FORSCHUNGSBERICHTE DER ABTEILUNG MEDIZINISCHE BIOMETRIE, UNIVERSITÄT HEIDELBERG

Nr. 39

METHODEN ZUR ENTSCHEIDUNGSUNTERSTÜTZUNG IN KLINISCHEN STUDIEN MIT ADAPTIVEM DESIGN

Mai 2002

RUPRECHT-KARLS-UNIVERSITÄT HEIDELBERG

INSTITUT FÜR MEDIZINISCHE BIOMETRIE UND INFORMATIK

Forschungsberichte der
Abteilung Medizinische Biometrie, Universität Heidelberg

Nr. 39

METHODEN ZUR ENTSCHEIDUNGSUNTERSTÜTZUNG IN KLINISCHEN STUDIEN MIT ADAPTIVEM DESIGN

MEINHARD KIESER

Institut für Medizinische Biometrie und Informatik (IMBI)
der Medizinischen Fakultät der Universität Heidelberg

Heidelberg, Mai 2002

Impressum:
Reihentitel: Forschungsberichte der Abteilung Medizinische Biometrie, Universität Heidelberg
Herausgeber: Prof. Dr. Norbert Victor
Anschrift: Im Neuenheimer Feld 305, 69120 Heidelberg
Druck: Hausdruckerei der Ruprecht-Karls-Universität Heidelberg
elektronischer Bezug: http://www.biometrie.uni-heidelberg.de
ISSN 1619-5833

Institut für Medizinische Biometrie und Informatik
Abteilung Medizinische Biometrie
(Direktor: Prof. Dr. N. Victor)

Methoden zur Entscheidungsunterstützung in klinischen Studien mit adaptivem Design

Habilitationsschrift
zur Erlangung der venia legendi
für das Fach Medizinische Biometrie
der Hohen Medizinischen Fakultät Heidelberg
der Ruprecht-Karls-Universität

vorgelegt von
Dr. Meinhard Kieser

aus
Buchen

2000

Inhaltsverzeichnis

Mediterranean climates have winter rains and summer droughts. To survive these annual droughts, plants have developed ways of retaining moisture over long periods of time. These are called adaptations.
Erläuterung auf einer Schautafel im Botanischen Garten des Golden Gate Parks, San Francisco

1. Einleitung und Übersicht

1.1 Einführung in die Thematik

Bei der Entwicklung und Verbesserung der Methodik klinischer Studien bilden häufig ethische Überlegungen den Ausgangspunkt. Ein wichtiges Beispiel hierfür ist das folgende Dilemma: Auf der einen Seite muss eine ausreichende Zahl von Patienten in eine Studie eingeschlossen werden, um die Effektivität und Verträglichkeit einer neuen Therapie mit ausreichender Sicherheit belegen zu können. Andererseits ist es geboten, die Patientenzahl so niedrig wie möglich zu halten, um Studienteilnehmer nicht unnötigerweise mit einer unverträglichen, unterlegenen oder unwirksamen Therapie zu behandeln. Dieser Zwiespalt lässt sich lösen, indem Zwischenauswertungen durchgeführt werden: Im Studienverlauf werden die akkumulierten Daten inspiziert, und die Studie wird beendet, wenn auffällige Unterschiede zwischen den Vergleichsgruppen beobachtet werden. Aus statistischer Sicht besteht das Problem, dass mit einem mehrfachen Testen der Daten die Wahrscheinlichkeit eines falsch-positiven Studienergebnisses anwächst. Die Anwendung derartiger Studiendesigns in der Therapieforschung setzt deshalb die Verfügbarkeit von Methoden voraus, die sicherstellen, dass das vorgegebene Signifikanzniveau nicht überschritten wird.

Bereits 1947 wurden von WALD sogenannte strikt-sequentielle Versuchspläne vorgeschlagen, die von ARMITAGE (1975) an die Notwendigkeiten klinischer Studien angepasst wurden. Diese Methoden sehen ein kontinuierliches Monitoring der Daten vor, wobei immer dann, wenn ein Patient die Studie abgeschlossen hat, ein Therapiegruppen-vergleich durchgeführt wird. Obwohl diese Designs das eingangs dargestellte Problem in scheinbar idealer Weise lösen, wurden sie in der Praxis kaum angewandt. Dies liegt in erster Linie daran, dass es die logistischen Voraussetzungen, unter denen klinische Studien durchgeführt werden, nur in den seltensten Fällen zulassen, die Entscheidung über Abbruch oder Fortführung einer Studie fortlaufend zu treffen. Vielmehr werden die meisten Studien mit geplanten Zwischenauswertungen von einem unabhängigen Data and Safety Monitoring

Committee (DSMC) begleitet, das sich zu festgelegten Zeitpunkten trifft und die dann verfügbare Information hinsichtlich Wirksamkeit und Sicherheit der untersuchten Behandlungen bewertet.

Die sogenannten gruppensequentiellen Designs sind genau für dieses Vorgehen konzipiert. Die grundsätzliche Idee besteht darin, mehrmals im Studienverlauf mit den dann jeweils verfügbaren Daten die Behandlungsgruppen zu vergleichen und die Studie zu beenden, wenn das Ergebnis signifikant ist. Wird der Test jeweils zum Niveau α durchgeführt, so ist die resultierende Wahrscheinlichkeit für einen Fehler 1. Art bei mehr als einer Auswertung größer als das vorgegebene α; für $\alpha = 0.05$ und fünf Zwischenauswertungen beträgt die tatsächliche Wahrscheinlichkeit für eine fälschliche Ablehnung der Null-Hypothese beispielsweise 0.14 (DEMETS und LAN, 1984). Erste *Ad-hoc*-Regeln für die Wahl geeigneter Signifikanzniveaus für die Zwischen- und Endauswertung, die ein Überschreiten der Fehlerwahrscheinlichkeit verhindern sollten, wurden von HAYBITTLE (1971) und PETO *et al.* (1976) vorgeschlagen. Der eigentliche Durchbruch in der methodischen Entwicklung gruppensequentieller Designs wurde von POCOCK (1977) erzielt. Hier wurde das konstante Signifikanzniveau, das bei Zwischen- und Endauswertung nach jeweils der gleichen Anzahl von Patienten anzuwenden ist, um insgesamt das Niveau α zu kontrollieren, exakt berechnet. Dieser Ansatz wurde nachfolgend in vielfacher Hinsicht erweitert und verallgemeinert, wie z.B. durch alternative Festlegungen der Signifikanzniveaus (siehe etwa O´BRIEN und FLEMMING, 1979; LAN und DEMETS, 1983; HWANG, SHIH und DECANI, 1990) oder der Möglichkeit des Beendens der Studie mit Beibehaltung der Null-Hypothese, falls das beobachtete Ergebnis keine Aussicht auf Erreichen des Studienziels verspricht (PAMPALLONA und TSIATIS, 1994). Für einen Überblick über das umfangreiche methodische Spektrum gruppensequentieller Designs sei auf die Lehrbücher von JENNISON und TURNBULL (1999) und WASSMER (1999b) und die Verfahrensübersicht von KIESER und KÖPCKE (1998) verwiesen.

Für Studien im gruppensequentiellen Ansatz ist es ebenso wie für Studien ohne Zwischenauswertung notwendig, die Design-Elemente, wie z.B. den maximalen Stichprobenumfang, *a priori* im Prüfplan festzulegen. Eine datenabhängige Veränderung dieser Definitionen im Studienverlauf ist nicht gestattet, weil ansonsten die Einhaltung der vorgegebenen Wahrscheinlichkeit α eines falsch-positiven Studienergebnisses nicht sichergestellt ist. Nun ist es aber in der Praxis mehr die Regel als die Ausnahme, dass in der Planungsphase einer Studie die Vorinformationen, die für eine optimale Wahl der Design-Spezifikationen notwendig sind, gänzlich fehlen oder mit großer Unsicherheit behaftet sind. Ein Beispiel hierfür ist die Streuung der Zielgröße, die bei normalverteilten Daten zusammen

mit dem klinisch relevanten Unterschied und den vorgegebenen Wahrscheinlichkeiten eines Fehlers 1. und 2. Art den notwendigen Stichprobenumfang determiniert. Wenn sich bei der Zwischenauswertung einer im gruppensequentiellen Design durchgeführten Studie herausstellt, dass die Streuung wesentlich größer ist als ursprünglich angenommen, so kann im Falle einer Fortsetzung der Studie nach der Interimanalyse der maximale Stichprobenumfang nicht entsprechend erhöht werden. Obwohl damit klar ist, dass die Studie eine unter Umständen wesentlich niedrigere Chance hat, ihr Ziel zu erreichen, lassen gruppensequentielle Designs eine Korrektur der falschen Planungsannahmen im Studienverlauf nicht zu.

Die in jüngster Zeit entwickelten adaptiven Designs (siehe z.B. BAUER und KÖHNE, 1994; PROSCHAN und HUNSBERGER, 1995; FISHER, 1998; LEHMACHER und WASSMER, 1999; WASSMER, 1999b; BRANNATH, POSCH und BAUER 1999; MÜLLER und SCHÄFER, 1999a, 1999b) sind dieser Restriktion nicht unterworfen. Im einfachsten Fall eines zweistufigen adaptiven Designs wird die Studie abhängig vom Ergebnis der Zwischenauswertung mit Ablehnung oder Beibehaltung der Null-Hypothese beendet oder mit einem zweiten Studienteil fortgesetzt. Für die Planung dieses zweiten Teils können alle Informationen aus dem ersten Studienteil (oder auch Ergebnisse aus mittlerweile abgeschlossenen anderen Studien) verwendet werden. Die Möglichkeiten von Design-Änderungen gehen dabei weit über eine Modifikation des ursprünglich festgelegten Stichprobenumfanges hinaus. Beispielsweise können die zu untersuchenden Behandlungsgruppen, die Zielgröße oder der für die Auswertung vorgesehene statistische Test auf der Basis der verfügbaren Daten abgeändert werden, wenn sich die Planungsannahme, auf der die entsprechende Festlegung ursprünglich aufbaute, im Studienverlauf als falsch erweist.

Das Prinzip dieses Vorgehens entspricht der Natur biomedizinischer Forschung, neue Fragestellungen auf der Grundlage des jeweils vorhandenen Wissenstandes zu untersuchen, und aus Erkenntnissen, die sich im Verlauf eines Experimentes ergeben, zu lernen und diese zur Optimierung der Versuchsdurchführung zu nutzen. Gleichzeitig wird die bis in die Gegenwart andauernde Dichotomisierung klinischer Studien in „learning versus confirming“ (SHEINER, 1997) aufgehoben: Innerhalb einer einzelnen Studie können beispielsweise im ersten Teil Arbeitshypothesen generiert oder konkretisiert werden, die dann in einem nachfolgenden zweiten Studienteil definitiv verifiziert werden können. Hierdurch ermöglicht die Klasse adaptiver Studiendesigns eine effizientere Ausnutzung der jeweils verfügbaren Information und somit insgesamt eine schnellere Entscheidungsfindung im Prozess der Therapieevaluierung.

Der Ablauf von Studien mit adaptivem Design ist identisch mit dem von Studien im gruppensequentiellen Ansatz. Im Unterschied zu den gruppensequentiellen Designs, bei denen die Signifikanztests in den Zwischenauswertungen mit der Gesamtheit der jeweils verfügbaren Studiendaten durchgeführt werden, wird bei den adaptiven Designs der statistische Test jeweils separat auf die Stichprobe der einzelnen Studienabschnitte angewendet. Die Verknüpfung der Studienteile und die Testentscheidung erfolgt über eine geeignete Kombinationsregel für die resultierenden *p*-Werte. Auf dieses Konstruktionsprinzip ist es letztlich zurückzuführen, dass im Studienverlauf Design-Veränderungen vorgenommen werden können, ohne die Wahrscheinlichkeit eines Fehlers 1. Art zu erhöhen (BRANNATH, POSCH und BAUER, 1999).

Der erste Vorschlag eines adaptiven Designs geht auf BAUER (1989) und BAUER und KÖHNE (1994) zurück. Dort wird als Kombinationsregel für die *p*-Werte der Produkttest nach FISHER (1932) betrachtet, es wird aber darauf hingewiesen, dass die *p*-Werte auch auf andere Art und Weise zusammengefasst werden können. LEHMACHER und WASSMER (1999) schlugen die Verwendung der inversen Normalverteilungs-Methode vor. Dies eröffnete die Möglichkeit, nach geringfügiger Modifikation der Teststatistik auch innerhalb der „klassischen“ gruppensequentiellen Ansätze datenabhängige Design-Veränderungen, z.B. der geplanten Fallzahl, vornehmen zu können. Alternative Vorschläge adaptiver Designs von PROSCHAN und HUNSBERGER (1995), CUI, HUNG und WANG (1999), FISHER (1998) und SHEN und FISHER (1999) beruhen, obwohl sie von jeweils gänzlich unterschiedlichen Ausgangspunkten her entwickelt wurden, ebenfalls auf dem Kombinations-Prinzips (POSCH und BAUER, 1999; WASSMER, 1999b; BAUER, BRANNATH und POSCH, 2000). Die Entscheidungsgrenzen für die bislang genannten Versuchspläne können für eine beliebige Anzahl von Zwischenauswertungen berechnet werden, wenn diese in der Planungsphase festgelegt wird (BAUER und KÖHNE, 1994; LEHMACHER und WASSMER, 1999; WASSMER, 1999a, 1999b). Eine weitere Flexibilisierung wird durch die jüngsten Vorschläge von BRANNATH, POSCH und BAUER (1999) und MÜLLER und SCHÄFER (1999a) erreicht, bei denen auch dieses Design-Element während der Studie ergebnisabhängig festgelegt werden kann. Die Methode von MÜLLER und SCHÄFER (1999b) schließlich erlaubt es, auch bei Studien, bei denen in der Planungsphase keine Zwischenauswertung vorgesehen war, zu einem beliebig gewählten Zeitpunkt die Daten zu inspizieren und falls notwendig Design-Änderungen vorzunehmen.

1.2 Überblick über die Arbeit

Aufgrund ihrer außerordentlichen Flexibilität haben sich adaptive Designs in kurzer Zeit in der Therapieforschung etabliert. Angesichts der weitreichenden Freiheiten hinsichtlich möglicher Design-Modifikationen sind aber Entscheidungshilfen notwendig, um deren Potential effektiv nutzen zu können. Ziel dieser Arbeit ist es, für einige wichtige Anwendungssituationen Regeln zur Entscheidungsunterstützung in adaptiven Designs anzugeben. Weiterhin wird untersucht, welche Eigenschaften die Anwendung dieser Methoden insbesondere im Vergleich zu nicht-adaptiven Designs besitzt. Die Verfahren werden anhand des Zwei-Stufen Designs von BAUER und KÖHNE (1994) dargestellt. Dies ist keine Einschränkung der Allgemeinheit der vorgestellten Methodik, da sich sämtliche Verfahren durch naheliegende Modifikationen auf andere adaptive Studiendesigns übertragen lassen.

