu messende Merkmal gibt oder daß ein

Außenkriterium existiert, für das der zu untersuchende klinische Score möglichst genaue Vorhersagen ermöglichen soll. Zunächst wird in Abschnitt 4.4.1 der Fall einer quantitativen Referenz- oder Vergleichsmethode behandelt. Er hat in der Praxis vor allem dann eine Bedeutung, wenn es bereits einen etablierten Score mit der gleichen Zielsetzung gibt. In der Regel ist dies aber keine echte Referenzmethode in dem Sinne, daß sie a-priori eine exaktere oder validere Messung des interessierenden Merkmals darstellt.

Ausführlich werden wir uns dann im folgenden Abschnitt 4.4.2 mit der Validität in Bezug auf ein dichotomes Außenkriterium befassen. Solche Validitätsuntersuchungen spielen eine große Rolle bei prädiktiven Scores, die zur Vorhersage eines gut meßbaren, aber in der Zukunft liegenden klinischen Ereignisses konstruiert wurden. Bezieht man die Zeit bis zum Eintreten des klinischen Ereignisses in die Betrachtung mit ein, so hat man grundsätzlich wieder ein quantitatives Außenkriterium, das sich aber vom obigen Fall dadurch unterscheidet, daß es in der Regel nicht bei allen Patienten beobachtet werden kann. Man spricht hier von zensierten Zeitvariablen (z.B. Überlebenszeit). Auf Validitätsmaße für solche Außenkriterien wird in Abschnitt 4.4.3 eingegangen.

4.4.1 Validitätsmaße für quantitative Kriterien

Als Ausgangspunkt wählen wir den Fall eines quantitativen Kriteriums, da sich hier die Parallelität zum Konzept der Richtigkeit eines physikalischen Meßverfahrens als hilfreich erweist. Es sind sowohl in jenem Bereich als auch bei klinischen Scores drei Fälle zu unterscheiden: 1.) das Kriterium ist als absolute oder relative Standardmethode zur Messung des gleichen Merkmals akzeptiert, 2.) das Kriterium ist eine (ebenfalls fehlerbehaftete) Vergleichsmethode zur Messung des gleichen Merkmals, 3.) das Kriterium bezieht sich auf ein anderes Merkmal.

Die ersten beiden Fälle sind die typischen Situationen im Bereich der Labormeßverfahren. Der erste Fall wird häufig als Kalibrationsproblem (vgl. etwa [104]) beschrieben und der zweite als Methodenvergleich. Gehen wir zunächst von der Idealvorstellung aus, daß jedes Objekt mit der zu evaluierenden Methode sowie mit einer fehlerfreien Referenzmethode (wahrer Wert T) gemessen wurde. In diesem Fall interessiert im üblichen Fehlermodell

$$X_{ij} = T_i + E_{ij}$$

vor allem der Erwartungswert μ_E der Meßfehlerverteilung, der den systematischen Fehler ausdrückt. Im einfachsten Modell geht man davon aus, daß μ_E für alle Objekte gleich ist, jedoch ist es oft realistischer anzunehmen, daß die systematische Abweichung sich in Abhängigkeit vom wahren Wert ändert. Um dies anhand der Daten erkennen zu können, stellt man die Meßwerte in einem Koordinatensystem mit T als Abszisse und X als Ordinate dar. Die folgende Abbildung stellt einige Prototypen solcher Graphiken dar, die allesamt lineare Beziehungen zwischen dem wahren Wert und dem systematischen Fehler zeigen.

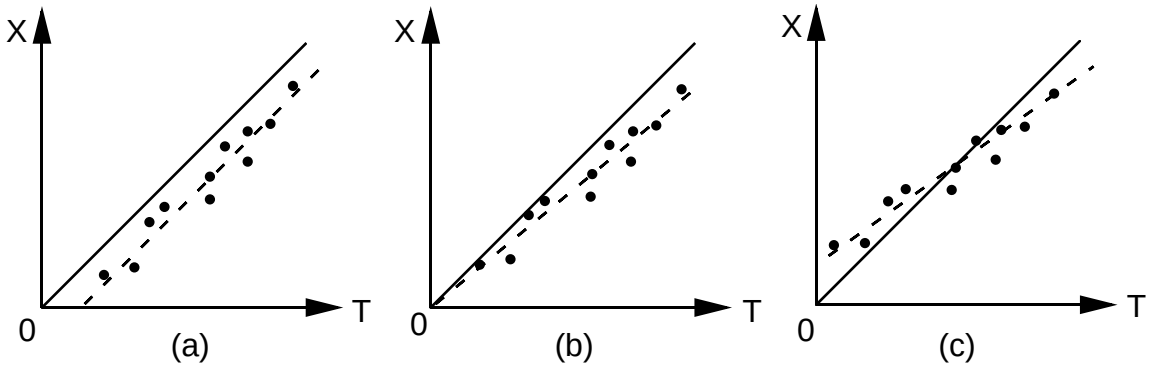

Abb. 5: Graphische Darstellung der Meßwertpaare mit Score (X) und Referenzmethode (T)

Im Falle einer solchen annähernd linearen Beziehung läßt sich diese als übliche Regressionsgerade von X auf T berechnen. Der systematische Fehler wird durch die Abweichung dieser Regressionsgeraden von der eingezeichneten Winkelhalbierenden gegeben. In den beiden ersten Fällen läßt sich der systematische Fehler durch einen einzigen einfachen Parameter ausdrücken. Im Falle einer Parallelverschiebung der Geraden (Abb. 5 a) ist dies der Achsenabschnitt, er entspricht dem oben genannten μ_E. Geht die Gerade zwar durch den Ursprung, aber mit einer anderen Steigung als 1 (Abb. 5 b), so ist der systematische Fehler proportional zu T und läßt sich durch einen konstanten Quotienten beschreiben. Wenn jedoch die Regressionsgerade sowohl in der Steigung als auch im Achsenabschnitt von der Einheitsgeraden abweicht (Abb. 5 c), so läßt sich der systematische Fehler nicht mehr einfach in einem Parameter ausdrücken.

Die Unverzerrtheit (bzw. der systematische Fehler) ist nur ein Teilaspekt des Validitätsbegriffes, eine erweiterte Vorstellung von der Richtigkeit einer Meßmethode bezieht die zufälligen Abweichungen mit ein. Ein globales Maß, welches dies leistet, ist die mittlere quadrierte Abweichung vom wahren Wert:

$$\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - t_i)^2$$

Die Berechnung dieses Parameters ist allerdings nur dann sinnvoll, wenn sowohl der systematische als auch der zufällige Fehler keine Abhängigkeit vom wahren Wert zeigen.

Der bisher geschilderte Fall ist bei klinischen Scores eigentlich unrealistisch, denn man muß meist annehmen, daß auch die Referenzmethode selbst meßfehlerbehaftet ist. In dieser Situation darf die Berechnung einer Regressionsgeraden für den Zusammenhang zwischen Score und Referenzmethode nicht auf der üblichen Kleinste-Quadrate-Schätzung parallel zur Ordinate erfolgen, da diese von einem meßfehlerfreien Merkmal auf der Abszisse ausgeht. Dies kann man sich ganz einfach wie folgt veranschaulichen: Untersucht man den systematischen Fehler der Referenzmethode in Bezug auf sich selbst, indem man Doppelmessungen durchführt und deren Werte, wie oben beschrieben, auf der Abszisse bzw. Ordinate abträgt, so

bekommt man bekanntlich bei Anwendung der üblichen Regressionsberechnung eine Gerade deren Steigung im Erwartungswert

$$\beta = \rho \cdot \frac{\sigma_{X'}}{\sigma_X}$$

ist, wobei ρ dem Korrelationskoeffizienten entspricht. Da es sich bei X und X´ um die gleiche Meßmethode handelt, sind die Standardabweichungen gleich. der Korrelationskoeffizient ist aber bei einem meßfehlerbehaften Merkmal kleiner als Eins und somit gilt dies auch für die Steigung der üblichen Regressionsgerade.

Die graphische Betrachtung und eventuelle Quantifizierung des systematischen Fehlers kann aber fast analog erfolgen, wenn man statt der üblichen linearen Regression die Ausgleichsgerade nach der Methode der "errors-in-variables"-Regression berechnet, die auch als Regressionsgerade zwischen den wahren Werten bezeichnet wird [105, 106]. Den Fall proportionaler, also nicht-konstanter Fehlervarianzen beschreibt Linnet [107].

Bland & Altman [108, 109] schlagen eine andere einfache Vorgehensweise für den Methodenvergleich vor. Als Form der graphischen Darstellung empfehlen sie, den Mittelwert der zwei Meßwerte als Abszisse und die Differenz als Ordinate aufzutragen. Dieser Art der Graphik ist noch besser anzusehen, ob die Meßfehlervarianz über den gesamten Meßbereich konstant ist oder eine Abhängigkeit vom wahren Wert zeigt. Ist ersteres der Fall, so schlagen Bland & Altman vor, aus den Differenzen d_i den Mittelwert $\bar{d}$ als Maß für die systematische Abweichung und die Standardabweichung s_d als Maß für die zufällige Abweichung der Methode Y in Bezug auf die Referenzmethode X zu wählen. Es ist zu bemerken, daß wir von einer ebenfalls meßfehlerbehafteten Referenzmethode ausgegangen waren und demgemäß die beiden Parameter symmetrisch in Bezug auf die Wahl von X und Y sind. Damit ist s_d eindeutig ein Reliabilitätsmaß, daß der in Abschnitt 4.3.1 definierten kritischen Differenz entspricht.

Der für Scores relevanteste Fall eines quantitativen Kriteriums ist gegeben, wenn als Standardverfahren ein anderer Score gewählt wird. Da die zwei Scores in der Regel keine einheitliche zugrundeliegende Meßeinheit haben, sondern unterschiedlich skaliert sind, kann als Validitätsparameter in diesen Fällen nur ein Korrelationskoeffizient herangezogen werden. Hier kommen die zwei bekannten Koeffizienten in Frage, der Pearson´sche Korrelationskoeffizient für die Beschreibung des linearen Zusammenhangs sowie der Spearman´sche Korrelationskoeffizient zur Beschreibung des monotonen Zusammenhangs.

4.4.2 Validitätsmaße für dichotome Kriterien

In Abschnitt 3.4.2 wurde begründet, warum die Verwendung dichotomer Außenkriterien häufig eine Rolle spielt, nicht nur beim Einsatz von Scores zu diagnostischen Zwecken, sondern auch im Zusammenhang mit Prognosescores. Im folgenden gehen wir zunächst von

dem einfachsten Fall aus, bei dem auch der Score dichotomisiert wurde. An diesem Fall können die statistischen Parameter am anschaulichsten eingeführt werden, um dann auf quantitative Scores erweitert zu werden.

Zur Untersuchung der Validität müssen mehrere Objekte mit bekanntem Kriteriumswert T (etwa 0 für gesund und 1 für krank) verfügbar sein. An jedem dieser Objekte sei zunächst genau eine Messung vorgenommen. Falls man voraussetzen könnte, daß sich die Übereinstimmung der Meßwerte mit den Kriteriumswerten nicht abhängig vom Kriteriumswert ändert, so würde ein einziger Parameter, nämlich die Trefferrate (entsprechend der Übereinstimmungsrate aus Abschnitt 4.2.2), als Maß für die Validität ausreichen. Diese Annahme ist in der Praxis meist unplausibel. Daraus ergibt sich, daß in der Regel zwei Parameter zu schätzen sind, nämlich die Wahrscheinlichkeiten p_0 und p_1 für eine richtige Messung oder Erkennung . Diese Werte entsprechen den Begriffen Spezifität und Sensitivität eines diagnostischen Tests. Liegen die Daten in der Form einer Vierfeldertafel vor

		Kriterium	
		0	1
Meßwert	0	a	b
	1	c	d

Abb. 6: Vierfeldertafel zur Bestimmung der Validität dichotomer Meßgrößen

so schätzt man die Sensitivität und Spezifität als

$$Se = a/(a+c) \qquad \text{und} \qquad Sp = d/(b+d).$$

Als zusammenfassendes Validitätsmaß wird häufig die Trefferrate

$$TR = (a+d)/(a+b+c+d)$$

berechnet. Für den hier betrachteten zweifach dichotomen Fall gibt es zahlreiche weitere Möglichkeiten der Zusammenfassung von Sensitivität und Spezifität in einem globalen Validitätsparameter, am bekanntesten davon ist der Youden-Index

$$Y = Se + Sp - 1.$$

Einen guten Überblick über die Vielzahl der Parameter sowie über Schätz- und Testverfahren gibt Abel [110].

Interessant für die Evaluierung von Scores ist die Tatsache, daß sich die gerade genannten Begriffe Sensitivität und Spezifität auch verwenden lassen, wenn man es mit einem ordinalen oder quantitativen Merkmal X zu tun hat. In diesem Fall kann man durch die Variation eines Grenzwerts x_0 eine ganze Serie dichotomisierter Versionen des Merkmals erzeugen, für die sich jeweils Sensitivität und Spezifität auf die oben beschriebene Weise berechnen lassen. Somit ergibt sich bei Variation von x_0 eine Reihe von Paaren $(Se(x_0), Sp(x_0))$, die sich als

Punkte in einem Diagramm graphisch darstellen lassen. Durch Interpolation dieser Punkte erhält man die sogenannte ROC-Kurve.

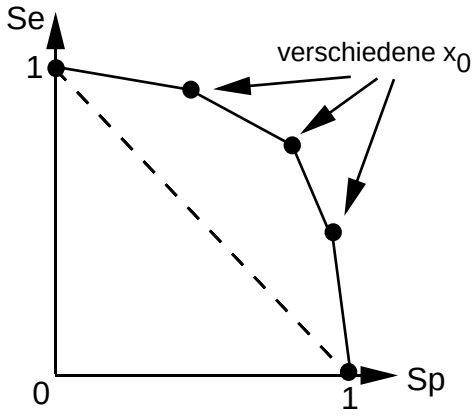

Abb. 7: ROC-Kurve

Dieses Vorgehen ist vielfach beschrieben (für einführende Darstellungen siehe [55, 71, 111]) und wird vor allem in diagnostischen Studien häufig angewandt. Dennoch ist es noch nicht dem festen Standardrepertoire der Biometrie zuzurechnen, da es in den meisten Lehrbüchern der Medizinischen Statistik nicht auftaucht [59, 102, 112, 113], und auch in den gängigen Statistik-Softwarepaketen nicht unterstützt wird. Aufgrund seiner zentralen Rolle für die Validitätsbeurteilung von Scores wird deshalb hierauf im weiteren näher eingegangen.

Die Darstellung der Validität eines quantitativen Scores in Bezug auf ein dichotomes Kriterium (z.B. Diagnose oder Verlauf) in Form der ROC-Kurve hat den wesentlichen Vorteil, daß sie leicht verständliche Parameter (Sensitivität und Spezifität) abbildet, ohne sich auf einen Schwellenwert zur Dichotomisierung des Scores festlegen zu müssen. Dies ist besonders relevant für den Vergleich der Validität zweier Scores: Das ROC-Diagramm mit zwei ROC-Kurven erlaubt eine leichte Differenzierung unterschiedlicher Situationen, die für die Anwendung des Scores von großer Bedeutung sind. Man unterscheidet insbesondere den Fall der gleichmäßigen Überlegenheit von dem der lokalen Überlegenheit eines Scores.

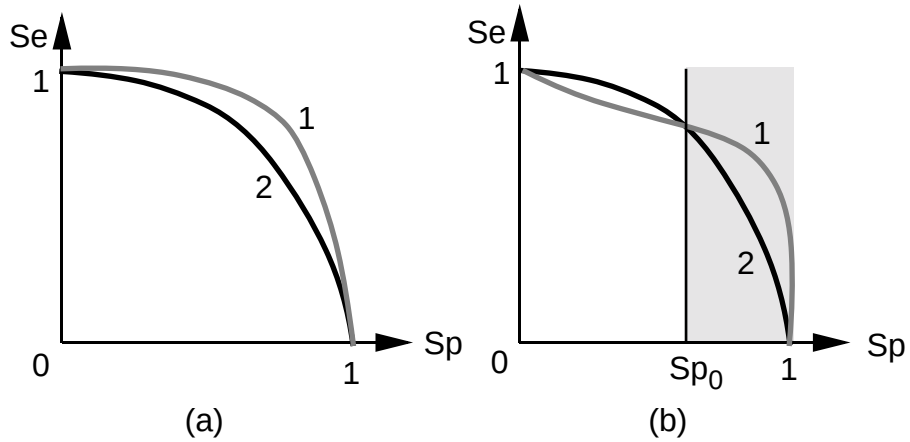

Abb. 8: ROC-Diagramm im Falle der globalen (a) vs. lokalen (b) Überlegenheit eines Scores

Im ersten Fall (Abb. 8a) ist die Sensitivität des einen Scores für jede fest vorgegebene Spezifität höher (bzw. an einigen Stellen gleich) als die des anderen, die ROC-Kurve verläuft oberhalb der anderen. Im zweiten Fall (Abb. 8b) überkreuzen sich die ROC-Kurven mindestens einmal, so daß in einem Bereich vorgegebener Spezifität der eine Score eine höhere Sensitivität aufweist, in einem anderen Bereich jedoch der andere. In diesem Fall der lokalen Überlegenheit scheint eine globale Aussage zum Validitätsvergleich der Scores nicht gerechtfertigt und erst der Bezug auf eine konkrete Einsatzsituation mit den damit verbundenen Kosten-Nutzen-Überlegungen macht eine rational begründete Entscheidung für einen Score möglich.

Diese eben dargestellte Möglichkeit des differenzierten Vergleichs der Validität zweier Scores geht verloren, wenn man zusammengefaßte Güteparameter verwendet. Dennoch ist man an solchen Parametern interessiert, da sie eine kompakte Beschreibung der Validität liefern. Eine Zusammenfassung ist hier in zweierlei Hinsicht möglich: Zum einen legt die graphische Darstellung im ROC-Diagramm den Gedanken nahe, daß die Fläche unter der ROC-Kurve einen geeigneten globalen Güteparameter darstellen könnte, der von der Festlegung eines Schwellenwertes unabhängig ist [114]. Interessanterweise ist dieses Validitätsmaß sogar als Wahrscheinlichkeit interpretierbar, und zwar als Wahrscheinlichkeit dafür, daß bei zufälliger Auswahl je eines Kranken und eines Gesunden der Kranke einen höheren Scorewert hat als der Gesunde. Der entsprechende Stichprobenschätzer für diese Wahrscheinlichkeit kommt auch als Wilcoxon-Statistik W im U-Test von Mann, Whitney & Wilcoxon vor [115].

$$W = \frac{1}{n_0 \cdot n_1} \cdot \sum_{x_0 \in K- \; x_1 \in K+} S(x_0, x_1) \;\; ,$$

wobei n_0 und n_1 die Stichprobengrößen der Gesunden (K-) bzw. der Kranken (K+) sind und die Zählfunktion S definiert ist als

$$S(x_0, x_1) = \begin{cases} 1, & \text{für } x_1 > x_0 \\ 0.5, & \text{für } x_1 = x_0 \\ 1, & \text{für } x_1 < x_0 \end{cases}.$$

Von Harrell [116] ist der gleiche Parameter als Konkordanzkoeffizient c bezeichnet worden und gezeigt worden, daß dieser eng verwandt ist mit dem Somer´schen Koeffizienten D. Während die Fläche unter der ROC-Kurve (und damit der als W oder c bezeichnete Parameter) im Falle eines völlig uninformativen Scores den Wert 0.5 annimmt, ist Somer´s D = 2 (c - 0.5) so skaliert, daß der Parameter in diesem Fall gleich Null ist.

Inferenzstatistische Verfahren, wie etwa die Berechnung von Konfidenzintervallen für die Fläche unter der ROC-Kurve sowie Tests zum Vergleich zweier Flächen aus unabhängigen oder abhängigen Stichproben sind in der Literatur beschrieben. Als Standard werden die Verfahren von Hanley & McNeil [114, 117] sowie von DeLong et al. [118] angesehen, die allerdings in den gängigen statistischen Auswertungssystemen bisher nicht implementiert sind. Metz [119] gibt einen Überblick über weitere Testverfahren, die teilweise in den von ihm erstellten Auswertungsprogrammen enthalten sind. Die Anwendung dieser Verfahren auf Scores ist unproblematisch, solange der Score viele Ausprägungen hat und näherungsweise als stetige Variable angesehen werden kann. Für Scores mit relativ wenigen diskreten Ausprägungen sind die genannten statistischen Verfahren nicht adäquat. Beam & Wieand [120] haben ein Testverfahren vorgeschlagen, welches den Vergleich eines diskreten mit einem quantitativen Prädiktor erlaubt. Dieses Verfahren kann von Nutzen sein, wenn man zwei Scores mit sehr unterschiedlicher Anzahl von Ausprägungen vergleichen möchte.

Die Möglichkeit des Einsatzes von ROC-Diagrammen zum Vergleich von Prognosescores, bzw. der ihnen zugrundeliegenden Prognosemodelle, wurde lange Zeit kaum beachtetet. Dabei bietet dieses Vorgehen eine einfache und anschauliche Möglichkeit des Validitätsvergleichs auch für Modelle unterschiedlichen Typs. Eines der frühesten Beispiele im Bereich der Diagnostik ist die Arbeit von Cook & Goldman [121], die multivariate Modelle zur Diagnose des Herzinfarkts nach dem Ansatz der logistischen Regression bzw. der hierarchischen Klassifikation mit ROC-Kurven einander gegenüberstellen. In eigenen Untersuchungen wurde die Anwendbarkeit der ROC-Methodologie auch auf prognostische Fragestellungen übertragen. An einer Studie an Patienten mit kleinzelligem Bronchialkarzinom wurde demonstriert, wie ein Vergleich von logistischer Regression und hierarchischer Klassifikation [122] sowie auch des Proportional-Hazards-Modells [123] möglich ist.

Die Anwendung der ROC-Darstellung auf logistische Regressionsmodelle ist einfach. Man betrachtet entweder die in dem Modell enthaltene Linearkombination der Prädiktoren, also den Score selbst, oder aber die damit in monotonem Zusammenhang stehende Vorhersagewahrscheinlichkeit und berechnet für jeden möglichen Grenzwert die Sensitivität und Spezifität in Bezug auf das Außenkriterium. Die Anzahl der Punkte der geschätzten

ROC-Kurve hängt von der Anzahl der Ausprägungen der Prädiktoren ab, bei quantitativen Prädiktoren ergibt sich eine relativ glatte Kurve. Im Falle eines hierarchischen Modells, das beispielsweise nach der CART-Methode [16] konstruiert wurde, hat man einen Prognosebaum mit k Endpunkten oder Blättern, denen jeweils eine Vorhersagewahrscheinlichkeit für das klinische Ereignis zugeordnet ist. Ein entsprechender Score läßt sich definieren durch eine geeignete Skalierung dieser Blätter. Verschiedene Dichotomisierungen mit variierendem Grenzwert ergeben eine ROC-Kurve, die aus k+1 Punkten besteht (wenn man die Punkte mit Se=0 und Sp=1 und umgekehrt mitzählt).

Eine andere Variante zur Analyse hierarchischer Modelle mit ROC-Methoden wenden Hadorn et al. [124] an, eine ausführlichere Darstellung dieses Vorgehens findet sich bei Raubertas et al. [125]. Diese Autoren gehen allerdings nicht von einem vorgegeben hierarchischen Modell aus, sondern betrachten verschiedene Prognosebäume, die aus der selben Menge von Prädiktoren hervorgehen. Dabei variieren sie nicht primär Grenzwerte für die Dichotomisierung quantitativer Prädiktoren, sondern sie variieren den Kostenfaktor, der die Gewichtung der Fehlprognosen ausdrückt. Dies führt dann zur Bildung ganz unterschiedlicher Bäume, aus denen sich keine übliche monotone ROC-Kurve mehr ableiten läßt. Mit dem dargestellten Vorgehen wird jedoch nicht die Bewertung eines Prognosemodells geleistet, sondern die einer Modellklasse, es ist im Hinblick auf die Evaluierung von Scores nicht brauchbar.

