

Beobachtungsstudien als Mittel der Erkenntnisgewinnung über die Wirksamkeit von Arzneimitteln

von

Berthold Schneider
Institut für Biometrie
Medizinische Hochschule Hannover
Carl-Neuberg-Str. 1
30625 Hannover

Inhalt

1	Einleitung	2
2	Therapeutische Wirksamkeit und der Wirksamkeitsnachweis.....	3
3	Kohortenstudien.....	5
3.1	Studienplan	5
3.2	Auswertungskonzept	7
4	Beispiel einer onkologischen Kohortenstudie.....	9
	Anhang: Grundbegriffe der Wahrscheinlichkeitsrechnung und Statistik.....	19

1 Einleitung

In einem Artikel aus dem Jahre 1864 schreibt der Leipziger Klinker C.A. Wunderlich: "Der Punkt, auf den zuletzt all unser Bestreben, alle unsere Untersuchung sich richten müssen, ist die Therapie. Sie ist nicht nur das letzte humane Ziel aller medizinischen Forschung, sondern weithin auch der wissenschaftlich interessanteste Teil derselben. Damit müssen alle Schulen, alle Richtungen, die in der Heilkunde bestehen, übereinstimmen. Der Unterschied ist nur der, daß die Einen eine rationelle Begründung der therapeutischen Regeln und Grundsätze zur eigenen wissenschaftlichen Befriedigung, wie zur größeren Garantie für die Behandelten verlangen, während die Anderen meinen, eine Anwendung des Erfahrenen reiche in der Therapie aus, oder sei gar das Höchste oder Einzige, was erwartet werden dürfe" [16]. Der Unterschied zwischen einer 'rationalen Medizin' und einer 'Erfahrungsmedizin' wird auch heute noch oft gemacht, wobei die Anhänger der ersteren sich als 'Schulmediziner' letzteren überlegen dünken, während die Anhänger der Erfahrungsmedizin auf die Erfolge in der Vergangenheit und auf die Akzeptanz in der Bevölkerung verweisen. Tatsächlich sind die Verhältnisse aber komplexer und die Unterschiede zwischen verschiedenen Schulen und Richtungen sind nicht einfach mit den beiden Schlagwörtern 'rational' und 'Erfahrung' zu charakterisieren. Zunächst ist festzustellen, dass jede Medizin (auch die sog. Schulmedizin) eine Erfahrungswissenschaft ist und daher immer von der Erfahrung ausgehen muss. Nach dem Satz: "Daß alle unsere Erkenntnis mit der Erfahrung anfangt, daran ist gar kein Zweifel", mit dem Immanuel Kant die 'Kritik der reinen Vernunft' beginnt, geht nicht nur die Medizin sondern jede menschliche Erkenntnis von der Erfahrung aus. Erkenntnis ist aber nicht nur Erfahrung. Dies betont Kant, wenn er weiter fortfährt: "Wenn aber gleich alle unsere Erkenntnis mit der Erfahrung anhebt, so entspringt sie darum doch nicht eben alle aus der Erfahrung. Denn es könnte wohl sein, daß selbst unsere Erfahrungserkenntnis ein Zusammengesetztes aus dem sei, was wir durch Eindrücke empfangen, und dem, was unser eigenes Erkenntnisvermögen ... aus sich selbst hergibt ..." [9]. Erkenntnisse sind also bewusste Vorstellungen oder Modelle, die wir auf Grund der Erfahrungseindrücke von uns und unserer Umwelt machen, um damit uns und die Welt zu verstehen und in der Welt bestehen zu können. Die Art, wie wir aus den Sineseeindrücken der Erfahrung Erkenntnisse gewinnen, ist primär genetisch fixiert. Ihre konkreten Ausprägung hängt aber von vielen historisch gewachsenen oder neu hinzugekommenen Einflüssen und Bedingungen ab, so dass es verschiedene Wege der Erkenntnisgewinnung und verschiedene Erkenntnisse aus denselben Erfahrungen gibt. Dies ist eine Erkenntnis, die wir selbst aus der Erfahrung machen; z.B. auch aus der Erfahrung, dass in der Medizin verschiedene Schulen und Richtungen bestehen. Die Bezeichnung 'komplementäre Medizin' (bzw. 'Komplementärmedizin') wird dieser Erfahrung eher gerecht als z.B. die Bezeichnungen 'Erfahrungsmedizin'. Dadurch wird einerseits der Tatsache Rechnung getragen, dass die Schulmedizin sich auch auf Erfahrung gründet und die Komplementärmedizin sich auch "rationeller Begründungen" bedient; andererseits wird zum Ausdruck gebracht, dass die Methoden und die hinter den Methoden stehenden Überlegungen und Modelle der Komplementärmedizin nicht im Widerspruch, sondern 'ergänzend' zu den anerkannten und bewährten Methoden und Überlegungen der Schulmedizin stehen. 'Komplementär' in diesem Sinne ist auch das Thema dieses Beitrags, nämlich die Verwendung von Beobachtungsstudien zum Nachweis der therapeutischen Wirksamkeit von Arzneimitteln

(bzw. allgemeiner von Behandlungsverfahren), die 'ergänzend' zur Verwendung kontrollierter klinischer Studien anzusehen ist. Um dies verständlich zu machen, werden zunächst der Begriff der therapeutischen Wirksamkeit und die Prinzipien des Wirksamkeitsnachweises erörtert. Anschließend wird das Konzept der Kohortenstudie als epidemiologische Beobachtungsstudie diskutiert und gezeigt, wie bei diesem Studientyp valide Erkenntnisse zur therapeutischen Wirksamkeit gewonnen werden können. Diese Verfahren werden schließlich am Beispiel einer onkologischen Kohortenstudie zum Nachweis der Wirksamkeit einer oralen Enzymtherapie in der Nachsorge von Brustkrebs-Patientinnen demonstriert.

2 Therapeutische Wirksamkeit und der Wirksamkeitsnachweis

Nach einem Leitsatz im Urteil des Bundesverwaltungsgerichts vom Oktober 1993 ist die therapeutische Wirksamkeit "unzureichend begründet", "wenn sich aus dem vorgelegten Material nach dem jeweils gesicherten Stand der wissenschaftlichen Erkenntnisse nicht ergibt, dass die Anwendung des Arzneimittels zu einer größeren Zahl an therapeutischen Erfolgen führt als seine Nichtanwendung" [4]. Demnach ist Wirksamkeit die Eigenschaft eines Arzneimittels, bei mehr Patienten eine Heilung oder zumindest Besserung oder Erleichterung ihrer Beschwerden hervorzurufen, als ohne die Anwendung des Mittels zu erwarten wäre. In dieser Begriffsbestimmung sind die beiden wichtigsten Säulen des Wirksamkeitsnachweises enthalten, nämlich die **Kausalität**, die durch den Vergleich der bei der Anwendung des Arzneimittels zu erwartenden Änderungen des Patientenzustandes (die 'Wirkungen' des Arzneimittels) mit den Änderungen, die ohne diese Anwendung (oder bei Anwendung eines anderen, bekannten Mittels) zu erwarten sind, zu belegen ist, und die **Universalität**, nach der die Wirksamkeit nicht nur für einzelne, selektierte Patienten, sondern für alle Patienten, insbesondere für alle zukünftigen Anwender des Mittels glaubhaft zu machen ist. Als dritte Säule ist die für wissenschaftliche Erkenntnisse selbstverständliche Forderung nach **Objektivität** hinzuzufügen, nach der die Verfahren, mit denen die Aussagen getroffen werden, klar darzulegen sind, so dass sie wiederholt und die Aussagen überprüft werden können.

Wenn auch die Wirksamkeit eines Arzneimittels alle, insbesondere auch die zukünftigen Anwendungen des Mittels betrifft, so müssen Aussagen über die Wirksamkeit doch mit den Erfahrungen und Beobachtungen bei vergangenen Anwendungen und Nichtanwendungen (bzw. Anwendungen anderer Mittel oder Behandlungen) gemacht werden. Aussagen über die Wirksamkeit verlangen also einen induktiven Schluss von bekannten Beobachtungen auf unbekanntes, zukünftige Ereignisse. Dies leisten die Methoden der Wahrscheinlichkeitsrechnung und Statistik.