In der Regel sollen in einer klinischen Studie mehrere Fragestellungen beantwortet werden. Beispielsweise lässt sich häufig die Wirksamkeit einer Therapie bei einem komplexen Krankheitsgeschehen nicht durch eine einzelne Zielgröße abbilden und es ist notwendig, bei der Auswertung multiple Endpunkte zu betrachten. Werden in einer Studie mehr als zwei Behandlungsgruppen untersucht, so sind selbst bei einer einzelnen Zielgröße mehrere Tests zum Vergleich der Therapiegruppen durchzuführen. Gerade für Studien, in denen mehrere Hypothesen analysiert werden sollen, ist die Implementierung eines adaptiven Designs attraktiv, da sich mit dem Komplexitätsgrad der Planung auch die Unschärfe bezüglich der Planungsannahmen erhöht. Für die Auswertung ist es notwendig, über Verfahren zu verfügen, die sicherstellen, dass auch bei multiplen Hypothesentests die Wahrscheinlichkeit einer falsch-positiven Testentscheidung kontrolliert wird. In Kapitel 2 werden solche multiplen Testprozeduren für adaptive Designs vorgestellt. Es wird ein allgemeines multiples Testverfahren angegeben, das auf dem sogenannten Abschlusstestprinzip beruht und aus dem entsprechende Prozeduren für spezielle Hypothesensysteme abgeleitet werden.

Eine besondere Bedeutung hat dabei ein Verfahren für die Situation, dass im zweiten Studienteil nur ein Teil der ursprünglich vorgesehenen Hypothesen getestet werden. Eine solche Reduktion der Hypothesenmenge kommt in adaptiven Designs beispielsweise dann zum Tragen, wenn bei Dosis-Findungs-Studien die Ergebnisse der Zwischenauswertung eine der untersuchten Dosierungen favorisieren und diese dann im zweiten Studienteil weiter untersucht werden soll, während die anderen Dosis-Gruppen gestoppt werden. Verfügt man über ein effektives Verfahren, um die aussichtsreichste Dosierung zu identifizieren, so bietet

diese Testprozedur die Basis, um im adaptiven Zwei-Stufen-Design innerhalb einer Stufe zwei Schritte zu kombinieren, die herkömmlicherweise in verschiedenen Phasen der Arzneimittelentwicklung bearbeitet werden: die explorative Untersuchung des Dosis-Wirkungs-Zusammenhangs (Phase II) und der konfirmatorische Wirksamkeitsnachweis (Phase III). In Kapitel 3 werden Methoden angegeben, mit denen für eine plateau-förmige Dosis-Wirkungs-Beziehung die minimale Dosis mit maximaler Wirksamkeit selektiert werden kann. Die Verfahren werden hinsichtlich verschiedener Gütekriterien verglichen, und es wird untersucht, wie die Anwendung dieser Strategie im Rahmen adaptiver Zwei-Stufen-Designs im Vergleich zum herkömmlichen Ein-Stufen-Design ohne Adaption bezüglich der statistischen Power abschneidet.

In Kapitel 4 werden Verfahren zur adaptiven Bestimmung des Stichprobenumfanges angegeben. Dabei werden zwei alternative Ansätze unterschieden. Anders als in adaptiven Designs mit Zwischenauswertung ist es auch möglich, die Planungsannahmen bezüglich der Streuung der Zielgröße zu überprüfen, ohne die Therapiegruppen-Zugehörigkeit offenzulegen. Falls ein vorzeitiger Abbruch der Studie nicht intendiert ist, sondern lediglich die Korrektheit der initialen Fallzahlplanung überprüft werden soll, ist dieser Ansatz hinsichtlich des logistischen Aufwandes wesentlich weniger aufwendig als die Durchführung einer Interimanalyse. Für beide Design-Typen werden Verfahren zur Fallzahladaption vorgeschlagen, und die Gemeinsamkeiten und Unterschiede der beiden Ansätze werden untersucht.

Die statistische Power eines Testverfahrens hängt von der Verteilung der zu analysierenden Zielvariablen ab, über die bei der Planung einer Studie in der Regel nur unsichere Informationen vorliegen. In Kapitel 5 wird ein auf dem Resampling-Prinzip aufbauendes Verfahren vorgestellt, mit dem für den zweiten Teil einer Studie mit adaptivem Design aus einer Klasse von Tests derjenige ausgewählt wird, der mit den Daten der Zwischenauswertung die maximale Power erzielt hätte. Die Eigenschaften einer solchen Selektionsstrategie werden für verschiedene Verteilungssituationen untersucht und mit dem Ein-Stufen-Design und dem Zwei-Stufen-Design ohne Wechsel der Teststatistik verglichen. Im sechsten Kapitel werden die wesentlichen Ergebnisse der Arbeit zusammengefasst und diskutiert.

2. Multiple Testverfahren für adaptive Designs

Die Anwendung multipler Testverfahren ermöglicht es, mehrere Fragestellungen innerhalb einer klinischen Prüfung zu beantworten. Die Multiplizität kann z.B. daher rühren, dass in der untersuchten Indikation die Wirksamkeit einer Therapie nicht anhand einer einzelnen Zielgröße beurteilt werden kann, sondern dass *multiple Endpunkte* betrachtet werden. Ein anderes Beispiel sind *mehrarmige klinische Studien*, bei denen mehr als zwei Behandlungsgruppen verglichen werden. Ein allgemeiner Überblick über Beispiele und Methoden für multiple Testprobleme in klinischen Studien wird in BAUER (1991) gegeben, spezifische Methoden für die Situation multipler Endpunkte sind in WASSMER, REITMEIR, KIESER und LEHMACHER (1999) dargestellt.

In diesem Kapitel werden Verfahren hergeleitet, die Mehrfach-Testen bei adaptiven und gruppensequentiellen Zwischenauswertungen erlauben und die als multiple Testprozeduren das experimentweise (multiple) Niveau kontrollieren. Die Verfahren werden im folgenden für das adaptive Zwei-Stufen-Design nach BAUER und KÖHNE (1994) beschrieben; mit offensichtlichen Modifikationen können die Methoden aber in beliebigen mehrstufigen adaptiven oder gruppensequentiellen Designs angewendet werden. Die Darstellung beruht auf den eigenen Vorarbeiten KIESER und LEHMACHER (1995) und KIESER, BAUER und LEHMACHER (1999). Die nach diesen Artikeln publizierte Arbeit von TANG und GELLER (1999) beschränkt sich auf den Spezialfall der Anwendung des Abschlusstest-Prinzips in Studien mit gruppensequentiellem Design. Die Theorie und Anwendung der Verfahren im Rahmen von Dosis-Findungs- und Dosis-Wirkungs-Studien wurde in den Publikationen BAUER und KIESER (1999) und LEHMACHER, KIESER und HOTHORN (2000) behandelt. Die Anwendung der multiplen Testprozeduren werden anhand von klinischen Studien illustriert, die im adaptiven Zwei-Stufen-Design durchgeführt wurden und multiple Endpunkte (MALSCH und KIESER, 2000) bzw. mehr als zwei Behandlungsgruppen (LAAKMANN, SCHÜLE, BAGHAI und KIESER, 1998) aufwiesen.

In Kapitel 2.1 wird zunächst das adaptive Zwei-Stufen-Design von BAUER und KÖHNE (1994) für den Fall einer zu testenden Null-Hypothese eingeführt. In Kapitel 2.2 wird eine allgemeine Abschlusstest-Prozedur für die Analyse multipler Testprobleme in adaptiven Designs angegeben. In Kapitel 2.3 werden als Spezialfälle der Abschlusstest-Prozedur eine multiple Testprozedur für *a priori* geordnete Hypothesen und eine Bonferroni-Holm-Prozedur abgeleitet. Bei diesen Betrachtungen wird zunächst vorausgesetzt, dass die Familie der zu testenden Hypothesen für die Zwischen- und die Endauswertung gleich ist. In Kapitel 2.4 wird ein Satz bewiesen, der die Durchführung der Abschlusstest-Prozedur und ihrer

Spezialfälle auch in der Situation erlaubt, dass die Hypothesenmenge nach der Zwischenauswertung reduziert wird. Wie wir sehen werden, birgt dieser Satz den Schlüssel zur testtheoretischen Behandlung weitergehender adaptiver Modifikationen der Hypothesenstruktur. Die praktische Anwendung der multiplen Testprozeduren wird in Kapitel 2.5 anhand zweier klinischer Studien dargestellt.

2.1 Adaptives Zwei-Stufen-Design nach BAUER und KÖHNE (1994)

In diesem Abschnitt wird das adaptive Zwei-Stufen-Design nach BAUER und KÖHNE (1994) für die Situation beschrieben, dass im Rahmen der Studie eine einzige Null-Hypothese H_0 untersucht wird. Die grundsätzliche Idee besteht darin, die beiden *p*-Werte p_1 und p_2 aus den disjunkten Stichproben des ersten und zweiten Studienteils mittels Niveau-α-Tests für H_0 zu berechnen und mit dem Fisher-Produkttest zu einer Teststatistik für die Endauswertung nach dem zweiten Studienteil zu kombinieren. Die Studie kann bereits nach der Zwischenauswertung beendet werden, wenn $p_1 \le \alpha_1$ (vorzeitige Ablehnung von H_0) oder $p_1 \ge \alpha_0$ (vorzeitige Beibehaltung von H_0). Für $\alpha_1 < p_1 < \alpha_0$ wird die Studie mit einem zweiten Teil fortgesetzt, wobei für dessen Planung alle Informationen der Zwischenauswertung verwendet werden können. Nach dem zweiten Studienteil kann H_0 abgelehnt werden, wenn $p_1 \cdot p_2 \le c_{\alpha_2}$ (siehe Abbildung 1).

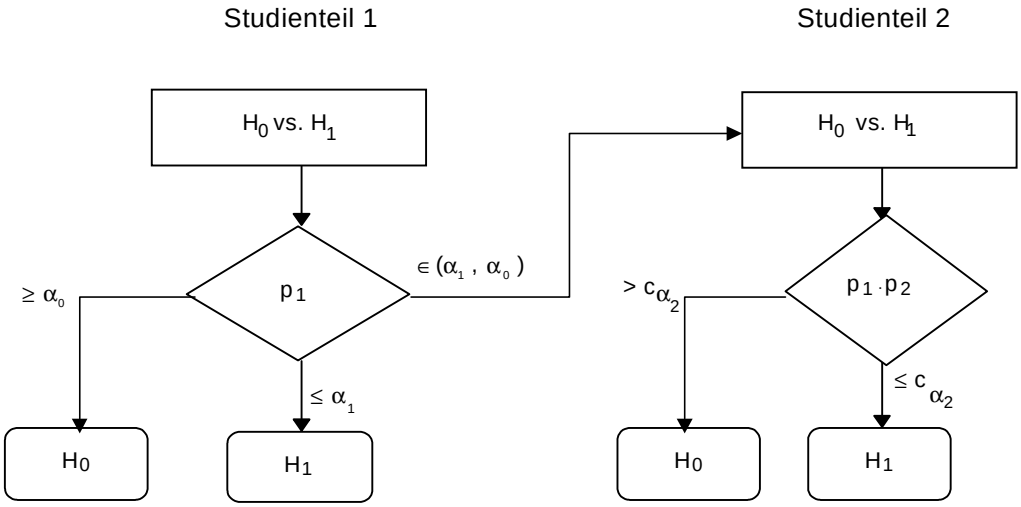

Abbildung 1: Adaptives Zwei-Stufen-Design nach BAUER und KÖHNE (1994) für den Test einer einzelnen Null-Hypothese H_0.

Aufgrund der Tatsache, dass p_1 und p_2 auf der Basis separater Stichproben berechnet werden, ist die auf p_1 bedingte Verteilung des *p*-Wertes des zweiten Studienteils stochastisch größer oder gleich der Gleichverteilung auf dem Intervall $[0,1]$. Diese Eigenschaft ist hinreichend dafür, dass im Rahmen der Zwischenauswertung datenabhängige Designänderungen vorgenommen werden können, ohne dass die Wahrscheinlichkeit eines Fehlers 1. Art erhöht wird (BRANNATH, POSCH und BAUER, 1999).

Damit für die Studie die Wahrscheinlichkeit eines Fehlers 1. Art durch α kontrolliert wird, müssen die Stop-Grenzen $\alpha_1 < \alpha$ und $\alpha_0 > \alpha$ und die kritische Schranke $c_{\alpha_2} \le \alpha_1$ für den Kombinationstest zum lokalen Niveau $\alpha_2 \le \alpha$ in der Planungsphase so gewählt werden, dass sie den folgenden Bedingungen genügen:

$$c_{\alpha_2} = \exp(-0.5 \cdot \chi^2_{4,1-\alpha_2}) \tag{2.1a}$$

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_{0}^{c_{\alpha_2}/p_1} dp_2 \; dp_1 = \alpha_1 + c_{\alpha_2} \cdot (\ln \alpha_0 - \ln \alpha_1) = \alpha. \tag{2.1b}$$

Dabei bezeichnet $\chi^2_{4,1-\alpha_2}$ das $(1-\alpha_2)$-Perzentil der Chi-Quadrat-Verteilung mit 4 Freiheitsgraden.

Der Spezialfall $\alpha_2 = \alpha$ entspricht dem ursprünglichen Vorschlag von BAUER und KÖHNE (1994), die Verallgemeinerung $\alpha_2 \le \alpha$ wurde erstmals von BAUER und RÖHMEL (1995) vorgeschlagen. Für $\alpha_0 = 1$ wird die Studie nur bei vorzeitiger Ablehnung der Null-Hypothese nach der Zwischenauswertung beendet (d.h. für $p_1 \le \alpha_1$). Unabhängig von den speziell gewählten Werten für α_0 und α_1 kann die Studie für $p_1 \le c_{\alpha_2}$ stets mit der Ablehnung von H_0 beendet werden, weil dann wegen $p_2 \le 1$ die Bedingung $p_1 \cdot p_2 \le c_{\alpha_2}$ bereits bei der Zwischenauswertung erfüllt ist (sog. „nonstochastic curtailment“). Entsprechende kritische Schranken für Designs mit mehr als zwei Stufen können durch Verallgemeinerung der Niveau-Bedingungen (2.1a) und (2.1b) hergeleitet werden (BAUER und KÖHNE, 1994; WASSMER, 1999a, 1999b).

2.2 Abschlusstest-Prozedur

Wir nehmen im folgenden an, dass die Null-Hypothesen $H_0^1,...,H_0^k$ mit zugehörigen Alternativ-Hypothesen $H_1^1,...,H_1^k, k \geq 1$, im adaptiven Zwei-Stufen-Design nach BAUER und KÖHNE (1994) getestet werden sollen. Die in den folgenden Kapiteln angegebenen multiplen Testprozeduren kontrollieren das experimentweise Niveau α, d.h., die Wahrscheinlichkeit für die Ablehnung mindestens einer der wahren Null-Hypothesen beträgt höchstens α, unabhängig davon, welche und wie viele der Null-Hypothesen $H_0^1,...,H_0^k$ tatsächlich wahr sind (HOCHBERG und TAMHANE, 1987).