Die Anwendung von ROC-Methoden auf Proportional-Hazards-Modelle geschieht ganz ähnlich wie bei der logistischen Regression. Als Score kann wieder die Linearkombination der Prädiktoren aus dem Modell gewählt werden oder aber die vorhergesagte Überlebenswahrscheinlichkeit zu einem beliebigen, aber festen Zeitpunkt. Die Berechnung von Sensitivität und Spezifität erfordert allerdings die Definition eines dichotomen Kriteriums. Dazu muß ein geeigneter Zeitpunkt festgelegt werden, wie dies in Abschnitt 3.4.2 besprochen wurde. Dieser Zeitpunkt sollte nicht wesentlich später als die Mindestnachbeobachtungsdauer liegen, denn Patienten mit zensierten Beobachtungen vor diesem Zeitpunkt können bei der Auswertung nicht berücksichtigt werden. Das vorgeschlagene Vorgehen kann sogar auf stratifizierte Proportional-Hazards-Modelle ausgedehnt werden, obwohl diese kaum noch als Scores formulierbar sind. Als Prognose-variable wird dann die laut Modell vorhergesagte Überlebenswahrscheinlichkeit gewählt, wobei es hier auf die Wahl des Zeitpunkts ankommt, so daß man sich auf denjenigen Zeitpunkt bezieht, der auch zur Dichotomisierung der Zeitachse gewählt wurde.

4.4.3 Validitätsmaße für zensierte Zeitvariablen als Kriterium

Die ausführliche Darstellung von Validitätsparametern für dichotome Außenkriterien war unter anderem dadurch motiviert worden, daß bei Prognosescores der Krankheitsverlauf häufig durch das Eintreten eines klinischen Ereignisses in einem bestimmten Zeitraum beschreiben wird. Das eigentlich interessierende Merkmal ist hier die Zeit bis zum Eintreten

des Ereignisses, so daß die durch die Dichotomisierung bedingte Vergröberung einen Informationsverlust darstellt. Dies ist bei Anwendungen von Scores, wie sie in Abschnitt 2.4.2 dargestellt sind, nicht so kritisch, in anderen Fällen jedoch, zum Beispiel für die Patientenberatung, möchte man genauere Vorhersagen treffen, die etwa in der Angabe geschätzter Überlebenskurven bestehen. Hierbei muß berücksichtigt werden, daß für einen Teil der Patienten meist nur zensierte Beobachtungen vorliegen.

Statistische Parameter zur Beschreibung der Güte der Vorhersage wurden von verschiedenen Autoren vorgeschlagen, einen Überblick geben Korn & Simon [126]. Die Vorschläge beziehen sich unterschiedlich stark auf das Proportional-Hazards-Modell, wir wollen hier nur einige einfache und allgemein verwendbare Parameter kurz erwähnen. Wenn man nur davon ausgeht, daß zu einer Stichprobe von Patienten die Scorewerte sowie nur unzensierte Überlebenszeitdaten vorliegen, so bietet sich eine Betrachtung der Rangkorrelation an, da diese beschreibt, ob Patienten mit höheren Scorewerten in der Regel längere (oder kürzere) Überlebenszeiten haben.

Für zensierte Daten schlagen Harrell et al. [116] eine Verallgemeinerung des Konkordanzkoeffizienten aus Abschnitt 4.4.2 vor. Dazu werden alle möglichen Paare von Patienten gebildet und für jedes Paar ein Vergleich der Scorewerte sowie ein Vergleich der Überlebenszeiten vorgenommen. Berücksichtigt werden alle Paare, die unterschiedliche Scorewerte haben und von denen entweder keiner zensiert ist oder der zensierte länger lebte als der verstorbene Patient. Der Konkordanzkoeffizient berechnet sich aus dem Anteil konkordanter Paare an allen Paaren, wobei ein Paar als konkordant angesehen wird, wenn der Patient mit dem besseren Scorewert auch länger gelebt hat. Dieses Maß wird aufgrund seiner Einfachheit und anschaulichen Interpretation gelegentlich angewendet.

> Ciampi et al. [127] vergleichen drei vierstufige Stadieneinteilungen für Patientinnen mit metastasiertem Mammakarzinom, die mit völlig unterschiedlichen statistischen Verfahren (Proportional Hazards Modell, Korrespondenzanalyse, Rekursive Partitionierung) gewonnen wurden. Dabei verwenden sie den Harrell´schen Konkordanzkoeffizienten, der zeigt, daß die Unterschiede der prädiktiven Validität der drei Stadieneinteilungen nur minimal sind (0.77 vs. 0.75 vs. 0.77). Ein Vergleich auf der Basis des Log-Likelihood Kriteriums liefert weniger anschauliche Ergebnisse.

Die Berechnung des Log-Likelihood Parameters erfordert die Einbeziehung des Scores in ein Modell. In dem zitierten Beispiel geschieht dies, indem die Stadien durch dichotome Dummy-Variablen in einem Proportional-Hazards-Modell repräsentiert werden. Der Nachteil dieses Parameters ist seine mangelnde Anschaulichkeit. Dies gilt erst recht für das Akaike Informationskriterium, das die Log-Likelihood mit der Modellkomplexität, ausgedrückt durch die Anzahl der Modellparameter, verrechnet. Dieses Kriterium mag bei der explorativen Modellsuche sinnvoll sein, um eine Selektion zu komplexer Modelle mit zufälliger Anpassung an die Daten zu vermeiden. Im Rahmen des Vergleichs vorgegebener Scores oder

Stadieneinteilungen sollte ein Güteparameter jedoch nicht von der Modellkomplexität abhängen.

4.5 Kalibration

Der Begriff der Kalibration wurde bereits im Zusammenhang mit der Validierung in Bezug auf eine quantitative Referenzmethode erwähnt. Wie in Abschnitt 3.4.4 angesprochen, wird das Konzept der Kalibration auch zunehmend als Gütekriterium für Prognosescores zur Vorhersage eines klinischen Ereignisses benutzt. Dies setzt voraus, daß die Scorewerte entweder bereits als geschätzte Ereigniswahrscheinlichkeiten skaliert sind oder eine solche Transformation zum Score angegeben ist, wie beim APACHE System. In diesem Fall kann die Übereinstimmung der geschätzten Vorhersagewahrscheinlichkeiten mit den beobachteten Auftretenshäufigkeiten durch geeignete statistische Parameter beschrieben werden, welche unter dem Begriff Kalibration zusammengefaßt werden. Der synonyme Gebrauch des Begriffs Reliabilität hierfür, der von Hilden et al. [128] aus der Meteorologie übernommen und in die biometrische Literatur eingeführt wurde, wird hier vermieden, da er Anlaß zu Mißverständnissen geben könnte.

Während in Abschnitt 3.4.4 bereits grundsätzliche Probleme des Konzepts der Kalibration diskutiert wurden, soll im folgenden ein Überblick gegeben werden, mit welchen statistischen Parametern dieses Konzept zu erfassen versucht wird. Grundsätzlich kann man zwei Zugänge unterscheiden: Der eine basiert auf einer Klassierung der Patienten in Gruppen mit ähnlichen Scorewerten und vergleicht die mittlere Vorhersagewahrscheinlichkeit jeder Gruppe mit der beobachteten Auftretenshäufigkeit des klinischen Ereignisses. Der andere geht von den einzelnen Patienten aus und beschreibt den Zusammenhang zwischen Vorhersage-wahrscheinlichkeit und dichotomem Zielkriterium. Es handelt sich hier um Probleme der Anpassungsgüte ("goodness-of-fit"), die im Zusammenhang mit statistischen Modellen bekannt sind. Wir interessieren uns jedoch für modellunabhängige Parameter, die generell auf Prognosescores angewendet werden können.

Hosmer & Lemeshow [129] stellen verschiedene Maße für die Güte der Anpassung dar, die aber allesamt Testsstatistiken sind und somit von der Patientenzahl abhängen. Davon hat die Vorgehensweise am meisten Verbreitung gefunden, bei der pro Scorewert die Anzahl b_j beobachteter und die Anzahl $e_j=np_j$ (aufgrund der Vorhersagewahrscheinlichkeit p_j) erwarteter Ereignisse verglichen werden und daraus nach

$$X^2 = \sum_{j=1}^{k} \frac{(b_j - e_j)^2}{n_j \cdot p_j \cdot (1-p_j)}$$

eine chi-quadrat-verteilte Teststatistik berechnet wird. Liegen zuviele verschiedene Scorewerte vor und sind diese daher zu gering besetzt, so empfehlen Hosmer & Lemeshow

eine Einteilung in zehn Klassen, indem man die Dezile nach den Vorhersagewahrscheinlichkeiten p_i als Klassengrenzen wählt.

Gelegentlich wird der (aus der Meteorologie übernommene) Brier-Score als statistischer Parameter für die Kalibration verwendet (z.B. [130]). Der Brier-Score ist die mittlere quadratische Abweichung zwischen der Vorhersagewahrscheinlichkeit und der mit 0 und 1 kodierten Ereignisvariablen X.

$$B = \frac{1}{n} \cdot \sum_{i=1}^{n} (p_i - x_i)^2$$

Eine graphische Veranschaulichung gibt Abb. 9. Der Brier-Score entspricht dem Mittelwert der Quadrate der gestrichelt eingezeichneten Abstände.

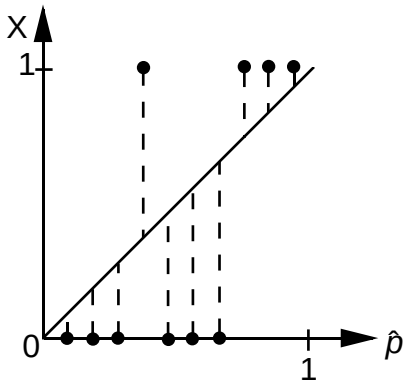

Abb. 9: Schematische Darstellung zur Veranschaulichung des Brier-Scores

Der Brier-Score ist zwar fallzahlunabhängig, jedoch erfaßt er nicht nur die Kalibration, sondern berücksichtigt auch einen Aspekt der Diskrimination, der von Ohmann [131] als "sharpness" bezeichnet wird. Ein Score hat dann eine hohe Sharpness, wenn für einen möglichst großen Teil der Patienten die aufgrund des Scores vorhergesagten Wahrscheinlichkeiten möglichst nahe bei Null und Eins liegen, also eine relativ sichere Prognose andeuten. Ohmann schlägt als Parameter für die so definierte Sharpness den Mittelwert

$$\frac{1}{n} \cdot \sum_{i=1}^{n} \max(p_i, 1 - p_i)$$

vor. In einer späteren Publikation [132] benutzt er eine vereinfachte Variante, nämlich den Anteil von Patienten, für die $p_i < 0.1$ bzw. > 0.9 ist.

Weder die Sharpness noch die Kalibration sind für sich allein betrachtet aussagekräftige Güteparameter für Prognosescores. In Abb. 10a ist ein gut kalibrierter Score mit sehr geringer Sharpness dargestellt, in Abb. 10b umgekehrt schlecht kalibrierter Score mit hoher Sharpness.

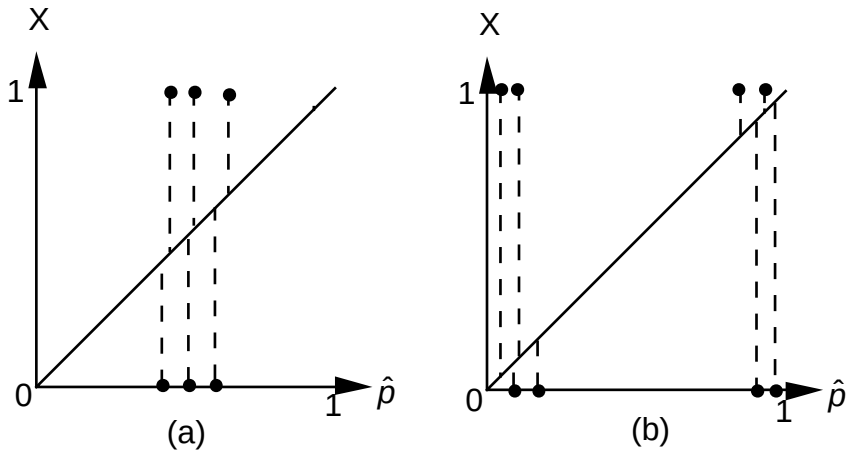

Abb. 10: Schematische Darstellung zur Veranschaulichung zweier Scores mit unterschiedlichen Eigenschaften (s.Text)

Nur wenn beide Eigenschaften erfüllt sind, kann man für einen großen Teil der Patienten relativ eindeutige und korrekte Prognosen stellen. In beiden abgebildeten Fällen ist der Brier-Score relativ weit vom optimalen Wert Null entfernt. Die genaue Beurteilung des Brier-Scores ist allerdings problematisch, da seine Skalierung ungünstig ist. Neben dem optimalen Wert Null ist nur noch der Wert Eins für den Fall einer perfekt falschen Vorhersage festgelegt. Günstiger wäre eine Verankerung des Falles einer zufälligen, nicht-informativen Vorhersage. Dies läßt sich beispielsweise durch den Vorschlag von Poses [133] erreichen, der die Transformation $(B_0 - B)/ B_0$ anwendet. Dabei wählt er für B_0 den Wert des Brier-Scores, den man maximal erreichen kann, wenn man für alle Patienten eine identische Vorhersagewahrscheinlichkeit rät. Es läßt sich leicht zeigen, daß dies der Fall ist, wenn $p_i = P(K+) = P_+$ für alle i ist, woraus sich dann für B_0 der Wert $P_+ - P_+^2$ ergibt .

Der Brier-Score mißt zwar die Kalibration und die Sharpness, jedoch nicht notwendig die Diskrimination. Dies läßt sich an den in Abb. 11 dargestellten Beispielen erkennen, wo der gleiche Brier-Score bei perfekter und bei weniger guter Diskrimination resultiert.

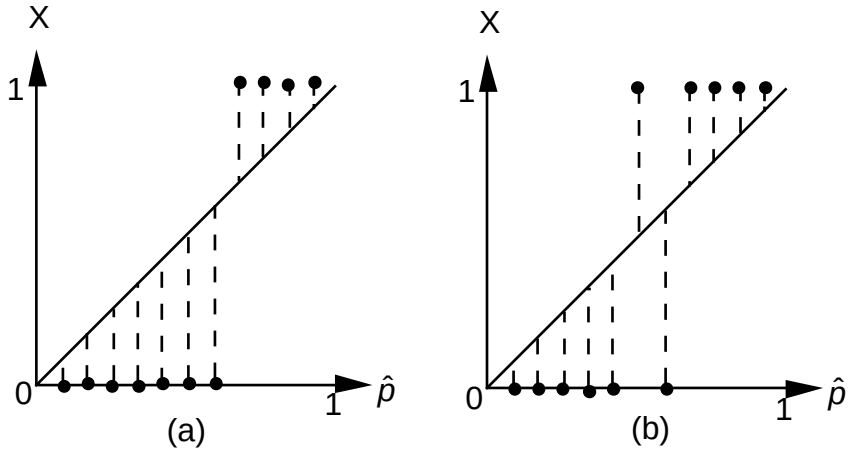

Abb. 11: Schematische Darstellung zur Veranschaulichung zweier Scores mit gleichem Brier-Score, aber unterschiedlicher Diskriniation

Diamond [81] versucht zu zeigen, daß eine perfekte Kalibration und eine perfekte Diskrimination unvereinbar sind. In seiner Beweisführung geht er allerdings von der unnötigen zusätzlichen Annahme aus, daß die laut Prognosemodell berechneten Wahrscheinlichkeiten in der Population gleichverteilt sind. Ein gut diskriminierendes und gleichzeitig gut kalibriertes Modell ist jedoch durchaus denkbar. Es liegt vor, wenn der Brier-Score den Wert Null annimmt.

Poses et al. [134] geben einen detaillierten Überblick über verschiedene Möglichkeiten der Evaluierung der Kalibration von Prognosescores. Miller et al. [135] weisen in ihrer Arbeit auf einen Vorschlag von Cox [136] hin, mit dem die verschiedenen Aspekte der Kalibration als unterschiedliche Parameter eines logistischen Regressionsmodells beschrieben werden können.

4.6 Zufallskorrigierte Validitätsparameter

Von verschiedenen Autoren wurden in den letzten Jahren Güteparameter für die Validität vorgeschlagen, die auf Sensitivität und Spezifität aufbauende sogenannte "zufallskorrigierte" Versionen dieser Parameter sein sollen. Diese Vorschläge sind einander ähnlich, indem sie alle eine strukturelle Verwandtschaft mit dem Übereinstimmungskoeffizienten Kappa aufweisen, jedoch führen sie zu zwei algebraisch unterschiedlichen Paaren von Parametern, da sie sich in ihrer Definition von "Zufall" unterscheiden. Von den Autoren werden zwar die mathematischen Eigenschaften dieser neuen Parameter ausführlich dargestellt, jedoch bleiben sie eine überzeugende Motivation für die Bedeutung dieser Maße schuldig. Trotzdem wurden die neuen Validitätsmaße und die darauf basierenden ROC-Methoden bereits in empirischen Untersuchungen verwendet [137]. Da sich in der biometrischen Fachliteratur noch kein Autor kritisch mit diesen zum Teil recht neuen Entwicklungen befaßt hat, soll im folgenden das Konzept der korrigierten Validitätsparameter kurz dargestellt und ihre Begründung diskutiert werden. Weitere Ausführungen hierzu enthält eine gerade fertiggestellte eigene Arbeit [138].

Die ersten Autoren, die eine Korrektur der Parameter Sensitivität und Spezifität vorschlugen, waren Kraemer und Bloch [139-141]. In der ersten Arbeit [139] werden die Parameter

$$RSe = \frac{Se - \pi_B}{1 - \pi_B} \quad \text{und} \quad RSp = \frac{Sp - (1 - \pi_B)}{\pi_B}$$

als "reskalierte Sensitivität bzw. Spezifität" eingeführt, wobei mit π_B die Wahrscheinlichkeit eines positiven Tests bezeichnet wird. Es muß erwähnt werden, daß diesem Vorschlag ein Populationsmodell zugrundeliegt, bei dem nicht, wie in der Diagnostik meist üblich, die zwei Populationen der Gesunden und Kranken unterschieden werden, sondern eine einzige Mischpopulation unterstellt wird, aus der zufällig Personen gezogen und in Bezug auf den dichotomen Test und den Krankheitsstatus beurteilt werden. Konsequenterweise werden Sensitivität und Spezifität hier auch nicht als als Merkmale des Tests in der Teilpopulation der

Gesunden bzw. der Kranken angesehen, sondern als Assoziationsparameter, die die Übereinstimmung von Test und Krankheitsstatus beschreiben sollen. Dies ist jedoch ein grundlegendes Mißverständnis der Bedeutung von Sensitivität und Spezifität. Zwar werden beide häufig als Validitätsparameter eines diagnostischen Tests beschrieben, jedoch sind sie eigentlich zwei Eigenschaften eines Tests, die eine Validitätsaussage nicht einzeln, sondern erst bei gemeinsamer Betrachtung erlauben. Kraemer´s Behauptung [139], daß durch die Reskalierung die inhaltliche Bedeutung der Parameter nicht verändert wird, ist offensichtlich falsch. Dies wird besonders deutlich, wenn man bedenkt, daß RSe und RSp sogar negative Werte annehmen können und damit im Gegensatz zu Se und Sp nicht mehr als Wahrscheinlichkeiten interpretierbar sind.

In einer weiteren Arbeit [141] führen Kraemer & Bloch als Argument für die Korrektur an, daß Sensitivität und Spezifität populationsabhängige Parameter seien. Dies ist sicherlich grundsätzlich der Fall, wenn auch nicht in dem gleichen Maße, wie beispielsweise die prädiktiven Werte von der Prävalenz der Kranken in der Population abhängen. Daß jedoch eine Selektion der Schwerkranken aus der Population der Kranken die Sensitivität in der Regel erhöht, ist unmittelbar plausibel und vielfach empirisch belegt, und ähnliches gilt entsprechend für die Spezifität (vgl. Begg [142]). Allerdings verschweigen Kraemer & Bloch, daß die reskalierten Parameter in gleicher Weise populationsabhängig sind. Sie sind zudem abhängig von der Krankheitsprävalenz, was in der Praxis eine wesentlich kritischere Eigenschaft ist.

Im Jahre 1992 erschienen unabhängig voneinander und ohne Bezug auf die Arbeiten Kraemers zwei Artikel, in denen die Parameter erneut präsentiert wurden, jeweils mit anderen Bezeichnungen, aber algebraisch identisch. Coughlin & Pickle [143] nennen die Parameter "sensitivity and specificity-like measures ... corrected for chance agreement", während Jamart [144] als erster den Begriff "chance-corrected sensitivity" gebraucht. Die erstgenannten Autoren stellen die neuen Parameter nicht als Alternativen zur herkömmlichen Sensitivität und Spezifität dar, sondern sehen sie eher als Konkurrenz zu anderen zusammenfassenden Validitätsmaßen, wie dem Youden-Index. Jamart wiederum bezieht sich auf die spezielle Situation, wo ein diagnostischer Test nicht nur positive und negative, sondern auch uneindeutige (grenzwertige) Ergebnisse liefert. Für diesen Fall schlägt er vor, die grenzwertigen Testbefunde zufällig zu dichotomisieren und begründet damit die Notwendigkeit einer Zufallskorrektur für die dann berechneten Werte von Sensitivität und Spezifität. Ein solches Vorgehen ist jedoch unüblich, denn in der Praxis wird man stattdessen die grenzwertigen Befunde entweder vollständig den positiven oder vollständig den negativen Befunden zuordnen, abhängig davon, ob man eine hohe Sensitivität oder eine hohe Spezifität anstrebt.

Der neueste Vorschlag einer zufallskorrigierten Sensitivität und Spezifität stammt von Gefeller und Brenner [145, 146], die die Grundidee der anderen Autoren aufgreifen, jedoch durch einen anderen Korrekturterm erreichen, daß die von ihnen vorgeschlagenen Parameter

nicht krankheitsprävalenzabhängig sind. In dem dem Kappa-Koeffizienten zugrundeliegenden Konstruktionsschema, das sich durch die Reskalierung

$$X_{neu} = \frac{X_{alt} - X_{zuf}}{1 - X_{zuf}}$$

beschreiben läßt, setzen sie im Falle der Sensitivität für X_{zuf} nicht die Wahrscheinlichkeit P(T+) des positiven Tests, sondern 1 - Sp ein.

Damit resultieren als zufallskorrigierte Maße

$$Se^* = 1 - \frac{1 - Se}{Sp} \text{ und } Sp^* = 1 - \frac{1 - Sp}{Se}.$$

Im Falle eines Tests, der völlig unabhängig vom "gold standard" ist, stimmen beide Definitionen zufallskorrigierter Parameter überein, da in diesem Fall die Wahrscheinlichkeit P(T+) des positiven Tests gleich 1 - Sp ist. Im Falle eines informativen Tests unterscheiden sich die Parameter jedoch, insbesondere sind Se* und Sp* nicht abhängig von der Krankheitsprävalenz.

Sowohl von Kraemer [140] als auch von Gefeller & Brenner [146] wird vorgeschlagen, im Falle eines auf einem quantitativen Merkmal basierenden diagnostischen Tests das Konzept der ROC-Kurven auf die korrigierten Parameter zu übertragen. Kraemer nennt das Ergebnis QROC, Gefeller & Brenner sprechen von der zufallskorrigierten ROC-Kurve. In beiden Versionen verlaufen die Populations-ROC-Kurven informativer Tests nun nicht mehr nur oberhalb der Diagonalen, sondern im gesamten Quadranten, während die zufallskorrigierte ROC-Kurve eines nicht-informativen Zufallstest auf den Koordinatenursprung beschränkt bleibt. Gefeller & Brenner versprechen sich davon eine deutlichere visuelle Diskrimination beim Vergleich zweier quantitativer Diagnosevariablen und führen zur Illustration die ursprünglichen und die korrigierten ROC-Kurven für den Fall zweier Normalverteilungen mit gleicher Varianz und unterschiedlichem Erwartungswert an, wie dies in Abbildung 12 dargestellt ist.