Die Wahrscheinlichkeit für ein bestimmtes Ergebnis bei einem Vorgang oder Ereignis (z.B. die Heilung eines Patienten nach Anwendung eines bestimmten Arzneimittels) wird interpretiert als die relative Häufigkeit, mit der das Ergebnis in der Grundgesamtheit aller möglichen Wiederholungen des Vorgangs oder Ereignisses vorkommt (s. Anhang). Sie ist somit ein Maß für die Zuverlässigkeit, mit der das Ergebnis bei zukünftigen Wiederholungen erwartet werden kann. Mit dem Wahrscheinlichkeitsbegriff kann die Wirksamkeit eines Arzneimittels folgendermaßen definiert werden: Ein Arzneimittel ist wirksam, wenn die Wahrscheinlichkeit eines therapeutischen Erfolgs bei Anwendung des Mittels größer ist als ohne Anwendung des Mittels. Der Nachweis der Wirksamkeit verlangt also einen Vergleich der Wahrscheinlichkeiten für den

Therapieerfolg. Bevor dies geschehen kann ist zunächst zu präzisieren, was unter Therapieerfolg zu verstehen ist. Dies geschieht durch die Festlegung einer 'primären Zielgröße'; d.h. eines Beobachtungsmerkmals, das primär die therapeutisch relevante Änderung des Gesundheitszustandes ausdrückt. Dies kann z.B. die völlige Heilung oder die Dauer bis zu einer definierten Besserung sein. Bei einer symptomatischen Behandlung wird man das zu bessernde Symptom als Zielgröße nehmen, z.B. die Blutdrucksenkung bei der Behandlung der Hypertonie. Da für die Wirksamkeit der Unterschied in der Wahrscheinlichkeitsverteilung der Zielgröße zwischen behandelten und nicht behandelten Patienten entscheidend ist, muss zusätzlich festgelegt werden, mit welchem Parameter dieser Unterschied ausgedrückt wird (Effektparameter, effect size). Bei dem binären Ereignis Heilung als Zielgröße kann dies die Differenz der Heilungswahrscheinlichkeit zwischen behandelten und nicht behandelten Patienten sein. Für die statistischen Aussagen vorteilhafter ist aber die relative Quote (odds ratio); d.i. der Quotient aus dem Verhältnis der Wahrscheinlichkeit für eine Heilung zur Wahrscheinlichkeit für keine Heilung zwischen den behandelten und den nicht behandelten Patienten. Bei der Dauer bis zu einer Besserung als Zielgröße kann die Differenz der medianen Dauern oder die relative Quote der Besserungen nach einer bestimmten Dauer Effektparameter sein (beim sog. 'proportional hazard rate' Modell ist die relative Quote unabhängig von der vorgegebenen Dauer). Bei quantitativen Zielgrößen, wie z. B. Blutdrucksenkung, kann die Differenz in den Erwartungswerten als Effektparameter genommen werden.

Sind Zielgrößen und Effektparameter festgelegt, dann sind Effektparameter und Konfidenzintervall zu schätzen und damit (bzw. mit einem äquivalenten Testverfahren) zu entscheiden, ob eine Wirksamkeit behauptet werden kann oder nicht. Dazu benötigt man Daten der Zielvariablen von Patienten, die mit dem zu prüfenden Arzneimittel behandelt wurden (Testgruppe), und von Patienten, die damit nicht behandelt wurden sondern z.B. keine spezifische Behandlung oder eine Behandlung mit einem Placebo oder Standardmittel erhalten haben (Kontrollgruppe). Mit diesen kann aber nur dann eine valide (unverzerrte) Schätzung erhalten werden, wenn die Patienten der Test- und Kontrollgruppe in den Ausgangs- und relevanten Behandlungsbedingungen vergleichbar (strukturgleich) sind; d.h. als zufällig aus derselben Grundgesamtheit dieser Bedingungen ausgewählt angesehen werden können. Dies ist dann gewährleistet, wenn die Patienten in einer **kontrollierten klinischen Studie** erfasst, behandelt und ihre Daten erhoben werden. Bei diesem Studientyp werden nach einem vorgegebenen Auswahlplan mit definierten Ein- und Ausschlusskriterien Patienten für die Studie ausgewählt und nach einem Zufallsmechanismus (randomisiert) der Test- oder Kontrollgruppe zugeteilt. Meist ist die Wahrscheinlichkeit für die Zuteilung zu beiden Gruppen gleich; d.h. es werden gleich viel Patienten der Test- wie der Kontrollgruppe zugeteilt. Es können aber auch unterschiedliche Zuteilungsraten vorgegeben werden. Die Patienten der Testgruppe werden mit dem zu prüfenden Arzneimittel, die der Kontrollgruppe mit dem Vergleichsmittel (Placebo oder Standardmittel) behandelt. Die sonstigen Behandlungsmaßnahmen und die Befunderhebung sind für beide Gruppen gleich und werden im Prüfplan festgelegt. Durch die randomisierte Zuteilung und die Vorgabe gleicher sonstiger Behandlungsmaßnahmen wird die Strukturgleichheit beider Gruppen garantiert. Um auch noch einen Einfluss durch die Kenntnis der zugeteilten Behandlung auszuschalten, kann (wenn dies sachlich möglich und ethisch vertretbar ist) die Zuteilung 'doppelblind' erfolgen, d.h. weder dem behandelnden Arzt noch dem Patienten bekannt gegeben werden.

Bei noch nicht zugelassenen Arzneimitteln, deren Wirksamkeit und Sicherheit völlig unbekannt sind, ist die kontrollierte Studie die einzig zulässige Studienart. Daneben gibt es aber vor allem in der komplementären Medizin viele Behandlungsverfahren, die seit vielen Jahren angewandt werden, ohne dass ihre Wirksamkeit und Sicherheit in kontrollierten Studien überprüft worden wäre. Über die mit diesen Behandlungen erzielten Wirkungen (d.h. Änderungen des Gesundheitszustandes) liegen in den Krankenakten Informationen vor. Es liegt nahe, diese Informationen zum Nachweis der Wirksamkeit und Sicherheit zu nutzen. Der dafür in Frage kommende komplementäre Studientyp ist die epidemiologische **Kohortenstudie**. Das Konzept dieser Studien und ihrer Auswertungsverfahren wird im nächsten Abschnitt besprochen.

3 Kohortenstudien

3.1 Studienplan

Kohortenstudien sind epidemiologische Populationsstudien, d.h. mit ihnen sollen die Zusammenhänge zwischen verschiedenen Maßnahmen und Faktoren, die die Gesundheit beeinflussen (z.B. Behandlungsmaßnahmen, Lebensgewohnheiten, Umwelteinflüsse), und dem Gesundheitszustand bzw. seinen Änderungen in realen Populationen untersucht werden. Die Maßnahmen oder Faktoren werden bei der Studie nicht festgelegt, sondern ergeben sich aus den praktischen Situationen. Hier besteht ein wesentlicher Unterschied zu den kontrollierten Studien, bei denen die bei den Patienten anzuwendenden Behandlungsmaßnahmen nach einem Zufallsschema festgelegt werden. Kontrollierte Studien sind 'Experimente' an Patienten, bei denen durch die Festlegung der Rahmenbedingungen und die systematische Vorgabe von Test- und Kontrollbehandlung die Natur zu einer Antwort "genötigt" wird (wie es Kant in der 'Kritik der reinen Vernunft' ausdrückte [9]). Demgegenüber sind Kohortenstudien 'Beobachtungsstudien', bei denen keine künstliche Situation geschaffen und die Natur nicht genötigt, sondern nur systematisch beobachtet wird.

Will man Kohortenstudien zum Nachweis der Wirksamkeit eines Arzneimittels (der Testbehandlung) gegen eine bestimmte Erkrankung verwenden, dann müssen hierfür Patienten mit dieser Erkrankung aus einer Population repräsentativ ausgewählt werden, in der neben der Testbehandlung auch andere Behandlungen gegen die Krankheit angewandt werden, die als Kontrollbehandlungen des Therapievergleichs dienen können. Da die Behandlungen nicht vorgegeben werden, ist es auch nicht erforderlich, dass nur neue Behandlungsfälle in die Studie eingeschlossen werden, deren Befunde 'prolektiv' (nach Feinstein [8]) erfasst und dokumentiert werden, sondern man kann auch auf die dokumentierten Krankenakten bereits abgeschlossener Fälle zurückgreifen und daraus die Daten 'retrolektiv' (nach Feinstein [8]) erfassen. Bei der zunehmenden Verwendung guter Arzt- und Klinik-Informationssystemen mit gut strukturierten Datenbanksystemen dürfte dies in Zukunft eine wesentlicher Erleichterung für die Studiendurchführung darstellen.

Für die Planung und Durchführung retrolektiver Kohortenstudien gelten die allgemeinen Empfehlungen für Anwendungsbeobachtungen (AWB), wie sie vom Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) erlassen wurden [3]. In den Empfehlungen des BfArM werden folgende generelle Anforderungen genannt: "AWB erfordern eine Planung, Durchführung, Aus- und Bewertung nach dem Stand der wissenschaftlichen Erkenntnis der beteiligten Disziplinen. Sie müssen eine medizinisch-

wissenschaftliche Zielsetzung ... verfolgen, die als präzise Fragestellung vorab formuliert sein muß. Das gewählte Design (Basis eines Vergleichs, zeitlicher Umfang und Untersuchungsumfang beim einzelnen Patienten, Patientenzahl) und die geplanten Methoden (Datenerhebung und Auswertung) müssen zur Beantwortung dieser Frage geeignet sein. Eine AWB ist prospektiv, ggf. mit zurückverlegtem Anfangspunkt, durchzuführen und orientiert sich in Anlage und Durchführung an einer Kohortenstudie. Sie kann auch auf geeigneten pharmakoepidemiologischen Datenbeständen basieren" [3]. Die letzten beiden Sätze schließen auch das Design einer 'retropektiven Kohortenstudie' ein, wie es von Feinstein [8] vorgeschlagen wurde.

