Planung und Auswertung von Dosis-Wirkungs-Studien unter Verwendung von multiplen Testprozeduren

Habilitationsschrift
zur Erlangung der Venia legendi
für das Fach Biometrie

Der Medizinischen Hochschule Hannover
vorgelegt von

Dr. Frank Bretz
aus Hannover

Hannover 2003

1 Einleitung

Die Bestimmung einer adäquaten Dosierung eines neuen, in den Markt einzuführenden Medikamentes beinhaltet einen komplexen Entscheidungsprozeß in der klinischen Forschung. In diesem Ablauf sind zwei wesentliche Komponenten hervorzuheben: Zum einen der Beleg für die globale Wirksamkeit des Medikamentes (*Proof of Concept*, PoC), zum anderen die Auswahl erfolgversprechender Dosierungen im Verlauf der fortschreitenden Entwicklung (*dose finding*).

Zur Auswertung von Dosis-Wirkungs-Studien stehen komplementäre Ansätze zur Verfügung – *modellbasierte Verfahren* und *multiple Testprozeduren* – denen unterschiedliche statistische Prinzipien zugrunde liegen. Modellbasierte Verfahren gehen von einem funktionalen Zusammenhang zwischen Dosis und Wirkung aus, wobei die Dosis als quantitativer Faktor in das vor der Studie zu spezifizierende parametrische Modell eingeht. Beispiele solcher Analysen sind die bekannten logistischen, $E_{\max}$ oder log-linearen Regressionsmodelle. Das nach der Datenerhebung angepaßte Modell wird dann mittels Interpolation zur Schätzung einer adäquaten Dosierung zu einem vorgegebenen Wirkungsniveau eingesetzt. Ein inhärentes und bisher ungelöstes Problem besteht in der Zuverlässigkeit der statistischen Ergebnisse als Funktion einer korrekten Spezifizierung der Dosis-Wirkungs-Beziehung vor Studienbeginn: Entspricht der postulierte Kurvenverlauf nicht dem (*a-priori* unbekannten) wahren funktionalen Zusammenhang, sind die abschließenden Resultate möglicherweise verzerrt und aufgrund mangelnder Fehlerkontrolle potentiell unbrauchbar. Dieses Dilemma ist in der streng regulierten Umgebung der Arzneimittelentwicklung jeweils in der finalen Auswertung mit in Betracht zu ziehen.

Demgegenüber stehen multiple Testprozeduren, welche die Dosis als qualitativen Faktor in eine varianzanalytische Auswertung mit aufnehmen und, wenn überhaupt, nur geringe Voraussetzungen an das zugrundeliegende Dosis-Wirkungs-Modell stellen. Das primäre Ziel dieser Verfahren ist häufig die Schätzung der minimal effektiven Dosis, welche statistisch signifikant und klinisch relevant gegenüber Placebo ist. Vor diesem Hintergrund werden die experimentellen Fragestellungen (etwa ein globaler Wirksamkeitsnachweis im Sinne einer PoC oder die Schätzung einer klinisch relevanten Dosis) als zu testende Hypothesen mathematisch formalisiert und anschließend unter Einhaltung einer vorgegebenen Irrtumswahrscheinlichkeit auf Signifikanz getestet. Naturgemäß sind die abschließenden Ergebnisse und Entscheidungen auf die in der Studie untersuchten Dosierungen beschränkt, da ein funktionaler Dosis-Wirkungs-Zusammenhang eben nicht in Betracht gezogen wird und Interpolationsmöglichkeiten *per definitionem* entfallen.

Die im nachfolgenden zusammengefaßten Artikel

(1) *Testing dose-response relationships with a priori unknown, possibly nonmonotone shapes*

(2) *Detecting dose-response using contrasts: Asymptotic power and sample size determination for binomial data*

(3) *Identifying effective and/or safe doses by stepwise confidence intervals for ratios*

(4) *Statistical analysis of monotone or non-monotone dose-response data from* in-vitro *toxicological assays*

befassen sich ausschließlich mit der Einführung und Weiterentwicklung adäquater multipler Testprozeduren im Rahmen der oben skizzierten Fragestellungen. Der Schwerpunkt der Arbeiten liegt auf *multiplen Kontrasttests*, welche in zu-

nehmenden Maße angewandt werden. Ein wesentlicher Gesichtspunkt ist die Herleitung geeigneter Gewichtsfunktionen sowohl bei monotonen als auch bei nichtmonotonen Kurvenverläufen, worauf in (1) eingegangen wird. In (2) werden erstmalig multiple Kontrasttests in allgemeiner Form für binomial verteilte Daten unter besonderer Berücksichtigung der Fallzahlbestimmung während der Planungsphase einer Studie eingeführt. In (3) wird auf das in der Literatur kaum beachtete, weil theoretisch schwierige Problem der Herleitung simultaner Konfidenzintervalle für schrittweise paarweise Kontrasttests am Beispiel von Wirkungsquotienten eingegangen. Der letzte Artikel (4) gibt eine Zusammenfassung der theoretischen Resultate zu multiplen Kontrasttests und befaßt sich vor allem mit den Besonderheiten prä-klinischer Dosis-Wirkungs-Studien anhand mehrerer Ames Mutagenitätsassays.

Die genannten Probleme werden im folgenden näher erläutert.

2 Beispiele für Dosis-Wirkungs-Studien

In diesem Abschnitt werden zwei Beispiele für Dosis-Wirkungs-Studien beschrieben, welche den Anwendungsbereich der hier dargestellten Verfahren umreißen.