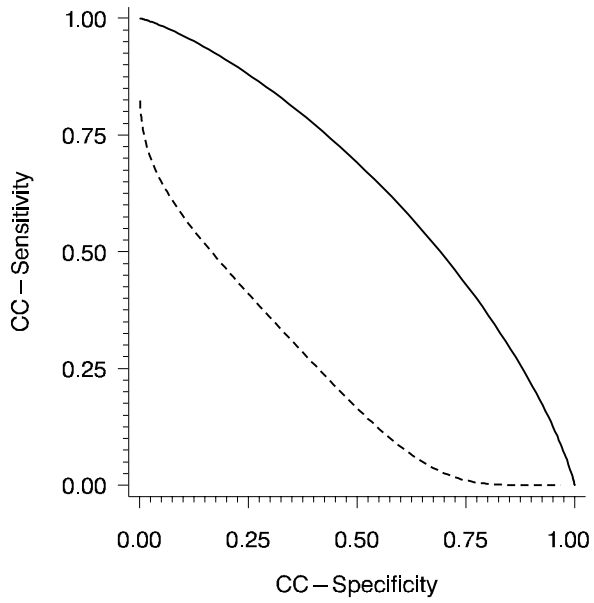

Abb. 12: Original (_____) und zufallskorrigierte (_ _ _ _) Populations-ROC-Kurve (Normalverteilungen mit $\mu_1=0$ und $\mu_2=0.5$ sowie $\sigma_1=\sigma_2=1$ zugrundegelegt)

Aus verschiedenen Gründen sind diese Vorschläge korrigierter ROC-Kurven weder sinnvoll noch praktisch brauchbar. Zum einen gibt es keinen Grund, das Prinzip der Zufallskorrektur von den Parametern auf die gemeinsame graphische Darstellung zu übertragen. Der vermeintliche Nachteil von Sensitivität und Spezifität, daß diese Parameter jeder für sich allein keine Information über die diskriminative Validität der Testvariablen zulassen, besteht ja bei gemeinsamer Betrachtung nicht mehr, dies ist gerade einer der Vorteile des ROC-Diagramms. Stattdessen gehen andere Vorteile des ROC-Diagramms in der zufallskorrigierten Version verloren, zum Beispiel die Interpretation der Fläche unter der Kurve als anschaulicher Validitätsparameter und der Zusammenhang der Steigung der ROC-Kurve mit dem sogenannten Likelihood-Ratio, der die Wahl eines optimalen Schwellenwerts für eine dichotome Entscheidungsregel unter Kosten-Nutzen-Aspekten erlaubt [71]. Ebenso verliert man die in der Praxis oft genutzte Möglichkeit des Vergleichs der Sensitivitäten verschiedener quantitativer Tests an einer Stelle oder in einem Bereich mit vorgegebener Spezifität.

Diesem Verlust sinnvoll interpretierbarer Eigenschaften steht nun aber keineswegs der postulierte Gewinn an visueller Transparenz entgegen, wie er in den von Gefeller & Brenner abgebildeten Graphiken suggeriert wird. Diese Graphiken stellen Populations-ROC-Kurven für den optimalen Fall zweier Normalverteilungen mit gleicher Varianz dar. In der Praxis muß man diese ROC-Kurven jedoch aus einer Stichprobe schätzen und hierbei erweist sich die zufallskorrigierte ROC-Kurve der üblichen ROC-Kurve als unterlegen (vgl. Windeler & Holle [138]). Während letztere nämlich, genau wie die Populations-Kurve, eine monoton fallende Funktion ist, gibt erstere bei linearer Interpolation der aus der Stichprobe berechneten Punkte ein bizarres Bild (vgl. Abb. 13). Dies gilt auch für die Kraemer´sche QROC-Kurve.

Vermutlich aus diesem Grund schlägt die Autorin vor, die konvexe Hülle der berechneten Punkte zu bilden. Besonders im Falle von nur schwach diskriminierenden Merkmalen erhält man aber dennoch stark zufällig fluktuierende Kurven, die zudem noch gelegentlich den ersten Quadranten verlassen (vgl Abb. 14). Gefeller & Brenner haben zu den auf ihren neuen Parametern basierenden zufallskorrigierten ROC-Kurven keine weiteren Hinweise über deren Eigenschaften und ihre Interpretation gegeben.

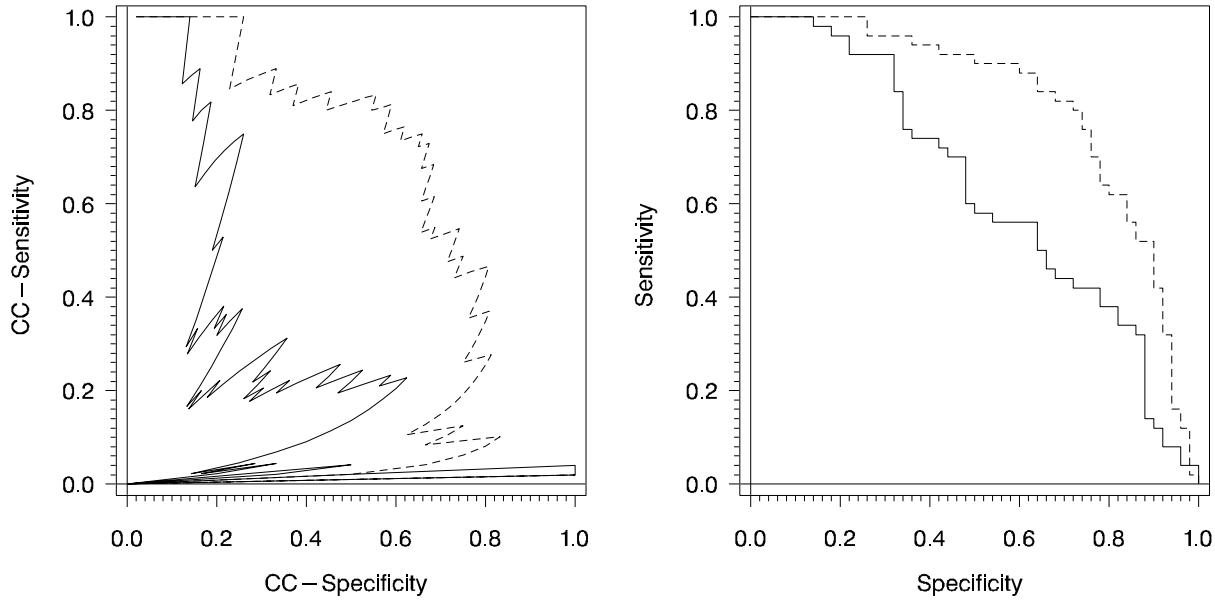

Abb. 13: Zufallskorrigierte (links) und original (rechts) ROC-Kurven zum Vergleich zweier Prognosescores. Schätzung aus Zufallsstichproben der Größe $n_1=n_2=50$ aus jeweils normalverteilten Populationen mit $\mu_1=0$ und $\mu_2=0.5$ sowie $\sigma_1=\sigma_2=1$ (_____) bzw. mit $\mu_1=0$ und $\mu_2=1$ sowie $\sigma_1=\sigma_2=1$ (_ _ _)

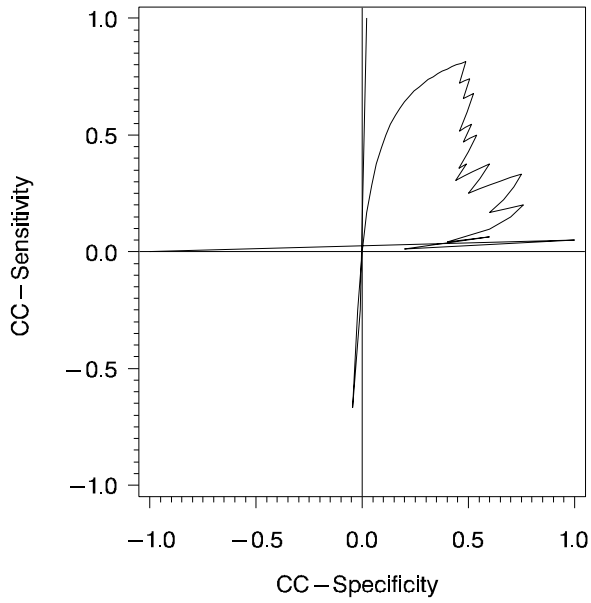

Abb. 14: Zufallskorrigierte ROC-Kurve. Schätzung aus Zufallsstichproben der Größe n_1=50, n_2=20 aus jeweils normalverteilten Populationen mit μ_1=0 und μ_2=0.5 sowie σ_1=σ_2=1

Zusammenfassend kann man sagen, daß die von verschiedenen Autoren eingeführten zufallskorrigierten Varianten von Sensitivität und Spezifität keinesfalls einen Ersatz für die herkömmlichen Parameter darstellen, sondern bestenfalls eine Alternative zu bereits existierenden zusammenfassenden Validitätsparametern, mit denen sie zum Teil algebraisch eng verwandt sind. Sie sind anschaulich nicht gut interpretierbar, insbesondere repräsentieren sie keine Wahrscheinlichkeiten und legen die Gefahr von Fehlinterpretationen aufgrund einer Verwechslung mit Sensitivität und Spezifität nahe. Sie sind daher überflüssig und von ihrer Verwendung ist abzuraten. Dies gilt aus den oben genannten Gründen erst recht für die darauf basierenden korrigierten ROC-Kurven.

4.7 Änderungssensitivität

Die definitorische Problematik dieses Begriffes wurde bereits oben diskutiert. Sie spiegelt sich auch in den unterschiedlichen Operationalisierungen für die Quantifizierung dieses Güteparameters wieder. Betrachtet man den Begriff zunächst gemäß seiner direkten sprachlichen Bedeutung, so bezieht er sich auf die Fähigkeit des Scores, tatsächliche Veränderungen im Merkmal zu erfassen. Dabei spielt es keine Rolle, ob der Score Änderungen direkt mißt oder ob sie durch den Vergleich der Scorewerte zu zwei Zeitpunkten erfaßt werden. Im letzteren Fall kann dieser Vergleich durch Differenzbildung, aber zum Beispiel auch durch Verhältnisse von Scorewerten beschrieben werden, obwohl die letztgenannte Möglichkeit in keiner Publikation explizit genannt ist.

Die Bestimmung einer Änderungs*sensitivität* ist aber eine Form der Kriteriumsvalidität und verlangt, daß ein Kriterium für die tatsächliche Veränderung vorhanden sein muß. Die hiermit

verbundenen Schwierigkeiten brauchen an dieser Stelle nicht noch einmal diskutiert zu werden (vgl. Abschnitt 3.4.3). Deyo & Centor [75] zeigen verschiedene Möglichkeiten, die nichts anderes als die Anwendung der oben dargestellten Methoden zur Bestimmung der Kriteriumsvalidität sind.

> In einer Therapiestudie an Patienten mit Rückenschmerzen untersuchten sie die Änderungssensitivität des Sickness Impact Profile (SIP [147]), seiner Unterskalen und einer Kurzform, indem sie zunächst Korrelationen mit Änderungen in physikalisch meßbaren klinischen Variablen (z.B. Rückenkrümmbarkeit) berechneten. Weiterhin teilten sie die Patienten gemäß eines dichotomen Außenkriteriums (Konsens zwischen Patienten- und Expertenrating) in solche mit Verbesserungen bzw. mit konstanten Verläufen oder Verschlechterungen ein und stellten dann die Fähigkeit der Skalen, diese zwei Patientengruppen mittels Scoredifferenzen zu differenzieren, durch ROC-Kurven graphisch dar.

Auch der mit Guyatt´s Namen belegte Güteindex [87, 148] benötigt ein Außenkriterium, um zu definieren, welche Patienten stabil blieben und welche sich änderten. Zur Darstellung des Guyatt´schen Index gehen wir von einer solchen Einteilung der Patienten aus. Eigentlich ist eine getrennte Betrachtung für den Fall positiver bzw. negativer Veränderungen erforderlich, obwohl dies meist ignoriert wird, hierauf wird in einem Beispiel nochmal zurückgekommen. Betrachten wir hier nur die Differenzierung von positiven Veränderungen vs. stabilen Verläufen, so seien die Mittelwerte der Scoredifferenzen in den Populationen als μ_+ und μ_0 und in den Stichproben als $\overline{D}_+$ und $\overline{D}_0$ bezeichnet. Die entsprechenden Standardabweichungen seinen in der Population σ_+ und σ_0 und in der Stichprobe s_+ und s_0. In dieser Notation läßt sich der Guyatt´sche Parameter R (für "responsiveness") in der Stichprobe schreiben als

$$R = \frac{\overline{D}_+}{s_0}$$, entsprechend in der Population als $\frac{\mu_+}{\sigma_0}$.

Die Responsiveness gemäß dieser Definition setzt also die mittlere Scoreänderung bei Patienten mit Verbesserung in Beziehung zur Standardabweichung der Scoredifferenzen in der Gruppe der stabilen Patienten.

Die genannte Definition der Responsiveness ist problematisch, denn es wird dabei unterstellt, daß die Scorewerte bei stabilen Patienten sich nicht systematisch ändern. Dieses Problem wird von Deyo et al. [87] erkannt, die vorschlagen, zur Korrektur die mittlere Differenz der Scorewerte bei stabilen Patienten im Zähler abzuziehen, um so einen Netto-Effekt zu haben.

Die Problematik gerade der letzten Annahme, nämlich daß bei den laut Außenkriterium stabilen Patienten keine Änderung des gemessenen Merkmals eintritt, wird in dem folgenden Beispiel deutlich.

> Marks et al. [149] entwickelten einen Lebensqualitätsfragebogen für Asthmatiker und untersuchten dabei die Responsiveness des Gesamtscores und der Unterskalen. Die Patienten wurden zweimal im Abstand von drei Monaten untersucht. Als Außenkriterium für die Veränderung wurde eine Kombination aus Lungenfunktion (FEV) und Histaminverbrauch herangezogen. Bei der Analyse der Responsiveness nach Guyatt´s Formel ergab sich beispielsweise für die Skala "Atemnot (breathlessness)" ein Wert von 1.49, fast doppelt soviel wie für die Skala "Stimmung (mood)" mit 0.82. Dem steht aber entgegen, daß ein statistischer Test für den Vergleich verbesserter vs. stabiler Patienten bzgl. Atemnot nicht signifikant (p=0.13) und bzgl. Stimmung signifikant war (p=0.03), bei gleichen Patientenzahlen. Der Grund hierfür lag darin, daß die "stabilen" Patienten in ihren Scorewerten zur Atemnot sich ebenfalls in der gleichen Richtung änderten, während sie bei der Skala zur Stimmung tatsächlich konstant blieben. Würde man nun, wie von Guyatt et al. propagiert, den Responsiveness Index als Anhaltspunkt für eine Fallzahlschätzung nehmen, so würde dies hier vermutlich zu Fehlschlüssen führen.

Guyatt et al. [70, 72, 73] schlagen einige Strategien vor, wie man auch ohne ein Referenzkriterium auskommt. So wird der Nenner beispielsweise anhand einer Gruppe unbehandelter Patienten geschätzt und der Zähler anhand einer Gruppe behandelter Patienten. Tuley et al. [148] verwenden diesen Ansatz und liefern hierfür Formeln, um approximative Konfidenzintervalle zu bestimmen und einen Test zum Vergleich zweier Responsiveness-Koeffizienten aus abhängigen Messungen durchzuführen. Dabei unterstellen aber auch sie, daß in der unbehandelten Gruppe keine systematische Änderung des Scores erfolgt. Mit der ebenso möglichen Situation, daß in der behandelten Gruppe im Mittel nur eine Stabilisierung des Scores erreicht wird, während in der unbehandelten Gruppe eine Verschlechterung eintritt, haben sich diese und andere Autoren nicht auseinandergesetzt.

Gelegentlich finden sich Versionen des Responsiveness-Parameters, die nochmals vereinfachend modifiziert sind, etwa indem alle Patienten mit einer als wirksam angenommenen Therapie behandelt werden und im Nenner auch die Standardabweichung der Scoredifferenzen dieser (anstatt stabiler oder unbehandelter) Patienten gesetzt wird [150].

Ein ausführliches Beispiel einer vergleichenden Untersuchung der Responsiveness verschiedener Skalen geben Guyatt et al. [78]. Dieses Beispiel zeigt insbesondere einige Probleme der Vorgehensweise auf.

> Im Rahmen der Entwicklung eines neuen Lebensqualitätsbogens (GQLQ) für gebrechliche ältere Patienten wurde dieser zusammen mit einigen etablierten Meßinstrumenten bei 100 geriatrischen Patienten in Abständen von drei Monaten beantwortet. Das Hauptproblem in dieser Studie besteht in der Wahl einer ungeeigneten Referenzmethode zur Beurteilung der tatsächlichen Veränderung. Indem man die Patienten mit einer einzigen Frage ("Generally speaking, how has your overall health been over the past two weeks?") zu jedem Meßzeitpunkt um eine Einschätzung auf

einer siebenstufigen Skala bat, hat man den schwächstmöglichen Standard gewählt. So wird denn auch in der Diskussion dieser Arbeit die unerwartet niedrige Änderungssensitivität der mangelnden Reliabilität des Außenkriteriums zugeschrieben. Ein zweites Problem beinhaltet die Vorgehensweise, die Patienten mit positiven bzw. negativen "wahren" Veränderungen zusammenzufassen, indem man die Scoredifferenzen der einen Gruppe einfach mit -1 multipliziert. Dies ist weder inhaltlich gerechtfertigt noch aufgrund der Daten plausibel (vgl. Tab. 2 der Publikation, dort speziell Barthel-Index) [78].

Zusammenfassend läßt sich sagen, daß die Änderungssensitivität als Aspekt der Validität am geeignetsten erfaßt wird, wenn man eine Referenzmethode als Maßstab heranzieht. Dies setzt aber voraus, daß man die Wahl dieser Methode überzeugend begründet. Akzeptable Vorgehensweisen zur Beurteilung der Änderungssensitivität, insbesondere im Vergleich zweier Meßinstrumente, sind dann die entsprechende Variante des Guyatt´schen Maßes sowie vor allem der Deyo´sche Vorschlag der Darstellung als ROC-Kurven. In beiden Fällen wird eine Dichotomisierung aufgrund des Kriteriums in positive Veränderungen vs. stabile Verläufe erwartet, die Einbeziehung negativer Veränderungen ist nur unbefriedigend gelöst. Eine pragmatische Alternative besteht darin, konkurrierende Scores als Zielgrößen in unkontrollierten oder kontrollierten Studien einzusetzen und zum Vergleich der Änderungssensitivität einfach die Teststatistik oder den entsprechenden p-Wert eines geeigneten statistischen Tests heranzuziehen. Mehrere Autoren (u.a. Pocock [151]) äußern die Ansicht, daß dieses Vorgehen auf einer breiten Basis den Selektionsprozeß geeigneter Skalen in die richtige Richtung steuern wird.

5 KONSTRUKTION KLINISCHER SCORES

Die Konstruktion eines neuen Scores sollte möglichst gut empirisch begründet sein. Dies war in der Vergangenheit nicht immer der Fall und etliche auch heute noch etablierte Scores sind eher auf einer intuitiven Basis entstanden. Zwar spricht es nicht zwangsläufig gegen die Qualität des Scores, jedoch macht dies nachträgliche Evaluierungsstudien notwendig, die eigentlich schon im Rahmen der Scorekonstruktion hätten stattfinden können. Insbesondere gibt es Möglichkeiten, bei der Konstruktion des Scores die Komponenten so auszuwählen und zu kombinieren, daß eine möglichst hohe Reliabilität bzw. Validität erreicht wird. In diesem Zusammenhang sollte auch beachtet werden, daß heute für sehr viele Anwendungsbereiche bereits Scores existieren, so daß die Konstruktion eines neuen nur gerechtfertigt ist, wenn die Unzulänglichkeit der verfügbaren Scores gezeigt und eine empirische Überlegenheit des neuen Scores demonstriert werden kann. Im folgenden beschreiben wir zunächst eine schrittweise Strategie der Entwicklung eines neuen Scores und gehen dann auf konkrete statistische Verfahren ein, die hierbei angewandt werden können. Dabei legen wir das Gewicht allerdings nicht auf die ausführliche Darstellung dieser Verfahren, die allesamt etabliert und auch in Lehrbüchern beschrieben sind, sondern beschränken uns auf einige Probleme der Anwendung dieser Verfahren auf die Scorekonstruktion.

Die Konstruktion eines klinischen Scores kann zunächst grob in vier Schritte aufgeteilt werden.

1. Zusammenstellung potentieller Komponenten oder Items
2. Selektion von geeigneten Komponenten
3. Suche einer geeigneten Kombination der Komponenten zu einem Score
4. Überprüfung des Scores auf seine Güteparameter

Leitlinie für die Kriterien, die bei den einzelnen Schritten anzuwenden sind, sollte eine Überlegung zur Zielsetzung des Scores sein. Vor allem die Frage, ob die spätere Anwendung des Scores der Prädiktion oder Evaluation dienen soll, beeinflußt das konkrete Vorgehen bei der Scorekonstruktion, wenn auch nicht so strikt, wie von Kirshner & Guyatt [18] behauptet wird. Jedoch auch spezifische Randbedingungen des späteren Einsatzes, wie die Beschränkung auf ein spezielles Setting oder die möglichst weitgehende Übertragbarkeit auf andere Kliniken, sind in jedem Schritt zu beachten.

Aus den folgenden Ausführungen wird erkennbar sein, daß die Konstruktion von Scores vor allem im Hinblick auf eine Validitätsoptimierung erfolgt, während Reliabilitätsbetrachtungen weniger berücksichtigt werden. Dies kann teilweise damit begründet werden, daß eine hohe Validität automatisch eine gute Reliabilität impliziert, wie in Abschnitt 3.5 festgestellt. Bei evaluativen Scores, vor allem in Bereichen ohne klares Validitätskriterium, stützt sich die Scorekonstruktion notgedrungen auf Reliabilitätsdaten sowie auf Analysen zur Änderungssensitivität.

5.1 Auswahl von Kandidatenvariablen

Der erste Schritt der Scorekonstruktion besteht in der Auswahl von Kandidatenvariablen, also von potentiell brauchbaren Komponenten oder Items. Dies beinhaltet zweierlei, nämlich die Auswahl der Merkmale und die Operationalisierung ihrer Erfassung. Mit dem letzteren ist zum Beispiel gemeint, durch wen und in welcher Formulierung eine Frage gestellt wird und welche Antwortalternativen angeboten werden.

Die Auswahl der Merkmale läßt sich besser strukturieren, wenn man zunächst analysiert, welche Merkmalsbereiche abgedeckt werden sollen. Hierzu ist ein konzeptueller Hintergrund zu dem betreffenden Krankheitsbereich erforderlich, der bei Evaluationsskalen eher theoretisch geprägt und bei Prognosescores vorwiegend empirisch begründet ist. Im Bereich der Lebensqualitätsmessung liegen beispielsweise theoretische Konzepte vor, die eine multidimensionale Struktur des Konstrukts beschreiben [152]. Bei Prognosescores für Intensivpatienten können Merkmalsbereiche allein aus der ärztlichen Erfahrung (z.B. betroffene Organsysteme) oder aus bereits vorliegenden empirischen Untersuchungen formuliert werden. Die Auswahl der möglichen Kandidatenvariablen aus diesen Merkmalsbereichen unterliegt oft einer erheblichen Willkür. Grundsätzlich ist deshalb zu empfehlen, in einem ersten Schritt eine möglichst große Zahl von potentiellen Kandidaten auszuwählen und die weitere Selektion mehrstufig empirisch geleitet und nachvollziehbar vorzunehmen. Die unterschiedlichen Stufen bei der Variablenselektion können sich in ihren Auswahlprinzipien grundlegend unterscheiden, wie im nächsten Abschnitt dargestellt. Zunächst sind jedoch noch einige Vorgehensweisen vorzustellen, wie man zu einer möglichst umfassenden Ausgangsmenge von Kandidatenvariablen kommen kann.