2.2.1 Abschlusstest-Prozedur für Studien ohne Zwischenauswertung

Das Abschlusstest-Prinzip ist eine allgemeine Methode zur Konstruktion multipler Testprozeduren, die das experimentweise Niveau kontrollieren. Hierzu betrachten wir die Familie von Null-Hypothesen $\Im = \bigcup_{J \subseteq \{1,...,k\}} H_0^J$, wobei $H_0^J = \bigcap_{i \in J} H_0^i, J \subseteq \{1,...,k\}$. $\Im$ ist nach Konstruktion durchschnittsabgeschlossen, d.h., der Durchschnitt von je zwei Elementen aus $\Im$ ist ebenfalls in $\Im$ enthalten. Weiterhin wird zur Anwendung des Abschlusstest-Prinzips für jede Null-Hypothese $H_0^J \in \Im$ ein Niveau-α-Test benötigt. Die folgende Abschlusstest-Prozedur kontrolliert dann das experimentweise Niveau α (MARCUS, PERITZ und GABRIEL, 1976): Eine Null-Hypothese $H_0^J \in \Im$ wird abgelehnt, wenn sie selbst zum Niveau α abgelehnt werden kann und alle Null-Hypothesen aus $\Im$, die H_0^J implizieren (d.h., alle $H_0^I \in \Im$ mit $I \supseteq J$) ebenfalls zum Niveau α abgelehnt werden können.

2.2.2 Abschlusstest-Prozedur für das adaptive Zwei-Stufen-Design

Das Abschlusstest-Prinzip kann auf klinische Studien mit Zwischenauswertung angewendet werden, indem die entsprechenden Regeln für eine vorzeitige Ablehnung oder Beibehaltung von Null-Hypothesen im Rahmen der Interimanalysen sowie die Regeln des Abschlusstest-Prinzips berücksichtigt werden. Für das adaptive Zwei-Stufen-Design nach BAUER und KÖHNE (1994) ergeben sich damit die im folgenden Satz angegebenen Entscheidungen. Mit

p_{Ij} sei der *p*-Wert eines Niveau-α-Tests der Hypothese $H_0^I, I \subseteq \{1,...,k\}$ in der Analyse $j, j = 1, 2,$ bezeichnet; dabei steht $j = 1$ für die Interimanalyse und $j = 2$ für die Endauswertung.

Satz 1:

Die folgende multiple Testprozedur kontrolliert das experimentweise Niveau α für die Hypothesenfamilie $\Im = \bigcup_{J \subseteq \{1,...,k\}} H_0^J$ im adaptiven Zwei-Stufen-Design nach BAUER und KÖHNE (1994).

Fall I: $p_{I1} \leq \alpha_1$ oder $(\alpha_1 < p_{I1} < \alpha_0$ und $p_{I1} \cdot p_{I2} \leq c_{\alpha_2})$ für alle $I \supseteq J$

- H_0^J wird abgelehnt.

Fall II: $p_{I1} > \alpha_1$ und nicht $(\alpha_1 < p_{I1} < \alpha_0$ und $p_{I1} \cdot p_{I2} \leq c_{\alpha_2})$ für alle $I \supseteq J$

- H_0^J wird beibehalten.

Beweis:

Fall I besagt, dass alle Null-Hypothesen, die H_0^J implizieren, entweder im Rahmen der Interimanalyse (1. Bedingung) oder bei der Endauswertung mit Fisher´s Kombinationstest (2. Bedingung) abgelehnt werden. Alle Null-Hypothesen werden zum (lokalen) Niveau α getestet, da die entsprechenden kritischen Schranken des adaptiven Zwei-Stufen-Designs angewendet werden. Weiterhin werden die Entscheidungsregeln des Abschlusstest-Prinzips angewendet. Damit ist die Behauptung bewiesen. ■

In Abbildung 2 ist die Abschlusstest-Prozedur für das durchschnittsabgeschlossene Hypothesensystem $\Im = \{H_0^1, H_0^2, H_0^{\{1,2\}}\}$ mit $H_0^{\{1,2\}} = H_0^1 \cap H_0^2$ dargestellt. Beispielsweise kann die Null-Hypothese H_0^1 nach dem ersten Schritt abgelehnt werden, falls $H_0^{\{1,2\}}$ und H_0^1 im Rahmen der Interimanalyse abgelehnt werden können, d.h., falls die zugehörigen *p*-Werte unter der kritischen Schranke α_1 liegen. Falls die Schnitthypothese $H_0^{\{1,2\}}$ nach dem ersten Schritt abgelehnt werden kann, aber für den zu H_0^1 gehörigen *p*-Wert p_{11} die Ungleichung $\alpha_1 < p_{11} < \alpha_0$ gilt, kann H_0^1 im Rahmen der Endauswertung abgelehnt werden, falls zusätzlich gilt $p_{11} \cdot p_{12} \leq c_{\alpha_2}$.

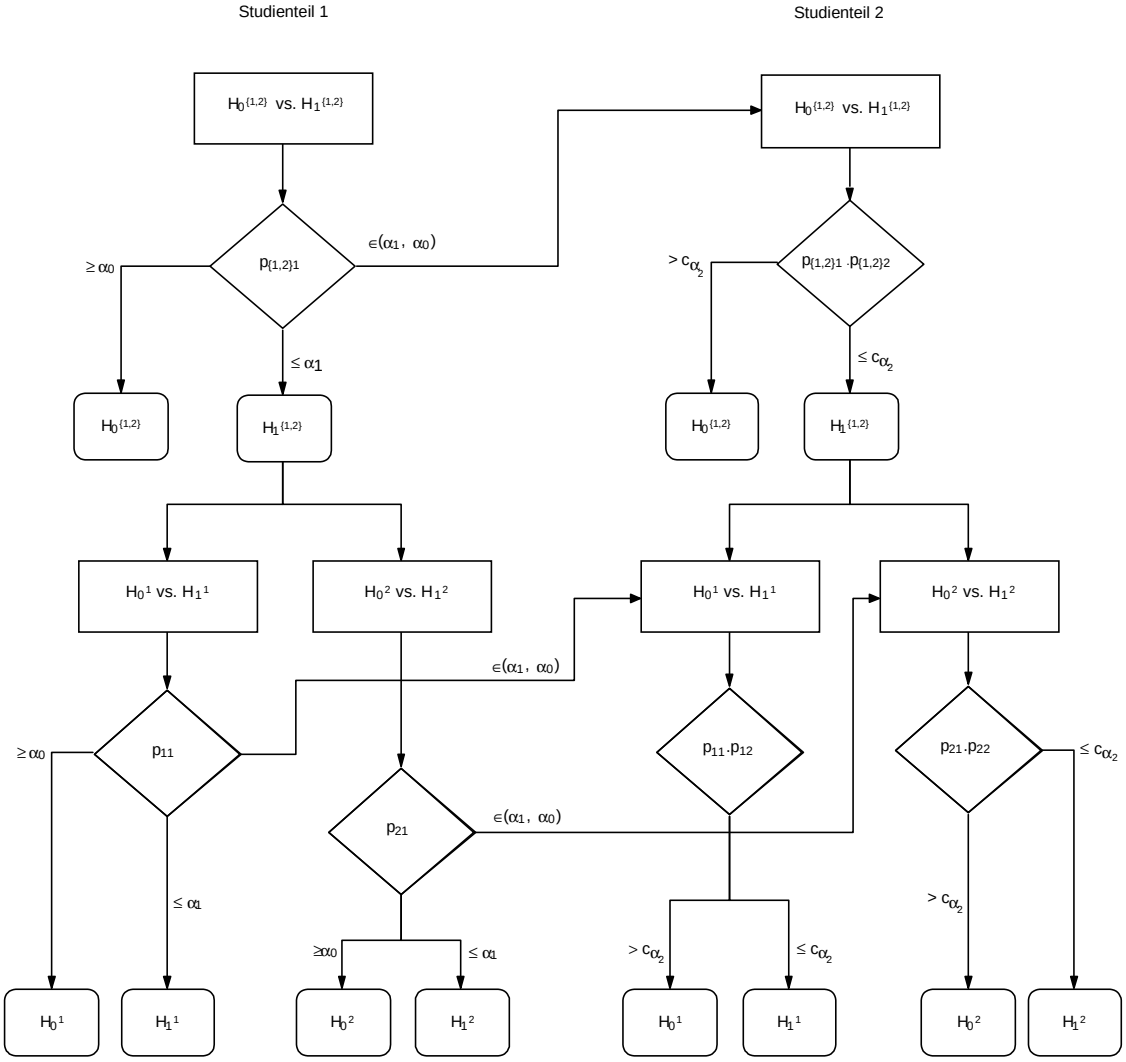

Abbildung 2: Abschlusstest-Prozedur für das adaptive Zwei-Stufen-Design nach BAUER und KÖHNE (1994) und die beiden Null-Hypothesen H_0^1 und H_0^2 mit Schnitt-Hypothese $H_0^{\{1,2\}} = H_0^1 \cap H_0^2$.

Es gibt eine Fülle von Verfahren zur Herleitung von Niveau-α-Tests für Schnitthypothesen $H_0^I, I \subseteq \{1,...,k\}$. Eine Klasse von Methoden basiert auf dem Prinzip, aus Tests zum Niveau α für die Einzelhypothesen $H_0^i, i \in I$, Tests für die Schnitthypothese H_0^I zu konstruieren, die ihrerseits das experimentweise Niveau α kontrollieren. Im nächsten Kapitel werden zwei wichtige Spezialfälle dieser Konstruktionsmethode betrachtet: Ein Testverfahren für *a priori* geordnete Hypothesen und eine Bonferroni-Holm-Prozedur.

2.3 Wichtige Spezialfälle der Abschlusstest-Prozedur

2.3.1 *A priori* geordnete Hypothesen

Häufig haben in klinischen Studien die Fragestellungen, die einem multiplen Testproblem zugrunde liegen, unterschiedliche Priorität im Hinblick auf die Zielsetzung der Studie. Dies ermöglicht es, die zugehörigen Null-Hypothesen in der Planungsphase (*a priori*) hierarchisch anzuordnen. Wir nehmen im folgenden an, dass die Reihenfolge der Wichtigkeit der Null-Hypothesen durch die Ordnung $H_0^1,...,H_0^k$ gegeben ist. In dieser Situation ist die Ablehnung einer Null-Hypothese $H_0^i, i \in \{1,...,k\}$ nur dann von Interesse, wenn alle Null-Hypothesen $H_0^1,...,H_0^{i-1}$ abgelehnt werden können. In der folgenden multiplen Testprozedur, die für Studien ohne Zwischenauswertung von MAURER, HOTHORN und LEHMACHER (1995) beschrieben wurde, werden die Null-Hypothesen zum lokalen Niveau α getestet, wobei für das Hypothesensystem $\Im$ das experimentweise Niveau α im strengen Sinne kontrolliert wird. Die Testprozedur ergibt sich aus der allgemeinen Abschlusstestprozedur, indem folgende Niveau-α-Tests für die Schnitthypothesen H_0^I verwendet werden: Lehne H_0^I ab, falls $p_I \underset{\text{def}}{=} p_{<I>} \leq \alpha$; dabei bezeichnet $< I >$ die kleinste Zahl in der Indexmenge I.

2.3.1.1 Testprozedur für Studien ohne Zwischenauswertung

In Schritt $i, i = 1,...,k,$ der multiplen Testprozedur wird ein Niveau-α-Test für das Testproblem H_0^i vs. H_1^i verwendet, der zugehörige *p*-Wert sei mit p_i bezeichnet. Falls $p_i \leq \alpha$ wird H_0^i abgelehnt; falls $i < k$ wird die Testprozedur mit Schritt $i+1$ fortgesetzt, falls $i = k$ stoppt die Testprozedur. Die Testprozedur stoppt ebenfalls für $p_i > \alpha$. Falls die Prozedur in Schritt $i, 1 \leq i \leq k,$ ohne Ablehnung der Null-Hypothese H_0^i stoppt, werden alle Null-Hypothesen $H_0^i,...,H_0^k$ beibehalten, während alle $H_0^1,...,H_0^{i-1}$ zum experimentweisen Niveau α abgelehnt werden.

2.3.1.2 Testprozedur für das adaptive Zwei-Stufen-Design

Die Konzepte adaptiver Zwischenauswertungen und *a priori* geordneter Hypothesen lassen sich zu einer Strategie verknüpfen, die das experimentweise Niveau α im strengen Sinne kontrollieren. Dabei nehmen wir an, dass die Hypothesen-Hierarchie im ersten und zweiten Studienteil gleich ist. (Aus dem in Kapitel 2.4 angegebenen Satz 6 kann gefolgert werden, dass diese Voraussetzung bei geeigneter Modifikation der Testprozedur auch aufgegeben werden kann; siehe hierzu Bemerkung 3 zu Satz 6.) Wie oben bezeichnet p_{ij} die *p*-Werte zur Hypothese H_0^i, $i = 1,...,k$, und Auswertung $j, j = 1, 2$, ($j = 1$: Zwischenauswertung, $j = 2$: Endauswertung).

Satz 2:

Die folgende multiple Testprozedur kontrolliert das experimentweise Niveau α für die *a priori* geordnete Hypothesen $H_0^1,...,H_0^k$ im adaptiven Zwei-Stufen-Design nach BAUER und KÖHNE (1994).

Zwischenauswertung:

Im Rahmen der Zwischenauswertung wird das lokale Niveau α_1 angewendet. In Schritt i der Testprozedur, $i = 1,...,k,$ sind die Null-Hypothesen $H_0^1,...,H_0^{i-1}$ bereits abgelehnt, und es sind folgende Testentscheidungen möglich:

Fall I: $p_{i1} \geq \alpha_0$

- Alle Null-Hypothesen $H_0^i,...,H_0^k$ werden beibehalten.
- Die Testprozedur stoppt und die Studie wird beendet.

Fall II: $p_{i1} \leq \alpha_1$

- H_0^i wird abgelehnt.
- Falls $i < k$: Die Testprozedur wird mit Schritt $i+1$ fortgesetzt.
- Falls $i = k$: Die Testprozedur stoppt und die Studie wird beendet.

Fall III: $\alpha_1 < p_{i1} < \alpha_0$

- Keine definitive Entscheidung über H_0^i im Rahmen der Zwischenauswertung.
- Die Testprozedur stoppt im Rahmen der Zwischenauswertung. Falls H_0^i abgelehnt werden soll, wird der zweite Studienteil geplant.

- Im Rahmen der Endauswertung wird die Testprozedur mit Schritt i fortgesetzt, wobei H_0^i dann mit Fisher´s Kombinationstest getestet wird.