Der dritte der vorgelegten Artikel (3) untersucht eine placebo-kontrollierte, multizentrische Kombinationsstudie hinsichtlich der Wirksamkeit und Sicherheit von Simvastatin und Colesevelam. Patienten mit Hypercholesterinämie wurden randomisiert einer von vier Behandlungen zugeordnet: Placebo, Simvastatin 10 mg, Simvastatin 20 mg oder Simvastatin 20 mg mit Colesevelam 2.3 g. Der LDL Cho-

Behandlung	Stichproben-umfang	Arithmetischer Mittelwert (LDL Cholesterin mg/dl)	Stichproben-varianz	Anteil an Patienten mit Nebenwirkungen in %
Placebo	33	177	30	71
Simvastatin 10 mg	35	136	31	61
Simvastatin 20 mg	39	119	26	59
Simvastatin 20 mg und Colesevelam 2.3 g	37	111	37	68

Tabelle 1: Einfluß verschiedener Simvastatintherapien auf den LDL Cholesterinspiegel und auf den prozentualen Anteil an Patienten mit Nebenwirkungen.

lesterinspiegel wurde als primäre Variable für den Wirksamkeitsnachweis ausgewählt. Für die Sicherheitsauswertung wurde der prozentuale Anteil an Patienten mit Nebenwirkungen festgestellt. Eine Zusammenfassung der Daten ist in Tabelle 1 gegeben.

Beachtenswert an diesem Beispiel ist zunächst, daß die Behandlungen logisch geordnet sind: Es kann vorausgesetzt werden, daß die Kombinationstherapie mindestens so wirksam ist wie die Einzeltherapien, welche wiederum nicht schlechter als Placebo wirken. Somit kann von einem monotonen Wirkungsverlauf ausgegangen werden. Dies wird auch anhand der arithmetischen Mittelwerte ersichtlich. Mit zunehmend höherer Dosierung von Simvastatin nimmt der LDL Cholesterinspiegel monoton ab. Die wirksamste Behandlung ist demnach die Kombinationstherapie. Allerdings scheint der prozentuale Anteil an Patienten mit Nebenwirkungen bei der Kombinationstherapie höher zu liegen als bei den beiden Einzeltherapien jeweils nur mit Simvastatin. Insbesondere ist die Dosis-Sicherheits-Kurve nicht mehr monoton, sondern folgt einem unimodalem Verlauf: Der Anteil an Patienten mit Nebenwirkungen fällt zunächst für die Simvastatintherapien, steigt dann aber für die Kombinationstherapie wieder auf das ursprüngliche Niveau der Placebogruppe an.

Aus dieser Studie ergeben sich mehrere Fragestellungen. Das primäre Ziel liegt in der Identifizierung therapeutisch geeigneter Behandlungen. Eine zweckmäßige Behandlung hat sowohl wirksam als auch sicher zu sein und soll somit eine relevante Senkung des LDL Cholesterinspiegels im Vergleich zu Placebo bewirken, ohne den prozentualen Anteil an Patienten mit Nebenwirkungen bedenklich zu erhöhen. Dabei stellt sich die Frage, wie die Vorinformation eines monotonen

Dosierung	Anzahl Revertanten	Arithmetischer Mittelwert
0	23, 22, 14	19.7
100	27, 23, 21	23.7
333	28, 37, 35	33.3
1000	41, 37, 43	40.3
3333	28, 21, 30	26.3
10000	16, 19, 13	16.0

Tabelle 2: Beispieldaten eines Ames Mutagenitätsassays.

Wirkungsverlaufes statistisch geeignet berücksichtigt bzw. wie mit dem nicht-monotonen Verlauf des Sicherheitsparameters umgegangen werden kann. Schließlich stellt sich noch das Problem nach einer adäquaten Festlegung der Sicherheits- bzw. der Relevanzgrenzen. Eine Angabe in absoluten Einheiten (etwa mg/dl Cholesterin) ist klinisch häufig schwer zu treffen. Ein alternative Möglichkeit beruht auf der relativen Angabe dieser Schranken zum Verhältnis der Mittelwerte der jeweiligen Behandlung oder Dosierung zu dem der Placebogruppe.

Auf diese Aspekte wird in den ersten drei der vorgelegten Artikeln (1) – (3) eingegangen. In der nachfolgenden Darstellung wird an geeigneter Stelle jeweils der Bezug zum obigen Beispiel hergestellt. Für die Auswertung dieses Beispiels mit einigen der hier besprochenen Verfahren sei auf (3) verwiesen. Zuletzt sei bemerkt, daß modellbasierte Verfahren im obigen Beispiel nicht angewendet werden können, weil die Behandlung lediglich als qualitativer Faktor vorliegt.

Ein Beispiel für Dosis-Wirkungs-Studien aus dem prä-klinischen Anwendungsbereich ist Artikel (4) entnommen. Tabelle 2 gibt die Anzahl der Revertanten in

einem Ames Assay wieder. Der Ames Assay dient dem Nachweis von Punktmutationen in *Salmonella typhimurium* bei Zugabe von mutagenen Agentien (his^--Mutanten revertieren zu his^+). Die Zahl der Revertanten ist in diesem Zusammenhang ein Maß für die potentielle Mutagenität des untersuchten Agens bei verschiedenen Dosierungen.

Tabelle 2 belegt, daß gerade bei diesen Anwendungen nicht-monotone Dosis-Wirkungs-Verläufe regelmäßig auftreten können. Die Anwendung von Regressionsverfahren in diesem Beispiel wird einerseits durch die polychotome Natur der Zähldaten limitiert. Andererseits ist der untersuchte Dosierungsbereich zu groß für eine effiziente modellbasierte Auswertung der Daten. Multiple Testprozeduren bieten sich hier an, da sie die Information eines potentiellen unimodalen Dosis-Wirkungs-Verlaufes mit in Betracht ziehen können und jede Einzeldosierung als Ausprägung eines qualitativen Faktors auffassen. Eine Auswertung des obigen Datenbeispiels findet sich in (4).

3 Kontrasttests für normalverteilte Daten

In diesem Abschnitt werden multiple Kontrasttests für normalverteilte Daten im Rahmen von Dosis-Wirkungs-Studien vorgestellt. Dabei sei k die Anzahl der zu untersuchenden Dosierungen $d_1 < \ldots < d_k$, wobei der Index 1 häufig eine Placebogruppe bezeichnet. Der Einfachheit halber beschränkt sich die nachfolgende Darstellung auf balanzierte einfaktorielle Anlagen der Form

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \ldots, k, j = 1, \ldots, n,$$

wobei Y_{ij} die j-te Beobachtung in der i-ten Dosisgruppe, μ der Gesamtmittelwert, α_i der Effekt (i.a. die Wirksamkeit) der i-ten Dosierung und n der Stichprobenumfang je Gruppe ist. Die Residualfehler $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ seien unabhängig normal verteilt mit Erwartungswert 0 und Varianz σ^2. Verallgemeinerungen der nachfolgenden Ergebnisse auf unbalanzierte allgemeine lineare Modelle, welche auch Kovariablen und faktorielle Designs erlauben, sind beispielsweise in Bretz *et al.* (2002) beschrieben.