Die hierzu anzuwendenden Strategien liegen eigentlich auf der Hand. Dennoch ist erst in einigen neueren Veröffentlichungen erkennbar, daß man sie bei der Konstruktion eines neuen Scores systematisch angewandt hat. Eine mögliche und heutzutage unbedingt zu empfehlende Strategie besteht in der umfassenden Literaturrecherche mit dem Ziel, alle bereits existierenden Scores zu dem interessierenden oder eng verwandten Anwendungsbereichen zu erfassen und die dort vorkommenden Merkmale als potentielle Kandidaten zu berücksichtigen. Eine alternative, oder besser zusätzliche, Möglichkeit ist die Befragung von Experten. Wenn es sich um eine klinische Skala zur Beschreibung des Zustands eines Patienten unter Berücksichtigung seiner subjektiven Befindlichkeit handelt, so ist auch der Patient selbst ein "Experte" und sollte daher ausführlich befragt werden, welche Teilaspekte seiner Befindlichkeit eine Rolle spielen und als potentielle Kandidatenmerkmale für eine Skala in Frage kommen. Hierzu gibt es bereits einige Beispiele in der Literatur, wo dieses Vorgehen, etwa in Form von Interviews oder schriftlichen Befragungen, gewählt wurde.

> Guyatt et al. [78] erstellten für ihr Meßinstrument zur Erhebung der Lebensqualität bei gebrechlichen alten Patienten zunächst eine Urliste von 131 Items, die aus zwei Quellen gewonnen wurde: erstens aus publizierten Fragebögen, sowohl zur Lebensqualität

allgemein als auch speziell aus dem gerontologischen Bereich, zweitens aus offenen Interviews mit verschiedenen Spezialisten (Ärzten, Pflegepersonal).

Cella et al. [12] bezogen 45 Krebspatienten und 15 Onkologen in die erste Phase ihrer Konstruktion eines Lebensqualitätsfragebogens für Krebspatienten ein, indem sie ihnen verschiedene existierende Fragebögen vorlegten und dann in einem Interview befragten, welche dieser und vor allem welche weiteren Fragen ihnen auf der Basis der eigenen Erfahrung wichtig erschienen.

Die Entscheidung über die geeignete Art der Erfassung der einzelnen Merkmale ist ein weiterer wichtiger Schritt in der ersten Phase der Scorekonstruktion. Hierhin gehören Entscheidungen wie etwa die inhaltliche Präzisierung bzw. Operationalisierung der Merkmale sowie die sprachliche Formulierung von Fragen. Im Bereich objektiver Merkmale ist eine genaue Spezifikation des Meßverfahrens erforderlich, einschließlich Angaben zum Zeitpunkt und zu den Randbedingungen der Messung.

Im APACHE Score [9] wird als ein Merkmal der Blutdruck des Patienten berücksichtigt. Dies wird spezifiziert durch die Angabe, daß der mittlere arterielle Blutdruck, also der Mittelwert aus systolischem und diastolischem Blutdruck, gemessen bzw. berechnet werden soll und daß der ungünstigste dieser Werte aus dem Zeitraum der ersten 24 Stunden nach Aufnahme auf die Intensivstation berücksichtigt werden soll.

Bei Merkmalen, die eine Einschätzung durch den Arzt beinhalten, ist eine Standardisierung der Erhebungssituation und der Beurteilungskriterien besonders wichtig.

Die Beurteilung von Patienten mit Hirnverletzungen anhand der Glasgow Coma Scale bzw. der Reaction Level Scale beinhaltet zum Teil die gleichen Merkmale, aber in unterschiedlicher Operationalisierung [50].

Bei Merkmalen, für die eine Beurteilung durch den Patienten selbst erforderlich ist, ist im Rahmen der Operationalisierung zu klären, ob die Angabe im Interview oder Fragebogen erhoben wird und wie die Frage und die möglichen Antwortalternativen zu formulieren sind.

Das bei der Untersuchung von Patienten mit Morbus Crohn wichtige Merkmal der Häufigkeit weicher oder flüßiger Stühle erfordert zum Beispiel einen Zeitbezug bei der Befragung (z.B. "in der vergangenen Woche") und läßt sich sowohl quantitativ [8] als auch qualitativ ("sehr häufig", "normal") [153] erfassen.

Für die hier angesprochenen und beispielhaft dargestellten Entscheidungen gibt es keine allgemeingültigen Empfehlungen, sondern sie müssen häufig auf der Basis von Erfahrung und gesundem Menschenverstand getroffen werden. In manchen Fällen, wie etwa im Falle des Blutdrucks im obigen Beispiel, kann es zweckmäßig sein, verschiedene Vorgehensweisen in einer Pilotuntersuchung nebeneinander zu realisieren und dann im nächsten Schritt eine Entscheidung auf der Basis von Daten zur Validität oder Reliabilität zu fällen.

Zur Formulierung der Fragen und der Antwortalternativen in Patientenfragebögen liegen zahlreiche empirische Untersuchungen vor, die teilweise systematisch angelegt waren und teilweise nur Erfahrungsberichte sind. Die Meinungen unterschiedlicher Autoren sind hier allerdings nicht stets konsensfähig und so halten sich auch Lehrbücher, die diese Thematik aufgreifen, mit klaren Empfehlungen eher zurück. Wir wollen hier nur einige Aspekte kurz ansprechen, um die Vielfalt der Möglichkeiten und Probleme anzudeuten.

Der erste Aspekt betrifft die Formulierung der Fragen sowie ihre Anordnung. Fragen können sowohl positiv als auch negativ oder neutral formuliert sein, und dies kann durchaus einen Effekt auf die Beantwortung haben.

> In einer eigenen Untersuchung zur Entwicklung eines Lebensqualitätsfragebogens [154] für Patienten mit kleinzelligem Bronchialkarzinom wurden verschiedene Versionen des Fragebogens getestet, bei denen bei einigen Fragen die Polung vertauscht wurde (also z.B. "Ich fühlte mich überwiegend gesund" anstatt "Ich fühlte mich überwiegend krank"), während bei anderen der über 30 Fragen die Reihenfolge geändert wurde. Diese Änderungen hatten bei einigen Items einen deutlichen Einfluß auf die Häufigkeit der Antwortalternativen und auf die Interkorrelationen, somit auch auf die Reliabilität im Sinne von interner Konsistenz der Unterskalen des Fragebogens.

Ein Argument für die unterschiedliche Polung von Items und für eine Durchmischung bei der Reihenfolge besteht darin, daß die einzelnen Fragen sorgfältiger beantwortet werden, weil sie mehr Konzentration erfordern und ein automatisches, einheitliches Ankreuzen nicht möglich ist. Überlegungen zu sogenannten Antwortstilen der Patienten spielen auch eine Rolle für die Entscheidung bezüglich der Zahl der angebotenen Antwortkategorien. Eine ungerade Anzahl unterstützt die Tendenz mancher Personen, bei bipolaren Fragen immer in der Mitte anzukreuzen. Empfehlungen in der Literatur [67, 155] schlagen fünf bis sieben Antwortkategorien vor, dort sind auch Hinweise auf die wenigen hierzu vorliegenden empirischen Untersuchungen zu finden.

Im Hinblick auf eine evaluative Verwendung einer Skala ist das Konzept der Änderungssensitivität von Bedeutung, das in Abschnitt 3.4.3 eingeführt wurde. Dabei wurde stets davon ausgegangen, daß mit dem Meßinstrument der aktuelle Status des Patienten erfaßt wird und daß sich die Veränderung durch die Differenz der Skalenwerte abbilden läßt. Es gibt jedoch auch die Möglichkeit, die Fragen von vornherein so zu formulieren, daß sie eine Einschätzung der Veränderung beinhalten. Dies kann zu einer verbesserten Änderungssensitivität der Skala führen, hat allerdings mehrere Nachteile [52]. Zum einen ist eine genauere Beurteilung nur dann zu erwarten, wenn der Patient sich in seiner Beurteilung auf den richtigen Referenzzeitraum bezieht. Zum anderen geht bei Veränderungsfragen die Möglichkeit verloren, das absolute Niveau des Merkmals zu bestimmen. Bei Mehrfachmessungen im Behandlungsverlauf, bei denen sich die vom Patienten angegebene Veränderung stets auf den vorangegangenen Meßzeitpunkt bezieht, wird die Rekonstruktion der Veränderungen über längere Zeitintervalle wegen der fehlenden Transitivität unmöglich.

Die sprachliche Formulierung der Fragen ist auch dann ein Problem, wenn ein Fragebogen übersetzt wurde und man von einer Gleichwertigkeit des Meßinstruments mit dem Original ausgehen möchte. Hierzu wird heutzutage ein aufwendiges System von Hin- und Rückübersetzungen angewendet, auf da hier aber nicht genauer eingegangen werden kann.

5.2 Variablenselektion

Die Reduktion der großen Menge der Kandidatenvariablen auf eine möglichst kleine Teilmenge von Scorekomponenten soll nicht zu einem Informationsverlust relevanten Ausmaßes führen. Dies läßt sich allerdings, in Abhängigkeit von der Größe der Ausgangsmenge, nur schwer sicherstellen. Drei Schritte können bei der Variablenselektion unterschieden werden: eine Vorselektion durch Experten, ein erster empirisch begründeter Auswahlschritt durch vergleichende univariate Betrachtungen der Gütekriterien jedes Merkmals und schließlich die endgültige Reduktion aufgrund multivariater statistischer Auswertungen. Der letztgenannte Schritt ist eigentlich schon Teil der Modellsuche, die dann im nächsten Abschnitt behandelt wird.

Die empirische Prüfung von Güteeigenschaften der einzelnen für den Score in Frage kommenden Merkmale ist sicherlich der wichtigste Schritt bei der Variablenreduktion. Eine vorangehende Grobselektion der Kandidatenvariablen ist aber dann sinnvoll, wenn mehr Merkmale zur Verfügung stehen als empirisch geprüft werden können. Dieser Selektionsschritt bezieht in der Regel entweder medizinische Experten ein oder aber, speziell bei Selbstbeurteilungsfragebögen, betroffene Patienten.

> Cella et al. [12] präsentierten ihren Itempool einer Gruppe von Krebspatienten mit der Bitte, jedes Item auf einer vierstufigen Skala bezüglich seiner Relevanz zu beurteilen und erreichten in diesem Schritt eine Reduktion der Merkmalszahl von über 100 auf 38.

Steht nun ein für eine empirische Prüfung geeigneter Merkmalssatz zur Verfügung, so kann eine vergleichende Untersuchung der Reliabilität und Validität der einzelnen Merkmale erfolgen. Hinweise zur Planung solcher empirischer Untersuchungen werden in den Abschnitten 6.1 und 6.2 gegeben. Hier wollen wir nur einige Aspekte erwähnen, die in den entsprechenden Abschnitten nicht behandelt werden, weil sie die Variablenauswahl betreffen.

Beziehen wir uns zunächst auf eine validitätsgesteuerte Variablenselektion, wie man sie insbesondere bei der Konstruktion von Prognosescores durchführen wird. Die erste Frage, die hier zu diskutieren ist, betrifft die Abfolge bzw. das Ineinandergreifen von Merkmalsselektion und Modellwahl. Grundsätzlich kann man der Auffassung sein, daß eine univariate Validitätsbetrachtung der Kandidatenvariablen überhaupt nicht erforderlich ist, da die bei der multivariaten Modellanpassung angewandten Algorithmen dies in der Regel mit übernehmen. Eher noch haben viele Forscher die Sorge, daß im univariaten Schritt Merkmale als unbrauchbar ausgesondert werden, die bei der multivariaten Modellierung doch zur

Verbesserung der Validität beitragen würden. Solche Merkmale werden in der psychometrischen Literatur als Moderatorvariablen bzw. Suppressorvariablen beschrieben.

Dem Problem des Übersehens wichtiger Merkmale für einen Prognosescore steht ein anderes Problem gegenüber, nämlich das der Überinterpretation zufälliger Datenkonstellationen in einer hoch multivariaten Stichprobe. Während schon bei den univariaten Validitätsanalysen das Problem des multiplen Testens und der damit verbundenen erhöhten Irrtumswahrscheinlichkeit für den Fehler 1. Art gesehen werden muß, ist dies bei multivariaten Analysen umsomehr der Fall und praktisch kaum noch kontrollierbar. Wir werden darauf später noch eingehen, möchten aber an dieser Stelle schon feststellen, daß eine vorgeschaltete radikale Merkmalsselektion eine mögliche Strategie in diesem Sinne ist. Es gibt zum Beispiel konkrete Empfehlungen derart, daß die Anzahl der Variablen bei multivariaten Regressionsmodellen nicht mehr als ein Fünftel [156] bzw. gar ein Zehntel [157] der verfügbaren Patientenzahl betragen soll. Damit ist nicht die Zahl der endgültig im Modell verbleibenden Variablen, sondern die der für die Modellsuche zur Verfügung gestellten Variablen gemeint. Ein zweites, ganz praktisches Problem entsteht bei multivariaten Analysen oft durch das Vorliegen von fehlenden Werten. Mögliche Strategien im Umgang mit dieser Problematik werden noch genannt, jedoch ist hier anzumerken, daß auch diese Schwierigkeit durch eine radikale Vorselektion der Kandidatenvariablen erheblich entschärft werden kann.

Wie die Beurteilung der Validität der einzelnen Merkmale zu erfolgen hat, hängt nun von der Art des Kriteriums und des Merkmals selbst ab. Wenn auch alle Merkmale auf ihre Validität in Bezug auf dasselbe Außenkriterium untersucht werden, so wird man bei unterschiedlichen Merkmalsarten doch unterschiedliche Validitätsparameter verwenden müssen. Eine Möglichkeit der Vereinfachung besteht darin, jeweils den p-Wert des Tests der Nullhypothese "Es gibt keinen Zusammenhang zwischen Merkmal und Außenkriterium" heranzuziehen. Dies ist zwar kein Validitätsparameter im eigentlichen Sinne und er war in anderem Zusammenhang auch bereits kritisiert worden, bei der Variablenselektion ist dieses Vorgehen jedoch durchaus adäquat. Es wird in der Praxis häufig angewandt und dabei werden zum Beispiel alle Merkmale ausgesondert, die einen vorgegebenen p-Wert nicht unterschreiten. Für die Wahl dieses kritischen p-Werts gibt es unterschiedliche Vorschläge, die bei 0.05 oder darüber liegen, bis hin zu 0.25 [129]. Da man der multivariaten Modellsuche nicht zuviel vorwegnehmen möchte, werden wesentlich strengere Signifikanzschranken selten angelegt [158, 159]. Am ehesten geschieht dies bei Vorliegen sehr großer Stichproben und sehr vieler Kandidatenvariablen (vgl. Bemerkungen am Ende dieses Abschnitts).

Bei der Beurteilung der Validität eines mehrkategoriellen oder quantitativen Merkmals stehen grundsätzlich mehrere Möglichkeiten zur Verfügung, da der Zusammenhang mit dem Außenkriterium unterschiedliche Formen annehmen kann. Hier kann es sinnvoll sein, nach einer geeigneten Transformation des Merkmals zu suchen, zum Beispiel einer Dichotomisierung oder einer Logarithmierung der Meßwerte. Die Art der Transformation hat

einen erheblichen Einfluß auf den Validitätsparameter im Rahmen des Screenings der Kandidatenvariablen, wie an einer eigenen Untersuchung [160] illustriert wurde:

> In einer Untersuchung zur Prognose des kleinzelligen Bronchialkarzinom sollten verschiedene Tumormarker in ihrer prädiktiven Aussagekraft im Hinblick auf die Überlebenszeit verglichen werden [161]. Als Validitätsparameter im Variablenscreening diente der p-Wert im Proportional-Hazards-Modell. Neben der untransformierten Variablen wurde eine logarithmische Transformation sowie drei verschiedene Dichotomisierungen verglichen. dabei zeigten sich beispielsweise beim Tumormarker CEA p-Werte zwischen 0.017 und 0.32, die völlig konträre Aussagen zur Validität beinhalten.

Zwar ist es möglich, die Suche nach einer geeigneten Transformation erst im multivariaten Schritt und damit unter Berücksichtigung der anderen Merkmale durchzuführen, jedoch stößt dies in der Regel an die Grenzen der Praktikabilität. Daß man überhaupt nach solchen Transformationen sucht, hat zwei Gründe: erstens will man eine Validitätsoptimierung erreichen und zweitens, im Falle der Dichotomisierung, strebt man nach einer Vereinfachung der Scoremerkmale. Die datengesteuerte Suche nach einer optimalen Transformation bringt natürlich wiederum das Risiko einer Überanpassung mit sich, wenn nicht entsprechende Vorsichtsmaßnahmen getroffen werden. So ist etwa im Falle der Dichotomisierung eines quantitativen Merkmals davon abzuraten, alle möglichen Schwellenwerte durchzuprobieren und denjenigen mit dem höchsten daraus resultierenden Wert des Validitätsparameters zu wählen. Für dieses Vorgehen, das teilweise auch im Rahmen der multivariaten Strategie (z.B. bei CART) angewandt wird, läßt sich zwar eine Korrektur der berechneten p-Werte vornehmen [162], jedoch ist es zusätzlich geraten, von vornherein eine Beschränkung auf wenige, entweder bereits etablierte oder zweckmäßige (z.B. gerundete) Schwellenwerte vorzunehmen. Ob man überhaupt eine Klassierung bzw. speziell Dichotomisierung für quantitative Merkmale vornimmt, hängt auch davon ab, welche Modellklasse man im multivariaten Schritt verwenden möchte und wie stark man an einem einfach zu handhabenden Score interessiert ist.

Von einigen Autoren, z.B. Hosmer & Lemeshow [129], wird empfohlen, die Suche nach der besten Skalierung der Variablen erst nach der multivariaten Modellsuche vorzunehmen. Daß dieses Vorgehen nicht immer optimal ist, zeigt das folgende Beispiel.

> In einem gemeinsamen Projekt mit der Chirurgischen Universitätsklinik Heidelberg [163, 164] ging es um die Konstruktion eines Scores für Intensivpatienten, der in Anlehnung an den APACHE II Score eine Risikovorhersage leisten sollte, hier aber nicht für Neuaufnahmen sondern für Langzeitintensivpatienten (mind. 7 Tage Liegedauer). Es wurden dafür die im APACHE II enthaltenen Merkmale sowie weitere Kandidatenvariablen berücksichtigt, ihre Skalierung jedoch nicht von vornherein übernommen, da die Anwendungssituation sich von der des APACHE II unterschied. Von den aus dem APACHE II übernommenen Merkmalen war bekannt, daß sie fast alle

einen nicht-linearen Zusammenhang mit dem Außenkriterium (Überleben vs. Tod auf ICU) hatten. Die Einbeziehung der untransformierten Variablenwerte in ein logistisches Regressionsmodell ohne nähere Inspektion des Zusammenhangs hätte dazu geführt, daß die Validität dieser Merkmale aufgrund einer unpassenden Skalierung verwischt worden wäre. Es wurde daher ein einfacher Algorithmus zur nicht-parametrischen Regressionsschätzung angewandt, der eine vereinfachte Version des Ansatz von Cleveland [165] darstellt. Der U-förmige Zusammenhang einiger Merkmale mit dem Sterberisiko wurde dabei sehr deutlich. Die Wahl einer geeigneten Skalierung für das Merkmal konnte dann an dieser Regressionskurve orientiert werden.

Bei anschließender Modellierung im logistischen Modell ist es naheliegend, eine dem Modell entsprechende Transformation (in diesem Fall: logit-Transformation log p/(1-p)) des Sterberisikos zu wählen und diese in Abhängigkeit vom quantitativen Prädiktor darzustellen. Neben der nicht-parametrischen Regression wird hierfür vor allem die Anwendung von Spline-Funktionen empfohlen [166], mit denen eine stückweise Beschreibung mittels kubischer Funktionen zu glatten Anpassungen führt. Weitere Möglichkeiten sind bei Hosmer & Lemeshow [129] dargestellt.

Die Anzahl benötigter Patienten für ein sinnvolles Variablenscreening hängt davon ab, ob der Score nur wenige Merkmale mit hoher prädiktiver Validität beinhalten soll oder ob man auch Variablen mit mäßiger Validität berüchsichtigen möchte, um die Aussagekraft des Scores zu stärken. Fallzahlschätzungen können hier mit den gleichen Methoden erfolgen wie es bei Therapiestudien üblich ist. Aufgrund der üblicherweise großen Zahl von Kandidatenvariablen spielt hier allerdings das Problem des multiplen Testens eine erhebliche Rolle. So besteht bei einer statistischen Prüfung von 50 "unsinnigen" Prädiktoren (z.B. Zufallszahlen) bei Anwendung des üblichen zweiseitigen Tests mit Signifikanzniveau 5% bereits eine Wahrscheinlichkeit von 46%, daß drei oder mehr dieser "Prädiktoren" fälschlicherweise einen signifikanten Zusammenhang mit dem Kriterium zeigen [167]. Um sich in dieser Phase gegen eine Einbeziehung vieler nur zufällig "signifikanter" Prädiktoren zu schützen, ist entweder eine Verschärfung des p-Wertes als Auswahlkriterium oder eine Kreuzvalidierung erforderlich.

5.3 Suche des "besten" multivariaten Modells

Nach der im vorangegangenen Abschnitt beschriebenen Einschränkung der Menge der Kandidatenvariablen auf ein sinnvolles Maß wird man eine weitere Reduktion der Merkmalsmenge unter Berücksichtigung der multivariaten Zusammenhänge im Rahmen eines geeigneten Modells vornehmen. Um inhaltlich direkt an die Ausführungen des vorigen Abschnitts anzuschließen, beziehen wir uns zunächst wieder auf eine validitätsgesteuerte Scorekonstruktion.

Die Modellsuche, deren Ergebnis anschließend in eine Scoredefinition übersetzt wird, beinhaltet drei Schritte:
- die Wahl einer geeigneten Modellklasse,
- die Variablenselektion innerhalb der Modellklasse, sowie
- die Modellspezifikation mit den ausgewählten Variablen.

Die beiden letzten Schritte sind bei der praktischen Durchführung eng miteinander verknüpft, während der erste Schritt völlig unabhängig davon erfolgt. Bevor diese drei Schritte weiter ausgeführt werden, müssen allerdings zwei grundlegende Mißverständnisse ausgeräumt werden. Zum einen handelt es sich um die Annahme, daß es ein eindeutig bestes Modell gibt und daß dieses anhand eines beschränkten Datensatzes identifiziert werden kann. Zum zweiten begegnet man dem noch fundamentaleren Irrglauben, daß die Suche nach dem "besten" Modell durch einen geeigneten automatischen Computeralgorithmus eine richtige und objektive Lösung bringt.

Obwohl die Wahl der Modellklasse durch die Vorgabe des Außenkriteriums eingeschränkt wird, bleiben doch stets mehrere Alternativen bestehen. Bei dichotomen Außenkriterien kommen beispielsweise neben der häufig angewandten logistischen Regression auch verschiedene Diskriminanzanalysevarianten (z.B. robuste Diskriminanzanalyse, log-logistisches Modell [17]) sowie hierarchische Modelle (CART, [16]) in Frage. Die meisten Forscher entscheiden sich von vornherein für eine Modellklasse, wobei die Verfügbarkeit von Statistik-Software und die vorliegende persönliche Erfahrung vermutlich ausschlaggebend sind. Aus den relativ wenigen Publikationen [121, 124, 127, 168], in denen unterschiedliche Modellklassen direkt an den gleichen Daten verglichen werden, sind keine klaren Leitlinien zur Wahl der Modellklasse zu entnehmen, sondern es ergibt sich insgesamt eher der Eindruck der globalen Gleichwertigkeit, wobei sich das logistische Regressionsmodell grundsätzlich als robust erwiesen hat. Es ist nicht ausgeschlossen, daß für einen konkreten Datensatz eine der Modellklassen überlegen ist, jedoch läßt sich dies in der Regel erst durch Anwendung auf die Daten erkennen.

Die Variablenselektion im Rahmen von multivariaten Modellen dient einer weiteren Reduktion der Menge der Prognosevariablen, indem sie die Redundanz aufgrund multivariater Abhängigkeiten zwischen den Daten berücksichtigt. Dieser Schritt der Auswertung wird von gängigen statistischen Auswertungsprogrammen automatisch nach Wahl eines vorgegebenen Algorithmus (z.B. Vorwärts- vs. Rückwärtsselektion) geleistet, was durch die einfache Durchführbarkeit schon häufig zu Mißbrauch und Fehlinterpretationen geführt hat. Gerade bei diesem Teil der Auswertung ist jedoch die enge Verbindung von statistischem und klinischem Wissen besonders wichtig, wenn man zu einer sachgerechten Lösung kommen möchte. Feinstein [1] spricht von "clinico-statistical judgment" als Grundlage der empirisch gestützten Konstruktion von klinischen Skalen. Hosmer & Lemeshow [129] betonen, daß " Successful modeling of a complex data set is part science, part statistical methods, and part experience and common sense".