Endauswertung:

Um eine Null-Hypothese H_0^i im Rahmen der Endauswertung ablehnen zu können, muss der *p*-Wert bei der Zwischenauswertung die notwendige Bedingung $p_{i1} < \alpha_0$ erfüllen. Definiert man den Index s als $s = \max_{i=1,\ldots,k}\{\, i : p_{j1} < \alpha_0 \text{ für alle } j \le i \,\}$, so können nur Hypothesen H_0^i mit $i \le s$ im Rahmen der Endauswertung abgelehnt werden. (Falls das Maximum nicht existiert, gilt definitionsgemäß $s = 0$.)

In Schritt i der Endauswertung wird das Produkt $p_{i1} \cdot p_{i2}$ mit der kritischen Schranke c_{α_2} verglichen:

Fall I: $p_{i1} > \alpha_1$ und $p_{i1} \cdot p_{i2} > c_{\alpha_2}$

- Alle Null-Hypothesen $H_0^i, \ldots, H_0^k$ werden beibehalten. Die Testprozedur stoppt.

Fall II: $p_{i1} \le \alpha_1$ oder $p_{i1} \cdot p_{i2} \le c_{\alpha_2}$

- H_0^i wird abgelehnt.
- Falls $i < s$: Die Testprozedur wird mit Schritt $i+1$ fortgesetzt.
- Falls $i = s$: Die Testprozedur stoppt.

Beweis:

In jedem Schritt der Testprozedur wird das lokale Niveau α kontrolliert, weil die kritischen Schranken des adaptiven Zwei-Stufen-Designs angewendet werden. Darüber hinaus werden die Entscheidungen entsprechend den Regeln der multiplen Testprozedur für *a priori* geordnete Hypothesen getroffen. Deshalb kontrolliert die oben beschriebene Testprozedur insgesamt das experimentweise Niveau α. ∎

Aus der Aussage in Fall II der Endauswertung folgt insbesondere, dass Hypothesen H_0^i mit $p_{i1} \le \alpha_1$, die bei der Zwischenauswertung nicht abgelehnt werden konnten, weil Hypothesen H_0^j, $j < i$, mit höherer Priorität nicht abgelehnt werden konnten, ungeachtet der *p*-Werte p_{i2} im zweiten Studienteil (H_0^i muss dort nicht einmal getestet werden) abgelehnt werden können, wenn alle H_0^j, $j < i$, zumindest bei der Endauswertung abgelehnt werden können.

Es sei noch angemerkt, dass sich die Testprozedur für den Fall $\alpha_0 = 1$ (keine vorzeitige Beendigung der Studie mit der Beibehaltung von Null-Hypothesen) wesentlich vereinfacht: Im Rahmen der Zwischenauswertung müssen dann lediglich die Fälle II und III (dort nur die Bedingung $\alpha_1 < p_{i1}$) berücksichtigt werden. In der Endauswertung kann Fisher´s Produkttest auf alle Null-Hypothesen angewandt werden, die nicht bereits bei der Zwischenauswertung abgelehnt wurden ($s = k$).

2.3.2 Bonferroni-Holm-Prozedur

Das Bonferroni-Holm-Verfahren (HOLM, 1979) ist eine einfache multiple Testprozedur für die Situation, dass keine Hierarchie zwischen den zu testenden Null-Hypothesen besteht. Die Prozedur kontrolliert das experimentweise Niveau.

2.3.2.1 Testprozedur für Studien ohne Zwischenauswertung

Mit $p_{(1)} \le p_{(2)} \le ... \le p_{(k)}$ seien die in aufsteigender Größe geordneten *p*-Werte bezeichnet und mit $H_0^{(1)}, H_0^{(2)}, ..., H_0^{(k)}$ die zugehörigen Null-Hypothesen. Beginnend mit dem kleinsten *p*-Wert $p_{(1)}$ werden schrittweise die *p*-Werte $p_{(i)}$ mit den kritischen Schranken $\alpha/(k-i+1)$ verglichen; die zugehörigen Null-Hypothesen werden abgelehnt, solange gilt $p_{(i)} \le \alpha/(k-i+1), i = 1,...,k$.

Die Bonferroni-Holm-Prozedur folgt aus dem allgemeinen Abschlusstest-Prinzip, indem die folgenden sogenannten Bonferroni-Globaltests zum lokalen Niveau α zum Test der Schnitthypothesen $H_0^I \in \Im$ verwendet werden: Lehne H_0^I ab, falls mindestens eine der Null-Hypothesen $H_0^i, i \in I,$ zum Niveau $\alpha/|I|$ abgelehnt werden kann; dabei bezeichnet $|I|$ die Anzahl der Elemente in der Menge I .

2.3.2.2 Testprozedur für das adaptive Zwei-Stufen-Design

Die Ergebnisse aus Kapitel 2.2 zur allgemeinen Abschlusstest-Prozedur können dazu genutzt werden, die Bonferroni-Holm-Prozedur auf Studien mit Zwischenauswertungen zu verallgemeinern. Zu diesem Zweck sind geeignete kritische Schranken für die Bonferroni-

Globaltests der Zwischen- und Endauswertung zu definieren. Für das adaptive Zwei-Stufen-Design zum Niveau α bezeichnen wir hierzu für einen festen Wert von α_0 mit $\alpha_1(\alpha)$ bzw. $\alpha_2(\alpha)$ die kritischen Niveaus, die bei der Zwischen- bzw. der Endauswertung zugrundegelegt werden; mit $c_{\alpha_2(\alpha)}$ wird der zugehörige kritische Wert für Fisher´s Kombinationstest bezeichnet.

Satz 3:

Die folgende multiple Testprozedur kontrolliert das experimentweise Niveau α für die Hypothesen $H_0^1,...,H_0^k$ im adaptiven Zwei-Stufen-Design nach BAUER und KÖHNE (1994).

Zwischenauswertung:

Mit $p_{(1)1} \le p_{(2)1} \le ... \le p_{(k)1}$ seien die geordneten *p*-Werte in der Zwischenauswertung bezeichnet, mit $H_0^{(1)}, H_0^{(2)},..., H_0^{(k)}$ die zugehörigen Hypothesen.

Fall I: $p_{(1)1} \ge \alpha_0$

- Alle Null-Hypothesen $H_0^{(i)}, i = 1,..., k,$ werden beibehalten. Die Studie wird beendet.

Fall II: $p_{(1)1} \le \alpha_1$

- $H_0^{(i)}$ wird abgelehnt für $i = 1,..., r$, wobei
  $r = \max_{i=1,...,k} \{ i : p_{(i)1} \le (\alpha_1(\alpha/(k - j + 1))) \text{ für alle } j = 1,..., i \}.$
  (Falls das Maximum nicht existiert, gilt definitionsgemäß $r = 0$ und $H_0^{(0)} = \varnothing$.)
- $H_0^{(i)}$ wird beibehalten für alle i mit $p_{(i)1} \ge \alpha_0$.
- Keine definitive Entscheidung über $H_0^{(i)}$ im Rahmen der Zwischenauswertung für $i = r + 1,..., s,$ wobei $s = \max_{i=1,...,k} \{ i : p_{(i)1} < \alpha_0 \}$. (Falls das Maximum nicht existiert gilt definitionsgemäß $s = 0$ und $H_0^{(0)} = \varnothing$.)

Im Rahmen der Endauswertung können die Null-Hypothesen $H_0^{(i)}, i = r + 1,..., s,$ erneut getestet werden.

Endauswertung:

Der einfacheren Bezeichnung halber nehmen wir an, dass die Null-Hypothesen $H_0^1,..., H_0^r$ bereits abgelehnt wurden und dass $H_0^{r+1},..., H_0^s, r + 1 \le s \le k,$ in der Zwischenauswertung weder abgelehnt noch beibehalten wurden und deshalb im Rahmen der Endauswertung erneut getestet werden können. Wir bezeichnen mit $(p_1 \cdot p_2)_{(i)}$ die geordneten Produkte der *p*-Werte

für die Null-Hypothesen $H_0^i, i = r+1,...,s,$ und nehmen an, dass die Null-Hypothesen so geordnet sind, dass $(p_1 \cdot p_2)_{(r+1)}$ zur Null-Hypothese H_0^{r+1} gehört. Es sei angemerkt, dass sich diese Ordnung, die auf den Produkten der *p*-Werte basiert, im allgemeinen von der Ordnung der Null-Hypothesen $H_0^{(1)},...,H_0^{(k)}$, die auf den *p*-Werten der Zwischenauswertung beruht, unterscheidet.

Fall I: $(p_1 \cdot p_2)_{(r+1)} > c_{\alpha_2(\alpha/(k-r))}$

- Alle Null-Hypothesen $H_0^i, i = r+1,...,s,$ werden beibehalten

Fall II: $(p_1 \cdot p_2)_{(r+1)} \le c_{\alpha_2(\alpha/(k-r))}$

- H_0^{r+1} wird abgelehnt.
- Falls $r+1 < s$: Die Testprozedur wird mit Schritt $r+2$ fortgesetzt.
- Falls $r+1 = s$: Die Testprozedur stoppt.

Schritt $r+2$ der Testprozedur:

a) $p_{(r+2)1} > \alpha_1(\alpha/(k-r-1))$ und $(p_1 \cdot p_2)_{(r+2)} > c_{\alpha_2(\alpha/(k-r-1))}$

- Alle Null-Hypothesen $H_0^i, i = r+2,...,s,$ werden beibehalten. Die Testprozedur stoppt.

b) $p_{(r+2)1} \le \alpha_1(\alpha/(k-r-1))$ oder $(p_1 \cdot p_2)_{(r+2)} \le c_{\alpha_2(\alpha/(k-r-1))}$

- H_0^i wird abgelehnt, wobei i den Index bezeichnet, für den gilt $p_{i1} = p_{(r+2)1}$ oder $(p_{i1} \cdot p_{i2}) = (p_1 \cdot p_2)_{(r+2)}$. (Falls i nicht eindeutig bestimmt ist, so ist dies für das Resultat unerheblich, weil jede Null-Hypothese H_0^i, die zumindest eine der beiden Bedingungen erfüllt, in diesem oder einem der folgenden Schritte abgelehnt wird.)
- Falls $r+2 < s$: Die Testprozedur wird mit Schritt $r+3$ fortgesetzt.
- Falls $r+2 = s$: Die Testprozedur stoppt.

Beweis:

Der folgende Bonferroni-Globaltest kontrolliert das Niveau α für das adaptive Zwei-Stufen-Design: Lehne eine Schnitthypothese $H_0^I \in \Im$ ab, falls zumindest eine der Null-Hypothesen $H_0^i, i \in I,$ im Rahmen der Zwischenauswertung zum Niveau $\alpha_1(\alpha/|I|)$ abgelehnt werden kann, oder falls für zumindest eine der Null-Hypothesen $H_0^i, i \in I$, die Bedingung $\{\alpha_1(\alpha/|I|) < p_i < \alpha_0$ und $p_{i1} \cdot p_{i2} \le c_{\alpha_2(\alpha/I)}\}$ gilt. Die zugehörige Bonferroni-Holm-Prozedur folgt nun direkt aus der in Satz 1, Kapitel 2.2, beschriebenen allgemeinen Abschlusstest-

Prozedur, indem die Zwei-Stufen Bonferroni-Globaltests auf alle Null-Hypothesen des durchschnittsabgeschlossenen Systems $\Im$ angewendet werden. ■

Wie bei der Testprozedur für *a priori* geordnete Hypothesen ergibt sich eine deutliche Vereinfachung für den Fall, dass eine vorzeitige Beibehaltung von Null-Hypothesen nicht vorgesehen ist ($\alpha_0 = 1$).

Abschließend sei noch auf eine Eigenschaft der Bonferroni-Holm-Prozedur in Studien mit Zwischenauswertung hingewiesen, die in ähnlicher Form bereits für *a priori* geordnete Hypothesen aufgetreten war. Aufgrund der Tatsache, dass die kritischen Schranken für die Zwischen- und die Endauswertung im Falle der Ablehnung von Null-Hypothesen anwachsen, kann es passieren, dass Null-Hypothesen, die nicht bei der Zwischenauswertung abgelehnt werden konnten, ungeachtet des zugehörigen *p*-Wertes im zweiten Studienteil im Rahmen der Endauswertung abgelehnt werden können. Zur Illustration nehmen wir an, dass $k = 3$ Hypothesen getestet werden und dass für die bei der Zwischenauswertung erzielten *p*-Werte gilt $\alpha_1(\alpha/3) < p_{11} = p_{(1)1} \le \alpha_1(\alpha/2)$, $p_{21} < \alpha_0$ und $p_{31} < \alpha_0$. Keine der Null-Hypothesen kann deshalb in der Zwischenauswertung abgelehnt werden, aber alle Null-Hypothesen können im Rahmen der Endauswertung nochmals getestet werden. Falls nun beispielsweise bei der Endauswertung gilt $p_{21} \cdot p_{22} \le c_{\alpha_2(\alpha/3)}$, dann kann H_0^2 abgelehnt werden, und H_0^1 kann ebenfalls abgelehnt werden, egal welches Ergebnis für diese Fragestellung im zweiten Studienteil erzielt wird (H_0^1 muss dort nicht einmal untersucht worden sein). H_0^3 kann abgelehnt werden, falls gilt $p_{31} \le \alpha_1(\alpha)$ oder $p_{31} \cdot p_{32} \le c_{\alpha_2(\alpha)}$.

2.3.2.3 Wahl der lokalen Signifikanzniveaus

Für die Testentscheidung im Rahmen der Bonferroni-Holm-Prozedur sollten die lokalen Signifikanzniveaus $\alpha_1(\alpha)$ und $\alpha_2(\alpha)$ des adaptiven Zwei-Stufen-Designs sinnvollerweise als monoton wachsende Funktionen des Gesamt-Signifikanzniveaus α gewählt werden. Dies ist gegeben, falls die folgenden Monotonie-Bedingungen erfüllt sind:

$$\alpha_1(\alpha') \le \alpha_1(\alpha'') \text{ für } \alpha' \le \alpha'' \tag{2.2a}$$

$$\alpha_2(\alpha') \le \alpha_2(\alpha'') \text{ für } \alpha' \le \alpha''. \tag{2.2b}$$

Für den Zwei-Stufen Fisher-Kombinationstest und $\alpha_2 = \alpha$, was dem Vorschlag von BAUER und KÖHNE (1994) entspricht, ist Bedingung (2.2b) offensichtlich erfüllt. Die folgenden Überlegungen zeigen, dass auch die Monotonie-Bedingung (2.2a) erfüllt ist.

Satz 4:

Für die Wahl $\alpha_2 = \alpha$ sind die Monotonie-Bedingungen (2.2a) und (2.2b) für die lokalen Signifikanzniveaus α_1 und α_2 der Zwischen- und Endauswertung erfüllt.