Die zu testende Nullhypothese lautet $H_0 : \alpha_1 = \ldots = \alpha_k$, d.h., die Dosierungen unterscheiden sich nicht voneinander in ihrer Wirksamkeit. In herkömmlichen Mehrstichprobenproblemen, zu denen keine Vorinformationen über die zu vergleichenden Stichproben vorliegen, wird auf die Omnibusalternative beliebiger Stichprobenunterschiede getestet (F-Test, Tukey-Test, usw.). Im Gegensatz dazu liegen bei Dosis-Wirkungs-Studien möglicherweise relevante Vorinformationen vor: Häufig kann davon ausgegangen werden, daß mit steigender Dosierung die Wirksamkeit ebenfalls ansteigt. Unter dieser Monotonievoraussetzung kann die klassische Alternativhypothese eines beliebigen Unterschiedes auf die Alternative $H_A : \alpha_1 \leq \ldots \leq \alpha_k, \alpha_1 < \alpha_k$, eingeschränkt werden. Spezielle *Trendtests* wurden somit entwickelt, die diese Restriktion des Alternativraumes mit in Betracht ziehen und somit mächtiger als herkömmliche Omnibustests sind (Robertson *et al.*, 1988; Chuang-Stein und Agresti, 1997). Bekannte Trendtests sind vor allem der Likelihood-Quotienten-Test von Bartholomew (1961) sowie der Trendtest nach Williams (1971). Beide Verfahren sind jedoch aufgrund der bis heute noch ungelösten numerischen Probleme in der Berechnung der jeweiligen Verteilungsfunktionen *in praxi* auf balanzierte, zumeist einfaktorielle Versuchsanlagen beschränkt. Des weiteren sind für beide Verfahren keine Konfidenzintervalle oder

geschlossene Güteformeln verfügbar. Aus diesen Gründen werden in zunehmenden Maße Kontrasttests untersucht, welche sich als vielseitig einsetzbar erwiesen und die Nachteile der vorhergehenden Tests überwinden.

Einzelkontrasttests sind – gerade auch im Rahmen von Dosis-Wirkungs-Problemen – seit langem bekannt (Abelson und Tukey, 1963; Schaafsma und Smid, 1966; Ruberg, 1989; Tamhane *et al.*, 1996). Aber erst mit der systematischen Untersuchung des multiplen Analogons, den multiplen Kontrasttests, werden diese Verfahren auch auf praktische Probleme angewandt (Mukerjee *et al.*, 1987; Hothorn *et al.*, 1997; Stewart und Ruberg, 2000). Ein Einzelkontrast $T(\mathbf{c})$ wird definiert durch

$$T(\mathbf{c}) = \sqrt{n}\frac{\sum_{i=1}^{k} c_i \overline{Y}_i}{\widehat{\sigma}\sqrt{\sum_{i=1}^{k} c_i^2}} \sim t_\nu,$$

wobei $\overline{Y}_i = n^{-1}\sum_{j=1}^{n} Y_{ij}$ der arithmetische Mittelwert der i-ten Gruppe und $\widehat{\sigma}^2 = \nu^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij} - \overline{Y})^2$ der gepoolte Varianzschätzer mit $\nu = k(n-1)$ Freiheitsgraden sind. Die Gewichte c_i – in Vektorschreibweise entsprechend $\mathbf{c}' = (c_1, \ldots, c_k)$ – werden als *Kontrastkoeffizienten* bezeichnet und sind unter der Bedingung $\sum_i c_i = 0$ frei wählbar. Die Teststatistik $T(\mathbf{c})$ ist als Quotient einer $\mathcal{N}(0,1)$ und einer $\sqrt{\chi^2_\nu}$ Variablen definitionsgemäß univariat t_ν verteilt.

Die Flexibilität der Einzelkontrasttests zeigt sich in der freien Bestimmung der c_i. Im folgenden sei eine spezielle Wahl der Kontrastkoeffizienten jeweils durch ein "$*$" gekennzeichnet. Wird beispielsweise $\mathbf{c}_1^* = (-1, 1, 0, 0)'$ gesetzt, d.h. $c_1 = -1, c_2 = 1, c_3 = c_4 = 0$ für $k = 4$, reduziert sich obige Teststatistik auf den Zweistichprobenvergleich $T(\mathbf{c}_1^*) = \sqrt{n}(\overline{Y}_2 - \overline{Y}_1)\Big/\sqrt{2\widehat{\sigma}^2}$. Analog zeigt sich, daß die Wahl $\mathbf{c}_2^* = (-3, -1, 1, 3)'$ zu einem linearen Vergleich der Dosierungen

führt, d.h., der Test ist besonders empfindlich, wenn der Effektunterschied zweier benachbarter Dosierungen konstant bleibt. In gleicher Weise ist ersichtlich, daß $\mathbf{c}_3^* = (-3, 1, 1, 1)'$ einem konkaven und $\mathbf{c}_4^* = (-1, -1, -1, 3)'$ einem konvexen Dosis-Wirkungs-Verlauf entsprechen.