Die Variablenauswahl wird durch zwei Arten von Problemen stark erschwert:
– vom Vorliegen fehlender Werte
– von Abhängigkeiten in den Variablen.

Im Falle des Vorliegens zahlreicher fehlender Werte wird die Durchführung der Variablenselektion im multivariaten Modell umständlich. Man kann entweder die fehlenden Werte durch "plausible" Werte ersetzen, hierzu gibt es eine große Zahl von Vorschlägen in der Literatur [169]. Die Alternative besteht darin, daß man die Fälle mit fehlenden Werten aus der Auswertung herausläßt, was jedoch schnell zu einer Reduktion der verfügbaren Patientenzahl um weit mehr als die Hälfte führen kann, wenn bei verschiedenen Variablen unterschiedliche Patienten betroffen sind. Folgt man dennoch dieser Strategie, so kann die Variablenselektion nicht mehr in einem Auswertungslauf durchgeführt werden, sondern es muß mit jeder Änderung des ausgewählten Datensatzes auch die Teilmenge der Patienten mit vollständigen Daten neu generiert werden. Es empfiehlt sich dabei, Merkmale mit vielen fehlenden Werten möglichst frühzeitig auszusondern, wenn sie nicht ganz erheblich zur prognostischen Validität beitragen, zumal zu befürchten ist, daß diese Merkmale auch später bei der Anwendung des Scores eventuell häufiger nicht verfügbar sein werden.

In den meisten Datensätzen sind die Kandidatenvariablen nicht unabhängig voneinander, sondern es liegen zum Teil erhebliche Kollinearitäten vor, zum Beispiel wenn sowohl der systolische als auch der diastolische Blutdruckwert berücksichtigt werden. In solchen Fällen ist das Ergebnis von Variablenselektionsverfahren in höchstem Maße zufallsabhängig, da beide Merkmale praktisch gleichgute Prädiktoren darstellen und die Auswahl einer der beiden von der spezifischen Datenkonstellation in der Stichprobe abhängt. Mit Hilfe der Bootstrap-Technik läßt sich das Ausmaß dieser Zufälligkeit gut sichtbar machen, wie Sauerbrei & Schumacher [170] zeigen. Dabei werden viele sogenannte Bootstrap-Stichproben aus dem Datensatz generiert und an jeder wird der Prozeß der Variablenselektion wiederholt und das Ergebnis festgehalten. Man kann dann sehen, wie sehr die Teilmengen der ausgewählten Variablen variieren und kann außerdem die Variablenauswahl darauf basieren lassen, welche Merkmale in den Replikationen am häufigsten ausgewählt wurden. Weitere Informationen zum Bootstrap-Verfahren sind bei Efron & Tibshirani [171] zu finden.

Von statistischer Seite sind Vorschläge zum Umgang mit Kollinearitäten gemacht worden, die auf eine Zusammenfassung der abhängigen Merkmale in Form von Hauptkomponenten hinauslaufen. Dies ist jedoch für klinische Scores ein völlig ungeeignetes Vorgehen, da es die Scoreerhebung bzw. die Berechnung erheblich erschwert. Stattdessen ist es angemessen, die Entscheidung für die Bevorzugung eines der Merkmale von dem klinischen Vorwissen oder von Praktikabilitätsaspekten abhängig zu machen.

Wir möchten kurz auf einige weitere Aspekte der Modellsuche hinweisen, ohne daß näher darauf eingegangen werden kann. Hierzu gehört die Modellierung von Wechselwirkungen zwischen Prädiktorvariablen, die durch die Definition neuer kombinierter Merkmale berücksichtigt werden können. Ebenfalls nur erwähnt sei das Problem von sogenannten

Ausreißern oder einflußreichen Beobachtungen im Datensatz ("influential observations"). Abhängig von der Art der Ausreißer existieren verschiedene Möglichkeiten, damit umzugehen. Insgesamt muß man feststellen, daß der Prozeß der Variablenselektion stark anfällig gegen Zufallseinflüße zweierlei Art ist: solche aufgrund des Stichprobenfehlers und solche aufgrund von Entscheidungen des Auswerters.

Zur Kalkulation der erforderlichen Fallzahl für eine Modellanpassung kann man die im vorigen Abschnitt zitierten Aussagen, die sich auf eine sinnvolle Variablenzahl bezogen, umkehren. Demnach empfiehlt es sich, etwa zehnmal so viele Patienten zur Verfügung zu haben als Variablen für die multivariate Modellbildung zur Verfügung stehen. Daraus folgt, daß man in den meisten Fällen Stichprobengrößen der Ordnung 100 oder wesentlich darüber benötigt.

Der letzte Schritt im Rahmen der Scorekonstruktion ist die Vereinfachung des gefundenen statistischen Modells zu einem Score. Bei verallgemeinerten linearen Regressionsmodellen geschieht dies meist durch Rundung der geschätzten Koeffizienten. Wenn man bedenkt, daß viele klinische Scores unter Praxisbedingungen eingesetzt werden, wo eine einfache Berechenbarkeit, maximal unter Zuhilfenahme eines Taschenrechners, gewährleistet sein soll, so scheint dies sinnvoll. Darüberhinaus hat sich in vielen Anwendungsfällen gezeigt, daß eine Rundung der Koeffizienten fast keinen Validitätsverlust mit sich bringt, sondern im Gegenteil in manchen Fällen bei einer Validierung an unabhängigen Daten noch besser abschnitt, weil das ursprüngliche Modell überangepaßt war.

Bisher hatten wir uns überwiegend auf eine Scorekonstruktion mit dem Ziel der Validitätsoptimierung bezogen und diese Priorität auch begründet. Zum Abschluß dieses Abschnitts soll noch kurz auf die reliabilitätsgesteuerte Merkmalsauswahl und Scorekonstruktion eingegangen werden. Diese spielt eine Rolle bei klinischen Skalen, für die kein überzeugendes Referenzkriterium zur Validitätsbestimmung existiert. Man geht dabei aus von einer Vorauswahl von Merkmalen bzw. Fragebogenitems, die auf Überlegungen zur inhaltlichen Validität basiert, wie in Abschnitt 5.1 dargestellt. Univariate, d.h. für jedes Merkmal getrennte Reliabilitätsbetrachtungen erfordern ein spezielles Studiendesign, eine einmalige Erhebung an einer Stichprobe reicht hierzu nicht aus (vgl. Abschnitt 6.1). Im Falle von Selbstbeurteilungsskalen wird sich die univariate Reliabilitätsuntersuchung in der Regel auf eine Testwiederholung beziehen, bei auf Fremdeinschätzung basierenden Scores auf einer Inter-Rater-Vergleich. Eine Variablenselektion wird man einfach anhand eines Vergleichs der univariaten Reliabilitätsparameter durchführen, hierzu sind keine weiteren Erläuterungen erforderlich. Dieses Vorgehen wird aber aufgrund des erhöhten Aufwands bei der Studiendurchführung selten praktiziert.

Stattdessen werden häufig Reliabilitätsbetrachtungen im Sinne einer internen Konsistenz der Skala durchgeführt. Diese multivariaten Auswertungen können an einer einfachen Stichprobe ohne jegliche Meßwiederholung gemacht werden. Hierzu wird die Zusammenhangsstruktur der Items oder Scorekomponenten mit einem geeigneten statistischen Verfahren analysiert

und solche Items eliminiert, die sich aufgrund der Interkorrelationen als Fremdkörper im Score erweisen. Man verwendet als statistisches Verfahren entweder die Faktorenanalyse oder die Itemanalyse mit Berechnung von Cronbach´s α. Die zur Verfügung stehenden Auswertungsprogramme berechnen die Werte des Koeffizienten α jeweils unter Auslassung eines Items, so daß erkennbar wird, welche Items die interne Konsistenz herabsetzen und deshalb aus der Skala eliminiert werden sollten.

6 ASPEKTE DER STUDIENPLANUNG

Von Seiten der Medizinischen Biometrie wird oft darauf hingewiesen, daß die entscheidende Voraussetzung für das Erreichen gut interpretierbarer Studienergebnisse nicht eine aufwendige statistische Auswertung, sondern eine optimale Versuchsplanung ist. Im Rahmen des Studiendesigns muß dafür gesorgt werden, daß sowohl die interne als auch die externe Validität der Untersuchung gewährleistet ist. Die interne Validität bezieht sich auf die Schlüssigkeit der Studienergebnisse, die in möglichst geringem Ausmaß von systematischen und zufälligen Fehlern gestört sein soll. Maßnahmen wie die Fallzahlplanung und Maskierungstechniken gehören in diesen Bereich. Die externe Validität betrifft die Übertragbarkeit der Studienergebnisse auf andere, vergleichbare Anwendungssituationen.

Die Priorität der Studienplanung vor der Auswertungsmethodik erklärt sich daraus, daß einmal gemachte Planungsfehler in der Regel nicht mehr reparierbar sind, während eine inadäquate statistische Auswertung prinzipiell korrigiert werden kann. Im Bereich der klinischen Therapieforschung wurde dies längst erkannt und es sind dort inzwischen mehr als zwanzig Lehrbücher, die sich nur mit der Methodik klinischer Therapiestudien, und dabei überwiegend mit Aspekten der Studienplanung, beschäftigen. Daß dem Gebiet der Versuchsplanung im Rahmen von Therapiestudien so viel Interesse gewidmet wurde, ist sicher auch darin begründet, daß die Konsequenzen schlecht geplanter und damit eventuell zu falschen Ergebnissen führender Therapiestudien klar erkennbar und vor allem nicht unerheblich sind. Inzwischen ist auch auf dem Gebiet der Diagnosestudien die Bedeutung der Studienplanung erkannt worden [55].

Im folgenden soll dargestellt werden, welche Probleme bei der Planung von Studien zur Konstruktion und Evaluierung von klinischen Skalen und Scores zu beachten sind und wie man diese vermeiden kann. Dabei ergibt sich hier keine so einheitliche Darstellung wie etwa im Falle der Planung von Therapiestudien, da die Probleme unterschiedlich sind in Abhängigkeit von der Zielsetzung des Scores. In der folgenden Darstellung wird primär danach unterschieden, ob das Studienziel die Untersuchung der Reliabilität oder der Validität eines Scores ist. In der Praxis wird dies tatsächlich fast immer getrennt betrachtet, jedoch sei hier schon der Vorschlag gemacht, in Zukunft die Untersuchung der beiden Gütekriterien nach Möglichkeit in einer Studie zu integrieren und damit auch die Voraussetzung für weitergehende Erkenntnisse zu schaffen.

6.1 Planung von Reliabilitätsstudien

Studien zur Untersuchung der Reliabilität klinischer Messungen sind in der wissenschaftlichen Literatur der letzten zwanzig Jahre immer häufiger zu finden. Dies gilt insbesondere auch für das Gebiet klinischer Skalen, da hier das zum Teil große Angebot an

konkurrierenden Meßinstrumenten eine Auswahl anhand von Aspekten der Meßqualität nahelegt. So gibt es zahlreiche Publikationen, in denen Vergleiche verschiedener Skalen in Bezug auf ihre Reliabilität vorgenommen werden. Methodische Aspekte der Studiendurchführung und der Auswertung scheinen aber bei Reliabilitätsstudien weit weniger standardisiert zu sein als etwa im Bereich von Therapiestudien. In diesem Abschnitt wollen wir eine Reihe von Planungsprinzipien für Reliabilitätsstudien kurz erläutern und zusammenstellen.

Bei Therapiestudien gilt es seit vielen Jahren als unabdingbar, daß im Rahmen der Studienplanung prospektiv alle relevanten Aspekte der Durchführung und Auswertung in einem Studienprotokoll schriftlich festgehalten werden. Dies hat einerseits Kontrollfunktion und soll den prospektiven Charakter der Studie gewährleisten. Andererseits hat es sich aber auch als qualitätssichernd herausgestellt, daß im Laufe der Zeit Standards für Studienprotokolle entstanden, die dafür sorgten, daß bei der Studienplanung an alle wichtigen Aspekte gedacht wurde. Gliederungsvorschläge für Studienprotokolle sind in zahlreichen Lehrbüchern und Publikationen zu finden, auch wir haben dazu in der Vergangenheit einen sehr ausführlichen Vorschlag erarbeitet [172]. Die folgenden Ausführungen entsprechen grob der Grundstruktur einer solchen Gliederung.

6.1.1 Ziele von Reliabilitätsstudien

Die Zielsetzung von Reliabilitätsstudien, speziell die Konsequenzen, die aus möglichen Studienergebnissen gezogen werden sollen, ist in vielen Publikationen nicht klar erkennbar. So wurden bisher überwiegend nicht-vergleichende Studien durchgeführt, und dabei nicht einmal ein absoluter Maßstab für eine Beurteilung der Reliabilität genannt. Grundsätzlich sind Reliabilitätsstudien in verschiedenen Phasen der Entwicklung eines klinischen Scores angebracht und sie werden dementsprechend unterschiedliche Zielsetzungen haben. In einer frühen Phase der Entwicklung des Meßinstruments können Reliabilitätsuntersuchungen dazu dienen, unreliable Teile (Komponenten, Items) des Meßinstruments frühzeitig zu erkennen, um sie entweder zu verbessern oder zu eliminieren. Darüberhinaus kann die Zielsetzung darin bestehen, die reliabilitätsbeeinflußenden Faktoren (Varianzkomponenten) zu untersuchen und in ihrer quantitativen Auswirkung zu vergleichen. Die Umsetzung solcher Erkenntnisse würde darin bestehen, durch geeignete Maßnahmen der Meßwiederholung (zum Beispiel Vergrößerung der Itemzahl oder erneute Befragung) die Reliabilität auf die ökonomisch vorteilhafteste Art zu erhöhen (siehe Beispiel zu D-Studien bei Streiner & Norman [67]).

Die gerade genannten Zielsetzungen von Reliabilitätsstudien sollten eigentlich noch während der Entwicklung einer klinischen Skala konkretisiert werden bzw. obligatorischer Teil der Skalenentwicklung sein. Sobald die Entwicklung des Meßinstruments als abgeschlossen betrachtet werden kann, rücken andere Fragestellungen in den Vordergrund. Dies sind zum einen die Übertragbarkeit von Reliabilitätsaussagen auf andere Settings oder Einsatzbereiche des Scores sowie die Untersuchung von Maßnahmen zur Verbesserung der Reliabilität. Die

Anwendung auf andere Patientengruppen oder die Einbeziehung anders ausgebildeter Beurteiler kann einen erheblichen Einfluß auf die Reliabilität von klinischen Scores haben.

> Wernick et al [173] ließen einen histologischen Score zur Beurteilung von Lupus nephritis durch fünf nicht-universitäre Pathologen zweimal im Abstand von ca. 9 Monaten auf die gleichen 25 Präparate anwenden, wobei eine ausführliche schriftliche Anleitung zur Punktevergabe bei den einzelnen Komponenten gegeben wurde. Während frühere Studien mit erfahrenen Universitätspathologen eine sehr gute Übereinstimmung berichtet hatten, ergaben sich in dieser Untersuchung nur sehr mäßige Übereinstimmungen sowohl zwischen als auch innerhalb der Beurteiler.
>
> Klinkhoff et al. [174] untersuchten den Effekt eines Trainings zur Standardisierung der Untersuchungstechnik auf die Reliabilität eines Gelenk-Scores bei Patienten mit rheumatoider Arthritis. Sie fanden eine Reduktion des relativen Anteils der Beobachtervarianz an der Gesamtvarianz von 13.8% auf 3.2% aufgrund des Trainings.

Die Fragestellung, die in Zukunft wegen des zunehmenden Angebots an klinischen Scores jedoch immer mehr in den Vordergrund rücken wird, beinhaltet den Vergleich konkurrierender Scores. Hier wird der Reliabilitätsvergleich von erheblichem Interesse sein, insbesondere wenn für Validitätsstudien kein überzeugendes Kriterium zur Verfügung steht. Aus statistischem Blickwinkel führen die unterschiedlichen Zielsetzungen zu einer der drei folgenden Möglichkeiten:

- Schätzung der Reliabilität eines Scores oder seiner Komponenten
- relative Bewertung von Einflußfaktoren auf die Reliabilität
- Vergleich der Reliabilität konkurrierender Scores

6.1.2 Design von Reliabilitätsstudien

Zur Konkretisierung des Studiendesigns gehört die Auswahl der Patientengruppe, der Scores sowie der Variationsfaktoren und ihrer Ausprägungen, also zum Beispiel der Beurteiler.

Die Auswahl der Patientengruppe für die Studie sollte sich nach dem intendierten Einsatzbereich des Scores richten, also für die spätere Zielpopulation möglichst repräsentativ sein. Während es durchaus sinnvoll sein kann, einen für eine bestimmte Patientenpopulation konstruierten Score zu Vergleichszwecken auch einmal an einer Gruppe von Gesunden zu erheben, so ist hiervon im Rahmen der Reliabilitätsuntersuchung abzuraten. Insbesondere die Vermischung von gesunden und kranken Personen führt zu einer artifiziellen Vergrößerung der Merkmalsvarianz und damit zu einer zu optimistischen Schätzung einiger Reliabilitätsparameter (z.B. der Intraklass-Korrelation). Eventuell kann eine Einschränkung der Patientengruppe erforderlich sein, wenn die Reliabilitätsstudie besondere Anforderungen an den Untersuchungsaufwand der Patienten stellt, wie etwa die Verfügbarkeit für Wiederholungsmessungen. Typische Ausschlußkriterien beziehen sich auf Patienten, bei denen der Score aus bestimmten Gründen nicht sinnvoll anwendbar ist, z.B. bei

Sprachproblemen im Falle von Selbstbeurteilungsskalen. Die Häufigkeit des Vorliegens solcher Gründe muß dann im Rahmen der Studie unbedingt dokumentiert werden, da sie Aussagen über die Anwendbarkeit des Scores in der Praxis zuläßt

Die Auswahl der Scores in vergleichenden Reliabilitätsstudien hängt in erster Linie davon ab, wieviele Scores für das interessierende Einsatzgebiet zur Verfügung stehen. Unter diesen wird man solche nicht berücksichtigen, die einen unverhältnismäßig hohen Aufwand zu ihrer Erhebung erfordern oder im Rahmen der geplanten Studie nicht praktisch durchführbar sind. Die Anzahl auszuwählender Scores hängt weiterhin davon ab, ob diese unabhängig und gleichzeitig an den gleichen Patienten erhoben werden können, oder ob man sie an unterschiedlichen Patienten evaluieren muß. Letzteres würde die für die Studie benötigte Patientenzahl erheblich beeinflussen.

Zur Operationalisierung der untersuchten Einflußfaktoren gehört die Festlegung geeigneter Zeitabstände für Wiederholungsmessungen und vor allem die Auswahl der Beurteiler bei Studien zur Inter-Rater-Reliabilität. Gerade der letztgenannte Aspekt wurde bei vielen Reliabilitätsstudien bisher vernachlässigt, indem nur zwei Beurteiler beteiligt und über deren Auswahl keine Aussagen gemacht wurden. Eine Verallgemeinerung der Studienergebnisse wäre in einem solchen Fall sehr problematisch und würde häufig zu überoptimistischen Aussagen führen, da sicherlich oft zwei Beurteiler an einer Studie beteiligt sind, die den gleichen Erfahrungshintergrund und somit ähnliche Beurteilungskriterien haben. Es sollte deshalb die für die praktische Anwendung des Scores relevante Grundgesamtheit von Beurteilern definiert werden und daraus eine zufällige, bei wenigen Ratern besser eine repräsentative Auswahl genommen werden. Dies impliziert in der Regel, daß die Beurteiler aus verschiedenen Zentren kommen und eine gewisse Varianz an Berufserfahrung aufweisen sollten.

Die Beteiligung von Beurteilern aus verschiedenen Zentren kann Durchführungsprobleme mit sich bringen, wenn die Scoreerhebung direkt am Patienten geschieht und die Eingangsinformation nicht konservierbar ist. Hier wird ein erheblicher organisatorischer Aufwand erforderlich, um Rater zur gleichzeitigen Beurteilung einer ausreichenden Zahl von Patienten zusammenzubringen und diese in geeigneter Weise durchführen zu lassen (vgl. Bemerkungen zur Unabhängigkeit in Abschnitt 6.1.3).

Bezüglich der Anzahl der zu beteiligenden Rater liegen keine statistisch begründeten Aussagen vor, jedoch erscheint eine Zahl von fünf bis sieben geeignet zu sein. Bei einer geringeren Anzahl ist die erforderliche Repräsentativität praktisch nicht realisierbar. Hat man gar nur zwei oder drei Rater, wie man es in vielen publizierten Studien findet, so ist es nicht möglich, einen Beurteiler mit klar diskrepanten Einschätzungen zu identifizieren. Dies sollte jedoch obligatorischer Teil der Auswertung von Studien zur Interrater-Reliabilität sein, da es leicht vorkommen kann, daß einer der Beurteiler einen Teil der Scoredefinition völlig anders versteht bzw. mißversteht. Eine einfache Strategie des statistischen Vorgehens zur Erkennung solcher Beurteiler besteht darin, die Berechnung von Reliabilitätsmaßen jeweils unter

Auslassung eines Raters zu wiederholen und nachzusehen, ob sich dadurch die Ergebnisse wesentlich ändern.

Für die Festlegung geeigneter Zeitintervalle von Meßwiederholungen können hier nur grundsätzliche Aussagen gemacht werden, da dies sehr stark vom konkreten Beispiel abhängt. Es ist darauf zu achten, daß das Zeitintervall etwa die Schwankungsbreite ausschöpft, die für den Score in der praktischen Anwendungssituation vorgesehen ist. So käme für einen neurologischen Score für Intensivpatienten ein Zeitintervall von 12 Stunden in Frage, wenn dieser vor allem als Prognosescore am Tag der Aufnahme herangezogen werden soll. Hat man es hingegen mit einer Lebensqualitätsskala zu tun, die als Zielgröße in klinischen Studien verwendet werden soll, so interessiert man sich eigentlich für die durchschnittliche Lebensqualität in einem z.B. vierwöchigen Zeitintervall zwischen zwei Erhebungen und wird deshalb Meßwiederholungen eher im Abstand von zwei Wochen machen [87].

Weitere mögliche Einflußfaktoren auf die Messung der Scorewerte, die nicht als Varianzkomponente im Rahmen der Studie untersucht werden sollen, sind im Studienprotokoll aufzuführen und es ist festzulegen, ob sie zufällig variieren dürfen, ob sie durch Standardisierungsmaßnahmen möglichst ausgeschaltet oder ob sie in ihrem systematischen Effekt untersucht werden sollen. Eine gewisse Standardisierung der Erhebung des Scores ist meistens sinnvoll, hierzu gehören Angaben zum Zeitpunkt und zu den Durchführungsbedingungen (Umgebung, beteiligte Personen). So können bei vielen klinischen Scores die Eingangsinformationen prinzipiell sowohl per Fragebogen erhoben werden als auch im persönlichen Gespräch oder telefonisch abgefragt werden. Während hier in vielen Studien auf konstante Durchführungsbedingungen Wert gelegt wurde, haben Chambers et al. [68] dies als systematischen Faktor mit untersucht.

Ein anderer wichtiger Einflußfaktor ist die Erfahrung der Beurteiler im Umgang mit dem Score. Neben der Möglichkeit, dieses als Teil der zufälligen Beurteilerheterogenität anzusehen, kann es aber auch interessant sein, hierzu Aussagen über systematische Effekte machen zu können, indem man die Studie in zwei Durchgänge aufteilt, zwischen denen ein Ratertraining stattfindet. Speziell bei Scores mit unzureichender Interrater-Reliabilität sollte überprüft werden, ob diese durch solche Trainingsmaßnahmen verbessert werden kann.