Im Studienplan sind die Verantwortlichkeiten (Leiter der Studie, Koordination, Monitoring, Biometrie, Sponsor) anzugeben. Die Zielsetzung und die präzise Fragestellung sind zu formulieren. Es ist weiter die Auswahl der Patienten (bzw. Patientenakten) festzulegen. Hierzu ist zunächst aus den in Frage kommenden Behandlungseinrichtungen (z.B. Praxen, Kliniken, Nachsorgeeinrichtungen), in denen Patienten mit der festgelegten Erkrankung sowohl die Testbehandlung als auch Kontrollbehandlungen erhalten, eine repräsentative Auswahl zu treffen. Die Maßnahmen zum Erreichen der Repräsentativität sind zu beschreiben. Es sind ferner genaue Ein- und Ausschlusskriterien für die Patienten anzugeben, deren Daten erfasst werden sollen. Bei den Einschlusskriterien sind die Zeiträume für die Durchführung der Behandlungen und die Behandlungsgründe (Diagnosen, Indikationen, Ausgangszustand) festzulegen. Es ist ferner festzulegen, welche Befunde (demographische Daten, Anamnesen, diagnostische Befunde, durchgeführte Maßnahmen, Ausgangsbefunde, Verlaufsbeurteilungen, spezielle Ereignisse, Behandlungsergebnisse) aus den Krankenakten zu erfassen sind. Die Relevanz dieser Befunde für die Fragestellung sollte erläutert werden. Es ist dabei auch anzugeben, welche dieser Befunde primäre oder sekundäre Zielgrößen sind (bzw. zur Ermittlung der Zielgrößen dienen) und welche als Begleit- oder Störgrößen anzusehen sind. Ferner ist anzugeben und zu begründen, welche Behandlungsmaßnahmen als Testbehandlung, welche als Kontrollbehandlungen und welche als Zusatzbehandlungen angesehen werden. Umfang und Art der zu dokumentierenden Angaben (z.B. Bezeichnung der Präparate, Darreichungsform, Dosierung, Dauer und Art der Behandlung (Dauertherapie, intermittierend, bei Bedarf)) sind festzulegen. Schließlich sind im Studienplan auch die geplanten Auswertungskonzepte (s.u.) und die Regelungen für die Berichterstattung anzugeben. Die vorge-sehen Patientenzahl ist zu begründen.

Die Daten aller Patienten, die den Einschlusskriterien genügen und keine Ausschlusskriterien aufweisen sind zu erfassen und zu dokumentieren. Falls die Daten der Krankengeschichten in einer auf Vollständigkeit, Richtigkeit und Plausibilität überprüften Datenbank gespeichert sind, macht dies keine Probleme. Es sind lediglich die entsprechenden Daten aus der Datenbank in eine Auswertungsdatei zu übertragen. Falls die Krankengeschichten aber nur in Papierform vorliegen, müssen die Daten in strukturierte Erfassungsbogen (Case Report Forms (CRF)) übertragen werden. Die Korrektheit der Übertragung sollte durch unabhängige Monitore kontrolliert werden. Die Daten der Erfassungsbogen müssen dann in ein Datenbanksystem eingegeben und auf Vollständigkeit, Richtigkeit und Plausibilität überprüft werden. Neben den Patientendaten sind auch die für die Fragestellung relevanten Angaben zu den Behandlungseinrichtungen (z.B. Fachrichtung der behandelnden Ärzte, Spezifikationen der Behandlungseinrichtungen u.ä.) zu erfassen und zu speichern. Die Art der

Erfassung und Dokumentation (z.B. das verwendete Datenbanksystem) sind im Studienplan anzugeben.

3.2 Auswertungskonzept

Bei Kohortenstudien geschieht die Zuteilung der Behandlungen zu den Patienten ausschließlich nach Entscheidungen des Arztes oder Patienten. Sie kann als ein Zufallsereignis angesehen werden, dessen Verteilung von zahlreichen Kenngrößen der Behandlungseinrichtung und des Patienten (Einflussvariablen, Kovariablen) abhängt. Diese Variablen werden aber im allgemeinen auch den Therapieerfolg beeinflussen. Dadurch ist der unmittelbare Vergleich zwischen Prüf- und Kontrollgruppe nicht mehr möglich. Prüf- und Kontrollgruppe können nicht mehr als strukturgleich angesehen werden. Deshalb kann man aus den Daten nicht unmittelbar den Effektparameter, der den Unterschied in den Verteilungen der Zielgröße zwischen Prüf- und Kontrollgruppe charakterisiert, schätzen. Es ist eines der Hauptprobleme der Auswertung, den Einfluss dieser Variablen auf das Behandlungsergebnis auszugleichen und so einen unverzerrten Vergleich des Therapieerfolgs zwischen den Behandlungsgruppen zu ermöglichen.

Hierfür gibt es im Prinzip zwei Ansätze: Stratifikation und Kovarianzanalyse. Bei der **Stratifikation** werden Untergruppen (strata) mit ähnlichen Werten der Kovariablen gebildet. Die Ergebnisse innerhalb der Untergruppen werden zwischen Prüf- und Kontrollgruppe verglichen (Schätzung des Effektparameters) und die Ergebnisse in geeigneter Form über die Gruppen zusammengefasst (z.B. nach der Methode von Mantel und Haenszel [9]). Eine Sonderform der Stratifikation bildet die Matched-Pairs-Technik, bei der die Untergruppe jeweils aus einem Patientenpaar mit ähnlichen Werten der Kovariablen besteht, von denen ein Patient die Prüf- der andere die Kontrollbehandlung erhält. Bei der **Kovarianzanalyse** wird die Abhängigkeit des Therapieerfolgs von den Kovariablen durch eine geeignete Funktion erfasst (meist eine lineare Funktion der entsprechend skalierten bzw. transformierten Kovariablen). Mit dieser Funktion werden die beobachteten Therapieergebnisse auf gemeinsame Referenzwerte der Kovariablen umgerechnet und diese bereinigten (adjusted) Ergebnisse zwischen den Behandlungsgruppen verglichen.

Selbstverständlich kann dieser Ausgleich nur für die Kovariablen durchgeführt werden, die bei der Studie auch erfasst wurden. Zu einem guten Ausgleich ist daher erforderlich, möglichst viele Kovariablen zuverlässig zu erfassen. Damit treten aber bei beiden Ausgleichsmethoden praktische Probleme auf. Eine Stratifikation nach 10 oder mehr Kovariablen ist praktisch kaum durchzuführen. Wenn jede der Kovariablen nur in 2 Ausprägungen vorliegt (z.B. vorhanden - nicht vorhanden), dann kommen bei 10 Kovariablen bereits 1024 Kombinationen vor, nach denen stratifiziert werden müsste. Auch Ausgleichsfunktionen mit sehr vielen Kovariablen können zu Problemen führen, zumal wenn noch gegenseitige Beeinflussungen der Kovariablen (sogenannte Wechselwirkungen) berücksichtigt werden sollen.

Diese Probleme können mit einem 'Ausgleichsscore' (balancing score) überwunden werden. Darunter ist eine Funktion aller Kovariablen zu verstehen, die die Zuteilung der Behandlungen beeinflussen können, so dass bei einem gegebenen Wert dieser Funktion die Zuteilung unabhängig von den Kovariablen ist. Zum Ausgleich braucht man dann nicht mehr nach allen möglichen Kombinationen sondern nur noch nach

den Werten des Ausgleichsscores zu stratifizieren bzw. den Therapieerfolg nur mit einer Funktion des Ausgleichsscores zu bereinigen. Ein nahe liegender Ausgleichsscore ist die Wahrscheinlichkeit für die Zuteilung der Prüfbehandlung als Funktion der Kovariablen. Diese Funktion wurde von Rosenblatt und Rubin zum Ausgleich bei Beobachtungsstudien eingeführt und 'propensity score' genannt [12], [13]. Dies soll hier mit 'Zuteilungsscore' übersetzt werden. Rosenblatt und Rubin haben gezeigt, dass bei der Stratifikation bzw. Kovarianzanalyse nach dem Zuteilungsscore ein optimaler Ausgleich erzielt wird, wenn diese Funktion alle Kovariablen enthält, die die Behandlungszuteilung beeinflussen. In diesem Fall ist nach dem Ausgleich mit dem Zuteilungsscore der Therapievergleich genau so unverzerrt und valide wie bei einer randomisierten Zuteilung. Darüber hinaus bietet aber der Zuteilungsscore noch wertvolle Information über die Bedingungen und Variablen, die in der Praxis die Ärzte veranlassen, die Prüfmedikation anzuwenden. Überhaupt besteht der Vorteil der Kohortenstudien gegenüber den randomisierten kontrollierten Studien darin, dass sie ein unverfälschtes Bild der praktischen Anwendung der Arzneimitteln geben und so auch neue Indikationen für das Arzneimittel aber auch Risiken aufzeigen können, die in kontrollierten Studien nicht erfasst werden (vgl. [5]).

Die Praktikabilität und Effektivität eines Ausgleichs mit dem Zuteilungsscore konnte mit mehreren Beobachtungsstudien demonstriert werden. Es sei hier auf die Arbeiten von Rubin [14], D'Agostino Jr [6], Perkins et al. [11], Wittenborg et al. [15] und Beuth et al. [2] verwiesen. Diese Publikationen zeigen, dass mit dem Zuteilungsscore ein sehr guter Ausgleich der Inhomogenitäten zwischen den Zuteilungsgruppen erreicht wird und so mit Kohortenstudien ein unverzerrter Therapievergleich möglich ist. Interessant ist in diesem Zusammenhang auch ein Artikel von Benson und Hartz [1], die in einem umfangreichen Vergleich kontrollierter Klinischer Studien und Anwendungsbeobachtungen gezeigt haben, dass die mit Anwendungsbeobachtungen geschätzten Therapieeffekte weder konsistent größer noch qualitativ verschieden von den in randomisierten, kontrollierten Studien erhaltenen Ergebnissen sind. Die Autoren kommen zu dem Schluss: "Our results suggest that observational studies usually do provide valid information. They could be used to exploit the many recently developed, clinically rich data bases. Only with a greater willingness to analyze these data bases is it possible to achieve a realistic understanding of how observational studies can best be used".