Somit bieten sich dem Anwender mannigfaltige Möglichkeiten zur Formulierung der letztlich zu verwendenden Teststatistik. Wenn beispielsweise die wahre Dosis-Wirkungs-Kurve für eine beliebige Konstante c einem linearen Verlauf mit $\alpha_{i+1} - \alpha_i = c, i = 1, \ldots, k-1$, entspricht, ist der *lineare Kontrast* $T(\mathbf{c}_2^*)$ optimal. Wird jedoch (unwissentlich) zu hoch dosiert, so daß das Sättigungsniveau frühzeitig erricht ist, entspricht die wahre Dosis-Wirkungs-Kurve einem konkaven Profil und der Kontrasttest $T(\mathbf{c}_3^*)$ wäre optimal. Allerdings ist die Wahl der Kontrastkoeffizienten vor Studienbeginn zu treffen, zu einem Zeitpunkt also, wenn der wahre Kurvenverlauf noch unbekannt ist. Die Güte eines Einzelkontrasttests hängt wiederum stark von der korrekten Wahl seiner Gewichte c_i ab. Im Falle einer fälschlichen Spezifikation des wahren Profils durch die Gewichte c_i kann die Wahrscheinlichkeit, einen vorhandenen Trend auch als solchen zu entdecken, erheblich sinken. Neuhäuser (1998) hat beispielsweise anhand der asymptotischen relativen Effizienz nach Pitman gezeigt, daß unter gewissen Regularitätsbedingungen ein auf $\mathbf{c}_3^*$ basierender Kontrasttest den neunfachen Stichprobenumfang wie ein auf $\mathbf{c}_4^*$ beruhender Test benötigt, wenn tatsächlich ein konvexes Profil vorliegt.

Das Problem *a-priori* unbekannter Dosis-Wirkungs-Profile wird auch am Hypercholesterinämiebeispiel im zweiten Abschnitt deutlich. Während der LDL Cholesterinspiegel einem monotonen, jedoch nicht linearen Kurvenverlauf folgt, gilt

dies nicht für den Sicherheitsparameter. Indessen ist im allgemeinen davon auszugehen, daß die wahren Profile, zumal für mehr als einen Parameter, vor Beginn der Studie unbekannt sind. Trendtests jedoch, die mit hoher Wahrscheinlichkeit einen vorhandenen Dosiseffekt entdecken sollen, müssen robust gegenüber unterschiedliche Dosis-Wirkungs-Profile sein.

Eine Möglichkeit zur Abschwächung dieser Profilabhängigkeit besteht in der Kombination mehrerer Einzelkontrasttests zu einer gemeinsamen Teststatistik. Dadurch wird erreicht, daß die Kombinationsstatistik die maximale Information aus den Einzelstatistiken erhält und somit selber robust gegenüber eine ungünstige Wahl der Kontrastkoeffizienten ist. Aus den unterschiedlichen Kombinationsmöglichkeiten bieten sich Maximumtests wegen der vereinfachten *post-hoc* Interpretation an (*diagnostic property*, s. Cox, 1977). Seien q Einzelkontrasttests $T(\mathbf{c}_1), \ldots, T(\mathbf{c}_q)$ gegeben mit geeigneten Kontrastvektoren $\mathbf{c}_l$, $l = 1, \ldots, q$. Dann wird mit $T_{\max}(\mathbf{c}_1, \ldots, \mathbf{c}_q) = \max\{T(\mathbf{c}_1), \ldots, T(\mathbf{c}_q)\}$ ein multipler Kontrasttest definiert. Der Vorteil dieser Vorgehensweise ist offensichtlich: Indem mehrere Kontraste (d.h. mehrere potentielle Dosis-Wirkungs-Profile) gleichzeitig berücksichtigt werden, nimmt das Maximum den Wert der besten Statistik an und reduziert dadurch die Gefahr einer Mißspezifikation. Bretz *et al.* (2001) zeigen, daß $T(\mathbf{c}_1), \ldots, T(\mathbf{c}_q)$ gemeinsam multivariat t verteilt sind unter Einbeziehung der Korrelation zwischen den Teststatistiken. Effiziente hochdimensionale Integrationverfahren zur Berechnung der Verteilungsfunktion werden von Genz und Bretz (2002) beschrieben. Folglich ist die numerische Verfügbarkeit multipler Kontrasttests sichergestellt, und P-Werte, Quantile, Gütewerte und somit Fallzahlen können mit gängigen Softwareimplementierungen berechnet werden.

Die vorgestellten Kontrasttests verdeutlichen einen wesentlichen Vorteil multipler Testprozeduren gegenüber modellbasierten Auswertungsverfahren. Die Eignung sowohl herkömmlicher Regressionsverfahren als auch von Einzelkontrasttests hängt maßgeblich von der korrekten Modellierung des Kurvenverlaufes ab. Multiple Kontrasttests überwinden dieses Problem bei anhaltender Kontrolle der Irrtumswahrscheinlichkeit. Wie eingangs dieser Arbeit ausgeführt, steht ein entsprechender Lösungsansatz für Regressionsverfahren allerdings noch aus. Selbstverständlich müssen multiple Kontrasttests für diesen Vorteil einen Preis in Form einer Multiplizitätsadjustierung bezahlen. Durch die hohe Korrelation zwischen den Einzelstatistiken und die Verwendung der exakten multivariaten t Verteilung fällt die Strafe jedoch deutlich geringer aus als eine krude Bonferroniadjustierung es vermuten läßt. Auf diese Gesichtspunkte gehen Bretz *et al.* (2001) näher ein.

4 Nicht-monotone Dosis-Wirkungs-Verläufe

Ein wesentlicher Aspekt von $T_{\max}(\mathbf{c}_1, \ldots, \mathbf{c}_q)$ ist die adäquate Wahl der Kontrastkoeffizienten. Wie bereits beschrieben, hat diese einen erheblichen Einfluß auf die Güte der sich ergebenden Tests. Auf diese Problematik geht der erste der vorgelegten Artikel (1) detailliert ein. Insbesondere wird die Diskussion um den Gesichtspunkt möglicher nicht-monotoner Kurvenverläufe erweitert. Die *in praxi* häufigste Verletzung der Ordnungsannahme H_A tritt mit einer Umkehr der Wirksamkeit bei hohen Dosierungen, beispielsweise durch sicherheitsauffällige Effekte bedingt, ein (*umbrella shape*). Derartige Verletzungen der Monotonieannahme verringern merklich die Güte eines jeden Trendtests, da diese speziell zum Aufdecken von