Die Frage, ob retrospektive Reliabilitätsstudien methodisch akzeptabel sind, stellt sich in der Regel nicht. Dies liegt darn, daß Reliabilitätsstudien fast immer ein Design erfordern, welches nicht dem klinischen Ablauf entspricht und somit nicht einer klinischen Dokumentation entnommen werden kann. Weder ist es in der klinischen Routine üblich, subjektive Fremdeinschätzungen durch mehr als einen Beurteiler vornehmen zu lassen und zu dokumentieren, noch werden Meßwiederholungen in so kurzen Abständen durchgeführt, daß man von einer Stabilität des Merkmals ausgehen kann.

6.1.3 Ausschaltung systematischer und zufälliger Fehler

Zur biometrischen Studienplanung gehören auch Überlegungen, wie systematische Fehlerquellen zu vermeiden sind und welche Fallzahlen für statistisch sichere Aussagen erforderlich sind.

Zur Gewährleistung der Unabhängigkeit der Beurteiler gehören Maßnahmen wie die gegenseitige Verblindung der Beurteiler, auf die besonders geachtet werden muß, wenn die Beurteilung zur gleichen Zeit im gleichen Raum vorgenommen wird. Wird die Intrarater-Reliabilität untersucht, so müssen Wiedererkennungseffekte ausgeschlossen werden. Hierzu ist es sinnvoll, eine große Zahl von Patienten einzubeziehen und das Zeitintervall zwischen den zwei Beurteilungen möglichst groß zu halten.

> Damiano et al. [51] stellen eine Reliabilitätsuntersuchung zum APACHE II Score vor. Dabei haben sie die Studie zwar multizentrisch mit einer recht großen Fallzahl von 196 Patienten durchgeführt, aber pro Patient nur zwei Beurteiler herangezogen. Da einer der zwei Beurteiler jeweils ein Experte in Bezug auf die Scoreerhebung war und vorher ein dreitägiges Training stattfand, waren gute Bedingungen zur Erreichung einer hohen Reliabilität gegeben. Zusätzlich wurde den Beurteilern die Möglichkeit der telefonischen Rückfrage bei Zweifelsfällen gegeben, was in Bezug auf die Unabhängigkeit etwas fragwürdig erscheint.

Zur Fallzahlplanung von Reliabilitätsstudien können kaum allgemeine Aussagen gemacht werden und dementsprechend wenig Hinweise finden sich in der Literatur. Dies liegt zum einen daran, daß die inferenzstatistische Auswertung hier grundsätzlich komplexer ist, weil sie sich auf Varianzmaße statt auf Mittelwerte bezieht, was durch die unterschiedlichen Designs noch erschwert wird. Zum anderen wird die Fallzahlplanung dadurch erschwert, daß die verwendeten Reliabilitätsparameter nicht sehr anschaulich sind und deshalb die Formulierung eines relevanten Effekts oder Unterschieds recht willkürlich bleibt. An dieser Stelle sei hier deshalb nur auf einige Literaturstellen verwiesen, die Nomogramme als Hilfsmittel zur Fallzahlplanung für einfache Designs präsentieren [57, 103].

6.2 Planung von Validierungsstudien

Im vorangegangenen Abschnitt wurde auf die verschiedenen Aspekte der Planung von Reliabilitätsstudien eingegangen, jetzt soll dies in analoger Weise für Validierungsstudien erfolgen. Auch in diesem Bereich ist die wissenschaftliche Literatur lange nicht so zahlreich wie für die Planung von vergleichenden Therapiestudien. Dennoch gab es in den letzten Jahren einige Beiträge, in Deutschland vor allem zur Validierung diagnostischer Tests [55], international mehr im Bereich der Medizinischen Informatik, die sich mit der Evaluierung computergestützter Entscheidungshilfen befassen [175, 176]. Für den Bereich der Prognose wurde in einer kürzlich erschienenen Publikation [177] beklagt, daß Richtlinien für die Planung und Auswertung solcher Studien fast völlig fehlen. Im folgenden beziehen wir uns

stets auf Studien zur Validierung klinischer Scores, und zwar nur auf die Untersuchung der Kriteriumsvalidität.

6.2.1 Ziele von Validierungsstudien

Das bekannte Schema der vier Phasen der Arzneimittelprüfung, die vor allem durch ihre unterschiedlichen Zielsetzungen definiert sind, hat verschiedene Autoren angeregt, eine ähnliche Phaseneinteilung auch für Studien zur Evaluierung von Diagnoseverfahren [55] bzw. von prognostischen Faktoren [177] vorzuschlagen. Die hier zu diskutierenden Validierungsstudien würden in den Schemata der beiden Autorengruppen als Phase III Studien bezeichnet werden. Sie lassen sich charakterisieren durch eine in der Regel prospektive Durchführung mit dem Ziel der konfirmativen Analyse einiger weniger, auf die Validität eines vorgegebenen Scores bezogener Fragestellungen. Explorative Auswertungen, die der Konstruktion und Optimierung eines Scores dienen, sind den Phasen I und II zuzurechnen, hierauf wurde im Kapitel 5 ausführlich eingegegangen. Besonders bedeutsam ist die von Köbberling et al. [55] definierte Phase IV, in die eine Bewertung des klinischen Nutzens eines diagnostischen oder prognostischen Scores fällt. Die Planung solcher Studien wird in einem späteren Abschnitt beschrieben.

Bei Validierungsstudien der Phase III, im weiteren einfach als Validierungsstudien bezeichnet, gehen wir davon aus, daß ein Score in seiner endgültigen Form vorliegt, einschließlich einer genauen Definition seiner Komponenten und ihrer Kombination zu einem Gesamtwert. Die Notwendigkeit der Validierung eines neuen Scores ergibt sich aus zweierlei Gründen: Erstens ist bei vielen Scores die Konstruktion so erfolgt, daß eine Optimierung der Kriteriumsvalidität in dem für die Konstruktion zur Verfügung stehenden Datensatz erreicht wurde. Wie bereits dargelegt wurde, führt dies zu einer verzerrten, überoptimistischen Schätzung der Validität. Eine unverzerrte Schätzung kann nur an einem unabhängigen Datensatz gewonnen werden. Zweitens stellt sich auch bei Vorliegen einer unverzerrten Validitätsschätzung die Frage, ob diese auch bei Anwendung des Scores in einem vergleichbaren, aber anderen Umfeld (z.B. in einer anderen Klinik) ihre Gültigkeit behält. Hier gibt es zahlreiche Beispiele in der Literatur, in denen sich erhebliche Validitätseinbußen bei Anwendung auf ein anderes Patientengut und durch anderes Personal ergaben [178, 179].

Der nächste Schritt der Validierung schließt den Vergleich mit anderen Scores ein. Diese Art von Studien hat in den vergangenen Jahren an Bedeutung gewonnen, seitdem es für einige Indikationsgebiete bereits fünf oder mehr konkurrierende Scores gibt [180]. Aus statistischem Blickwinkel sind also als mögliche Zielsetzungen von Validitätsstudien zu unterscheiden die Schätzung der Validität eines Scores und der Vergleich der Validität konkurrierender Scores.

6.2.2 Design von Validierungsstudien

Die Festlegung des Patientenkollektivs für eine Validierungsstudie hängt vor allem vom Verwendungszweck des Scores ab. In dieser Phase der Validitätsprüfung sollte besonders darauf geachtet werden, daß eine exakt beschriebene Stichprobe von Patienten möglichst vollständig in die Studie aufgenommen wird, da durch den Ausschluß gewisser Patienten die Ergebnisse der Studie systematisch beeinflußt werden können. Eine Überschätzung der Validität kann leicht zustandekommen, wenn statt einer repräsentativen Stichprobe eine stratifizierte Stichprobe ausgewählt wird, bei der die Extremgruppen an den Rändern des Krankheitsschwerespektrums überrepräsentiert sind.

Die Vollständigkeit ist auch deshalb von Bedeutung, damit untersucht werden kann, ob in der Praxis der Score auch tatsächlich bei allen Patienten angewandt werden kann. Im Gegensatz zu Reliabilitätsstudien ist es hier von großer Bedeutung, daß es sich bei den Patienten um eine in Bezug auf den Zeitpunkt der Aufnahme in die Studie homogene Gruppe handelt ("inception cohort" [111]). Die Wahl eines geeigneten Zeitpunktes, also beispielsweise die Diagnosestellung oder die Aufnahme auf die Intensivstation, hängt ebenfalls vom späteren Einsatzzweck des Scores ab.

Die Auswahl der Scores bei vergleichenden Studien wird im wesentlichen durch den Erhebungsaufwand eingeschränkt. Oft haben konkurrierende Scores einen erheblichen Anteil gemeinsamer Komponenten, so daß Studien mit fünf oder mehr Scores durchaus durchführbar sind. Erfordert ein Score jedoch bestimmte, zum Beispiel apparative Voraussetzungen, die nicht immer gegeben sind, so wird dies eher gegen seine Berücksichtigung sprechen.

Die Wahl des Kriteriums, an dem die Validität der Scores gemessen wird, ist eine wichtige Entscheidung im Rahmen der Studienplanung. Üblicherweise geht man davon aus, daß die Referenzmethode bzw. das Außenkriterium exakt, also meßfehlerfrei, bestimmt werden können. Natürlich ist es in der Praxis oft schwierig, ein tatsächlich fehlerfreies Kriterium zu finden. Es sollte aber darauf geachtet werden, daß das Kriterium deutlich sicherer zu bestimmen ist als der Score und somit auch als Maßstab gelten kann. Problematisch wird es erst, wenn das Kriterium so unreliabel oder invalide ist, daß in einer nennenswerten Zahl von Fällen der Score eine richtige und das Kriterium gleichzeitig eine falsche Aussage liefert. Ist dies der Fall, so lassen sich Ergebnisse von Validitätstudien nur schwer interpretieren. Eine Gefahr besteht jedoch darin, ein relevantes, aber nicht vollständig reliables Kriterium durch ein reliables, aber weniger relevantes zu ersetzen. Im Zweifelsfalle sollte ein in der klinischen Praxis etabliertes Kriterium gewählt werden.

In manchen Fällen wird die Experteneinschätzung das einzige mögliche Referenzkriterium sein. Da allerdings auch Expertenmeinungen nicht absolut valide sind, empfiehlt sich die Beteiligung mehrerer Fachleute in Form eines Expertenpanel oder "peer review committee" [176], von dem alle Patienten beurteilt werden. Dies erfordert jedoch einen hohen organisatorischen Aufwand, wie schon im Zusammenhang mit Interrater-Reliabilitätsstudien

angesprochen; tatsächlich würde eine solche Studie bei diesem Vorgehen automatisch enthalten sein.

Bei prognostischen Scores ist das Kriterium selbst meist wenig problematisch. Oft handelt es sich um das Eintreten eines klinisch klar definierten Ereignisses, im Extremfall z.B. des Todes. Will man aus Gründen, die oben dargestellt wurden, jedoch eine Dichotomisierung des Krankheitsverlaufes vornehmen, so gehört zur Definition des Kriteriums vor allem die Festlegung eines Zeitraumes für das Auftreten des Ereignisses. Diese Festlegung wird fast immer willkürlich sein und sie hat Unschärfen bei der Abgrenzung von günstigen und ungünstigen Verläufen zur Folge. Je mehr Patienten es gibt, für die das Zielereignis in der Nähe der zeitlichen Trennlinie liegt, desto schwieriger wird zwangsläufig die prognostische Diskriminierung sein. Aus diesem Grunde ist es günstig, bei der Wahl des Trennpunktes für die zeitliche Eingrenzung des Zielereignisses darauf zu achten, daß eine relativ "natürliche" Trennung der Patienten in zwei Gruppen erfolgt, indem man den Zeitpunkt in einem Bereich geringer Dichte wählt.

Im Zusammenhang mit der Wahl des Referenzkriteriums muß auch kurz das Problem des Verifikationsbias (auch: "work-up bias") genannt werden, das in der Praxis eine große Rolle spielt. Dieses Problem wird vor allem im Zusammenhang mit der Evaluierung von Diagnosetests beschrieben und betrifft daher auch Scores, wenn sie im Rahmen der Diagnostik eingesetzt werden. Der Verifikationsbias kommt dadurch zustande, daß die Erhebung des Referenzkriteriums nicht bei allen Patienten durchgeführt wird. Wenn dies abhängt vom Scorewert des Patienten, zum Beispiel weil im Falle eines günstigen Scorewertes manche Patienten die eventuell unangenehme oder invasive Referenzuntersuchung ablehnen, dann kann hierdurch eine erhebliche Verzerrung bei der Bestimmung der Validität resultieren [142, 181]. Im Rahmen der Versuchsplanung ist deshalb darauf zu achten, daß das Referenzkriterium vollständig bei allen Patienten erhoben wird.

Studien zur Validierung von Scores sollen in der Regel prospektiv angelegt werden. Dies ist bei Querschnittstudien mit gleichzeitiger Erhebung des Referenzkriteriums kein Problem, während es bei Längsschnittstudien an Prognosescores einen erheblichen Aufwand bedeutet. Aus den gerade genannten Forderungen, insbesondere nach einer repräsentativen Erfassung einer klar definierten Patientenpopulation, ergibt sich jedoch, daß retrospektive Validierungsstudien nur in absoluten Ausnahmefällen sinnvoll sind. Sie lassen sich bestenfalls auf der Basis eines vollständigen und alle relevanten Informationen enthaltenden Patientenregisters durchführen [182]. Levine et al. [183] sprechen von "historical prospective cohort studies", die sich zum Beispiel auf Patienten aus multizentrischen randomisierten Therapiestudien beziehen können. Studien nach dem Fall-Kontroll-Ansatz lassen keine zuverlässigen Aussagen über die Validität von Prognosescores zu.

6.2.3 Ausschaltung systematischer und zufälliger Fehler

Wendet man sich den systematischen Fehlerquellen bei Validitätsuntersuchungen zu, so ist ein Hauptproblem die Unabhängigkeit der Beurteilung des Außenkriteriums bzw. der Referenzmethode vom zu evaluierenden Score. Insbesondere wenn das Kriterium einen Ermessensspielraum für subjektive Einschätzungen enthält, ist unbedingt eine Beurteilung durch eine Person, die den Scorewert des Patienten nicht kennt, erforderlich. Dies ist bei Prognosescores meist leicht realisierbar, da ein erheblicher zeitlicher Abstand zwischen der Erhebung von Score und Kriterium liegt und dadurch auch in der Regel unterschiedliche Personen involviert sind. Bei der Validierung an einer auf den gleichen Zeitpunkt bezogenen Referenzmethode sollte diese unbedingt durch einen zweiten Beurteiler und in Unkenntnis des Scorewertes erhoben werden. Die Forderung von Wasson [184], daß bei der Erhebung des Kriteriums keine Information einfließen darf, die auch im Score enthalten ist, ist allerdings unnötig und oft unpraktikabel, und in vielen Fällen würde sie die Qualität der Erhebung des Kriteriums einschränken.

Wenn es möglich ist, dem behandelnden Arzt die Scorewerte der Patienten vorzuenthalten, dann kann dies auch eine Maßnahme zur Vermeidung des oben angesprochenen Verifikationsbias sein. In den meisten Fällen werden aber die für den Arzt unmittelbar zugänglichen Informationen eine gewisse Abschätzung des Scorewertes erlauben, auch wenn dieser nicht explizit bekanntgegeben wird.

Ein besonderes Problem ergibt sich speziell im Falle von Prognosescores, nämlich der Einfluß therapeutischer Maßnahmen im Zeitraum zwischen Scoreerhebung und Beobachtung des Außenkriteriums. Diese Störgröße muß im Rahmen der Studienplanung möglichst weitgehend konstant gehalten werden. Wenn die Therapie einen meßbaren Einfluß auf den Krankheitsverlauf hat, und davon sollte man in der Regel ausgehen, dann sind Aussagen über einen Prognosescore stets nur in Bezug auf diese Therapie gültig. Eine Übertragung von Validitätsaussagen auf andere Patienten außerhalb der Studie erfordert eine genaue Kenntnis des therapeutischen Vorgehens. Dies ist unproblematisch, solange alle Patienten in der Studie die gleiche Therapie erhalten haben. Damit ist nicht notwendig eine identische Behandlung gemeint, sondern die Behandlung nach einem operationalisierbaren, für alle Patienten gleichen Konzept. Häufig wird die Evaluierung von Prognosemodellen als Nebenfragestellung im Rahmen randomisierter Therapiestudien durchgeführt, so daß zwei Gruppen von Patienten mit unterschiedlicher Therapie vorliegen. Hier ist die Auswertung zur prädiktiven Validität zunächst stratifiziert durchzuführen und nur bei weitgehend identischen Ergebnissen dürfen diese zusammengefaßt werden. Konkurrierende Therapien sind in ihrer Wirksamkeit oftmals einander so ähnlich, daß sie nur einen vernachlässigbaren Einfluß auf die Beurteilung der Validität haben.

Unbedingt zu vermeiden ist es, daß die Kenntnis der Scorewerte die Behandlung mit beeinflußt. Wenn die Scorewerte dem behandelnden Arzt bekannt sind und dieser therapeutische Alternativen mit unterschiedlichem Effekt auf das Hauptkriterium für den

Krankheitsverlauf zur Hand hat, so kann es sowohl zu einer Über- als auch Unterschätzung der prädiktiven Validität kommen.

> Beispiel 1: Ein Prognosescore zur Vorhersage der Überlebenswahrscheinlichkeit von Intensivpatienten wird untersucht. Wenn bereits in der Validierungsstudie Patienten mit besonders ungünstigen Scorewerten wegen der geringen Erfolgschancen nicht mehr im vollen Umfang intensivtherapiert werden, so wird dies ihre Überlebenschance weiter reduzieren. Damit kommt es zu einer Überschätzung der prädiktiven Validität.
>
> Beispiel 2: Für die Behandlung von Krebspatienten gibt es meist unterschiedlich wirksame Behandlunsgmöglichkeiten, unter denen sich bei Inkaufnahme höherer Nebenwirkungen auch bessere Wirkungen erreichen lassen. Bei Kenntnis der Werte eines zu evaluierenden Prognosescores wird der behandelnde Arzt versuchen, Patienten mit einer ungünstigeren Prognose durch eine intensivere Therapie doch noch erfolgreich behandeln zu können, während er bei prognostisch günstigen Patienten eher die Behandlung mit geringerer Belastung des Patienten wählt, man denke etwa an die adjuvante Chemotherapie bei Patientinnen mit nicht-metastasiertem Mammakarzinom. Der Effekt dieses Vorgehens wird in einer Unterschätzung der prädiktiven Validität liegen.

Aus den aufgeführten Beispielen wird klar, daß bei einer Evaluierung von Prognosescores den behandelnden Ärzten die Kenntnis der Scorewerte unbedingt vorenthalten werden muß. Das Problem besteht allerdings auch dann, wenn der Arzt aufgrund seiner klinischen Erfahrung unterschiedlich therapiert.

Im Rahmen der Studienplanung sind auch Maßnahmen zur Kontrolle des zufälligen Fehlers erforderlich. Hierzu gehören erstens die Fallzahlplanung und zweitens Bemühungen zur Standardisierung der Scoreerhebung. Eine statistische Fallzahlschätzung erfordert eine klare Formulierung des Studienziels, Vorwissen über die zu erwartende Streuung des Validitätsparameters und eine Vorgabe der angestrebten Genauigkeit und Sicherheit der statistischen Aussage. Als statistischer Ansatz für die Fallzahlschätzung bietet sich entweder die Vorgabe der Breite eines zu bestimmenden Konfidenzintervalls für den Validitätsparameter an, oder aber die Vorgabe eines in Bezug auf den Validitätsparameter formulierten kleinsten relevanten Unterschieds zwischen zwei zu vergleichenden Scores. Dies ist in der Praxis nur möglich, wenn man einen anschaulich interpretierbaren Validitätsparameter gewählt hat, oder wenn bereits Erfahrungen aus vorangegangenen Studien vorliegen.

Für Fallzahlschätzungen im Zusammenhang mit ROC-Kurven ist eine ungefähre Kenntnis der Auftretenswahrscheinlichkeit des klinischen Ereignisses erforderlich. Tritt das interessierende klinische Ereignis relativ selten auf, zum Beispiel nur bei 10% aller Patienten, so erhöht dies die notwendige Gesamtfallzahl, wenn nicht eine stratifizierte Stichprobenziehung vorgenommen wird. Dies ist aber bei Prognosestudien nicht einfach möglich und sollte auch

vermieden werden, wie im vorigen Abschnitt begründet wurde. Beim Vergleich von ROC-Kurven aus abhängigen Daten spielt außerdem der statistische Zusammenhang der Scores eine Rolle und muß bei der Fallzahlberechnung berücksichtigt werden.

De facto wird in den publizierten Validitätsstudien zu klinischen Scores fast nie über eine statistische Fallzahlplanung berichtet, was sicherlich auch an den gerade dargestellten Schwierigkeiten liegt. Die meisten Publikationen beschränken sich ohnehin auf eine deskriptive Auswertung, die eine Bewertung der gefundenen Ergebnisse schwierig macht. Die Arbeiten von Hanley & McNeil [114, 117] enthalten Hinweise für Fallzahlplanungen für den Vergleich zweier ROC-Kurven aus unabhängigen bzw. abhängigen Stichproben. Geht man davon aus, daß der wahre Unterschied zwischen zwei Scores, gemessen als Differenz der Flächen unter den ROC-Kurven, in der Regel höchsten 0.1 beträgt (siehe zum Beispiel Ohmann et al. [132]), so sind bei einseitigem Test mit Irrtumswahrscheinlichkeiten $\alpha = 0.05$ und $\beta = 0.1$ laut der Tabelle von Hanley & McNeil bereits 432 Patienten erforderlich. Diese Zahl vergrößert sich erheblich, wenn das klinische Ereignis entweder sehr selten oder sehr häufig auftritt oder wenn noch kleinere Unterschiede (z.B. 0.05) als relevant erachtet werden. Die genannten Zahlen beziehen sich auf den Vergleich von ROC-Kurven aus unabhängigen Stichproben. Bei Erhebung der Scores an den gleichen Patienten kann sich die Fallzahl, je nach Ausmaß der Abhängigkeit, durchaus auf die Hälfte verringern. Diese Zahlen geben nur einen groben Anhaltspunkt, sie zeigen jedoch, daß für aussagekräftige Validierungsstudien in der Regel Fallzahlen von mehreren Hundert erforderlich sind, die meist nur in multizentrischen Studien erbracht werden können.

Der Forderung nach multizentrischen Studien ist auch aus Gründen der Verallgemeinerbarkeit der Ergebnisse relevant. Insbesondere bei multizentrischen Studien ergibt sich aber das Problem der Standardisierung. Dieses Problem kann oft durch eine genaue Operationalisierung der Scorekomponenten gelöst werden. Je größer der subjektive Anteil bei der Scoreerhebung ist, um so dringlicher sind weitergehende Maßnahmen, wie beispielsweise ein Training in der Anwendung des Scores.

Ein weiteres Problem der Studienplanung stellt die Vermeidung von Drop-outs dar. Es ist unbedingt anzustreben, daß bei allen Patienten, an denen der Score erhoben wurde, auch das Kriterium ermittelt werden kann. Bei prognostischen Scores ist dies manchmal schwierig, wenn zwischen der Erhebung von Score und Kriterium ein größeres Zeitintervall liegt, in dem Patienten aus organisatorischen Gründen oder wegen mangelnder Motivation aus der Studie ausscheiden können.

6.3 Bewertung des klinischen Nutzens von Scores

6.3.1 Bedeutung von Feldstudien

Nach der Erörterung von Studien zur Evaluierung der Reliabilität und Validität von Scores muß zum Abschluß nochmals die Frage nach dem klinischen Nutzen aufgeworfen werden. Die Antwort darauf wird, je nach Anwendung des Scores, ganz unterschiedlich sein. Der Nutzen im Rahmen wissenschaftlicher Anwendungen, zum Beispiel als Zielgröße oder Kontrollvariable in klinischen Studien, ist evident und braucht hier nicht weiter erörtert zu werden. In Abschnitt 2.4.2 wurde bereits ausführlich dargestellt, wie Scores als Grundlage für diagnostisches oder therapeutisches Handeln dienen können. Der Nutzen ärztlichen Handelns ist grundsätzlich bewertbar, jedoch ist er meist multidimensional und in einigen seiner Dimensionen nur schwer meßbar. Dies ist bei klinischen Therapiestudien nicht anders, wenn man etwa an die Dimensionen Überlebenszeit und Lebensqualität bei onkologischen Studien denkt.