Beweis:

Die kritische Schranke $\alpha_1(\alpha)$ erhält man durch Auflösung der Gleichung

$$F(\alpha, \alpha_0, \alpha_1) = \alpha_1 + c_\alpha \cdot (\ln \alpha_0 + \ln \alpha_1) - \alpha = 0. \tag{2.3}$$

Aus dem Satz über implizite Funktionen folgt

$$\begin{aligned} \left(\frac{\partial \alpha_1(\alpha)}{\partial \alpha} \right) &= -\left(\frac{\partial F}{\partial \alpha_1} \right)^{-1} \cdot \left(\frac{\partial F}{\partial \alpha} \right) \\ &= \frac{\alpha_1(\alpha)}{c_\alpha - \alpha_1(\alpha)} \cdot \left(\frac{\partial c_\alpha}{\partial \alpha} \cdot (\ln \alpha_0 - \ln \alpha_1(\alpha)) - 1 \right) \end{aligned} \tag{2.4}$$

Benutzt man die Definition von c_α in (2.1a) und den Zusammenhang zwischen den Quantilen und der Dichte der Chi-Quadrat-Verteilung mit 4 Freiheitsgraden, so erhält man

$$\frac{\partial c_\alpha}{\partial \alpha} = \frac{c_\alpha}{\alpha - c_\alpha}. \tag{2.5}$$

Aus der Niveau-Bedingung (2.1b) für den Fisher Kombinationstest im Zwei-Stufen-Design folgt für $\alpha_2 = \alpha$

$$\alpha_1 + c_\alpha \cdot (\ln \alpha_0 - \ln \alpha_1) = \alpha,$$

und damit gilt

$$c_\alpha \cdot (\ln \alpha_0 - \ln \alpha_1(\alpha)) = \alpha - \alpha_1(\alpha). \tag{2.6}$$

Setzt man (2.5) und (2.6) in (2.4) ein, so erhält man

$$\left(\frac{\partial \alpha_1(\alpha)}{\partial \alpha} \right) = \frac{\alpha_1(\alpha)}{\alpha_1(\alpha) - c_\alpha} \cdot \left(1 - \frac{\alpha - \alpha_1(\alpha)}{\alpha - c_\alpha} \right) > 0. \tag{2.7}$$

Dies beweist die Monotonie-Bedingung (2.2a) ∎

Für $\alpha_2 < \alpha$ ist die Situation wesentlich komplexer, weil es nun beliebig viele Möglichkeiten für die Wahl von α_1 und α_2 gibt. Darüber hinaus kann die Wahrscheinlichkeit für einen

Fehler 1. Art für unterschiedliche Niveaus $\alpha/|I|$ auf verschiedene Art und Weise auf die Zwischen- und die Endauswertung verteilt werden. Eine naheliegende und sinnvolle Möglichkeit besteht darin, das Verhältnis zwischen den Niveaus der Zwischen- und der Endauswertung für alle Signifikanzniveaus $\alpha/|I|, |I| = 1,...,k$, konstant zu halten:

$$\frac{\alpha_1(\alpha/|I|)}{c_{\alpha_2(\alpha/|I|)} \cdot (\ln\alpha_0 - \ln\alpha_1(\alpha/|I|))} = \kappa \quad \text{für} \quad |I| = 1,...,k,\ \kappa > 0. \tag{2.8}$$

Die folgenden Überlegungen zeigen, dass für jedes $\kappa > 0$ die Monotonie-Bedingungen (2.2a) und (2.2b) für die aus dieser Festlegung resultierenden Niveaus α_1 und α_2 erfüllt sind.

Satz 5:

Für ein konstantes Verhältnis $\kappa > 0$ zwischen den lokalen Signifikanzniveaus α_1 und α_2 der Zwischen- und der Endauswertung sind die Monotonie-Bedingungen (2.2a) und (2.2b) erfüllt.

Beweis:

Ein konstantes Verhältnis κ zwischen den Niveaus der Zwischen- und der Endauswertung bedeutet

$$\frac{\alpha_1(\alpha)}{c_{\alpha_2(\alpha)} \cdot (\ln\alpha_0 - \ln\alpha_1(\alpha))} = \kappa \quad \text{für alle} \quad 0 \le \alpha \le 1,\ \kappa > 0. \tag{2.9}$$

Daraus folgt

$$\alpha_1(\alpha) = \frac{\kappa}{(1+\kappa)} \cdot \alpha \tag{2.10}$$

und

$$c_{\alpha_2(\alpha)} = \frac{\frac{\alpha}{(1+\kappa)}}{\ln\frac{\alpha_0}{\kappa} - \ln\frac{\alpha}{(1+\kappa)}}. \tag{2.11}$$

Daraus ersieht man, dass sowohl $\alpha_1(\alpha)$ als auch $c_{\alpha_2(\alpha)}$ monoton wachsend in α sind. Berücksichtigt man nun noch, dass $\alpha_1(\alpha)$ monoton wachsend von $c_{\alpha_2(\alpha)}$ abhängt, so ist die Behauptung bewiesen. ∎

In Tabelle 1 sind Werte für $\alpha_1(\alpha/k)$ und $c_{\alpha/k}$ für $\alpha = 0.025$ und 0.05 und für $k = 1,...,5$ und $\alpha_0 = 0.3, 0.4, 0.5, 0.6, 0.7, 1.0$ angegeben; bei der Endauswertung wird jeweils Fisher´s Kombinationstest zum Niveau α/k angewendet. In Tabelle 2 sind die kritischen Niveaus

$\alpha_1(\alpha/k)$ und $\alpha_2(\alpha/k)$ für die Zwischen- und Endauswertung sowie die kritische Schranke $c_{\alpha_2(\alpha/k)}$ für das Produkt der *p*-Werte für die gleichen Werte für α, k und α_0 wie in Tabelle 1 angegeben; hier wird die Situation betrachtet, dass das Signifikanzniveau α/k gleichmäßig auf die Zwischen- und die Endauswertung aufgeteilt wird ($\kappa = 1$).

Tabelle 1: Lokales Signifikanzniveau $\alpha_1(\alpha/k)$ für die Zwischenauswertung und kritische Schranke $c_{\alpha/k}$ für Fisher´s Kombinationstest bei der Endauswertung für das adaptive Zwei-Stufen-Design nach BAUER und KÖHNE (1994) und das Gesamt-Niveau α/k. Dies entspricht der Situation, dass das lokale Signifikanzniveau $\alpha_2(\alpha/k) = \alpha/k$ bei der Endauswertung angewendet wird. α_0 bezeichnet die kritische Schranke für vorzeitige Beibehaltung.

			$\alpha = 0.025$		
	k	1	2	3	4	5
	α/k	0.025	0.0125	0.00833	0.00625	0.005
	$c_{\alpha/k}$	0.00380	0.00169	0.00106	0.00076	0.00059
α_0				$\alpha_1(\alpha/k)$		
0.3		0.0131	0.0058	0.0037	0.0026	0.0020
0.4		0.0115	0.0051	0.0032	0.0023	0.0018
0.5		0.0102	0.0045	0.0028	0.0020	0.0016
0.6		0.0090	0.0040	0.0025	0.0018	0.0014
0.7		0.0080	0.0036	0.0022	0.0016	0.0012
1.0		0.00380	0.00169	0.00106	0.00076	0.00059
			$\alpha = 0.05$		
	k	1	2	3	4	5
	α/k	0.05	0.025	0.0167	0.0125	0.01
	$c_{\alpha/k}$	0.00870	0.00380	0.00237	0.00169	0.00131
α_0				$\alpha_1(\alpha/k)$		
0.3		0.0299	0.0131	0.0081	0.0058	0.0045
0.4		0.0263	0.0115	0.0071	0.0051	0.0040
0.5		0.0233	0.0102	0.0063	0.0045	0.0035
0.6		0.0207	0.0090	0.0056	0.0040	0.0031
0.7		0.0183	0.0080	0.0050	0.0036	0.0027
1.0		0.00870	0.00380	0.00273	0.00169	0.00131

Tabelle 2: Lokale Signifikanzniveaus $\alpha_1(\alpha/k) = \alpha/k$ und $\alpha_2(\alpha/k)$ für die Zwischen- und die Endauswertung und kritische Schranken $c_{\alpha_2(\alpha/k)}$ für Fisher´s Kombinationstest bei der Endauswertung für das adaptive Zwei-Stufen-Design nach BAUER und KÖHNE (1994) und das Gesamt-Niveau α/k. Dies entspricht der Situation, dass das lokale Signifikanzniveau α/k gleichmäßig auf die Zwischen- und Endauswertung aufgeteilt wird. α_0 bezeichnet die kritische Schranke für vorzeitige Beibehaltung.

$\alpha = 0.025$					
	k	1	2	3	4	5
	α/k	0.025	0.0125	0.00833	0.00625	0.005
	$\alpha_1(\alpha/k)$	0.0125	0.00625	0.00417	0.003125	0.0025
α_0				$\alpha_2(\alpha/k)$ $c_{\alpha_2(\alpha/k)}$		
0.3		0.0257 0.00393	0.0120 0.00161	0.00773 0.00097	0.00567 0.00068	0.00447 0.00052
0.4		0.0239 0.00361	0.00113 0.00150	0.00730 0.00091	0.00538 0.00064	0.00424 0.00049
0.5		0.0227 0.00339	0.0108 0.00143	0.00700 0.00870	0.00517 0.00062	0.00409 0.00047
0.6		0.0217 0.00323	0.0104 0.00137	0.00678 0.00084	0.00501 0.00059	0.00397 0.00046
0.7		0.0210 0.00311	0.00101 0.00132	0.00660 0.00081	0.00488 0.00058	0.00387 0.00044
1.0		0.0196 0.00285	0.00948 0.00123	0.00622 0.00076	0.00462 0.00054	0.00366 0.00042
$\alpha = 0.05$					
	k	1	2	3	4	5
	α/k	0.05	0.025	0.0167	0.0125	0.01
	$\alpha_1(\alpha/k)$	0.00870	0.00380	0.00237	0.00169	0.00131
α_0				$\alpha_2(\alpha/k)$ $c_{\alpha_2(\alpha/k)}$		
0.3		0.0563 0.01006	0.0257 0.00393	0.0164 0.00233	0.0120 0.00161	0.00941 0.00122
0.4		0.0515 0.00902	0.0239 0.00361	0.0154 0.00215	0.0113 0.00150	0.00887 0.00114
0.5		0.0483 0.00835	0.227 0.00339	0.0146 0.00204	0.0108 0.00143	0.00850 0.00109
0.6		0.0460 0.00787	0.0217 0.00323	0.0141 0.00195	0.0104 0.00137	0.00821 0.00104
0.7		0.0442 0.00750	0.0210 0.00311	0.0137 0.00188	0.00101 0.00132	0.00799 0.00101
1.0		0.0406 0.00678	0.0196 0.00285	0.0128 0.00174	0.00948 0.00123	0.00752 0.00094

Den Tabellen kann man entnehmen, dass mit steigender Anzahl k der zu testenden Hypothesen die kritischen Niveaus für die Zwischen- und die Endauswertung ganz erheblich abfallen. Beispielsweise ist für $\alpha = 0.05, \alpha_0 = 0.5, k = 5$ und $\alpha_2(\alpha / k) = \alpha / k$ bei Anwendung der Bonferroni-Holm-Prozedur zumindest ein p-Wert $p_{i1} \leq 0.0035$ notwendig, um eine der Null-Hypothesen bereits bei der Zwischenauswertung ablehnen zu können.

Es sei angemerkt, dass es Situationen gibt, für die das lokale Niveau α_2 für die Endauswertung größer als das Gesamtniveau α ist (z.B. für $\alpha = 0.05, \kappa = 1$ und $(\alpha_0 = 0.3, k = 1, 2)$ sowie für $(\alpha_0 = 0.4, k = 1))$. Wenn die Bedingung $\alpha_2 \leq \alpha$ aufrecht erhalten und gleichzeitig das Gesamtniveau α ausgeschöpft werden soll, muss in diesen Fällen bei der Zwischenauswertung weniger als die Hälfte des gesamten Niveaus „verbraucht“ werden.

Bislang haben wir α_0 für alle Schritte der Bonferroni-Holm-Prozedur konstant gehalten. Im Prinzip könnte man α_0 abhängig von der Anzahl a bereits abgelehnter Null-Hypothesen wählen. Ein Grund für die Wahl einer in a wachsenden Schranke $\alpha_0(a)$ wäre die Überlegung, eine frühzeitige Beendigung der Studie mit der Beibehaltung der Global-Hypothese zu erleichtern, dagegen aber bei Ablehnung der Global-Hypothese eine vorzeitige Beibehaltung von Einzelhypothesen $H_0^i, i = 1,..., k,$ zu erschweren. Da unter allen bislang betrachteten Szenarien α_1 mit α_0 abnimmt, müsste dann $\alpha_0(a)$ so gewählt werden, dass weiterhin α_1 und α_2 mit a wachsen.

2.4 Reduktion der Hypothesenmenge nach der Zwischenauswertung

Bislang haben wir angenommen, dass bei Zwischen- und Endauswertung die gleiche Hypothesenmenge $\left\{ H_0^1, H_0^2, ..., H_0^k \right\}$ getestet wird. Adaptive Designs erlauben es, die Menge der zu testenden Null-Hypothesen auf der Basis von Informationen aus der aktuellen oder parallel laufenden bzw. abgeschlossenen Studie zu verändern. Wir behandeln im folgenden Satz 6 den Fall, dass im Rahmen der Endauswertung nur ein Teil der bei der Zwischenauswertung untersuchten Null-Hypothesen getestet wird. (Konsequenzen von Satz 6 auf andere Arten der Hypothesen-Adaption sind in den Bemerkungen 3 und 4 beschrieben.) Diese Option adaptiver Designs kommt beispielsweise in Dosis-Findungs-Studien zum Tragen, wenn die Ergebnisse der Zwischenauswertung für eine oder mehrere Dosis-Gruppe(n)

ein ungünstiges Nutzen-Risiko-Verhältnis indizieren und diese Dosierungen im zweiten Schritt der Studie nicht weiter untersucht werden (siehe BAUER, BAUER und BUDDE, 1998; BAUER und KIESER, 1999; LEHMACHER, KIESER und HOTHORN, 2000).

Ein weiteres Anwendungsfeld liegt im Bereich multipler Endpunkte, wenn sich eine der Zielgrößen, deren Erhebung invasive Maßnahmen erfordert oder kostenintensiv ist, bei der Zwischenauswertung als wenig sensitiv für den Nachweis eines Therapieeffektes erweist und deshalb im zweiten Studienabschnitt nicht mehr dokumentiert wird (siehe KIESER, BAUER und LEHMACHER, 1999).