monotonen Effektanstiegen entwickelt wurden. Selbst im Falle einer statistischen Signifikanz bleibt die Interpretation der Resultate schwierig. Streng genommen darf nach Ablehnung der Nullhypothese H_0 lediglich die Alternative H_A, d.h. ein monotoner Anstieg der Wirksamkeit in den Dosierungen, gefolgert werden – im Widerspruch zu einem möglicherweise ausgeprägten Effektabfall im oberen Dosierungsbereich. Aber auch bei fehlender Signifikanz, wenn im Extremfall der Effektabfall zu groß ist, ergibt sich ein Widerspruch zwischen dem statistisch nicht nachgewiesenem Dosiseffekt und einer deutlich sichtbaren Dosis-Wirkungs-Beziehung. Spezielle *Umbrellatests* werden daraufhin in der Literatur vorgeschlagen, wenn vor Studienbeginn die Monotonieannahme H_A nicht vorausgesetzt werden kann. Typischerweise wird in diesem Zusammenhang die monotone Trendalternative H_A ersetzt durch die eingeschränkte Alternative $H_A' : \alpha_1 \leq \ldots \leq \alpha_h \geq \ldots \geq \alpha_k$ – mit mindestens einer echten Ungleichung – wobei der Index h den (unbekannten) Umkehrpunkt der Wirksamkeit bezeichnet. Derartige Testprobleme finden sich nicht nur in klinischen sondern auch in prä-klinischen Studien wieder, worauf insbesondere (4) näher eingeht. Beide im Abschnitt 2 vorgestellten Beispiele weisen nicht-monotone Verläufe auf und unterstreichen die Notwendigkeit, geeignete Testverfahren für diese Anwendungen zu entwickeln.

In (1) wird der obige Sachverhalt ausführlich diskutiert, und Schwachpunkte in der traditionellen Formulierung H_A' werden dargestellt. Ausgehend von der Beobachtung, daß der abfallende Teil der Dosis-Wirkungs-Kurve nicht von eigentlichem Interesse ist, regt der Artikel eine adäquate Neuformulierung der Alternativhypothese an. Um die Güte zu erhöhen, einen ansteigenden Trend bis zum Umkehrpunkt zu entdecken, wird die Alternative $H_A'' : \alpha_1 \leq \ldots \leq \alpha_h, \alpha_1 < \alpha_h$, vorgeschlagen. Somit wird lediglich auf Effektanstiege bis zum Umkehrpunkt ge-

testet und der anschließende Kurvenverlauf bleibt unberücksichtigt, weil nicht relevant. In diesem Kontext werden in (1) mehrere multiple Kontrasttests mit unterschiedlichen Zielsetzungen für die Testprobleme H'_A, H''_A und H_A hergeleitet (letztere Alternative ergibt sich unmittelbar als Spezialfall aus einer der beiden vorherigen Alternativen, wenn $h = k$ als bekannt vorausgesetzt wird). In einer ausführlichen Simulationsstudie werden für verschiedene Parameterbedingungen die Güten der verschiedenen multiplen Kontrasttests untereinander sowie mit zwei aus der Literatur bekannten Verfahren (der Umbrellatest nach Simpson und Margolin, 1986, sowie der many-to-one Test nach Dunnett, 1955) verglichen. Die folgenden Schlußfolgerungen können gezogen werden. (i) Der Dunnett-Test hat sowohl für monotone als auch für die untersuchten nicht-monotonen Profile eine deutlich geringere Güte als die konkurrierenden Verfahren. (ii) Das nichtparametrische Simpson-Margolin-Verfahren ist schlechter als die parametrischen Kontrasttests, wobei ein Vergleich mit den nichtparametrischen Rangversionen der Kontrasttests noch aussteht. (iii) Auf der Restriktion H''_A basierende Kontrasttests sind zumeist mächtiger als diejenigen, die auf H'_A beruhen. Lediglich, wenn die Effektumkehr bereits bei niedrigen Dosierungen stattfindet, sind die letzteren Tests vergleichbar mächtig. (iv) Die Wahl optimaler Kontrastkoeffizienten ist nicht eindeutig zu beantworten. Zunächst hängt die Bestimmung maßgeblich von der unbekannten, situationsspezifischen Parameterkonstellation ab, weil jeder Kontrasttest seine ihm eigene Gütefunktion aufweist. Einzelkontrasttests sollten aber vermieden werden, sie sind im Durchschnitt deutlich schlechter als die multiplen Varianten. Weitere Faktoren, wie beispielsweise die Interpretierbarkeit der Kontraste und die Verfügbarkeit schrittweiser simultaner Konfidenzintervalle, sind bei der Auswahl der Koeffizienten ebenfalls mit zu berücksichtigen.

5 Kontrasttests für andere Verteilungsannahmen

Der zweite der vorgelegten Artikel (2) führt multiple Kontrasttests für binomialverteilte Daten ein und leitet asymptotische Güte- bzw. Fallzahlformeln her. Die erzielten Ergebnisse sind eine direkte Verallgemeinerung der obigen Betrachtungen und erweitern den Anwendungsbereich der Kontrasttests auf die Planung und Auswertung klinischer Studien mit dichotomen Endpunkten wie beispielsweise Responseraten. Ein klassisches Anwendungsbeispiel ergibt sich aus der Fragestellung, ob der Anteil an Patienten mit Nebenwirkungen mit höherer Dosierung zunimmt. Zur Vereinfachung der Notation sei wieder der gleiche Stichprobenumfang n je Dosisgruppe vorausgesetzt. Seien zu diesem Zweck $X_1, \ldots, X_k$ unabhängig binomialverteilte Zufallsvariablen mit Erfolgswahrscheinlichkeit π_i in Gruppe $i = 1, \ldots, k$, d.h. $X_i \sim \text{Bin}(n, \pi_i)$. Das Testproblem lautet in diesem Fall $\widetilde{H}_0 : \pi_1 = \ldots = \pi_k$, d.h. Gleichheit aller Erfolgswahrscheinlichkeiten, gegen eine geordnete Alternative, beispielsweise $\widetilde{H}_A : \pi_1 \leq \ldots \leq \pi_k, \pi_1 < \pi_k$.