Klinische Scores sind in diesem Zusammenhang als Entscheidungshilfen anzusehen und die Evaluierung ihres Nutzens hat daher ähnlich zu erfolgen wie bei anderen Entscheidungshilfen. Ansätze hierzu finden sich in der neueren Literatur im Bereich diagnostischer Tests [55, 181] sowie in Arbeiten zur Evaluierung von medizinischen Expertensystemen [175, 176, 185]. Studien mit der hier angesprochenen Zielsetzung werden dort als Phase-IV-Studien [55] oder auch als Feldstudien [176] bezeichnet.

Bereits in Abschnitt 2.4.2 hatten wir unterschieden zwischen Anwendungen von Scores, in denen der Einfluß auf die Therapieentscheidung genau festgelegt ist und solchen, wo dies nicht geregelt und somit dem behandelnden Arzt überlassen ist. In beiden Fällen ergibt sich eine Aussage über den klinischen Nutzen des Scores erst aus dem Vergleich mit einer Behandlung unabhängig vom Score, also möglichst ohne dessen Kenntnis. Studien, die einen solchen Vergleich anstreben, sind in der Literatur zu klinischen Scores äußerst rar. Wasson et al. [184] fanden unter 33 publizierten Studien zu klinischen Prognoseregeln überhaupt nur zwei, in denen der (potentielle) Effekt der Entscheidungsregel auf die Patientenversorgung beschrieben wurde.

Der Hauptgrund für das Fehlen solcher Studien mag darin liegen, daß im Falle der nachgewiesenen Validität viele Kliniker bereits vom Nutzen des Scores überzeugt sind. Ob dieser jedoch meßbar ist, entweder in Form besserer Heilungsraten oder aber geringerer Nebenwirkungen, wird ganz entscheidend von drei Komponenten abhängen: Erstens von der Größe der Patientengruppe, bei der der Score zu einer Änderung der Behandlung führt, zweitens vom tatsächlichen Behandlungseffekt in dieser Gruppe und drittens von der Akzeptanz des Scores durch die Ärzte, die sich in der Vollständigkeit und Korrektheit der Erhebung des Scores und in seiner Umsetzung ausdrückt. Während die erste Komponente bereits aus einer Validierungsstudie abgeschätzt werden kann, wenn diese an einer

repräsentativen Patientenstichprobe durchgeführt wurde, lassen sich die beiden anderen Faktoren erst in einer Feldstudie quantifizieren.

6.3.2 Design von Feldstudien

Feldstudien sollten als randomisierte Studien durchgeführt werden, sofern dies möglich ist. Hiergegen können vor allem organisatorische Gründe sprechen, aber auch ethische Vorbehalte existieren. In Abhängigkeit von der Art der Erhebung des Scores kann es dabei manchmal schwierig sein, eine Randomisation auf der Ebene der Patienten durchzuführen. Dies würde erfordern, daß bei einem Teil der Patienten der Score entweder gar nicht bestimmt oder zumindest nicht therapeutisch umgesetzt würde. Grundsätzlich ist es jedoch auch möglich, eine Randomisation auf der Ebene des Arztes oder der Abteilung vorzunehmen, Gründe hierfür gibt Spiegelhalter [185] an.

> D´Agostino & Pozen [186, 187] entwickelten einen prädiktiven Score zur Diagnose der akuten koronaren Herzerkrankung (ACHD) bei Patienten, die mit der entsprechenden Verdachtssymptomatik in einer Notfallambulanz erschienen. Der Score basierte auf einem logistischen Regressionsmodell und wurde den Ärzten in Form einer geschätzten Wahrscheinlichkeit für die Diagnose ACHD mitgeteilt. Zur Evaluierung des klinischen Nutzens führten die Autoren eine nicht-randomisierte, vergleichende Studie durch, bei der in einigen der teilnehmenden Kliniken über 10 Monate hinweg pro Monat jeweils einem Team von Ärzten die Wahrscheinlichkeiten zur Verfügung gestellt wurden und dem nächsten Team nicht. Die Teams wurden im monatlichen Rhythmus neu besetzt. Es wurde den Ärzten überlassen, ob und in welcher Form sie die Information des Scores therapeutisch umsetzen würden. Die Autoren berichten, daß nicht nur die Richtigkeit der Erstdiagnosen signifikant erhöht wurde, sondern daß auch der Anteil der unnötigen Aufnahmen in die kostenintensive "coronary care unit" verringert werden konnte.

In dem eben geschilderten Beispiel wurde zwar keine echte Randomisation, aber immerhin ein quasi-experimenteller Ansatz durchgeführt. Der Unterschied ist bei Studien, in denen ein Team die Beobachtungseinheit darstellt, auch weniger relevant, da es sich zwangsläufig um eine offene, d.h. nicht-maskierte Studie handelt. Stattdessen ist bei der Auswertung solcher Studien auf die Abhängigkeit der patientenbezogenen Ergebnisse innerhalb der Teams zu achten, was bei der genannten Studie versäumt wurde.

Ein wesentlich schwächeres Studiendesign stellt der historische Vergleich dar, bei dem eine Studienphase ohne Kenntnis des Scores mit einer Phase nach Einführung des Scores verglichen wird. Die Vor- und Nachteile von historisch kontrollierten Studien im Bereich der Entscheidungsstützung sowie methodische Anforderungen an diesen Studientyp diskutieren Guyatt et al. [74].

Beziehen wir uns jetzt auf den Fall in dem ein Score, oder allgemein eine Entscheidungshilfe, zu einer festgelegten individualisierten Behandlung führt, wie in Abschnitt 2.4.2 beschrieben.

Hier sind Studien mit Randomisierung auf Patientenebene durchführbar, wobei mehrere Vorgehensweisen zur Auswahl stehen. Drei mögliche Studiendesigns werden von Lange et al. [188] am Beispiel der Osteoporose behandelt. Die dort beschriebenen Vor- und Nachteile sind jedoch nicht verallgemeinerbar, sondern beispielspezifisch [189]. Dies soll im folgenden ausgeführt werden.

Wir gehen aus von einem Prognosescore P, auf dessen Basis eine dichotome Entscheidungsregel mit den Ausprägungen P+ und P- formuliert wird. Die prädiktive Validität des Scores sei unter einer einheitlichen Standardtherapie T ermittelt worden. Als Behandlungsalternativen mögen eine intensivierte Therapie T+ sowie eine reduzierte Therapie T- in Frage kommen. Beide Therapien sollen grundsätzlich erprobt sein, ihre Wirksamkeit im Vergleich zu T ist für über den Score definierte Teilkollektive jedoch nicht bekannt. Bei T+ wird eine im Vergleich zu T bessere Wirkung bei gleichzeitig stärkeren Nebenwirkungen erwartet, während T- als etwas weniger wirksam, aber besser verträglich angesehen wird. Es hängt nun ganz vom konkreten Anwendungsbeispiel ab, ob für prognostisch ungünstige Patienten (P+) im Rahmen einer differenzierten Therapie T+ oder T- oder weiterhin T angeboten wird und analog für P-.

> Das bereits besprochene Beispiel der Behandlung des hochmalignen Non-Hodgkin-Lymphoms sieht eine Therapieintensivierung für prognostisch ungünstige Patienten vor: $P+ \rightarrow T+$, $P- \rightarrow T$.
>
> In einer Studiengruppe zur Behandlung von Patienten mit kleinzelligem Bronchialkarzinom waren die Ergebnisse einer intensiven Chemotherapie insgesamt unbefriedigend, da nur bei einem sehr kleinen Teil der Patienten längeranhaltende Heilungserfolge erzielt werden können. Es wurde nach einem Prognosemodell gesucht, das zur Identifizierung aussichtsloser Fälle geeignet ist, um diesen im Rahmen einer palliativen Behandlung die nebenwirkungsreiche Therapie zu ersparen: $P+ \rightarrow T-$, $P- \rightarrow T$.

Ein mögliches Studiendesign sieht einen randomisierten Vergleich zwischen der einheitlichen und der individualisierten Therapie vor. Als Nachteil dieses Designs nennen Lange et al. [188] Interpretationsschwierigkeiten im Fall eines negativen Ergebnisses, weil ein fehlender Effekt sowohl der mangelnden prädiktiven Validität der Entscheidungsregel als auch der fehlenden therapeutischen Effektivität der Behandlungsalternative zuzuschreiben sein könnte. Dieses Argument trifft nicht zu, wenn eine ausreichende Validierung des Scores vorab geschehen ist. Stattdessen hat dieses Design jedoch einen anderen Nachteil, nämlich daß eine eventuell erhebliche Zahl von Patienten in beiden Gruppen identisch behandelt wird (mit der Standardtherapie T), was sowohl eine Erhöhung des organisatorischen Auswands als auch einen Verlust an Power beim statistischen Vergleich mit sich bringt.

Als Konsequenz läßt sich ein vereinfachtes Design anwenden, bei dem der Score bzw. die Entscheidungsregel zur Vorauswahl der Patienten dient und dann eine reguläre randomisierte

Therapiestudie für diejenige Patientengruppe durchgeführt wird, für die sich das therapeutische Vorgehen unterscheidet. Dieses Vorgehen wurde in beiden oben genannten Beispielen gewählt.

> Für eine Hochrisiko-Gruppe von Patienten mit hochmalignem Non-Hodgkin-Lymphom wird aktuell eine randomisierte Studie zum Vergleich der bisherigen Standard-Chemotherapie gegen eine Hochdosis-Behandlung unter autologer Knochenmarkstransplantation durchgeführt.
>
> In einer prognostisch ungünstigen Untergruppe von Patienten mit kleinzelligem Bronchialkarzinom wurde in einer randomisierten Studie die bisherige aggressive Chemotherapie gegen eine palliative Kombination verglichen. Aufgrund der als besser angenommenen Verträglichkeit der Therapie wurde die Untersuchung der therapeutischen Effektivität als einseitige Äquivalenzstudie geplant [190].

Das letztgenannte Beispiel zeigt die Problematik solcher Studien auf: Häufig ist der Nutzen der individualisierten Therapie nicht an einer einzigen Zielgröße festzumachen, da sich die Vergleichstherapien sowohl bezüglich Wirkung als auch bei den Nebenwirkungen unterscheiden. Lange et al. bemängeln an diesem Design, daß im Falle eines positiven Ergebnisses unklar bleibt, ob dieser Therapieeffekt nicht sogar global gilt, also unabhängig von der Entscheidungsregel. Sie schlagen deshalb ein drittes Design vor, welches auch in der bisher ausgenommenen Patientengruppe den gleichen Therapievergleich mittels einer randomisierten Studie untersucht. Dieses Vorgehen entspricht nun aber der Durchführung einer üblichen randomisierten Studie im gesamten Patientenkollektiv, mit der einzigen Modifikation, daß der Score (bei Lange et al. der diagnostische Test) bei allen Patienten zwar erhoben, aber den behandelnden Ärzten nicht bekannt gegeben wird.

Gegen dieses einfache Vorgehen sprechen zwei Gründe: Zum einen ist häufig die neue Behandlungsalternative nach der medizinischen Lehrmeinung aufgrund des unterschiedlichen Nutzen-Risiko-Profils nicht bei allen Patienten ethisch vertretbar, wie es zum Beispiel in den beiden oben dargestellten Studien der Fall war. Zum anderen erhält man in einer Studie, in der die Ärzte wegen der randomisierten Therapieentscheidung die Entscheidungshilfe nicht aktiv in Anspruch nehmen müssen, keine Hinweise auf die Akzeptanz einer solchen Entscheidungshilfe in der Praxis.

7 ZUSAMMENFASSUNG

Die Bedeutung klinischer Scores für wissenschaftliche Zwecke und in der ärztlichen Routine ist vielfach belegt und findet Ausdruck in verstärkten Forschungsbemühungen in diesem Bereich. Publizierte Studien zur Evaluierung klinischer Scores zeigen jedoch, daß es hier noch an methodischen Standards fehlt.

Die vorliegende Arbeit gibt einen Überblick über bekannte und neuere Ansätze zur Konstruktion und Evaluierung klinischer Scores. Dabei wird besonderer Wert gelegt auf ein einheitliches Konzept, das sich aus der Betrachtung von Scores als klinische Meßinstrumente ergibt.

Im Zentrum der Arbeit steht die Darstellung der wichtigsten Gütekriterien und der statistischen Parameter, mit denen eine Quantifizierung der Güte auf der Basis empirischer Ergebnisse versucht wird. Neben der Erörterung der klassischen Reliabilitäts- und Validitätskonzepte wird eine Einordnung und kritische Diskussion einiger neuerer Gütekriterien gegeben. Hierzu gehören die Änderungssensitivität evaluativer Scores sowie die Kalibration von Prognosescores. Es wird gezeigt, daß einigen dieser neuen Ansätze begriffliche Unklarheiten zugrundeliegen, die ihre Verwendung nicht sinnvoll erscheinen lassen.

Für eine empirisch geleitete Konstruktion klinischer Scores stellen die Gütekriterien Maßstäbe dar, die einen wesentlichen Einfluß auf die Auswahl und Zusammensetzung der Scorekomponenten haben. Im Rahmen der dargestellten Strategie einer schrittweisen Scorekonstruktion hat das Ziel der Validitätsoptimierung eine vorrangige Rolle. Es werden aber auch zahlreiche praktische Probleme bei der Anwendung der hierfür zur Verfügung stehenden Verfahren beschrieben.

Der Planung und Durchführung empirischer Studien mit dem Ziel der Evaluierung klinischer Scores widmet sich der letzte Teil der Arbeit. Im Bereich von Reliabilitätsstudien wird auf die Notwendigkeit der Formulierung klarer und relevanter Zielsetzungen hingewiesen und gezeigt, welche Anforderungen an das Studiendesign daraus resultieren. Auch an Validitätsuntersuchungen von klinischen Scores wird eine Reihe von Anforderungen formuliert, allen voran die nach Unabhängigkeit der Validierungsphase von der Konstruktionsphase.

Die Arbeit endet mit Überlegungen zu Studiendesigns, anhand derer der klinische Nutzen, speziell von Prognosescores, erfaßt und geprüft werden kann. In diesem Bereich liegt bisher kaum praktische Erfahrung vor. Das gleiche gilt für die Verbindung der Untersuchung von Reliabilitäts- und Validitätsaspekten im Rahmen gemeinsamer Studien, für die diese Arbeit Anregungen geben sollte.

LITERATUR

1. Feinstein, A.R., *Clinimetrics*. 1987, New Haven: Yale University Press.
2. Holle, R., *Klinische Epidemiologie - Versuch einer Standortbestimmung aus der Sicht der Medizinischen Biometrie*, in *Proceedings Dresden GMDS*, H. Kunath and R. Koch, Editors. 1995.
3. *Webster's Ninth New Collegiate Dictionary*. 1984, Springfield: Merriam-Webster Inc.
4. *Pschyrembel Klinisches Wörterbuch*. 255 ed. 1986, Berlin: de Gruyter.
5. Apgar, V., *A Proposal for a New Method of Evaluation of the Newborn Infant.* Anesth.Analg., 1953. **32**: p. 260-267.
6. *Is the Apgar Score Outmoded? (Editorial).* Lancet, 1989. **1**: p. 591-592.
7. Karnofsky, D.A. and J.H. Burchenal, *The Clinical Evaluation of Chemotherapeutic Agents in Cancer*, in *Evaluation of Chemotherapeutic Agents*, C.M. MacLeod, Editor. 1949, Columbia University Press: New York. p. 191-205.
8. Best, W.R., et al., *Development of a Crohn¦s Disease Activity Index.* Gastroenterology, 1976. **70**: p. 439-444.
9. Knaus, W.A., et al., *APACHE - Acute Physiology and Chronic Health Evaluation: A Physiologically Based Classification System.* Crit.Care Med., 1981. **9**: p. 591-597.
10. Knaus, W.A., et al., *APACHE II: A Severity of Disease Classification System.* Crit.Care Med., 1985. **13**: p. 818-829.
11. Knaus, W.A., et al., *The APACHE III Prognostic System. Risk Prediction of Hospital Mortality for Critically Ill Hospitalized Adults.* Chest, 1991. **100**: p. 1619-1636.
12. Cella, D.F., et al., *The Functional Assessment of Cancer Therapy Scale: Development and Validation of the General Measure.* J.Clin.Oncol., 1993. **11**: p. 570-579.
13. Gerber, P. and O. Wicki, *Stadien und Einteilungen in der Medizin*. 1990, Stuttgart: George Thieme Verlag.
14. Bowling, A., *Measuring Health - A Review of Quality of Life Measurement Scales*. 1991, Milton Keynes: Open University Press.
15. Norris, R.M., et al., *A New Coronary Prognostic Index.* Lancet, 1969. **1**: p. 274-278.
16. Breiman, L., et al., *Classification and Regression Trees*. 1984, Belmont: Wadsworth.

17. Wacha, H., et al., *Mannheim Peritonitis Index - Prediction of Risk of Death from Peritonitis: Construction of a Statistical and Validation of an Empirically Based Index.* Theor.Surg., 1987. **1**: p. 169-177.

18. Kirshner, B. and G. Guyatt, *A Methodological Framework for Assessing Health Indices.* J.Chron.Dis., 1985. **38**: p. 27-36.

19. Williams, J.I. and C.D. Naylor, *How Should Health Status Measures be Assessed? Cautionary Notes on Procrustean Frameworks.* J.Clin.Epidemiol., 1992. **45**: p. 1347-1351.

20. Schag, C.C., R.L. Heinrich, and P.A. Ganz, *Karnofsky Performance Status Revisited: Reliability, Validity, and Guidelines.* J.Clin.Oncol., 1984. **2**: p. 187-193.

21. Kane, R.A. and R.L. Kane, *Assessing the Elderly: A Practical Guide to Measurement*. 1981, Toronto: Lexington Books.

22. Neugebauer, E. and B. Bouillon, *Was können Scoresysteme leisten?* Unfallchirurg, 1994. **97**: p. 172-176.

23. van Gijn, J. and C.P. Warlow, *Down With Stroke Scales!* Cerebrovasc.Dis., 1992. **2**: p. 244-246.

24. Annas, G.J., *Informed Consent, Cancer, and Truth in Prognosis.* N.Engl.J.Med., 1994. **330**: p. 223-225.

25. Celani, M.G., et al., *Comparability and Validity of Two Clinical Scores in the Early Differential Diagnosis of Acute Stroke.* Br.Med.J., 1994. **308**: p. 1674-1676.

26. Coiffier, B., et al., *Prognostic Factors in Aggressive Malignant Lymphomas: Description and Validation of a Prognostic Index that Could Identify Patients Requiring a More Intensive Therapy.* J.Clin.Oncol., 1991. **9**: p. 211-219.

27. Köppler, H., et al., *Randomised Comparison of CHOEP Versus Alternating hCHOP/IVEP for High-Grade Non-Hodgkin¦s Lymphomas: Treatment Results and Prognostic Factor Analysis in a Multi-Centre Trial.* Ann.Oncol., 1994. **5**: p. 49-55.

28. The International Non-Hodgkin's Lymphoma Prognostic Factors Project, *A Predictive Model for Aggressive Non-Hodgkin¦s Lymphoma.* N.Engl.J.Med., 1993. **329**: p. 987-994.

29. Ménard, S., et al., *Prognosis Based on Primary Breast Carcinoma instead of Pathological Nodal Status.* Br.J.Cancer, 1994. **70**: p. 709-712.

30. Selker, H.P., et al., *How Do Physicians Adapt When the Coronary Care Unit is Full?* JAMA, 1987. **257**: p. 1181-1185.

31. Bombardier, C., et al., *Auranofin Therapy and Quality of Life in Patients with Rheumatoid Arthritis.* Am.J.Med., 1986. **81**: p. 565-578.

32. McDowell, I. and C. Newell, *Measuring Health - A Guide to Rating Scales and Questionnaires*. 1987, new York: Oxford University Press.

33. Blum, A.L., *Clinical Evaluation of Success*, in *The Randomized Clinical Trial and Therapeutic Decisions*, N. Tygstrup, J.M. Lachin, and E. Juhl, Editors. 1982, Marcel Dekker: New York.

34. Bulpitt, C.J., *Randomised Controlled Clinical Trials*. 1983, The Hague: Martinus Nijhoff.

35. Pocock, S.J., *Clinical Trials - A Practical Approach*. 2 ed. 1987, Chichester New York: John Wiley & Sons.

36. Spriet, A. and P. Simon, *Methodology of Clinical Drug Trials*. 1985, Basel: Karger Verlag.

37. Abel, U. and J. Windeler, *Comprehensive Blinded Prognostic Rating - A New Procedure for Nonrandomized Treatment Comparisons*. 1994.

38. von Kummer, R., et al., *Does Arterial Recanalization Improve Outcome in Carotid Territory Stroke?* Stroke, 1995.

39. Holle, R. and M. Pritsch, *Problems of Covariate Adjustment in Clinical Trials (Abstract).* Control.Clin.Trials, 1991. **12**: p. 652.

40. Rosenbaum, P.R. and D.B. Rubin, *Reducing Bias in Observational Studies Using Subclassification on the Propensity Score.* JASA, 1984. **79**: p. 516-524.

41. Knaus, W.A., et al., *An Evaluation of Outcome from Intensive Care in Major Medical Centers.* Ann.Intern.Med., 1986. **104**: p. 410-418.

42. Boyd, C.R., M.A. Tolson, and W.S. Copes, *Evaluating Trauma Care: The TRISS Method.* J.Trauma, 1987. **27**: p. 370.

43. Guirguis, E.M., et al., *Trauma Outcome Analysis of Two Canadian Centres Using the TRISS Method.* J.Trauma, 1990. **30**: p. 426-429.

44. Holle, R., *Möglichkeiten und Grenzen des QALY-Konzepts in der onkologischen Therapieforschung*, in *Lebensqualität in der Onkologie II*, R. Schwarz, et al., Editors. 1995, W. Zuckschwerdt Verlag: München.

45. Ohmann, C. and O. Horstmann, *Formale Entscheidungshilfen in der Gastroenterologie.* Z.Gastroenterol., 1992. **30**: p. 558-564.

46. Ohmann, C., et al., *Prognostic Scores in Oesophageal or Gastric Variceal Bleeding.* Scand.J.Gastroenterol., 1990. **25**: p. 501-512.

47. Stevens, S.S., *On the Theory of Scales of Measurement.* Science, 1946. **103**: p. 677-680.

48. MacKenzie, C.R. and M.E. Charlson, *Standards for the Use of Ordinal Scales in Clinical Trials.* Br.Med.J., 1986. **292**: p. 40-43.

49. Hutchinson, T.A., N.F. Boyd, and A.R. Feinstein, *Scientific Problems in Clinical Sales, as Demonstrated in the Karnofsky Index of Performance Status.* J.Chron.Dis., 1979. **32**: p. 661-666.

50. Starmark, J.E., et al., *A Comparison of the Glasgow Coma Scale and the Reaction Level Scale (RLS85).* J.Neurosurg., 1988. **69**: p. 699-706.

51. Damiano, A.M., et al., *Reliability of a Measure of Severity of Illness: Acute Physiology of Chronic Health Evaluation - II.* J.Clin.Epidemiol., 1992. **45**: p. 93-101.

52. Wright, J.G. and A.R. Feinstein, *A Comparative Contrast of Clinimetric and Psychometric Methods for Constructing Indexes and Rating Scales.* J.Clin.Epidemiol., 1992. **45**: p. 1201-1218.

53. Flechtner, H., et al., *Quality of Life Assessment - Results and Comparison of Two Small Cell Lung Cancer Clinical Trials (EORTC - German Cooperative Group BMFT). Conference of the European Society of Psychosocial Oncology, Amsterdam 24.-25.10.1988.*

54. Büttner, J., E. Hansert, and D. Stamm, *Auswertung, Kontrolle und Beurteilung von Messergebnissen*, in *Methoden der enzymatischen Analyse. Band I*, H.U. Bergmeyer, Editor. 1970, Verlag Chemie: Weinheim. p. 281-364.

55. Köbberling, J., et al., *Methodologie der medizinischen Diagnostik*. 1991, Berlin: Springer Verlag.

56. Barry, B.A., *Errors in Practical Measurement in Science, Engineering and Technology*. 1978, New York: John Wiley & Sons.