Inzwischen wird auch die Verwendung von epidemiologischen Kohortenstudien zumindest im Europäischen Recht unter bestimmten Bedingungen als gültiger Nachweis der Wirksamkeit und Sicherheit von Arzneimitteln anerkannt. In den bereits erwähnten Empfehlungen des BfArM [3] wird zwar festgestellt: "Ein Nachweis der Wirksamkeit allein durch AWB ist bis auf besonders begründete Ausnahmefälle nicht möglich". Diese Einschränkung wird aber durch die Fußnote: "Soweit bei bekannten Arzneimitteln umfangreiches und nachvollziehbar dokumentiertes, plausibles Erfahrungswissen vorliegt, kann eine sorgfältig geplante AWB allerdings die Akzeptanz von Indikationsaussagen ermöglichen. Über die Möglichkeit der Verwendung von Ergebnissen aus AWB in den Sonderfällen, in denen die Durchführung klinischer Prüfungen nicht möglich ist, muß im jeweiligen Einzelfall entschieden werden" relativiert. Einen noch größeren Stellenwert beim Wirksamkeitsnachweis erhalten Anwendungsbeobachtungen durch die Richtlinie 1999/83/EG der EG-Kommission [7], nach der " 'bibliographischer Verweis' auf andere Informationsquellen (beispielsweise Untersuchungen nach dem Inverkehrbringen, epidemiologische Studien, mit ähnlichen

Erzeugnissen durchgeführte Prüfungen) und nicht nur Versuche und Prüfungen als gültiger Nachweis für die Sicherheit und Wirksamkeit eines Erzeugnisses dienen können, wenn der Antragsteller hinreichend erläutert und begründet, warum er diese Informationsquellen anführt". Diese Richtlinie hat inzwischen Gesetzeskraft.

Anwendungsbeobachtungen sind aber nicht nur zum Wirksamkeitsnachweis bei der Nachzulassung von Arzneimitteln von Bedeutung. Sie spielen eine große Rolle in der Pharmakoepidemiologie. In einem Editorial stellt Cepeda fest [5]: "Observational studies are very important in pharmacoepidemiologic research. They provide information that is difficult to collect during randomized controlled trials (RCT's)".

4 Beispiel einer onkologischen Kohortenstudie

Mit einer retrolektiven Kohortenstudie sollte die Wirksamkeit und Verträglichkeit einer zusätzlichen oralen Enzymtherapie (Wobe Mugos E, Herst.: MUCOS Pharma, Gertsried) in der postoperativen Behandlung von Patientinnen mit Brustkrebs untersucht werden (vgl. [2]). Hierzu wurden in 128 Behandlungseinrichtungen (Praxen, Krankenhäuser, Nachsorgeeinrichtungen, onkologische Praxen) die Krankenakten aller Patientinnen herausgesucht, die zwischen 1991 und 1997 eine postoperative Nachsorge bei primärem, nicht-metastasierendem Brustkrebs erhalten haben. Als Nachsorgetherapien wurde in diesen Einrichtungen neben antineoplastischen Therapien (Bestrahlung, adjuvante Chemotherapie, systemische Hormontherapie) auch teilweise eine zusätzliche orale Enzymtherapie mit Wobe Mugos E angewandt. Das Alter der Patientinnen bei Nachsorgebeginn sollte zwischen 18 und 80 Jahre liegen.

Aus den Krankenakten wurden von allen Patientinnen, die die Einschlusskriterien erfüllten (nicht-metastasierender Brustkrebs, Nachsorge zwischen 1991 und 1997), die anamnestischen Daten, Angaben zu den durchgeführten Behandlungen und zum Behandlungsergebnis herausgesucht, auf standardisierte Befundbogen übertragen und in ein Datenbanksystem eingegeben. Die Plausibilität und Korrektheit der Daten wurde überprüft. Patientinnen, die in der Nachsorge zusätzlich zur Standardbehandlung mit Wobe Mugos E behandelt wurden, bildeten die Testgruppe, Patientinnen, die diese zusätzliche Behandlung nicht erhielten, die Kontrollgruppe. Es wurden die Daten von 649 Patientinnen zur Auswertung erfasst, von denen 239 (37%) der Testgruppe und 410 (63%) der Kontrollgruppe zuzuordnen waren.

Da die Zuteilung zur Test- und Kontrollgruppe nicht randomisiert erfolgte, sind zwischen beiden Gruppen in den Patientencharakteristika und den Charakteristika der Behandlungseinrichtungen Unterschiede zu erwarten. Tabelle 1 zeigt die Kenngrößen der Verteilungen bzw. die Häufigkeitsverteilung einiger relevanter Charakteristika für beide Gruppen. Das Alter der Frauen zu Beginn der Nachsorge war in beiden Gruppen ähnlich verteilt. Die mittlere Dauer der Nachsorge unterscheidet sich aber zwischen beiden Gruppen beträchtlich (609 Tage in der Testgruppe und 441 Tage in der Kontrollgruppe). Der Zustand nach der Operation (response) und das postoperative UICC-Stadium sind in beiden Gruppen ähnlich verteilt. Bei den angewandten Nachsorgemaßnahmen zeigen sich aber größere Unterschiede. So wurde eine Hormontherapie in der Testgruppe bei 39,3% und in der Kontrollgruppe bei 56,1% der Patientinnen angewandt. Besonders groß sind die Unterschiede in den Charakteristika der Behandlungseinrichtungen, d.h. im Fachgebiet und Alter des behandelnden Arztes. Von den Patientinnen der Testgruppe wurden 64,9% und von denen der Kon-

trollgruppe nur 21,2% von Allgemeinpraktikern betreut. Der Prozentsatz der von Onkologen betreuten Patientinnen betrug in der Testgruppe 0,4% (1 Patientin) und in der Kontrollgruppe 36,8%. Von den Patientinnen der Testgruppe wurden 63,7% von einem Arzt über 45 Jahre behandelt, von den Patientinnen der Kontrollgruppe 37,3%.

	Testgruppe	Kontrollgruppe
Alter zu Beginn (Jahre)	Mittelwert: 59 (\pm 10)	Mittelwert: 60 (\pm 12)
Dauer der Nachsorge (Tage)	Mittelwert: 609 (\pm 476)	Mittelwert: 441 (\pm 462)
Zustand nach Operation		
komplette Remission	224 (94,9%)	352 (94,9%)
partielle Remission	11 (4,7%)	16 (4,3%)
minimale Besserung	1 (0,4%)	3 (0,8%)
UICC-Stadium		
0 oder I	81 (35,4%)	148 (37,9%)
IIa	92 (40,2%)	129 (33,0%)
IIb	43 (18,8%)	77 (19,7%)
IIIa oder höher	13 (5,7%)	37 (9,4%)
Bestrahlung		
nein	84 (35,1%)	113 (27,6%)
ja	155 (64,9%)	297 (72,4%)
Chemotherapie		
nein	188 (78,7%)	284 (69,3%)
ja	51 (21,3%)	126 (30,7%)
Hormontherapie		
nein	145 (60,7%)	180 (43,8%)
ja	94 (39,3%)	230 (56,1%)
Physikalische Therapie		
nein	211 (88,3%)	325 (79,3%)
ja	28 (11,7%)	85 (20,7%)
Behandelnder Arzt		
Allgemeinpraktiker	155 (64,9%)	87 (21,2%)
Internist	11 (4,6%)	40 (9,8%)
Gynäkologe	20 (8,4%)	47 (11,5%)
Onkologe	1 (0,4%)	151 (36,8%)
Radiologe	49 (20,5%)	85 (20,7%)
Alter des behandelnden Arztes		
bis 45 Jahre	65 (36,3%)	245 (62,7%)
über 45 Jahre	114 (63,7%)	146 (37,3%)

Tabelle 1:

Häufigkeitsverteilung (bzw. Mittelwerte und Standardabweichungen) verschiedener Patienten- und Behandlungscharakteristika in beiden Gruppen

Wegen dieser Inhomogenitäten zwischen beiden Gruppen ist ein direkter Vergleich der Therapieergebnisse nicht möglich, da sie von den unterschiedlichen Ausgangs- und Behandlungsbedingungen beeinflusst und der Vergleich damit verzerrt sein kann. Um einen unverzerrten Therapievergleich zu erhalten, müssen daher die Therapieergebnisse vom Einfluss möglicher Störgrößen bereinigt werden. Dies kann mit Hilfe des Zuteilungsscores (propensity score) geschehen; d.h. mit der Wahrscheinlichkeit für die Zuteilung eines Patienten zur Testgruppe als Funktion aller relevanten Patienten- und Behandlungscharakteristika.