Die Teststatistik eines binomialen Einzelkontrasttests ist definiert als

$$\widetilde{L}(\mathbf{c}) = \sum_{i=1}^{k} c_i \widehat{\pi}_i,$$

wobei $\widehat{\pi}_i = X_i/n$ die empirischen Häufigkeiten bezeichnen. Die c_i sind wiederum Kontrastkoeffizienten, die ähnlich wie zuvor im Normalmodell bei geeigneter Wahl verschiedene Dosis-Wirkungs-Verläufe abbilden. Die Teststatistik $\widetilde{L}(\mathbf{c})$ kann entweder mittels geeigneter Permutationsverfahren (Pesarin, 2001) oder, wie in (2) geschehen, asymptotisch ausgewertet werden. Letzteres hat den Vorteil, daß geschlossene Güteformeln existieren und dieser Ansatz auch bei umfangreichen Studien numerisch verfügbar ist. Sei hierzu, unter $\widetilde{H}_0$, nach entsprechender Stan-

dardisierung ein asymptotischer Einzelkontrasttest gegeben durch

$$\widetilde{T}(\mathbf{c}) = \frac{\widetilde{L}(\mathbf{c})}{\sqrt{\widehat{\pi}(1-\widehat{\pi}) \sum_{i=1}^{k} c_i^2}},$$

wobei $\widehat{\pi} = \sum_{i=1}^{k} X_i/(kn)$ der über die Dosisgruppen gepoolte Gesamtschätzer ist. Unter der Alternative $\widetilde{H}_A$ gilt eine entsprechende Darstellung unter Beachtung der Einzelhäufigkeiten $\widehat{\pi}_i$. Basierend auf diesen Statistiken werden in (2) geschlossene Ausdrücke zur Bestimmung des notwendigen Stichprobenumfanges n für ein vorgegebenes Dosis-Wirkungs-Profil unter Einhaltung der vor der Studie festzusetzenden Fehlerraten 1. bzw. 2. Art hergeleitet.

Auf analoge Weise lassen sich multiple Kontrasttests im oben skizzierten Rahmen durch simultane Berücksichtigung von q Einzelkontrasttests $\widetilde{T}(\mathbf{c}_1), \ldots, \widetilde{T}(\mathbf{c}_q)$ einführen. Es läßt sich zeigen (2), daß $\widetilde{T}_{\max}(\mathbf{c}_1, \ldots, \mathbf{c}_q) = \max\{\widetilde{T}(\mathbf{c}_1), \ldots, \widetilde{T}(\mathbf{c}_q)\}$ asymptotisch multivariat normal verteilt ist mit einer gegebenen Korrelationsmatrix, die die stochastischen Abhängigkeiten zwischen den Statistiken beschreibt. Allerdings existiert keine Fallzahlformel in geschlossener Form, weil die für deren Herleitung benötigte Umkehrfunktion der multivariaten Normalverteilung im mehrdimensionalen Fall $q > 1$ nicht eindeutig definiert ist. Eine iterative Lösung mittels der in (2) angegebenen Güteformel ist jedoch möglich, indem der Stichprobenumfang durch "Versuch und Irrtum" solange iteriert wird, bis die gewünschte Güte erreicht wird. Hierzu können die numerischen Integrationsverfahren von Genz und Bretz (2002) benutzt werden. In diesem Zusammenhang beschreiben Hothorn und Bretz (2000) eine Simulationsstudie zum Gütevergleich von Kontrasttests für typische Designs in *in vivo* Kanzerogenitätsstudien.

In diesem Abschnitt wurden bisher Kontrasttests für dichotome Daten vorgestellt.

In vielen prä-klinischen Dosis-Wirkungs-Studien, wie beispielsweise den Mutagenitätsassays (siehe Abschnitt 2 für ein Datenbeispiel), sind die Daten polychotom (d.h. Zähldaten mit diskreten Werten) oder unterliegen einem heteroskedastischem Modell ungleicher Varianzen je Dosisgruppe. Spezielle Trendtests werden hierfür benötigt, worauf in (4) anhand konkreter Beispiele eingegangen wird. Im heteroskedastischen Fall wird der gepoolte Varianzschätzer durch die einzelnen Gruppenvarianzen ersetzt. Anpassungen der im Zweistichprobenfall bekannten Verfahren auf die vorliegende Mehrstichprobensituation (beispielsweise die Adjustierung der Freiheitsgrade nach Satterthwaite oder die gewichtete Mittelung der Einzelquantile nach Cochran) garantieren eine approximative Kontrolle der Irrtumswahrscheinlichkeit. Im zweiten Fall, wenn keine parametrischen Verteilungsannahmen getroffen werden, können lineare Rangstatistiken als Erweiterung der multiplen Kontrasttests eingeführt werden, so daß die bisherigen Konzepte auch hier beibehalten werden. Details und weiterführende Literatur hierzu sind in (4) angegeben.

6 Simultane Konfidenzintervalle

Im dritten der vier eingangs erwähnten Artikel (3) werden weiterführende Gesichtspunkte multipler Testprozeduren im Rahmen von Dosis-Wirkungs-Studien untersucht. Die wichtigsten Ergebnisse betreffen: (i) Schätzung des therapeutischen Fensters, (ii) Herleitung simultaner Konfidenzintervalle für schrittweise Hypothesentests und (iii) Untersuchung multipler Quotiententests. Im folgenden werden diese drei Aspekte näher erläutert.