57. Healy, M.J.R., *Measuring Measuring Errors.* Stat.Med., 1989. **8**: p. 893-906.

58. Bablok, W., et al., *Biometrical Methods*, in *Evaluation Methods in Laboratory Medicine*, R. Haeckel, Editor. 1993, VCH Verlag: Weinheim. p. 203-241.

59. Daly, L.E., G.J. Bourke, and J. McGilvray, *Interpretation and Uses of Medical Statistics*. 4 ed. 1991, Oxford: Blackwell Scientific Publications.

60. Lienert, G.A., *Testaufbau und Testanalyse*. 4 ed. 1989, München: Psychologie Verlags Union.

61. *Noten in 20 Fächern.* Spiegel, 1994. **2**: p. 158-159.

62. Eisenwiener, H.G., et al., *Präzisionsangaben beim Methodenvergleich.* Lab.Med., 1983. **7**: p. 273-281.

63. Stark, F.M. and G. Buchkremer, *Das Fünfminuteninterview - Eine Kurzfassung zur Erfassung des Expressed Emotion Status.* Nervenarzt, 1992. **63**: p. 42-45.

64. Elmore, J.G. and A.R. Feinstein, *A Bibliography of Publications on Observer Variability (Final Installment).* J.Clin.Epidemiol., 1992. **45**: p. 567-580.

65. Cronbach, L.J., N. Rajaratnam, and G.C. Gleser, *Theory of Generalizability: A Liberalization of Reliability Theory.* Br.J.Statist.Psychol., 1963. **16**: p. 137-163.

66. Shavelson, R.J., N.M. Webb, and G.L. Rowley, *Generalizability Theory.* Am.Psychol., 1989. **44**: p. 922-932.

67. Streiner, D.L. and G.R. Norman, *Health Measurement Scales - A Practical Guide to their Development and Use*. 1989, Oxford: Oxford University Press.

68. Chambers, L.W., et al., *Sensitivity to Change and the Effect of Mode of Administration on Health Status Measurement.* Med.Care, 1987. **25**: p. 470-480.

69. Evans, W.J., C.G. Cayten, and P.A. Green, *Determining the Generalizability of Rating Scales in Clinical Settings.* Med.Care, 1981. **19**: p. 1211-1220.

70. Guyatt, G.H., B. Kirshner, and R. Jaeschke, *Measuring Health Status: What Are the Necessary Measurement Properties?* J.Clin.Epidemiol., 1992. **45**: p. 1341-1345.

71. Metz, C.E., *Basic Principles of ROC Analysis.* Semin.Nucl.Med., 1978. **8**: p. 283-298.

72. Guyatt, G., S. Walter, and G. Norman, *Measuring Change over Time: Assessing the Usefulness of Evaluative Instruments.* J.Chron.Dis., 1987. **40**: p. 171-178.

73. Guyatt, G.H., et al., *Responsiveness and Validity in Health Status Measurement: A Clarification.* J.Clin.Epidemiol., 1989. **42**: p. 403-408.

74. Guyatt, G.H., et al., *The Role of Before-After Studies of Therapeutic Impact in the Evaluation of Diagnostic Technologies.* J.Chron.Dis., 1986. **39**: p. 295-304.

75. Deyo, R.A. and R.M. Centor, *Assessing the Responsiveness of Functional Scales to Clinical Change: An Analogy to Diagnostic Test Performance.* J.Chron.Dis., 1986. **39**: p. 897-906.

76. Cella, D.F. and D.S. Tulsky, *Measuring Quality of Life Today: Methodological Aspects.* Oncology, 1990. **4**: p. 29-38.

77. van Knippenberg, F.C.E. and J.C.J.M. de Haes, *Measuring the Quality of Life of Cancer Patients: Psychometric Properties of Instruments.* J.Clin.Epidemiol., 1988. **41**: p. 1043-1053.

78. Guyatt, G.H., et al., *Measuring Quality of Life in the Frail Elderly.* J.Clin.Epidemiol., 1993. **46**: p. 1433-1444.

79. Stelzl, I., *Fehler und Fallen der Statistik für Psychologen, Pädagogen und Sozialwissenschaftler*. 1982, Wien: Verlag Hans Huber.

80. Ohmann, C. and W. Gross-Weege, *Scoring-Systeme auf der chirurgischen Intensivstation. I.* Chirurg, 1992. **63**: p. 1021-1028.

81. Diamond, G.A., *What Price Perfection? Calibration and Discrimination of Clinical Prediction Models.* J.Clin.Epidemiol., 1992. **45**: p. 85-89.

82. Fleiss, J.L., *The Design and Analysis of Clinical Experiments*. 1986, New York: John Wiley & Sons.

83. Lehmann, G., *Testtheorie: Eine systematische Übersicht*, in *Forschungsmethoden der Psychologie. Band 3: Messen und Testen*, H. Feger and J. Bredenkamp, Editors. 1983, Verlag für Psychologie Hogrefe: Göttingen.

84. Burdick, R.K. and F.A. Graybill, *Confidence Intervals on Variance Components*. 1992, New York: Marcel Dekker.

85. Dunn, G., *Design and Analysis of Reliability Studies*. 1989, New York: Oxford University Press.

86. Lin, L.I., *A Concordance Correlation Coefficient to Evaluate Reproducibility.* Biometrics, 1989. **45**: p. 255-268.

87. Deyo, R.A., P. Diehr, and D.L. Patrick, *Reproducibility and Responsiveness of Health Status Measures.* Control.Clin.Trials, 1991. **12**: p. 142S-158S.

88. Müller, R., *Intraklassenkorrelationsanalyse - ein Verfahren zur Beurteilung der Reproduzierbarkeit und Konformität von Meßmethoden (unveröffentlichte Dissertation)*. 1993.

89. Chinn, S., *The Assessment of Methods of Measurement.* Stat.Med., 1990. **9**: p. 351-362.

90. Cohen, J., *A Coefficient of Agreement for Nominal Scales.* Educ.Psychol.Meas., 1960. **20**: p. 37-46.

91. Bortz, J., G.A. Lienert, and K. Boehnke, *Verteilungsfreie Methoden in der Biostatistik*. 1990, Berlin: Springer Verlag.

92. Landis, J.R. and G.G. Koch, *The Measurement of Observer Agreement for Categorical Data.* Biometrics, 1977. **33**: p. 159-174.

93. Cicchetti, D.V. and A.R. Feinstein, *High Agreement but Low Kappa: II. Resolving the Paradoxes.* J.Clin.Epidemiol., 1990. **43**: p. 551-558.

94. Feinstein, A.R. and D.V. Cicchetti, *High Agreement but Low Kappa: I. The Problems of Two Paradoxes.* J.Clin.Epidemiol., 1990. **43**: p. 543-549.

95. Byrt, T., J. Bishop, and J.B. Carlin, *Bias, Prevalence and Kappa.* J.Clin.Epidemiol., 1993. **46**: p. 423-429.

96. Guggenmoos-Holzmann, I., *How Reliable Are Chance-corrected Measures of Agreement?* Stat.Med., 1993. **12**: p. 2191-2205.

97. Walter, S.D. and L.M. Irwig, *Estimation of Test Error Rates, Disease Prevalence and Relative Risk from Misclassified Data: A Review.* J.Clin.Epidemiol., 1988. **41**: p. 923-937.

98. Cohen, J., *Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit.* Psychol.Bull., 1968. **70**: p. 213-220.

99. MacLure, M. and W.C. Willett, *Misinterpretation and Misuse of the Kappa Statistic.* Am.J.Epidemiol., 1987. **126**: p. 161-169.

100. Fleiss, J.L., J. Cohen, and B.S. Everitt, *Large Sample Standard Errors of Kappa and Weighted Kappa.* Psychol.Bull., 1969. **72**: p. 323-327.

101. Tesseris, J., et al., *A Comparative Study of the Reaction Level Scale (RLS85) with Glasgow Coma Scale (GCS) and Edinburgh-2 Coma Scale (Modified) (E2CS(M)).* Acta Neurochir.(Wien), 1991. **110**: p. 65-76.

102. Armitage, P. and G. Berry, *Statistical Methods in Medical Research*. 2 ed. 1987, Oxford: Blackwell.

103. Donner, A. and M. Eliasziw, *Sample Size Requirements for Reliability Studies.* Stat.Med., 1987. **6**: p. 441-448.

104. Snedecor, G.W. and W.G. Cochran, *Statistical Methods*. 6 ed. 1967, Ames: Iowa State University Press.

105. Feldmann, U., *Robust Bivariate Errors-in-Variables Regression and Outlier Detection.* Eur.J.Clin.Chem.Clin.Biochem., 1992. **30**: p. 405-414.

106. Feldmann, U., B. Schneider, and H. Klinkers, *A Multivariate Approach for the Biometric Comparison of Analytical Methods in Clinical Chemistry.* J.Clin.Chem.Clin.Biochem., 1981. **19**: p. 121-137.

107. Linnet, K., *Estimation of the Linear Relationship Between the Measurements of Two Methods with Proportional Errors.* Stat.Med., 1990. **9**: p. 1463-1473.

108. Bland, J.M. and D.G. Altman, *Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement.* Lancet, 1986: p. 307-310.

109. Bland, J.M. and D.G. Altman, *A Note on the Use of the Intraclass Correlation Coefficient in the Evaluation of Agreement between Two Methods of Measurement.* Comput.Biol.Med., 1990. **20**: p. 337-340.

110. Abel, U., *Die Bewertung diagnostischer Tests*. 1993, Stuttgart: Hippokrates Verlag.

111. Sackett, D.L., et al., *Clinical Epidemiology - A Basic Science for Clinical Medicine*. 2 ed. 1991, Boston Toronto London: Little, Brown and Company.

112. Lorenz, R.J., *Grundbegriffe der Biometrie*. 1984, Stuttgart: Gustav Fischer.

113. Rosner, B., *Fundamentals of Biostatistics*. 3 ed. 1990, Boston: PWS-Kent.

114. Hanley, J.A. and B.J. McNeil, *The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve.* Radiology, 1982. **143**: p. 29-36.

115. Bamber, D., *The Area above the Ordinal Dominance Graph and the Area below the Receiver Operating Characteristic Graph.* J.Math.Psychol., 1975. **12**: p. 387-415.

116. Harrell, F.E., et al., *Evaluating the Yield of Medical Tests.* JAMA, 1982. **247**: p. 2543-2546.

117. Hanley, J.A. and B.J. McNeil, *A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases.* Radiology, 1983. **148**: p. 839-843.

118. DeLong, E.R., D.M. DeLong, and D.L. Clarke-Pearson, *Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach.* Biometrics, 1988. **44**: p. 837-845.

119. Metz, C.E., *Some Practical Issues of Experimental Design and Data Analysis in Radiological ROC Studies.* Invest.Radiol., 1989. **24**: p. 234-245.

120. Beam, C.A. and H.S. Wieand, *A Statistical Method for the Comparison of a Discrete Diagnostic Test with Several Continuous Diagnostic Tests.* Biometrics, 1991. **47**: p. 907-919.

121. Cook, E.F. and L. Goldman, *Empiric Comparison of Multivariate Analytic Techniques: Advantages and Disadvantages of Recursive Partitioning Analysis.* J.Chron.Dis., 1984. **37**: p. 721-731.

122. Holle, R., *Using ROC Methodology for the Comparison of Prognostic Models. 11th International Conference on Clinical Biostatistics, Nimes 18.-21.9.1990*.

123. Holle, R., *Graphische Methoden zum explorativen Vergleich unterschiedlicher Prognosemodelle (Poster). 36.GMDS-Jahrestagung München,16.- 18.9.1991*.

124. Hadorn, D.C., et al., *Cross-Validation Performance of Mortality Prediction Models.* Stat.Med., 1992. **11**: p. 475-489.

125. Raubertas, R.F., et al., *ROC Curves for Classification Trees.* Med.Decis.Making, 1994. **14**: p. 169-174.

126. Korn, E.L. and R. Simon, *Measures of Explained Variation for Survival Data.* Stat.Med., 1990. **9**: p. 487-503.

127. Ciampi, A., et al., *Stratification by Stepwise Regression, Correspondence Analysis and Recursive Partition: A Comparison of Three Methods of Analysis for Survival Data with Covariates.* Comp.Stat.Data Anal., 1986. **4**: p. 185-204.

128. Hilden, J., J.D.F. Habbema, and B. Bjerregaard, *The Measurement of Performance in Probabilistic Diagnosis. II. Trustworthiness of the Exact Values of the Diagnostic Probabilities.* Methods Inf.Med., 1978. **17**: p. 227-237.

129. Hosmer, D.W.J. and S. Lemeshow, *Applied Logistic Regression*. 1989, New York: John Wiley & Sons.

130. Lee, K.L., et al., *Predicting Outcome in Coronary Disease.* Am.J.Med., 1986. **80**: p. 553-560.

131. Ohmann, C., *Preoperative Prediction of Peri- and Postoperative Complications: Criteria for Statistical Evaluation of Predictive Systems.* Theor.Surg., 1991. **6**: p. 3-18.

132. Ohmann, C., et al., *Prospective Evaluation of Prognostic Scoring Systems in Peritonitis.* Eur.J.Surg., 1993. **159**: p. 267-274.

133. Poses, R.M., et al., *Are Two (Inexperienced) Heads Better Than One (Experienced) Head?* Arch.Intern.Med., 1990. **150**: p. 1874-1878.

134. Poses, R.M., R.D. Cebul, and R.M. Centor, *Evaluating Physicians' Probabilistic Judgments.* Med.Decis.Making, 1988. **8**: p. 233-240.

135. Miller, M.E., S.L. Hui, and W.M. Tierney, *Validation Techniques for Logistic Regression Models.* Stat.Med., 1991. **10**: p. 1213-1226.

136. Cox, D.R., *Two Further Applications of a Model for Binary Regression.* Biometrika, 1958. **45**: p. 562-565.

137. Clarke, D.M. and D.P. McKenzie, *Screening for Psychiatric Morbidity in the General Hospital: Methods for Comparing the Validity of Different Instruments.* Int.J.Methods Psychiat.Res., 1991. **1**: p. 79-87.

138. Holle, R. and J. Windeler, *Is There a Gain From "Chance-corrected" Measures of Diagnostic Validity?* J.Clin.Epidemiol., 1997. **50**: p. 117-120.

139. Kraemer, H.C., *The Robustness of Common Measures of 2 x 2 Association to Bias Due to Misclassifications.* Am.Statist., 1985. **39**: p. 286-290.

140. Kraemer, H.C., *Assessment of 2x2 Associations: Generalization of Signal-Detection Methodology.* Am.Statist., 1988. **42**: p. 37-49.

141. Kraemer, H.C. and D.A. Bloch, *Kappa Coefficients in Epidemiology: An Appraisal of a Reappraisal.* J.Clin.Epidemiol., 1988. **41**: p. 959-968.

142. Begg, C.B., *Biases in the Assessment of Diagnostic Tests.* Stat.Med., 1987. **6**: p. 411-423.

143. Coughlin, S.S. and L.W. Pickle, *Sensitivity and Specificity-like Measures of the Validity of a Diagnostic Test that Are Corrected for Chance Agreement.* Epidemiology, 1992. **3**: p. 178-181.

144. Jamart, J., *Chance-Corrected Sensitivity and Specificity for Three-Zone Diagnostic Tests.* J.Clin.Epidemiol., 1992. **45**: p. 1035-1039.

145. Brenner, H. and O. Gefeller, *Chance-Corrected Measures of the Validity of a Binary Diagnostic Test.* J.Clin.Epidemiol., 1994. **47**: p. 627-633.

146. Gefeller, O. and H. Brenner, *How to Correct for Chance Agreement in the Estimation of Sensitivity and Specificity of Diagnostic Tests.* Methods Inf.Med., 1994. **33**: p. 180-186.

147. Bergner, M., et al., *The Sickness Impact Profile: Development and Final Revision of a Health Status Measure.* Med.Care, 1981. **19**: p. 787-805.

148. Tuley, M.R., C.D. Mulrow, and C.A. McMahan, *Estimating and Testing an Index of Responsiveness and the Relationship of the Index to Power.* J.Clin.Epidemiol., 1991. **44**: p. 417-421.

149. Marks, G.B., S.M. Dunn, and A.J. Woolcock, *An Evaluation of an Asthma Quality of Life Questionnaire as a Measure of Change in Adults with Asthma.* J.Clin.Epidemiol., 1993. **46**: p. 1103-1111.

150. Wiklund, I. and J. Karlberg, *Evaluation of Quality of Life in Clinical Trials.* Control.Clin.Trials, 1991. **12**: p. 204S-216S.

151. Pocock, S.J., *A Perspective on the Role of Quality-of-Life Assessment in Clinical Trials.* Control.Clin.Trials, 1991. **12**: p. 257S-265S.

152. Aaronson, N.K. and J. Beckmann, *The Quality of Life of Cancer Patients*. 1987, New York: Raven Press.

153. Guyatt, G., et al., *A New Measure of Health Status for Clinical Trials in Inflammatory Bowel Disease.* Gastroenterology, 1989. **96**: p. 804-810.

154. Holle, R. and H. Flechtner, *Ergebnisse der Lebensqualitätserhebung der BMFT-Studien zur Therapie des kleinzelligen Bronchilakarzinoms*, in *Lebensqualität in der Onkologie*, R. Schwarz, et al., Editors. 1991, W. Zuckschwerdt Verlag: München. p. 89-100.

155. Cox, D.R., et al., *Quality-of-life Assessment: Can We Keep It Simple? (With Discussion).* J.Roy.Statist.Soc.Ser.A Stat., 1992. **155**: p. 353-393.

156. Kleinbaum, D.G., L.L. Kupper, and K.E. Muller, *Applied Regression Analysis and Other Multivariable Methods*. 2 ed. 1988, Belmont: Duxbury Press.

157. Harrell, F.E., et al., *Regression Models for Prognostic Prediction: Advantages, Problems, and Suggested Solutions.* Cancer Treat.Rep., 1985. **69**: p. 1071-1077.

158. Bendel, R.B. and A.A. Afifi, *Comparison of Stopping Rules in Forward Regression.* JASA, 1977. **72**: p. 46-53.

159. Mickey, R.M. and S. Greenland, *The Impact of Confounder Selection Criteria on Effect Estimation.* Am.J.Epidemiol., 1989. **129**: p. 125-137.

160. Holle, R., *Strategies for the Analysis of Prognostic Factors. Workshop on "Design and Analysis of Clinical Trials", Oberwolfach 22.-28.2.1987*. 1987.

161. Jaques, G., et al., *Prognostic Value of Pretreatment Carcinoembryonic Antigen, Neuron-Specific Enolase, and Creatine Kinase-BB Levels in Sera of Patients with Small Cell Lung Cancer.* Cancer, 1988. **62**: p. 125-134.

162. Altman, D.A., et al., *Dangers of Using "Optimal" Cutpoints in the Evaluation of Prognostic Factors.* J.Natl.Cancer Inst., 1994. **86**: p. 829-835.

163. Foitzik, T., et al., *The Heidelberg Intensive Ward Scoring System (HDWS): Development of a Computer Assisted Scoring System for Daily Evaluation of Physiological Parameters and Therapeutical Interventions and Prediction of Outcome in a Surgical ICU. (in Vorbereitung).* 1995.

164. Holle, R., *Scores in der Intensivmedizin: Konstruktion und Validierung. 39. GMDS Jahrestagung, Dresden 18.-22.9.1994*.

165. Cleveland, W.S., *Robust Locally Weighted Regression and Smoothing Scatterplots.* JASA, 1979. **74**: p. 829-836.

166. Smith, P.L., *Splines as a Useful and Convenient Tool.* Am.Statist., 1979. **33**: p. 57-62.

167. Holle, R., *Methodische Probleme bei der Planung und Auswertung von Prognosestudien*. 1991. p. 48-60.

168. Feinstein, A.R., C.K. Wells, and S.D. Walter, *A Comparison of Multivariable Mathematical Methods for Predicting Survival I. Introduction, Rationale, and General Strategy.* J.Clin.Epidemiol., 1990. **43**: p. 339-347.

169. Little, R.J.A., *Regression With Missing X¦s: A Review.* JASA, 1992. **87**: p. 1227-1237.

170. Sauerbrei, W. and M. Schumacher, *A Bootstrap Resampling Procedure for Model Building: Application to the Cox Regression Model.* Stat.Med., 1992. **2093**: p. 2109.

171. Efron, B. and R.J. Tibshirani, *An Introduction to the Bootstrap*. 1993, New York: Chapman & Hall.

172. Holle, R. and N. Victor, *Gliederungsvorschlag für Studienprotokolle bei prospektiven vergleichenden Therapiestudien (unveröffentlichtes Manuskript).* 1994.

173. Wernick, R.M., et al., *Reliability of Histologic Scoring for Lupus Nephritis: A Community-based Evaluation.* Ann.Intern.Med., 1993. **119**: p. 805-811.

174. Klinkhoff, A.V., et al., *An Experiment in Reducing Interobserver Variability of the Examination for Joint Tenderness.* J.Rheumatol., 1988. **15**: p. 492-494.

175. Hilden, J. and J.D.F. Habbema, *Evaluation of Clinical Decision Aids - More to Think About.* Med.Inform., 1990. **15**: p. 275-284.

176. Wyatt, J. and D. Spiegelhalter, *Evaluating Medical Expert Systems: What to Test and How?* Med.Inform., 1990. **15**: p. 205-217.

177. Simon, R. and D.G. Altman, *Statistical Aspects of Prognostic Factor Studies in Oncology.* Br.J.Cancer, 1994. **69**: p. 979-985.

178. Centor, R.M., B. Yarbrough, and J.P. Wood, *Inability to Predict Relapse in Acute Asthma.* N.Engl.J.Med., 1984. **310**: p. 577-580.

179. Rose, C.C., J.G. Murphy, and J.S. Schwartz, *Performance of an Index Predicting the Response of Patients with Acute Bronchial Asthma to Intensive Emergency Department Treatment.* N.Engl.J.Med., 1984. **310**: p. 577-580.

180. Bouillon, B., et al., *Traumascoresysteme als Instrumente der Qualitätskontrolle.* Unfallchirurg, 1993. **96**: p. 55-61.

181. Windeler, J., *Zur Methodik der Effektivitätsbeurteilung diagnostischer Tests,*, B. Habilitationsschrift, Editor. 1992.

182. van Ruiswyk, J., et al., *A Measure of Mortality Risk for Elderly Patients with Acute Myocardial Infarction.* Med.Decis.Making, 1993. **13**: p. 152-160.

183. Levine, M.N., et al., *When Is a Prognostic Factor Useful?: A Guide for the Perplexed.* J.Clin.Oncol., 1991. **9**: p. 348-356.

184. Wasson, J.H., et al., *Clinical Prediction Rules: Applications and Methodological Standards.* N.Engl.J.Med., 1985. **313**: p. 793-799.

185. Spiegelhalter, D.J., *Evaluation of Clinical Decision-Aids, with an Application to a System for Dyspepsia.* Stat.Med., 1983. **2**: p. 207-216.

186. D'Agostino, R.B. and M.W. Pozen, *The Logistic Function as an Aid in the Detection of Acute Coronary Disease in Emergency Patients (A Case Study).* Stat.Med., 1982. **1**: p. 41-48.

187. Pozen, M.W., et al., *A Predictive Instrument to Improve Coronary-Care-Unit Admission Practices in Acute Ischemic Heart Disease.* N.Engl.J.Med., 1984. **310**: p. 1273-1278.

188. Lange, S., J. Windeler, and H.J. Trampisch, *Diagnose: Entscheidungsfindung oder Selbstzweck? Beispiel Osteoporose (unveröffentlichtes Manuskript).* 1994.

189. Holle, R., *Prognose und Entscheidungsfindung - Klinische Bedeutung und methodische Probleme. 21.Sitzung der GMDS-Arbeitsgruppe "Methoden der Prognose und Entscheidungsfindung" , Heidelberg 29.4.94*.

190. Pritsch, M. and R. Holle, *Testing Therapeutic Equivalence of Cancer Treatments with Survival Time as Major Outcome Variable (Poster). 15th Conference of the International Society of Clinical Biostatisticians, Basel 25.-28.7.94*.