Dieser Zuteilungsscore kann mit einer logistischen Regression aus den Studiendaten geschätzt werden. Bei dieser Regression wird angenommen, dass der Logarithmus der Quote (odds) für die Testgruppe (d.i. der Quotient aus der Wahrscheinlichkeit P für die Zuteilung zur Testgruppe zur Wahrscheinlichkeit 1-P für die Zuteilung zur Kontrollgruppe) eine lineare Funktion der relevanten Charakteristika ist. In Formeln lautet diese logistische Funktion:

$$\log \frac{P}{1-P} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

Dabei symbolisieren die x_1, \dots, x_k die Werte der Patienten- oder Behandlungscharakteristika (die Einflussvariablen) und $\beta_0, \beta_1, \dots, \beta_k$ die entsprechenden Koeffizienten der linearen Funktion. Für eine Einflussvariable x_i gibt der Koeffizient β_i an, um wie viel sich der Logarithmus der Quote ändert, wenn der Wert dieser Variablen um eine Einheit erhöht wird. Quantitative Einflussvariablen wie z.B. das Alter oder die Dauer der Nachsorge können unmittelbar verwendet werden. Kategoriale Variable wie z.B. das Fachgebiet des behandelnden Arztes oder die Durchführung einer Bestrahlung müssen zu quantitativen Variablen umkodiert werden. Dies ist bei binären kategorialen Variablen, wie z.B. die Durchführung einer Bestrahlung, bei denen die Kategorie entweder vorliegt oder nicht vorliegt, einfach. Liegt die Kategorie vor, dann wird die Variable mit 1 kodiert, sonst mit 0. Bei kategorialen Variablen mit mehr als 2 Kategorien wird eine Kategorie (z.B. beim Fachgebiet des Arztes die Kategorie 'Allgemeinpraktiker') als Referenzkategorie ausgewählt und es werden für die restlichen Kategorien Hilfsvariablen (dummy variables) eingeführt, die den Wert 1 erhalten, wenn die Kategorie vorliegt, sonst den Wert 0. Die Referenzkategorie liegt vor, wenn alle Hilfsvariablen den Wert 0 haben. Sie wird nicht eigens mit einer Variablen charakterisiert.

Für die logistische Regression mit den in Tabelle 1 angegebenen Einflussvariablen wurden folgende quantitative Variablen benutzt: x_1 = Alter in Jahren; x_2 =Dauer der Nachsorge in Tagen. Beim Zustand nach Operation wurde 'komplette Remission' als Referenzkategorie gewählt. Die übrigen Kategorien wurden mit Hilfsvariablen folgendermaßen kodiert: x_3 =1 bei partieller Remission, x_4 =1 bei minimaler Erholung. Beim UICC-Stadium wurde Stadium 0 oder I als Referenzkategorie genommen. Die übrigen Kategorien wurden kodiert: x_5 =1 bei Stadium IIa, x_6 =1 bei Stadium IIb, x_7 =1 bei Stadium IIa oder höher. Bestrahlung (x_8), Chemotherapie (x_9), Hormontherapie (x_{10}) und physikalische Therapie (x_{11}) wurden mit 1 kodiert, wenn die betr. Therapie durchgeführt wurde, sonst mit 0. Beim Fachgebiet des behandelnden Arztes wurde 'Allgemeinpraktiker' als Referenzkategorie gewählt. Die übrigen Kategorien wurden folgendermaßen kodiert: x_{12} =1 für Internist, x_{13} =1 für Gynäkologe, x_{14} =1 für Onkologe, x_{15} =1 für Radiologe. Das Alter des behandelnden Arztes (x_{16}) wurde bei einem Alter bis 45 Jahre mit 0 und bei einem Alter über 45 Jahre mit 1 kodiert.

Aus den Studiendaten werden die Koeffizienten β_i nach der Maximum-Likelihood-Methode geschätzt. Die Likelihood ist die Wahrscheinlichkeit für das beobachtete Stichprobenergebnis als Funktion der unbekannt Parameter. Als Stichprobenergebnis interessiert hier, ob der Patienten der Testgruppe oder der Kontrollgruppe zugeordnet wurde. Die Wahrscheinlichkeit, dass ein Patient mit den Einflussvariablen x_1, \dots, x_k der Testgruppe zugeordnet wird, ist: $P(x_1, \dots, x_k; \beta_0, \beta_1, \dots, \beta_k)$, und die Wahrscheinlichkeit, dass er der Kontrollgruppe zugeordnet wird, ist: $1-P(x_1, \dots, x_k; \beta_0, \beta_1, \dots, \beta_k)$. Die genaue Form dieser Funktionen ist durch die logistische Gleichung bestimmt. Die Wahrscheinlichkeiten hängen von den unbekannt Parametern $\beta_0, \beta_1,$

... β_k ab. Da die Zuordnung zu den Gruppen für verschiedenen Patienten unabhängig ist, ist die Likelihood das Produkt dieser Wahrscheinlichkeiten über alle Patienten der Studie, wobei bei einem Patienten der Testgruppe P, bei einem der Kontrollgruppe 1-P einzusetzen ist. Schätzwerte b_0, b_1, \dots, b_k der Koeffizienten sind diejenigen Werte, bei denen die Likelihood ein Maximum annimmt. Die Maximierung kann mit einem geeigneten Statistik-Paket, z.B. SPSS oder SAS durchgeführt werden. Wir benutzten das Paket SPSS 10 for Windows.

Einflussvariable	b_i	exp(b_i)	95% Konf.Intervall	Sign. p
x_1 =Alter (Jahre)	-0,030	0,971	0,950-0,992	0,008
x_2 =Dauer der Nachsorge (Tage)	0,000	1,000	0,999-1,000	0,991
Zustand nach OP				
x_3 = partielle Remission	-0,250	0,779	0,175-1,455	0,205
x_4 = minimale Erholung	0,536	1,709	0,097-30,140	0,714
UICC-Stadium				
x_5 =IIa	0,124	1,132	0,654-1,960	0,657
x_6 =IIb	0,217	1,242	0,606-2,548	0,554
x_7 =IIIa oder höher	-0,026	0,974	0,368-2,581	0,958
x_8 =Bestrahlung	0,064	1,066	0,633-1,796	0,810
x_9 =Chemotherapie	-0,810	0,445	0,239-0,827	0,010
x_{10} =Hormontherapie	-1,370	0,254	0,148-0,435	<0,001
x_{11} =physikalische Therapie	-0,852	0,427	0,221-0,823	0,011
behandelnder Arzt				
x_{12} =Internist	-1,292	0,275	0,113-0,668	0,004
x_{13} =Gynäkologe	-1,282	0,278	0,138-0,557	<0,001
x_{14} =Onkologe	-5,244	0,005	0,001-0,041	<0,001
x_{15} =Radiologe	-0,293	0,746	0,376-1,483	0,403
x_{16} =Alter des Arztes >45 Jahre	0,570	1,768	1,040-3,003	0,035
Konstante b_0	2,651	14,163		

Tabelle 2

Koeffizienten b_i und relative Quoten exp(b_i) (mit 95%-Konfidenzintervall und Signifikanzwahrscheinlichkeit p) der Einflussvariablen des Zuteilungsscores (propensity score)

Die geschätzten Koeffizienten b_i für die in Tabelle 1 aufgeführten Einflussvariablen sind in Tabelle 2 angegeben. Zur Interpretation eignet sich der Ausdruck exp(b_i) bei binären Variablen, die mit 0 oder 1 kodiert sind, besser als der Koeffizient b_i . Die Größe exp(b_i) entspricht in diesem Fall der relativen Quote (odds ratio) für die Zuteilung zur Testgruppe; d.h. sie gibt an, um wie viel sich die Zuteilungsquote ändert, wenn die betreffende Variable vorhanden ist (im Vergleich zur Quote bei Fehlen der Variable). Bei einer relativen Quote >1 ist die Quote (und damit auch die Wahrscheinlichkeit) für die Zuteilung zur Testgruppe größer als die für die Zuteilung zur Kontrollgruppe, bei einer relativen Quote <1 ist sie kleiner. In Tabelle 2 sind daher neben den Größen b_i auch die Größen exp(b_i) und ihre 95%-Konfidenzintervalle angegeben. Überdeckt das Konfidenzintervall den Wert 1, dann ist der Einfluss der betreffenden Variablen auf die Zuteilung nicht signifikant. Zusätzlich ist noch die Signifikanzwahrscheinlichkeit p angegeben, d.i. die Wahrscheinlichkeit, mit der der geschätzte Wert exp(b_i) oder ein größerer (bzw. bei Schätzwerten <1 ein kleinerer) zu erwarten sind, wenn in der Grundgesamtheit die relative Quote genau 1 ist. Bei Konfidenzintervallen, die den Wert 1 nicht enthalten, ist $p < 0,05$. Die Größe b_0 ist der Schätzwert des Koeffizienten β_0 , der den Wert von $\log(P/(1-P))$ für den Fall angibt, dass alle Variablen x_i gleich 0 sind. Diese Größe ist für die Berechnung der Quoten $P/(1-P)$ bei gegebenen x_i -Werten erforderlich.