Die Schätzung relevanter Dosierungen in einem eigenen, vom PoC-Nachweis entkoppelten Dosis-Findungs-Schritt ist ein wesentlicher Bestandteil klinischer Dosis-Wirkungs-Studien. Eine der wichtigsten zu schätzenden Dosierungen ist die minimal effektive Dosis (MED). Üblicherweise wird das *Abschlußprinzip* (Marcus *et al.*, 1976; Maurer *et al.*, 1995) als Begründung zur folgenden schrittweisen Testprozedur herangezogen. Im ersten Schritt wird zunächst auf Trend unter Einbeziehung aller Dosierungen getestet. Wenn dieser Test nicht signifikant ist, wird keine Dosierung als effektiv eingeschätzt, da bereits der globale Wirksamkeitsnachweis fehlt. Andernfalls wird im zweiten Schritt unter Auslassung der höchsten Dosierung auf Trend unter Einbeziehung der verbleibenden Dosierungen getestet. Ist dieser Test nicht signifikant, wird die höchste Dosierung als MED geschätzt und die Prozedur endet. Andernfalls wird im dritten Schritt unter Auslassung der beiden höchsten Dosierungen auf Trend getestet. Die Prozedur wird schrittweise jeweils unter Auslassung der höheren, signifikanten Dosierungen fortgeführt, bis entweder ein nicht-signifikantes Resultat erzielt wird (die zuletzt signifikant getestete Dosierung wird als MED geschätzt) oder alle Dosierungen effektiv sind (dann wird die niedrigste Dosierung als MED geschätzt). Auf jeder Stufe darf ein (Trend-)Test zum Niveau α durchgeführt werden, ohne daß der Gesamtfehler 1. Art α übersteigt. Diese Prozedur, erstmalig von Williams (1971) im Rahmen von Dosis-Wirkungs-Studien beschrieben, wird *in praxi* häufig aufgrund seiner einfachen Struktur sowie seiner hohen Güte (alle Einzeltests zum Niveau α) angewandt. Kontrasttests können erfolgreich in diese Prozedur eingebunden werden. Tamhane *et al.* (1996) beispielsweise vergleichen das Verhalten mehrerer Einzelkontrasttests miteinander für die obige und verwandte Testprozeduren. Die Verwendung multipler Kontrasttests wird beispielsweise in (1) untersucht.

Ein wesentliches Problem der obigen Prozedur sind fehlende Konfidenzintervalle für die relevanten Parameter. Häufig sind allerdings weitere Informationen zu dem Ausmaß der untersuchten Effekte von Interesse als lediglich ja/nein Entscheidungen zur Signifikanz. In (3) wird dieses Problem im breiteren Kontext der Schätzung des therapeutischen Fensters untersucht. Über die Schätzung der MED hinaus wird gleichzeitig eine maximal sichere Dosis (MSD) angestrebt. Zwischen der MED und der MSD liegende Dosierungen sind dann wirksam und sicher zugleich und stehen somit einem potentiellen therapeutischen Einsatz zur Verfügung. Um für beide Testprobleme (MED und MSD) korrekte schrittweise Konfidenzintervalle für die jeweilig relevanten Parameter herzuleiten, wird an Stelle des Abschlußprinzips das *Partitionsprinzip* benutzt (Finner und Straßburger, 2002). Dieses ist eine Verallgemeinerung des Abschlußprinzips und führt für die vorliegende Anwendung letztlich zur selben Entscheidungsprozedur, stellt aber Konfidenzintervalle zur Verfügung. Details zu diesem Verfahren sind in (3) gegeben.

Hsu und Berger (1999) bereits leiten in einer auf das MED Problem beschränkten Arbeit schrittweise Konfidenzintervalle für paarweise Kontrasttests ab (jeweils für den Vergleich Dosierung gegen Placebo). Die Autoren beziehen sich allerdings in einer für den Leser nicht ersichtlichen Weise auf das Partitionsprinzip. Dies wird in (3) nachgeholt und die notwendigen Voraussetzungen und Verbindungen werden klar dargelegt. Des weiteren werden die Testprobleme als multiple Quotiententests verallgemeinert, wenn klinische Relevanzschranken prozentual zum Verhältnis zweier Mittelwerte angegeben werden. Die Motivation dieser Umformulierung der klassischen auf Mittelwertdifferenzen basierenden Statistiken liegt in der häufig schwierigen Definition und Interpretation der Relevanzschranken in

absoluten Einheiten (Hauschke *et al.*, 1999). Beispielsweise kann es im Hypercholesterinämiebeispiel aus Abschnitt 2 schwierig sein, klinische Relevanzschranken im absoluten Maßstab mg/dl Cholesterin festzulegen. Eine prozentuale Festlegung erscheint intuitiv leichter zugänglich, beispielsweise in der Form "die Kombinationsbehandlung wirkt mindestens 10% besser als Placebo". Unter Verwendung von Fiellers Theorem zur Konstruktion der marginalen Konfidenzintervalle (d.h. bevor diese mit dem Partitionsprinzip zu simultanen schrittweisen Konfidenzintervallen verallgemeinert werden) sind in (3) Details sowohl für Quotienten normalverteilter Mittelwerte als auch für Quotienten von Erfolgswahrscheinlichkeiten gegeben. Insbesondere zeigt sich, daß diese Quotiententests wiederum als Spezialfälle von Kontrasttests angesehen werden können.

7 Ausblick

Die vorliegenden Artikel zeigen, daß sowohl prä-klinische als auch klinische Dosis-Wirkungs-Studien effizient unter Verwendung geeigneter multipler Testprozeduren ausgewertet werden können. Insbesondere multiple Kontrasttests erlauben eine hohe Flexibilität an Design und Datenkondition. Hierfür wurden erfolgreich untersucht: (i) geschlossene Güte- und Fallzahlformeln, (ii) Einbettung der Verfahren in die Theorie der linearen Modelle (so daß unbalanzierte faktorielle Anlagen mit etwaigen Kovariablen verwendet werden können), (iii) unterschiedliche Anpassungen der Kontrasttests sowohl für parametrische (normalverteilte und binomialverteilte Daten) als auch für nichtparametrische Modelle (verteilungsunabhängige, rangbasierte Kontrasttests) und (iv) verfügbare Softwareimplemen-

tierungen, so daß wesentliche statistische Größen wie $P-$Werte, Quantile, Güten und Stichprobenumfänge problemlos berechnet werden können. Viele dieser Ergebnisse können mit konkurrierenden Verfahren nicht erzielt werden.