Satz 6:

Im Rahmen des adaptiven Zwei-Stufen-Designs nach BAUER und KÖHNE (1994) werden nach dem ersten Studienabschnitt die Hypothesen $H_0^i, i \in K_1 = \{1,...,k_1\}$ untersucht, nach dem zweiten Studienabschnitt die Teilmenge $H_0^i, i \in K_2 \subseteq K_1$. Mit p_{I1} und p_{I2} seien die *p*-Werte zu definierten Niveau-α-Tests für die Null-Hypothesen $H_0^I = \bigcap_{i \in I} H_0^i, I \subseteq K_1,$ im ersten bzw. zweiten Studienabschnitt bezeichnet.

Fall I: $p_{K_1 1} \geq \alpha_0$

- Alle Null-Hypothesen $H_0^i, i \in K_1,$ werden beibehalten. Die Studie wird beendet.

Fall II: $p_{K_1 1} \leq \alpha_1$

- $H_0^{K_1}$ wird abgelehnt.
- Zusätzlich werden alle $H_0^i, i \in K_1,$ abgelehnt für die gilt: $p_{I1} \leq \alpha_1$ für alle $I \ni i,\ I \subseteq K_1$.

Fall III: $\alpha_1 < p_{K_1 1} < \alpha_0$

- Die Studie kann mit einem zweiten Abschnitt fortgesetzt werden.
- Im Rahmen der Endauswertung können die Null-Hypothesen $H_0^i, i \in K_2,$ abgelehnt werden für die gilt:

  $$\{p_{I1} \leq \alpha_1\} \cup \left(\{\alpha_1 < p_{I1} < \alpha_0\} \cap \{p_{I1} \cdot p_{I_2 2} \leq c_{\alpha_2}\}\right) \text{ für alle } I \ni i,\ I \subseteq K_1. \qquad (2.12)$$

  Dabei bezeichnet $I_2 = K_2 \cap I$.

Beweis:

Wir nehmen an, dass zum Test von H_0^I im zweiten Studienabschnitt abzählbar viele Testszenarien $T_{j2}, j \in J,$ zur Wahl stehen, die eine abzählbare Menge an verschiedenen

Stichprobenumfängen und Teststatistiken abdecken. Für jede Wahl T_{j2} ist der zugehörige *p*-Wert p_{I2} gleichverteilt auf dem Intervall $[0,1]$. Mit $f_0(p_{I_2 2}|T_{j2}, p_{I1})$ sei die Dichte von $p_{I_2 2}$ für eine gegebene Wahl von T_{j2} und gegebenen *p*-Wert p_{I1} unter H_0^I bezeichnet. Die bedingte Dichte existiert, falls die Studie fortgesetzt wird und ein geeignetes Testszenario T_{j2} ausgewählt wurde, und es gilt $f_0(p_{I_2 2}|T_{j2}, p_{I1}) = 1$ für jede Wahl T_{j2} (siehe BAUER und KIESER, 1999, S. 1846). Sei weiterhin $p_0(T_{j2}|p_{I1})$ die bedingte Wahrscheinlichkeit, bei gegebenem Wert von p_{I1} das Testszenario T_{j2} zu wählen. Wie $f_0(p_{I_2 2}|T_{j2}, p_{I1})$ muss auch $p_0(T_{j2}|p_{I1})$ nur definiert sein, wenn ein zweiter Studienabschnitt durchgeführt wird; die konkrete Form von $p_0(T_{j2}|p_{I1})$ ist in der Regel unbekannt und deren Kenntnis für die folgende Herleitung auch nicht erforderlich.

Es gilt

$$\Pr_{H_0^I}(\text{Ablehnung von } \mathrm{H}_0^{\mathrm{I}}) = \int_0^{\alpha_1} dp_{I1} + \int_{\alpha_1}^{\alpha_0} \Big[\sum_{j\in J} \int_0^{c_{\alpha_2 / p_{I1}}} \underbrace{f_0(p_{I_2 2}|T_{2j}, p_{I1})}_{=1} \cdot p_0(T_{2j}|p_{I1}) dp_{I_2 2} \Big] dp_{I1}$$

$$= \alpha_1 + \int_{\alpha_1}^{\alpha_0} \underbrace{\Big[\sum_{j\in J} p_0(T_{2j}|\ p_{11}) \Big]}_{\leq 1} \cdot (c_{\alpha_2} / p_{I1})\, dp_{I1}$$

$$\leq \alpha_1 + \int_{\alpha_1}^{\alpha_0} \left(c_{\alpha_2} / p_{I1}\right) dp_{I1} \underset{(2.1b)}{=} \alpha . \qquad \blacksquare$$

Bemerkungen:

1. Für $K_1 = K_2$, d.h., für den Fall, dass im ersten und zweiten Studienabschnitt die gleiche Hypothesenmenge getestet wird, resultiert aus der obigen allgemeinen Testprozedur die in Satz 1, Kapitel 2.2, beschriebene Abschlusstest-Prozedur.
2. Für $K_1 \setminus K_2 \neq \varnothing$ ist es möglich, dass auch Null-Hypothesen H_0^i, $i \in K_1 \setminus K_2$, die im zweiten Studienabschnitt nicht untersucht werden, im Rahmen der Endauswertung abgelehnt werden können. In diesem Fall kann eine Null-Hypothese H_0^i, $i \in K_1 \setminus K_2$, dann abgelehnt werden, wenn der zugehörige *p*-Wert im ersten Studienabschnitt unter der kritischen Schranke α_1 liegt, das gleiche gilt für die entsprechenden Schnitthypothesen $H_0^I = \bigcap_{i\in I} H_0^i$ mit $I \cap K_2 = \varnothing$. Insgesamt resultiert die folgende Entscheidungsregel: H_0^i, $i \in K_1 \setminus K_2$,

wird abgelehnt, falls Bedingung (2.12) gilt für alle $I \subseteq K_1$ mit $i \in I$ und $I \cap K_2 \neq \emptyset$, und falls gilt: $p_{I1} \leq \alpha_1$ für alle $I \subseteq K_1$ mit $i \in I$ und $I \cap K_2 = \emptyset$.

3. Aus dem Beweis folgt insbesondere, dass eine Schnitt-Hypothese H_0^I im zweiten Studienteil durch Test einer beliebigen Null-Hypothese $H_0^{I_2}, I_2 \subset I$, getestet werden kann und dass für die Auswahl dieser Hypothese $H_0^{I_2}$ die Ergebnisse der Zwischenauswertung verwendet werden können. KROPF, HOMMEL, SCHMIDT, BRICKWEDEL und JEPSEN (2000) und HOMMEL (2000) wiesen darauf hin, dass dies bei *a priori* geordneten Hypothesen und Fortsetzung der Studie nach dem ersten Studienteil einen Wechsel der Hypothesen-Hierarchie nach der Interimanalyse ermöglicht: Bezeichnet man beispielsweise mit $H_0^{i_1}$ bzw. $H_0^{i_2}$ die Null-Hypothesen, denen im ersten bzw. zweiten Studienteil die höchste Priorität zugewiesen wird, so können bei der Endauswertung im Rahmen der Abschlusstest-Prozedur alle Hypothesen H_0^I mit $i_2 \in I$ durch einen Niveau-α-Test von $H_0^{i_2}$ getestet und abgelehnt werden, falls gilt $p_{I1} \cdot p_{I2} = p_{i_1 1} \cdot p_{i_2 2} \leq c_{\alpha_2}$; diese Bedingung sichert dann die Ablehnung von $H_0^{i_2}$ bei der Endauswertung. Mit den gleichen Argumenten sieht man, dass nach der Zwischenauswertung beliebige Veränderungen in der Hypothesen-Hierarchie vorgenommen werden können.
4. HOMMEL (2000) argumentiert, dass auf der Grundlage des obigen Beweises nach der Zwischenauswertung auch neue Null-Hypothesen zu der bei Studienbeginn festgelegten Hypothesen-Familie hinzugenommen werden können: Sei $H_0^{i^*}$ eine Hypothese mit $i^* \notin \{1,...,k_1\}$, dann ist der Test von $H_0^{K_1}$ im ersten Schritt auch ein Test von $H_0^{K_1} \cap H_0^{i^*}$ gewesen, und diese Schnitt-Hypothese kann (wie alle anderen Schnitt-Hypothesen, die $H_0^{i^*}$ enthalten) bei der Endauswertung z.B. durch Test von $H_0^{i^*}$ getestet werden. Zur Ablehnung von $H_0^{i^*}$ ist insbesondere ein *p*-Wert aus der zweiten Stufe von $p_{i^* 2} \leq \alpha$ notwendig. Für die zugehörige Fragestellung entspricht damit die Hinzunahme der Null-Hypothese nach der Zwischenauswertung der Durchführung einer neuen Studie im Design mit festem Stichprobenumfang und Niveau α. Ob diese Strategie deshalb tatsächlich einen Vorteil verspricht, hängt von der konkreten Anwendungssituation ab.

Satz 6 und sein Beweis zeigen, dass im Rahmen des adaptiven Designs unter Einhaltung der Wahrscheinlichkeit für einen Fehler 1. Art auf der Basis der Ergebnisse der

Zwischenauswertung (und/oder von Informationen, die von anderen parallel laufenden oder abgeschlossenen Studien herrühren) folgende Design-Charakteristika für den zweiten Studienabschnitt frei gewählt werden können:

- die zu untersuchenden Hypothesen H_0^i, $i \in K_2 \subseteq K_1$,
- die zu verwendenden Teststatistiken,
- der Stichprobenumfang.

In den nachfolgenden Kapiteln 3, 4 und 5 werden Methoden vorgestellt, die es erlauben, diese Optionen effizient zu nutzen. Zuvor soll die Anwendung der in diesem Kapitel dargestellten multiplen Testprozeduren an konkreten klinischen Prüfungen illustriert werden.

2.5 Spezielle Anwendungssituationen mit Beispielen

In diesem Kapitel werden für zwei wichtige Anwendungssituationen multipler Testprozeduren in klinischen Studien die in den Abschnitten 2.2-2.4 vorgestellten Verfahren konkretisiert: Studien mit mehr als einer Zielgröße und Studien mit mehr als zwei Behandlungsgruppen. Es wird die Struktur der Hypothesenfamilie angegeben, und die praktische Anwendung der multiplen Testprozeduren wird anhand von zwei klinischen Studien illustriert, die im adaptiven Zwei-Stufen-Design nach BAUER und KÖHNE (1994) durchgeführt wurden.

2.5.1 Multiple Endpunkte

In zweiarmigen klinischen Studien, bei denen k Zielgrößen im Rahmen der konfirmatorischen Auswertung untersucht werden, können die zugrunde liegenden Testprobleme bei normalverteilten Endpunkten mit Erwartungswerten μ_1^i bzw. μ_2^i, $i = 1,...,k$, in Gruppe 1 bzw. 2 formuliert werden als $H_0^i : \mu_1^i = \mu_2^i$ vs. $H_1^i : \mu_1^i < \mu_2^i$, $i = 1,...,k$. Je nach inhaltlicher Fragestellung wird ein Testverfahren für *a priori* geordnete oder für gleichberechtigte Null-Hypothesen angewendet. Für die Anwendung der Abschlusstest-Prozedur zur Analyse multipler Endpunkte steht eine Vielzahl von Verfahren zum Test von Schnitthypothesen $H_0^I = \bigcap_{i \in I} H_0^i$, $I \subseteq \{1,...,k\}$, zur Verfügung (für eine Übersicht siehe z.B.

WASSMER, REITMEIR, KIESER und LEHMACHER, 1999). Die Eigenschaften dieser Testverfahren im Rahmen der Abschlusstest-Prozedur wurden untersucht in den Arbeiten von LEHMACHER, WASSMER und REITMEIR (1991), KIESER, REITMEIR und WASSMER (1995) und REITMEIR und WASSMER (1996).

Beispiel 1:

In einer randomisierten, doppelblinden, placebokontrollierten klinischen Studie wurde die Wirksamkeit und Verträglichkeit des Kava-Kava Spezialextraktes WS 1490 bei Patienten mit nervösen Angst-, Spannungs- und Unruhezuständen nicht-psychotischer Genese untersucht (MALSCH und KIESER, 2000). In die Studie wurden Patienten aufgenommen, die zuvor über mindestens zwei Wochen ununterbrochen mit Benzodiazepinen behandelt worden waren und bei denen eine medizinische Indikation zur Beendigung der Bezodiazepin-Behandlung und zu einem Wechsel auf ein alternatives anxiolytisches Medikament bestand. Während der ersten Behandlungswoche nach Randomisierung wurde die Tagesdosis der Studienmedikation allmählich von 50 mg auf 300 mg erhöht. Parallel dazu wurde die Benzodiazepin-Behandlung im Verlauf der ersten beiden Wochen abgesetzt. An diese Phase der Dosis-Anpassung schloss sich eine dreiwöchige alleinige anxiolytische Behandlung mit der Studienmedikation an. Da dies die erste klinische Studie zur Untersuchung der Wirksamkeit von WS 1490 bei einer Vorbehandlung mit Benzodiazepinen war, bot sich ein adaptives Design an. Die Studie wurde mit zwei Stufen nach dem Vorschlag von BAUER und KÖHNE (1994) mit Fisher´s Kombinationstest und mit den Spezifikationen $\alpha = 0.05$ (einseitig), $\alpha_2 = \alpha$ und $\alpha_0 = 0.6$ geplant. Aus diesen Definitionen resultierte ein lokales einseitiges Signifikanzniveau von $\alpha_1 = 0.0207$ für die Zwischenauswertung. Eine Interimanalyse war nach 40 abgeschlossenen Patienten vorgesehen. Die $k = 3$ Zielgrößen für die konfirmatorische Analyse waren die Hamilton-Angstskala (HAMA), (HAMILTON, 1976), die Befindlichkeitsskala (Bf-S) (VON ZERSSEN, 1976) und die Rate der Patienten mit Entzugssymptomen. Die entsprechenden Hypothesen wurden entsprechend ihrer Wichtigkeit *a priori* in dieser Reihenfolge angeordnet.

Die *p*-Werte in der Zwischenauswertung lauteten $p_{11} = 0.0103, p_{21} = 0.0032$ (jeweils Mann-Whitney U-Test, einseitig) und $p_{31} = 0.215$ (Chi-Quadrat-Test, einseitig). Damit konnten H_0^1 und H_0^2 nach dem ersten Studienabschnitt abgelehnt werden. Die Studie wurde mit diesem Ergebnis und der Beibehaltung von H_0^3 beendet, da die primären Studienziele erreicht waren. Zur Ablehnung von H_0^3 wäre im zweiten Studienabschnitt ein *p*-Wert

$p_{31} \le c_\alpha / p_{31} = 0.040$ notwendig gewesen. Die Fallzahlplanung für den zweiten Studienabschnitt wäre folglich zu diesem Signifikanzniveau durchgeführt worden (siehe Kapitel 4.2).