Tabelle 2 zeigt, dass die Gruppenzuteilung vor allem vom Fachgebiet und Alter des behandelnden Arztes, von den durchgeführten Behandlungen (vor allem Hormonbehandlung) und (in geringerem Maße) vom Alter der Patientin abhängt. Der Zustand nach OP (Remission, UICC-Stadium) und die Dauer der Nachsorge haben keinen Einfluss. Dies steht im Einklang mit den Daten von Tabelle 1, nach denen sich die Verteilungen des Fachgebiets und Alters des behandelnden Arztes und der angewandten Therapien zwischen beiden Gruppen besonders unterscheiden. Orale Enzyme als Zusatztherapie wurden vor allem von Allgemeinpraktikern und älteren Ärzten in der Nachsorge angewandt, während Internisten, Gynäkologen und vor allem Onkologen darauf verzichteten. Bei der Anwendung von Chemotherapie physikalischer Therapie und vor allem von Hormonen wurden seltener zusätzlich orale Enzyme gegeben als ohne diese Anwendungen.

Mit den in die Berechnung des Zuteilungsscores einbezogenen Variablen kann die Zuteilungsgruppe in 78% der Fälle richtig vorhergesagt werden, wenn man für eine Patientin mit einem Zuteilungsscore von 0,5 und größer die Zuteilung zur Testgruppe und für weniger als 0,5 zur Kontrollgruppe prognostiziert. Dies zeigt, dass mit den Variablen die Wahrscheinlichkeit für die Zuteilung zur Testgruppe sehr gut erfasst wird. Dies ist auch aus der kumulativen Häufigkeitsverteilung (empirische Verteilungsfunktion) des Zuteilungsscores in beiden Gruppen zu ersehen, die in Abb. 1 gezeigt sind. Von den Patientinnen der Kontrollgruppe haben 60% einen Zuteilungsscore kleiner oder gleich 0,3, während von den Patientinnen der Testgruppe nur etwa 5% einen so niedrigen Zuteilungsscore haben.

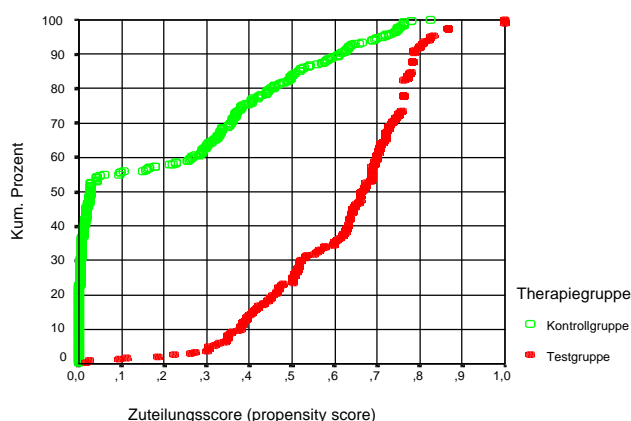


Abb. 1
Verteilung des Zuteilungsscores in beiden Gruppen

Um die Ausgewogenheit zwischen den Therapiegruppen innerhalb von Klassen des Zuteilungsscores zu überprüfen, wurden die Klassen von 0,3 bis 0,6 und von 0,6 bis 1,0 des Zuteilungsscores gebildet und die Häufigkeiten von ausgewählten Kategorien der Einflussvariablen für beide Gruppen innerhalb dieser Klassen verglichen. Die Klasse von 0 bis 0,3 ist für diese Betrachtung nicht geeignet, da nur 8 Patientinnen der Testgruppe einen so niedrigen Zuteilungsscore hatten und damit die Häufigkeiten große Zufallsschwankungen aufweisen. Die Ergebnisse sind in Tabelle 3 zu sehen. Die Klasse von 0,3 bis 0,6 umfasst insgesamt 182 Patientinnen, von denen 75 der Test- und 107 der Kontrollgruppe zugehören. Die Klasse von 0,6 bis 1,0 umfasst 200 Patientinnen, von denen 155 der Testgruppe und 45 der Kontrollgruppe zugehören. Die Häufigkeiten der ausgewählten Kategorien sind innerhalb der Zuteilungsscore-

klassen zwischen den beiden Gruppen deutlich homogener als in der Gesamtstichprobe. Die größeren Unterschiede zwischen den Gruppen in der Klasse 0,6-1,0 ist darauf zurückzuführen, dass zu dieser Klasse nur 45 Patientinnen der Kontrollgruppe gehören und damit die Häufigkeiten eine größere Zufallsschwankung aufweisen.

Kategorie	Zuteilungsscore	Testgruppe	Kontrollgruppe
Allgemeinpraktiker	0,3-0,6	39%	45%
	0,6-1,0	80%	62%
Alter des Arztes ≤45 Jahre	0,3-0,6	30%	41%
	0,6-1,0	39%	26%
Bestrahlung	0,3-0,6	71%	61%
	0,6-1,0	37%	42%
Chemotherapie	0,3-0,6	69%	77%
	0,6-1,0	85%	67%
Hormontherapie	0,3-0,6	75%	75%
	0,6-1,0	21%	5%
physikalische Therapie	0,3-0,6	19%	17%
	0,6-1,0	7%	11%

Tabelle 3

Häufigkeiten einiger Charakteristika in Test- und Kontrollgruppe innerhalb von Zuteilungsscoreklassen

Das primäre Ziel der Studie bestand darin, zu untersuchen, ob mit der zusätzlichen oralen Enzymtherapie krankheits- oder therapiebedingte Beschwerden stärker reduziert werden können als ohne diese Zusatztherapie. Als solche Beschwerden wurden angesehen: Gastrointestinale Beschwerden, Befindlichkeitsstörungen, Dyspnoe, Kopfschmerzen, Tumorschmerzen, Kachexie, Hautreaktionen, Infektionen. Aus den Krankenakten wurde ermittelt, ob ein Patientin im Verlauf der Nachsorge solche Beschwerden hatte und ob die Beschwerden am Ende der Nachsorge verschwunden waren. Jede Beschwerde wurde für sich bewertet. In die Bewertung wurden nur die Patientinnen einbezogen, bei denen die entsprechende Beschwerde aufgetreten ist. Als Erfolg wurde gewertet, wenn die Beschwerde am Ende der Nachsorge nicht mehr vorhanden war.

Beschwerden	Gruppe	N	Erfolg n (%)	rohe RQ	adjust. RQ	95%-Konfid. Intervall	Sign p
gastrointestinale Beschwerden	Test	140	59 (42%)	1,270	1,843	1,022-3,321	0,042
	Kontr.	203	74 (36%)				
Befindlichkeitsstö- rungen	Test	201	49 (24%)	0,992	1,113	0,636-1,948	0,707
	Kontr.	322	79 (24%)				
Dyspnoe	Test	52	16(31%)	2,222	3,105	0,972-9,918	0,056
	Kontr.	60	10 (17%)				
Kopfschmerz	Test	50	25 (50%)	2,400	1,568	0,652-3,773	0,315
	Kontr.	85	25 (29%)				
Tumorschmerz	Test	51	33 (65%)	1,244	0,705	0,266-1,871	0,483
	Kontr.	47	28 (60%)				
Kachexie	Test	23	15 (65%)	24,375	133,95	3,695-4855	0,008
	Kontr.	14	1 (7%)				
Hautreaktionen	Test	85	32 (38%)	0,397	3,028	1,371-6,685	0,006
	Kontr.	227	137 (60%)				
Infektionen	Test	52	25 (48%)	2,887	1,318	0,473-3,672	0,597
	Kontr.	70	17 (24%)				

Tabelle 4

Rohe und mit Zuteilungsscore adjustierte relative Erfolgsquote bei verschiedenen Beschwerden.

In Tabelle 4 sind in der 3. Spalte die Anzahlen N der Patientinnen angegeben, die in der Test- und Kontrollgruppe die jeweiligen Beschwerden hatten. In der 4. Spalte sind die Anzahlen und Prozentsätze der Erfolge (keine Beschwerden am Ende der Nachsorge) angegeben. Daraus wurde für jede Beschwerde als Vergleichsmaß die rohe relative Quote (odds ratio) berechnet; d.i. das Verhältnis der Erfolgsquote (d.i. Anzahl der Erfolge zu Anzahl der Misserfolge) zwischen Test und Kontrollgruppe. Wenn a die Anzahl der Erfolge und b die der Misserfolge in der Testgruppe sowie c die Anzahl der Erfolge und d die der Misserfolge in der Kontrollgruppe bezeichnet, dann errechnet sich die rohe relative Quote als: $(a*d)/(b*c)$. In Spalte 5 der Tabelle 4 sind diese rohen relativen Quoten (RQ) angegeben. Werte größer als 1 zeigen eine Überlegenheit der Testgruppe, Werte kleiner als 1 eine der Kontrollgruppe an.