Während mit der Anwendung multipler Kontrasttests diese klassischen Fragestellungen weitgehend und zufriedenstellend gelöst sind, werden mit dem Artikel (3) weitere interessante Forschungsfelder erschloßen. Dies schlägt sich auch in der gegenwärtigen Forschungsarbeit des Autors nieder. Beispielsweise werden im Rahmen eines seitens der DFG unterstützten Projektes simultane Konfidenzbereiche und Konfidenzintervalle für multiple Quotienten beliebiger Kontraststatistiken untersucht (Dilba *et al.*, 2003) – ein Projekt, welches den Ergebnissen zu den paarweisen Quotienten in (3) entsprang. Eine weitere Forschungsrichtung befaßt sich mit der Untersuchung sich aus dem Partitionsprinzip ergebender Konfidenzbereiche für beliebige Kontrasttests für Mittelwertdifferenzen. Erste Resultate sind in Straßburger *et al.* (2003) zusammengefaßt. Schließlich vereinigen Branson *et al.* (2003) die beiden eingangs erwähnten unterschiedlichen statistischen Prinzipien (modellbasierte bzw. modellfreie Ansätze), indem die Auswahl eines Dosis-Wirkungs-Modells aus einer Kandidatenmenge an Regressionsmodellen als multiples Testproblem aufgefaßt wird. Die Selektion geschieht zu einer gegebenen Fehlerkontrolle, so daß anschließend die mächtigen Schätzmethoden der modellbasierten Verfahren zur Verfügung stehen.

Danksagung

Mein Dank gilt Herrn Prof. Dr. L.A. Hothorn, der meinen wissenschaftlichen Werdegang förderte, mein Interesse für das Thema weckte und mir großzügige Arbeitsmöglichkeiten gewährte. Weiterhin danke ich Herrn Prof. Dr. H. Hecker für sein Interesse an dieser Arbeit sowie vielen wertvollen Hinweisen und kritischen Diskussionen. Schließlich danke ich meiner Frau Jiamei, die häufig auf mich verzichten mußte, für ihre Verbundenheit, ihr Verständnis und ihre Unterstützung.

Literatur

Abelson, R.P. und Tukey, J.W. (1963) Efficient utilisation of non-numerical information in quantitative analysis: General theory and the case of simple order. *The Annals of Mathematical Statistics*, **34**, 1347–1369.

Bartholomew, D.J. (1961) Ordered tests in the analysis of variance. *Biometrika*, **48**, 325–332.

Branson, M., Pinheiro, J.C. und Bretz, F. (2003) Searching for an adequate dose: Combining multiple comparisons and modeling techniques in dose-response studies. Novartis Biometrics Technical Report No. 2003-08-20.

Bretz, F., Genz, A. und Hothorn, L.A. (2001) On the numerical availability of multiple comparison procedures. *Biometrical Journal*, **43**, 645–656.

Bretz, F., Hothorn, T. und Westfall, P.H. (2002) On multiple comparisons in R. *R News*, **2**, 14–17.

Chuang-Stein, C. und Agresti, A. (1997) A review of tests for detecting a monotone dose-response relationship with ordinal response data. *Statistics in Medicine*, **16**, 2599–2618.

Cox, D.R. (1977) The role of significance tests. *Scandinavian Journal of Statistics*, **4**, 49–70.

Dilba, G., Bretz, F., Guiard, V. und Hothorn, L.A. (2003) Simultaneous confidence sets and confidence intervals for multiple ratios. In Bearbeitung.

Dunnett, C.W. (1955) A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, **50**, 1096–1121.

Finner, H. und Straßburger, K. (2002) The partitioning principle: A powerful tool in multiple decision theory. *The Annals of Statistics* **30**, 1194–1213.

Genz, A. und Bretz, F. (2002) Comparison of methods for the computation of multivariate t probabilities. *Journal of Computational and Graphical Statistics*, **11**, 950–971.

Hauschke, D., Kieser, M., Diletti, E. und Burke, M. (1999) Sample size determination for proving eqivalence based on the ratio of two means for normally distributed data. *Statistics in Medicine*, **18**, 93–105.

Hothorn, L.A. and Bretz, F. (2000) Evaluation of animal carcinogenicity studies: Cochran-Armitage trend test vs. multiple contrast tests. *Biometrical Journal*, **42**, 553–567.

Hothorn, L.A., Neuhäuser, M. und Koch, H.F. (1997) Analysis of randomised dose-finding studies: closure test modifications based on multiple contrast tests. *Biometrical Journal*, **39**, 467–479.

Hsu, J.C. und Berger, R.L. (1999) Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies. *Journal of the American Statistical Association* **94**, 468–482.

Neuhäuser, M. (1998) The asymptotic relative efficiency of contrast tests. *Allgemeines Statistisches Archiv*, **82**, 243–251.

Marcus, R., Peritz, E. und Gabriel, K.B. (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–660.

Maurer, W., Hothorn, L.A. und Lehmacher, W. (1995) Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In: *Biometrie in der chemisch-pharmazeutischen Industrie, Band 6*, Vollmar, J. (Hrgb.), 3–18.. Fischer Verlag, Stuttgart.

Mukerjee, H., Roberston, T. und Wright, F.T. (1987) Comparison of several treatments with a control using multiple contrasts. *Journal of the American Statistical Association*, **82**, 902–910.

Pesarin, F. (2001) *Multivariate permutation tests*. Wiley, New York.

Robertson, T., Wright, F.T. und Dykstra, R.L. (1988) *Order restricted statistical inference*. Wiley, New York.

Ruberg, S.J. (1989) Contrasts for identifying the minimum effective dose. *Journal of the American Statistical Association*, **84**, 816–822.

Schaafsma, W. und Smid, L.J. (1966) Most stringent somewhere most powerful tests against alternatives restricted by a number of linear inequalities. *The Annals of Mathematical Statistics*, **37**, 1161–1172.

Simpson, D.G. und Margolin, B.H. (1986) Recursive nonparametric testing for dose-response relationships subject to downturns at high doses. *Biometrika*, 1986, **73**, 589–596.

Stewart, W.H. und Ruberg, S.J. (2000) Detecting doseresponse with contrasts. *Statistics in Medicine*, **19**, 913–921.

Straßburger, K., Bretz, F. und Hochberg, Y. (2003) Compatible confidence intervals for intersection union tests involving two hypotheses. Zur Publikation eingereicht.

Tamhane, A.C., Dunnett, C.W. und Hochberg, Y. (1996) Multiple test procedures for dose finding. *Biometrics*, **52**, 21–37.

Williams, D.A. (1971) A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, **27**, 103–117.