Zur Illustration nehmen wir nun an, dass die Null-Hypothesen nicht *a priori* geordnet waren, sondern dass die Bonferroni-Holm-Prozedur angewendet werden soll. In diesem Fall wären folgende kritischen Niveaus im Rahmen der Interimanalyse zugrunde zu legen: $\alpha_1(\alpha/3) = 0.0056, \alpha_1(\alpha/2) = 0.0090, \alpha_1(\alpha) = 0.0207$. Dementsprechend könnte wegen $p_{(1)1} = p_{21} = 0.0032 < \alpha_1(\alpha/3)$ die Null-Hypothese H_0^2 nach dem ersten Studienabschnitt abgelehnt werden, aber wegen $\alpha_1(\alpha/2) < p_{(2)1} = p_{11} < \alpha_0$ und $p_{(3)1} = p_{31} < \alpha_0$ könnte keine vorzeitige Entscheidung über H_0^1 und H_0^3 getroffen werden. Um sowohl H_0^1 als auch H_0^3 nach dem zweiten Schritt ablehnen zu können, müsste einer der entsprechenden *p*-Werte unter die kritische Schranke $c_\alpha(\alpha/2) = 0.0038$ fallen und der andere unter $c_\alpha(\alpha/2) = 0.0087$. Nehmen wir an, dass nach dem zweiten Studienabschnitt das Produkt der *p*-Werte für H_0^3 die Bedingung $p_{31} \cdot p_{32} \le c_\alpha(\alpha/2)$ (d.h. $p_{32} \le 0.017$) erfüllt. Klarerweise könnte dann H_0^3 abgelehnt werden. Zusätzlich könnte dann aber auch H_0^1 abgelehnt werden, unabhängig vom Ergebnis dieser Null-Hypothese im zweiten Studienabschnitt (selbst wenn diese Hypothese dort überhaupt nicht untersucht worden wäre). Dies liegt daran, dass nach Ablehnung von H_0^3 die kritische Schranke für den *p*-Wert p_{11} gegeben ist durch $\alpha_1(\alpha) = 0.0207$ und folglich $p_{11} < \alpha_1(\alpha)$ erfüllt ist.

2.5.2 Mehrarmige Studien

Wir nehmen an, dass im Rahmen einer klinischen Studie k_1 Dosierungen eines Medikamentes und eine Kontrollbehandlung untersucht werden und dass die Wirksamkeit anhand einer normalverteilten Zielgröße beurteilt wird. Ein typisches multiples Testproblem ist dann gegeben durch die k_1 Vergleiche der Dosis-Gruppen gegen die Kontrolle („many to one“) $H_0^i : \mu_0 \ge \mu_i$ versus $H_1^i : \mu_0 < \mu_i, i = 1,...,k_1,$ wobei $\mu_0, \mu_1,..., \mu_{k_1}$ die Erwartungswerte für die Zielgröße unter Behandlung mit der Kontrolle und den k_1 Dosis-Gruppen bezeichnen. Die Ablehnung der globalen Schnitthypothese $H_0 = \bigcap_{i=1}^{k_1} H_0^i$ besagt, dass mindestens eine der k_1

Dosis-Gruppen der Kontroll-Gruppe überlegen ist. Die zusätzliche Ablehnung von Null-Hypothesen $H_0^i, i = 1,...,k_1$, im Rahmen einer multiplen Testprozedur erlaubt weitergehende Aussagen darüber, welche der Dosis-Gruppen sich von der Kontroll-Gruppe unterscheiden.

Falls eine Ordnungsrelation der Form $\mu_0 \le \mu_1 \le ... \le \mu_{k_1}$ für die Erwartungswerte der Zielgröße in den Behandlungsgruppen vorausgesetzt werden kann, reduziert sich die Menge der Schnitthypothesen $H_0^I = \bigcap_{i \in I} H_0^i, I \subseteq \{1,...,k_1\}$, erheblich, denn in diesem Fall gilt $H_0^i \cap H_0^j = H_0^{\max(i,j)}$. Damit vereinfacht sich auch die zugehörige Abschlusstest-Prozedur, denn eine Null-Hypothese $H_0^i, i = 1,...,k_1$, wird im Rahmen derselben genau dann abgelehnt, wenn alle H_0^j mit $j \ge i$ abgelehnt werden.

Die in den Kapiteln 2.2-2.4 beschriebenen multiplen Testprozeduren sind ebenso auf andere multiple Testprobleme bei mehrarmigen Studien anwendbar, wie z.B. alle paarweisen Vergleiche zwischen den Behandlungsgruppen. Aus Platzgründen soll auf die entsprechende Hypothesenstruktur hier nicht weiter eingegangen werden.

Beispiel 2:

In einer randomisierten, placebokontrollierten Doppelblindstudie wurde die Wirksamkeit und Verträglichkeit zweier *Hypericum*-Extrakte mit unterschiedlichem Hyperforin-Gehalt bei Patienten mit leichten bis mittelschweren depressiven Episoden untersucht (LAAKMANN, SCHÜLE, BAGHAI und KIESER, 1998). Obwohl die Wirksamkeit von *Hypericum*-Extrakten in zahlreichen klinischen Prüfungen nachgewiesen wurde, ist der Beitrag der mehr als 15 Inhaltsstoffe zum pharmakologischen Effekt noch nicht vollständig aufgeklärt. Jüngere Forschungsergebnisse legen nahe, dass neben Hypericin auch Hyperforin und Adhyperforin eine wichtige Rolle spielen. Die vorliegende klinische Studie war die erste ihrer Art, die die Relevanz des Hyperforin-Gehaltes des *Hypericum*-Extraktes für die klinische Wirksamkeit untersuchte. Diese Pilotstudien-Situation war geradezu prädestiniert für die Implementierung eines adaptiven Designs. Die Studie wurde im Zwei-Stufen-Design nach BAUER und KÖHNE (1994) durchgeführt. Im ersten Studienabschnitt wurden insgesamt $n = 147\ (3 \times 49)$ Patienten der Placebo-Gruppe ($j = 0$) oder eine der aktiven Behandlungsgruppen randomisiert zugewiesen: *Hypericum*-Extrakt mit einem Hyperforin-Gehalt von 0.5% ($j = 1$) oder 5% ($j = 2$). Nach einer Run-in-Phase von 3-7 Tagen erhielten die Patienten 42 Tage 3×300 mg der randomisierten Medikation. Primäre Zielgröße war die Veränderung des Gesamtscores der

Hamilton-Depressions-Skala (HAMD) (HAMILTON, 1986) zwischen Tag 0 und Tag 42. Aufgrund der präklinischen Untersuchungen konnte eine Ordnungsrelation mit einem mit dem Hyperforin-Gehalt wachsenden Behandlungseffekt angenommen werden. Als Auswertungsmethode war Fisher´s Kombinationstest mit den folgenden Charakteristika festgelegt: Die Wahrscheinlichkeit eines einseitigen Fehlers 1. Art war $\alpha = 0.05$, die nach dem ersten und zweiten Studienabschnitt zu verwendenden Niveaus waren festgelegt als $\alpha_1 = 0.0299$ bzw. $\alpha_2 = 0.05$ bei einer kritischen Schranke für die vorzeitige Beibehaltung von Null-Hypothesen $\alpha_0 = 0.30$. Zum Test von $H_0^{\{1,2\}} = \bigcap_{i=1}^{2} H_0^i$ wurde der nicht-parametrische Jonckheere-Terpstra-Test (JONCKHEERE, 1954) verwendet, zum Test von H_0^1 der Mann-Whitney U-Test.

Nach dem ersten Studienabschnitt ergab die Intention-to-treat-Auswertung einen *p*-Wert von $p_{\{1,2\}1} = 0.017$. Damit konnte $H_0^{\{1,2\}}$ im Rahmen der Zwischenauswertung abgelehnt werden, womit die Wirksamkeit des *Hypericum*-Extraktes mit dem höheren Hyperforin-Gehalt nachgewiesen wurde. Für den *p*-Wert p_{11} ergab sich $\alpha_1 < p_{11} = 0.19 < \alpha_0$. Die Null-Hypothese H_0^1 konnte damit nicht vorzeitig abgelehnt werden. Um diese Null-Hypothese nach einem zweiten Studienabschnitt ablehnen zu können, wäre ein *p*-Wert von $p_{12} \leq c_{\alpha_2} / p_{11} = 0.046$ notwendig gewesen. Aufgrund der Tatsache, dass für den *Hypericum*-Extrakt mit dem niedrigeren Hyperforin-Gehalt der Unterschied zur Placebo-Gruppe nicht klinisch relevant war, sondern sich diese Gruppe bezüglich aller Wirksamkeitskriterien nur unwesentlich von der Placebo-Gruppe unterschied, wurde die Studie mit der Ablehnung von $H_0^{\{1,2\}}$ und der Beibehaltung von H_0^1 nach der Zwischenauswertung beendet.

3. Adaptive Auswahl von Behandlungsgruppen

Die Untersuchung der Dosis-Wirkungs-Abhängigkeit eines Medikamentes in sogenannten Dose-Response-Studien gehört zu den komplexesten und gleichzeitig zu den wichtigsten Aufgaben im Rahmen der Arzneimittel-Entwicklung. In seinem Überblick nennt RUBERG (1995a, 1995b) vier Fragen, die es in Dose-Response-Studien zu beantworten gilt: „(1) Is there any drug effect? (2) What doses exhibit a response different from control? (3) What is the nature of dose response relationship? (4) What is the optimal dose?“ Für eine fundierte Planung einer Studie mit derart anspruchsvollen und vielfältigen Zielen sind detaillierte Vorinformationen erforderlich. Auf der anderen Seite werden Dose-Response-Studien gemäß ihrer Funktion in einer frühen Phase der Arzneimittel-Entwicklung durchgeführt, wo vergleichsweise wenig über die Charakteristika eines Medikamentes bekannt ist. Dieses Dilemma ist der Grund dafür, dass Dose-Response-Studien traditionell im explorativen Ansatz in der Phase II der klinischen Entwicklung durchgeführt wurden. Ziel der Studien dieser Phase ist es, die aussichtsreichste Dosierung zu identifizieren; mit dieser Dosis werden anschließend die für die Zulassung geforderten konfirmatorischen Phase III-Studien zum ultimativen Nachweis der Wirksamkeit durchgeführt. Dieser Ansatz ist insbesondere deshalb unbefriedigend, weil durch die ICH-Guideline E4 „Dose-Response Information to Support Drug Registration“ (ICH, 1994) die Möglichkeit eingeräumt wurde, durch den Nachweis einer Dosis-Wirkungs-Beziehung die Wirksamkeit eines Medikamentes zu zeigen. Damit kann eine im konfirmatorischen Rahmen durchgeführte Dose-Response-Studie mit positivem Ergebnis die Rolle einer pivotalen Phase III-Studie übernehmen.

Adaptive Mehr-Stufen-Designs sind geeignet, diese antagonistischen Anforderungen an Dose-Response-Studien zu versöhnen. Die grundlegende Idee besteht darin, im ersten Schritt die Dosis-Wirkungs-Abhängigkeit zu untersuchen. Falls bis zur Zwischenauswertung das Studienziel nicht erreicht ist, kann die zu diesem Zeitpunkt vorliegende Information (aus dem ersten Studienteil und gegebenenfalls auch aus anderen Studien, die zwischenzeitlich abgeschlossen wurden) für die Planung des folgenden Studienteils genutzt werden, insbesondere für die Wahl der Behandlungsgruppe, für die die Wirksamkeit nachgewiesen werden soll. Im adaptiven Zwei-Stufen-Design kann dies wie folgt bewerkstelligt werden. Wir nehmen im folgenden an, dass in einer klinischen Studie k Dosis-Gruppen und eine Kontrolle untersucht werden. Nach dem ersten Studienteil wird die Null-Hypothese $H_{01} = \bigcap_{i=1}^{k} H_0^i$ mit $H_0^i : \mu_0 \geq \mu_i, i = 1,...,k,$ getestet. Unter der Annahme, dass höhere Werte

für die Zielgröße einer besseren Wirksamkeit entsprechen, besagt die Null-Hypothese H_{01}, dass keine der k Dosierungen der Kontrolle überlegen ist. Falls die Null-Hypothese H_{01} im Rahmen der Zwischenauswertung nicht abgelehnt werden kann und die Studie mit einem zweiten Prüfungsteil fortgesetzt wird, wird dort die Null-Hypothese $H_0^{\hat{w}}$ getestet, wobei $\hat{w} \in \{1,...,k\}$ die Dosis-Gruppe bezeichnet, die bei der Zwischenauswertung zur weiteren Untersuchung ausgewählt wurde. Die Ablehnung von $H_0 = H_{01} (= H_{01} \cap H_0^{\hat{w}})$ zeigt, dass für mindestens eine der Dosierungen ein Therapieeffekt besteht, womit die Frage (1) von RUBERG (1995a) beantwortet ist. Die in BAUER und KIESER (1999) und in Kapitel 2 angegebenen multiplen Testprozeduren ermöglichen nach Ablehnung der globalen Null-Hypothese kontrollierte Inferenz für die Einzel-Hypothesen $H_0^i, i = 1,...,k,$ und damit die Beantwortung der Frage (2).

Es ist evident, dass für die Beantwortung der Fragen (3) und (4) sowie für die Effektivität der oben beschriebenen Prozedur im Rahmen des adaptiven Zwei-Stufen-Designs die Güte der Methode, die zur Auswahl der besten Dosis-Gruppe verwendet wird, essentiell ist. Wir werden Selektionsregeln vorschlagen sowie deren Charakteristika untersuchen. Dabei wird die Situation betrachtet, dass die Dosis-Wirkungs-Abhängigkeit durch ein Modell beschrieben wird, bei dem die Wirksamkeit bis zu einer Dosis-Gruppe $w \in \{1,...,k\}$ linear anwächst und für die höheren Dosierungen auf diesem Plateau verbleibt. Für dieses Plateau-Modell besteht das Ziel der Selektion darin, die niedrigste Dosis-Gruppe mit maximaler Wirksamkeit auszuwählen, die sogenannte niedrigste Plateau-Dosis. Unter der für die Mehrzahl von Medikamenten gültigen Annahme, dass eine höhere Dosierung mit einem ungünstigeren Verträglichkeitsprofil verbunden ist, beinhaltet diese Selektionsstrategie implizit auch den Aspekt der Anwendungssicherheit. Die statistische Formulierung des Plateau-Modells wird in Kapitel 3.1 dargestellt. In Kapitel 3.2 werden dann verschiedene Kriterien zur Beurteilung der Güte von Selektionsregeln angegeben. Anschließend werden in Kapitel 3.3 vier Schätzmethoden für die gesuchte Dosis-Gruppe w vorgestellt, und diese Verfahren werden in Kapitel 3.4 bezüglich der zuvor eingeführten Gütekriterien verglichen. In Kapitel 3.5 wird untersucht, wie die oben beschriebene Strategie unter Verwendung einer solchen Selektionsregel im Rahmen des adaptiven Zwei-Stufen-Designs im Vergleich zum „klassischen“ nicht-adaptiven Design unter verschiedenen Szenarien abschneidet.