Die rohen relativen Quoten können aber durch Inhomogenitäten der beiden Gruppen verzerrt sein. Um diese Verzerrungen auszugleichen wurde mit einer logistischen Regression die adjustierte relative Quote berechnet. Dabei wird angenommen, dass der Logarithmus der rohen Erfolgsquote $P_e/(1-P_e)$ in der Grundgesamtheit eine lineare Funktion der Gruppenzugehörigkeit x_1 (die für die Testgruppe mit 1 und für die Kontrollgruppe mit 0 kodiert wird) und des Zuteilungsscores x_2 ist:

$$\log(P_e/(1-P_e)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Der delogarithmierte Wert $\exp(\beta_1)$ des Koeffizienten β_1 gibt die von Inhomogenitäten der Gruppen bereinigte (adjusted) relative Erfolgsquote an; d.h. die relative Erfolgsquote von Test- zu Kontrollgruppe, die bei gleichem Wert des Zuteilungsscores für beide Gruppen gilt. Dies kann durch eine einfache mathematische Überlegung eingesehen werden: Da die Variable x_1 für die Testgruppe den Wert 1 und für die Kontrollgruppe den Wert 0 annimmt, ist der Logarithmus der relativen Quote:

$$\log\left(\frac{P_e(\text{Test})}{1-P_e(\text{Test})} / \frac{P_e(\text{Kontr.})}{1-P_e(\text{Kontr.})}\right) = \log\frac{P_e(\text{Test})}{1-P_e(\text{Test})} - \log\frac{P_e(\text{Kontr.})}{1-P_e(\text{Kontr.})} = \beta_1 + \beta_2(x_2(\text{Test}) - x_2(\text{Kontr.}))$$

Wenn der Wert des Zuteilungsscores $x_2(\text{Test})$ in der Testgruppe derselbe ist wie der Wert $x_2(\text{Kontr.})$ in der Kontrollgruppe, dann ist die Differenz zwischen beiden gleich 0 und der Logarithmus der relativen Quote gleich β_1 . Der delogarithmierte Wert $\exp(\beta_1)$ gibt also die auf gleiche Bedingungen (d.h. Wert des Zuteilungsscores) in Test- und Kontrollgruppe adjustierten und so von Inhomogenitäten in beiden Gruppen bereinigte relative Erfolgsquote an. Er ist somit ein unverzerrter Effektparameter für den Therapieerfolg der Testbehandlung.

Schätzwerte b_0 , b_1 und b_2 der Koeffizienten erhält man nach der Maximum-Likelihood-Methode aus den Kohortendaten. In Tabelle 4 sind in der 6. Spalte die Schätzwerte für die adjustierte relative Quote ($\exp(b_1)$), in Spalte 7 das 95%-Konfidenzintervall und in Spalte 8 die Signifikanzwahrscheinlichkeit p (d.h. die Wahrscheinlichkeit mit der die geschätzte oder eine größere (bzw. kleinere) adjustierte relative Quote zu erwarten ist, wenn in der Grundgesamtheit die adjustierte relative Quote gleich 1 ist) angegeben. Bei gastrointestinalen Beschwerden, Befindlichkeitsstörungen, Dyspnoe, Kachexie und Hautreaktionen ist die adjustierte relative Quote größer, bei Kopfschmerz, Tumorschmerz und Infektionen kleiner als die rohe relative Quote. Eine signifikante Überlegenheit der Testgruppe gegenüber der Kontrollgruppe zeigt sich bei gastrointestinalen Beschwerden, Kachexie und Hautreaktionen. Bemerkenswert ist, dass bei Hautreaktionen die rohe relative Quote deutlich kleiner als 1, die adjustierte aber signifikant größer als 1 ist. Die rohe relative Quote ist hier besonders stark von Inhomogenitäten beeinflusst.

Wie im Abschnitt 3 ausgeführt wurde, kann ein Ausgleich der Inhomogenitäten auch durch Stratifikation erreicht werden. Dabei werden Zuteilungsscoreklassen (Strata) gebildet und innerhalb jeder Klasse die relative Quote berechnet. Diese relativen Quoten werden anschließend zu einer gemeinsamen relativen Quote nach der Methode von Mantel und Haenszel [10] zusammengefasst, vorausgesetzt, dass die Strata homogen sind, d.h. für jedes Stratum in der Grundgesamtheit derselbe Wert der relativen Quote angenommen werden kann. Dieses Verfahren eignet sich weniger für diese Studie, da die Verteilung des Zuteilungsscores in beiden Gruppen sehr verschieden ist und so sehr große Klassen gebildet werden müssen, um in jedem Stratum eine hinreichend große Zahl von Fällen aus jeder Klasse zu haben. Es soll trotzdem für die gastrointestinalen Beschwerden zur Demonstration gebracht werden. Die Werte des Zuteilungsscores werden in 3 Klassen unterteilt: Klasse 1: 0 bis 0,3; Klasse 2: 0,3 bis 0,6; Klasse 3: 0,6 bis 1. Die Zahl der Patientinnen und Erfolge bezüglich gastrointestinaler Beschwerden innerhalb dieser Klassen sind für beide Gruppen in Tabelle 5 angegeben.

Zuteilungsscore	Gruppe	N	Erfolg n %	Relative Quote	95%-Konf. Intervall
0 - 0,3	Test	7	3 (43%)	0,960	0,958 - 3,056
	Kontrolle	114	50 (44%)		
0,3 - 0,6	Test	45	16 (36%)	1,336	0,593 - 3,007
	Kontrolle	65	19 (29%)		
0,6 - 1	Test	88	40 (45%)	3,167	1,085 - 9,239
	Kontrolle	24	5 (21%)		
adjustierte RQ (Mantel-Haenszel)				1,711	0,958 - 3,056

Tabelle 5

Häufigkeit der Erfolge bezüglich gastrointestinaler Beschwerden innerhalb von Klassen (Strata) des Zuteilungsscores.

Tabelle 5 zeigt auch für jede Klasse die berechnete relative Quote, die von 0,960 bis 3,167 variiert. In der Grundgesamtheit kann aber für alle 3 Klassen dieselbe relative Quote angenommen werden; der Test auf Homogenität (berechnet mit dem Programm StatXact 4.01) ergibt keine signifikante Abweichung von der Homogenitätshypothese ($p=0,332$). Als Schätzwert für die adjustierte gemeinsame relative Quote ergibt sich nach der Methode von Mantel-Haenszel der Wert 1,711 mit einem 95%-Konfidenzintervall von 0,958 bis 3,056; die Signifikanzwahrscheinlichkeit p ist 0,069. Bei Stratifikation mit verhältnismäßig großen Klassen ergibt sich für die bereinigte relative Quote ein ähnlicher Schätzwert (mit ähnlichem Konfidenzintervall) wie bei der logistischen Regression (bereinigte relative Quote: 1,843; KI: 1,022-3,321). Dies demonstriert die Äquivalenz der beiden Ausgleichsverfahren.

Als Ergebnis dieses Beispiels ist festzustellen, dass durch die Bereinigung der Therapieergebnisse von Kohortenstudien mit Hilfe des Zuteilungsscores valide Aussagen über die therapeutische Wirksamkeit eines Präparats erhalten werden können. Die Güte der Bereinigung von Inhomogenitäten der Zuteilung hängt davon ab, in welchem Maße der Zuteilungsscore alle relevanten Einflussvariablen erfasst. Als ein Kriterium kann die Häufigkeit genommen werden, mit der aus den Werten des Zuteilungsscores die tatsächliche Zuteilungsklasse richtig vorhergesagt wird. Im Beispiel betrug diese Häufigkeit fast 80%, wenn bei einem Zuteilungsscore von weniger als 0,5 die Kontrollgruppe und bei einem Zuteilungsscore von 0,5 oder größer die Testgruppe vorhergesagt wird. Man kann daher davon ausgehen, dass der Zuteilungs-

score alle relevanten Einflussfaktoren erfasst und so mit ihm deren Einfluss auf des Ergebnis bereinigt wird. Durch den Einschluss weiterer möglicher Einflussvariablen konnte auch keine Verbesserung des Scores erreicht werden. Die Analyse der Vollständigkeit des Zuteilungsscores (Sensitivitätsanalyse) ist ein wichtiger Schritt bei der Auswertung von Kohortenstudien.

Literatur

- 1 Benson K, Hartz AJ (2000) A comparison of observational studies and randomized controlled trials. *New England Journal of Medicine*, 342, 1878-1886
- 2 Beuth J, Ost B, Pakdaman A, Rethfeldt E, Bock PR, Hanisch J, Schneider B (2001) Impact of complementary oral enzyme application on the postoperative treatment results of breast cancer patients - results of an epidemiological multicenter retrospective cohort study. *Cancer Chemotherapy and Pharmacology*; 47, Suppl., S45-S54,
- 3 Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) (1998) Empfehlungen zur Planung und Durchführung von Anwendungsbeobachtungen. *Bundesanzeiger Jg. 50, Nr. 229, Seite 16884*
- 4 Bundesverwaltungsgericht, 3. Senat, Az: 3C 21/91, Urteil vom 14. Oktober 1993
- 5 Cepeda MS (2000) Editorial: The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and Drug Safety*, 9, 103-104
- 6 D'Agostino Jr. RB (1998) Tutorial in Biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17, 2265-2281
- 7 EG-Kommission, Richtlinie 199/83/EG vom 8. September 1999. *Amtsblatt der Europäischen Gemeinschaften vom 15. 9. 1999, L243/9*
- 8 Feinstein AR (1985) *Clinical epidemiology*. Saunders, Philadelphia
- 9 Kant I (1976) *Kritik der reinen Vernunft*. Philosophische Bibliothek Band 37a, Felix Meiner Verlag Hamburg
- 10 Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748
- 11 Perkins SM, Tu W, Underhill MG, Zhou X_H, Murray MD (2000) The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and Drug Safety*, 9, 93-101
- 12 Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41-55
- 13 Rosenbaum PR, Rubin DB (1984) Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516-524
- 14 Rubin DB (1997) Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127, 757-763
- 15 Wittenborg A, Bock PR, Hanisch J, Saller R, Schneider B (2000) Vergleichende epidemiologische Studie bei Erkrankungen des rheumatischen Formenkreises am Beispiel der Therapie mit nichtsteroidalen Antiphlogistika versus einem oralen Enzymkombinationspräparat. *Arzneimittel-Forschung* 50 (II), 728-738
- 16 Wunderlich CA (1864) *Die rationelle Therapie*. *Arch. für Physiologische Heilkunden*, 5, 1-16

Anhang: Grundbegriffe der Wahrscheinlichkeitsrechnung und Statistik

Die Gesamtheit der Ergebnisse oder Merkmalwerte, die man bei unbegrenzt häufiger Wiederholung der Beobachtung eines bestimmten Vorgangs oder Ereignisses erhält (z.B. des Therapieerfolgs bei allen möglichen Anwendungen eines Arzneimittels gegen eine bestimmte Krankheit), nennt man die Grundgesamtheit oder Population (R. von Mises hat hierfür auch die Bezeichnung 'Kollektiv' verwendet). Die relative Häufigkeit, mit der ein bestimmtes Einzelergebnis (d.i. ein bestimmter Merkmalwert) oder eine bestimmte Menge von Ergebnissen (Merkmalwerten) in der Grundgesamtheit vorkommen, ist die Wahrscheinlichkeit für das Ergebnis bzw. für die Menge von Ergebnissen. Es sei hier angemerkt, dass die Wahrscheinlichkeit auch ohne Bezug auf eine Grundgesamtheit als 'Erwartung' einer Person an ein unbekanntes Ereignis interpretiert werden kann. Da diese Interpretation einer 'personellen' oder 'subjektiven' Wahrscheinlichkeit in der Literatur z.Z. wenig verwendet wird, soll hier nicht weiter darauf eingegangen werden. Die Zuordnung der Wahrscheinlichkeiten zu den möglichen Ergebnis- oder Merkmalwerten ist die Wahrscheinlichkeitsverteilung (kurz auch nur 'Verteilung' genannt). Sie charakterisiert vollständig die Grundgesamtheit und ist daher das Ziel aller induktiven Schlüsse, d.h. aus beobachteten Ergebnissen (den Daten) sollen Aussagen über die den Beobachtungen zugrunde liegende Wahrscheinlichkeitsverteilung oder über bestimmte Kenngrößen (sog. Parameter) der Verteilung gemacht werden. Dies geschieht dadurch, dass aus den beobachteten Werten (den Daten) Schätzwerte oder Statistiken berechnet werden, die die Wahrscheinlichkeitsverteilung oder die interessierenden Parameter repräsentieren. Z.B. ist die relative Häufigkeit, mit der ein bestimmtes Ergebnis (z.B. eine Heilung) bei einer bestimmten Anzahl von Beobachtungen vorgekommen ist, ein Schätzwert für die Wahrscheinlichkeit dieses Ergebnisses. Man nennt die beobachteten Ergebnisse eine Stichprobe und nimmt an, dass sie zufällig und unabhängig aus der Grundgesamtheit entnommen wurden. Die Anzahl der Beobachtungen in einer Stichprobe nennt man den Stichprobenumfang und bezeichnet ihn meist mit n . Die mit den Daten einer Stichprobe berechneten Schätzwerte repräsentieren zwar die Verteilung, tun dies aber nur ungenau, da sie ja nur einen sehr kleinen Teil der Grundgesamtheit darstellen. Um die Genauigkeit der Repräsentanz durch Schätzwerte angeben zu können, nimmt man modellmäßig an, dass die Stichprobenentnahme beliebig oft wiederholt wird und aus jeder Stichprobe der Schätzwert berechnet wird. Diese Schätzwerte werden zufällig variieren. Sie bilden so eine neue Grundgesamtheit, in der den möglichen Schätzwerten eine Wahrscheinlichkeitsverteilung zugeordnet ist, die 'Schätzverteilung' genannt werden soll. Mit dieser Verteilung kann die Zuverlässigkeit der Schätzung charakterisiert werden. Der Mittelwert (der bei Grundgesamtheiten auch 'Erwartungswert' genannt wird) charakterisiert die 'Richtigkeit' der Schätzung. Eine Schätzung, bei der der Erwartungswert der Schätzverteilung stets mit dem tatsächlichen Wert des Parameters (der nicht bekannt ist) übereinstimmt, nennt man 'erwartungstreu' oder 'unverzerrt' (engl. unbiased). Die Standardabweichung der Schätzverteilung wird 'Standardfehler' der Schätzung genannt. Sie charakterisiert die Präzision (genauer Impräzision) der Schätzung; je kleiner der Standardfehler ist, desto präziser erfasst (bei erwartungstreuen Schätzungen) der Schätzwert den tatsächlichen Wert des Parameters. Eine weiter gehende Charakterisierung der Genauigkeit liefert das Konfidenzintervall zu einer vorgegebenen Konfidenzwahrscheinlichkeit γ , die meist auf 95% festgesetzt wird. Dies ist ein Intervall, das mit der vorgegebenen Wahrscheinlichkeit γ den tatsächlichen Wert des Parameter enthält; d.h., wenn man bei

der gedachten beliebig häufigen Wiederholung der Stichprobenentnahme jedes Mal das Konfidenzintervall nach derselben Regel bildet, dann wird in dieser Gesamtheit von Intervallen der Anteil γ (z.B. 95%) den tatsächlichen Parameterwert enthalten und der Anteil $1-\gamma$ (z.B. 5%) ihn nicht enthalten. Je schmaler das konkret aus den Daten einer gegebenen Stichprobe berechnete Konfidenzintervall ist, desto größer ist die Genauigkeit der Schätzung. Bei Vorgabe der Breite des Konfidenzintervalls kann der Stichprobenumfang n so festgelegt werden, dass diese Vorgaben eingehalten werden. Mit dem Konfidenzintervall lässt sich auch die Hypothese testen, dass der Parameter einen bestimmten Wert, z.B. den Wert 0 hat. Man nennt diese Hypothese die 'Nullhypothese'. Enthält das aus den Daten einer gegebenen Stichprobe berechnete Konfidenzintervall diesen Wert (z.B. 0), dann wird die Nullhypothese angenommen, sonst abgelehnt. Die Irrtumswahrscheinlichkeit (1. Art) dieses Tests ist $1-\gamma$; d.h., wendet man dieses Testverfahren bei der gedachten beliebig häufigen Wiederholung der Stichprobenentnahme jedes mal an und stimmt der tatsächliche Parameterwert mit dem in der Nullhypothese festgelegten Wert überein, dann beträgt der Anteil der fälschlichen Ablehnungen $1-\gamma$ (z.B. 5%).

Ein dazu äquivalentes Testverfahren besteht darin, aus den Stichprobenwerten eine Teststatistik zu berechnen, die den Unterschied zwischen den Stichprobendaten und dem in der Nullhypothese festgelegten Wert des Parameters charakterisiert. Bei wiederholter Stichprobenentnahme wird der Wert der Teststatistik zufällig variieren, wobei die Verteilung dieser Variationen von der Verteilung der Stichprobenwerte (und somit vom tatsächlichen Wert des Parameters) und vom Stichprobenumfang n abhängt. Die Nullhypothese wird abgelehnt, wenn die Wahrscheinlichkeit für den mit den beobachteten Stichprobenwerten berechneten Wert der Teststatistik oder für noch größere Werte bei Gültigkeit der Nullhypothese höchstens einen kleinen Wert α hat, für den meist 5% vorgegeben wird. α entspricht der Vorgabe $1-\gamma$ beim Test mit dem Konfidenzintervall; d.h. α ist die Irrtumswahrscheinlichkeit (1. Art), die Nullhypothese abzulehnen, obwohl sie richtig ist. Ist die Wahrscheinlichkeit für den berechneten Wert der Teststatistik oder einen größeren bei Gültigkeit der Nullhypothese (die 'Signifikanzwahrscheinlichkeit') größer als α , dann wird die Nullhypothese angenommen. Stimmt der tatsächliche Wert des Parameters nicht mit dem in der Nullhypothese festgelegten Wert überein und entscheidet man für die Annahme der Nullhypothese, dann begeht man einen Fehler 2. Art. Die Wahrscheinlichkeit für diesen Fehler hängt von der tatsächlichen Abweichung des Parameterwertes von dem in der Nullhypothese festgelegten Wert und vom Stichprobenumfang n ab. Je größer der Stichprobenumfang ist, desto geringer ist die Wahrscheinlichkeit, bei einem gegebenen alternativen Referenzwert des Parameters die Nullhypothese anzunehmen. Dies eröffnet die Möglichkeit, den Stichprobenumfang n so festzulegen, dass bei einem gegebenen alternativen Referenzwert des Parameters, die Annahme der Nullhypothese höchstens mit einer geringen Wahrscheinlichkeit β bzw. die Ablehnung der Nullhypothese mindestens mit Wahrscheinlichkeit $1-\beta$ erwartet werden kann. Diese zuletzt genannte Wahrscheinlichkeit nennt man die Stärke (power) des Testverfahrens. Man gibt häufig hierfür den Wert 80% vor.