Eine erste Selektionsregel für das Plateau-Modell wurde in der Arbeit von BAUER, BAUER und BUDDE (1998) angegeben, und dort, wie auch in BAUER und KIESER (1999) für Simulationsuntersuchungen verwendet. Die im folgenden vorgestellten alternativen

Vorschläge zur Wahl der Dosis-Gruppe und die Betrachtungen zur Leistungsfähigkeit der Verfahren basieren auf den Arbeiten FRIEDE und KIESER (1999) und FRIEDE, MILLER, BISCHOFF und KIESER (2000). Der Beweis, dass das Zwei-Stufen-Design für diese Art von Adaption und den Test von H_0 das Niveau kontrolliert (und im Rahmen der Abschlusstest-Prozedur das multiple Niveau) findet sich in Satz 6, Kapitel 2.4.

3.1 Statistisches Modell für die Dosis-Wirkungs-Abhängigkeit

Wir betrachten im folgenden mehrarmige klinische Studien mit $k+1$ Dosis-Gruppen und nehmen der Einfachheit halber an, dass in jeder Behandlungsgruppe m Beobachtungen vorliegen. Die Beobachtungen $X_{ij}, i = 0, ..., k, j = 1, ..., m$, nehmen wir als unabhängig und normalverteilt mit gleicher aber unbekannter Varianz σ^2 an. Eine Dosis-Wirkungs-Abhängigkeit, bei der die Wirksamkeit linear anwächst bis ein Plateau an der Stelle w erreicht wird, kann dann wie folgt modelliert werden:

$$X_{ij} = \beta_1 + \beta_2 \min(0, i - w) + \varepsilon_{ij}. \qquad (3.1)$$

Dabei bezeichnen

- β_1 und $\beta_2 \geq 0$ die unbekannten Regressionsparameter,
- $w \in \{1, ..., k\}$ die unbekannte Dosis-Gruppe, für die das Wirksamkeits-Plateau erreicht wird,
- ε_{ij} die unabhängigen und normalverteilten Fehler mit Erwartungswert 0 und unbekannter Varianz σ^2.

In Abbildung 3 sind für $k = 6$ und $w = 1, ..., 6$ die Verläufe der Erwartungswerte $\mu_i = \beta_1 + \beta_2 \min(0, i - w), i = 1, ..., 6,$ dargestellt.

Die Dosierung w kann als Change point im Regressionsmodell (3.1) angesehen werden. In Kapitel 3.3 werden Auswahlregeln vorgestellt, die auf der Schätzung des Change points unter diesem Modell basieren. Da wir zur Konstruktion einer Sub-Klasse solcher Schätzer Kriterien zu deren Bewertung benötigen, werden diese zuvor im folgenden Kapitel 3.2 eingeführt.

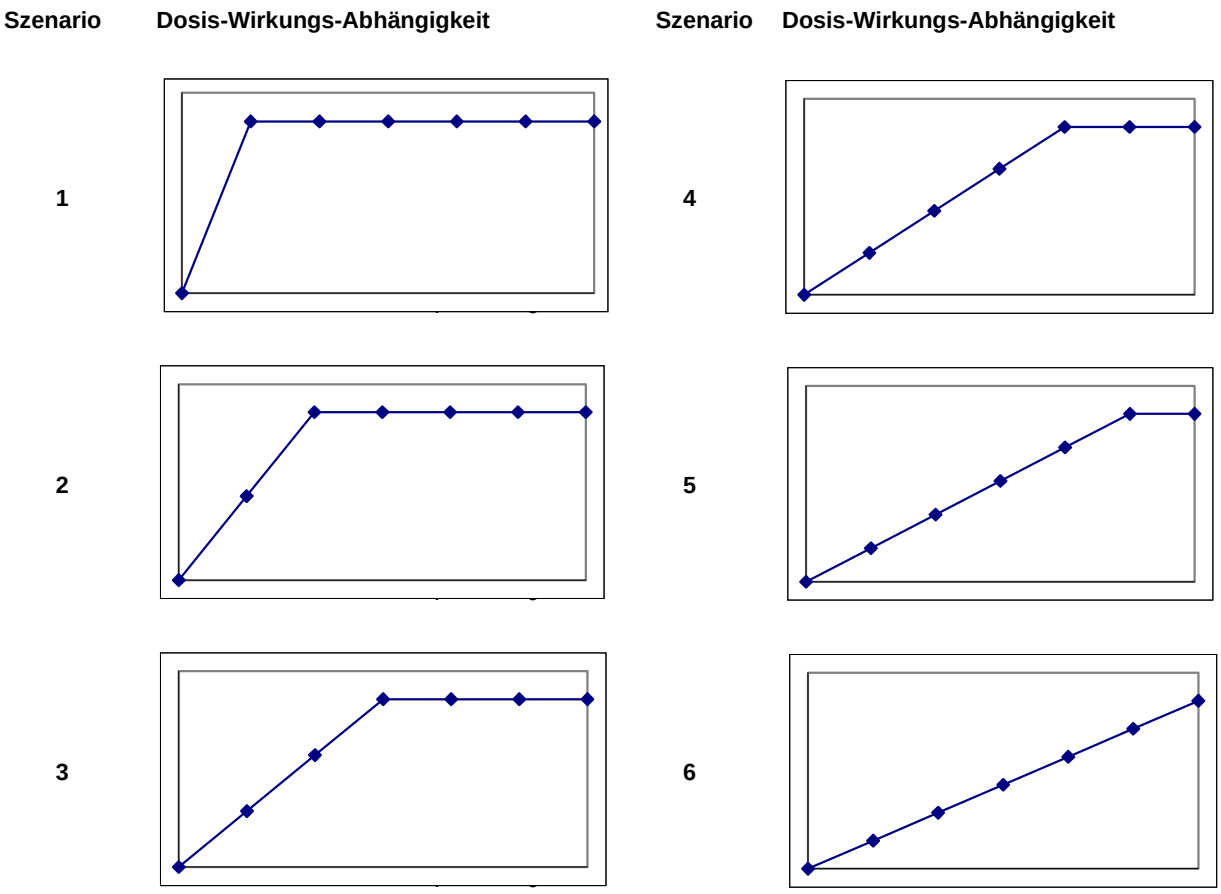

Abbildung 3: Verlauf der Dosis-Wirkungs-Abhängigkeit unter Modell (3.1) für $k = 6$. Szenario i, $i = 1, ..., 6$, entspricht der Situation, dass der Change point für die Gruppe $w = i$ vorliegt. Abszisse: Dosis-Gruppe, Ordinate: Erwartungswert.

3.2 Gütekriterien zur Bewertung von Auswahlregeln

Bei den im folgenden betrachteten Auswahlregeln wird so vorgegangen, dass unter dem Regressionsmodell (3.1) der unbekannte Change point w auf der Basis des Beobachtungsvektors $X = (X_{ij}), i = 0, ..., k, j = 1, ..., m,$ mit einer Schätzfunktion U geschätzt wird und die Dosis-Gruppe mit dem Index $\hat{w} = U(X)$ selektiert wird. Die naheliegendste Art und Weise, verschiedene Schätzfunktionen zu vergleichen, besteht darin, die *Rate korrekter Schätzungen* zu betrachten. Andere Vergleichskriterien können im Rahmen der Entscheidungstheorie formuliert werden. Dabei misst eine sogenannte *Verlustfunktion* $l(w, \hat{w})$ den Abstand zwischen dem zu schätzenden Parameter w (in unserem Fall dem Index der niedrigsten Plateau-Dosis) und der Schätzung $\hat{w} = U(X)$. Das sogenannte *Risiko* ist definiert als der

Erwartungswert der Verlustfunktion für gegebene Schätzfunktion U und gegebenen Parameterwert w:

$$R(w;U) = E\left[l(w, U(X)\right].$$

Die Abhängigkeit des Risikos vom betrachteten Parameterwert kann dazu führen, dass verschiedene Schätzfunktionen für unterschiedliche Parameterwerte in unterschiedlichem Größenverhältnis zueinander stehen und damit der Gütevergleich abhängig vom speziell betrachteten (unbekannten!) Parameterwert w ist. Eine Möglichkeit, dieses Manko zu eliminieren, besteht darin, das *maximale Risiko* über alle möglichen Parameter w zu betrachten:

$$R_{\max}(U) = \max_{w\in\{1,\ldots,k\}} R(w;U)$$

Das Risikoprofil einer Schätzfunktion U kann auch wie folgt zusammengefasst werden: Es wird eine *a priori* Verteilung a für den dann als Zufallsvariable aufgefassten Parameter W mit Realisationen w eingeführt und der Erwartungswert des Risikos bezüglich dieser Verteilung betrachtet. Man erhält dann das sogenannte *Bayes-Risiko*

$$R_{Bayes}(U) = E\left[R(W;U)\right]$$

$$= \sum_{w=1}^{k} R(w;U)a(w).$$

Wir werden für unsere Untersuchungen im folgenden als *a priori* Verteilung stets die Gleichverteilung annehmen

$$a(w) = P(W = w) = \frac{1}{k} \quad \text{für alle } w\in\{1,\ldots,k\}.$$

Als Verlustfunktionen betrachten wir zum einen die quadratische Verlustfunktion

$$l_2(w,\hat{w}) = (\hat{w} - w)^2,$$

die symmetrisch ist und damit eine Über- und eine Unterschätzung des wahren Parameterwertes mit dem gleichen Verlust belegt. Darüber hinaus betrachten wir die Klasse der sogenannten LINEX-Verlustfunktionen, die durch VARIAN (1975) eingeführt wurde (siehe auch BISCHOFF, FIEGER und WULFERT, 1995):

$$l_{LINEX}(w,\hat{w}\,|\,\lambda,s) = s\cdot(e^{\lambda(\hat{w}-w)} - \lambda(\hat{w}-w) - 1),$$

wobei $\lambda \neq 0$ und $s > 0$ Parameter bezeichnen, die die Form des Verlaufs der Verlustfunktion festlegen. Die LINEX-Verlustfunktion wird für $\hat{w} \leq w$ durch den linearen Anteil dominiert und für $\hat{w} > w$ durch den exponentiellen Term. In Abbildung 4 sind LINEX-Verlustfunktionen für verschiedene Parameterwerte λ und s dargestellt.

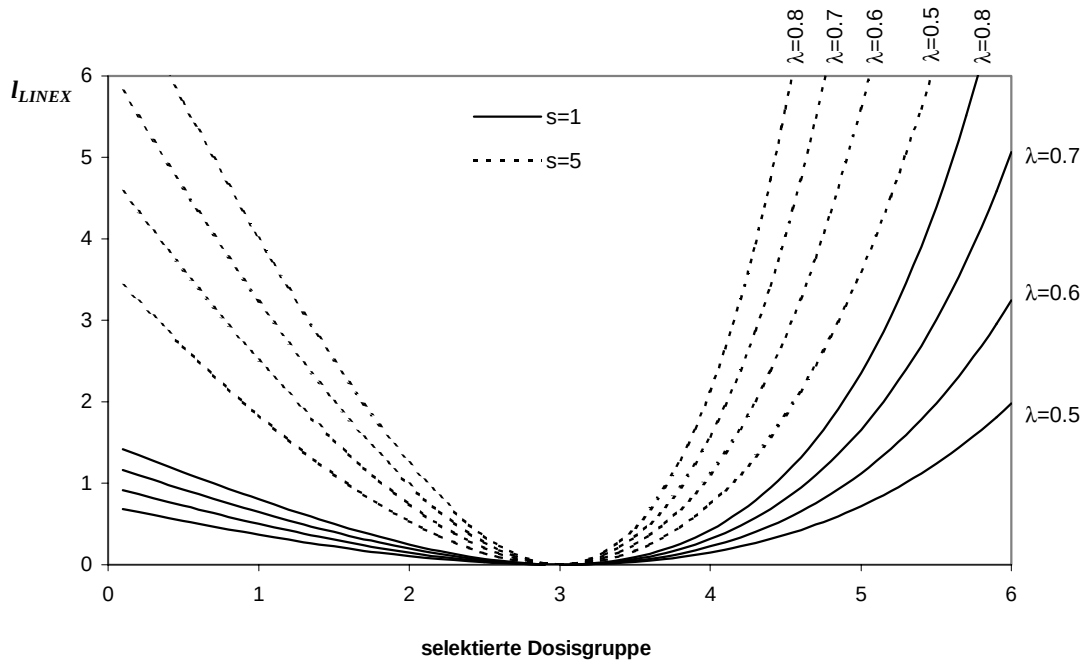

Abbildung 4: Verlauf der LINEX-Verlustfunktion l_{LINEX} für k = 6 und den Change point w = 3 für die Parameterwerte λ = 0.5, ..., 0.8 und s = 1, 5.

Im Unterschied zur quadratischen Verlustfunktion ist l_{LINEX} asymmetrisch bezüglich w. Diese Eigenschaft macht die Klasse der LINEX-Verlustfunktionen insbesondere attraktiv für das vorliegende Problem der Auswahl von optimalen Behandlungsgruppen: In dieser Situation hat eine Über- bzw. Unterschätzung in der Praxis unterschiedliche Auswirkungen, was sich entsprechend bei dem in Rechnung zu stellenden Verlust niederschlagen sollte. Beispielsweise sind höhere Dosierungen in der Regel mit einem ungünstigeren Sicherheitsprofil verbunden. Damit ist eine Überschätzung der Plateau-Dosis mit keinem Gewinn an Wirksamkeit verknüpft, wohingegen sich die Verträglichkeit unter Umständen erheblich schlechter darstellt als für die Plateau-Dosis. Im Unterschied dazu hat eine Unterschätzung, die „lediglich“ mit einer niedrigeren Wirksamkeit einhergeht, oftmals weniger dramatische Konsequenzen. Die Klasse der LINEX-Verlustfunktionen erlaubt es, über den Parameter λ die Form der Verlustfunktion so zu wählen, dass sie in der konkreten praktischen Situation die Konsequenzen einer Über- bzw. Unterschätzung entsprechend gewichtet. Es sei angemerkt, dass für sehr kleine Werte von $|\lambda|$ die LINEX-Verlustfunktion nahezu symmetrisch und die Form sehr ähnlich zur quadratischen Verlustfunktion ist.

3.3 Auswahlregeln, die auf der Schätzung eines Change points basieren

3.3.1 Helmert-Schätzer

Der Helmert-Schätzer (H) wurde bereits in der Arbeit von BAUER, BAUER und BUDDE (1998) verwendet und basiert auf einer Drei-Schritt-Prozedur, bei der in der zweiten Stufe der Berechnung Helmert-Kontraste der Behandlungsgruppen-Mittelwerte verwendet werden.

1. Schritt:

In diesem Schritt wird der Tatsache Rechnung getragen, dass sich die als monoton vorausgesetzte Dosis-Wirkungs-Abhängigkeit in der Praxis für höhere Dosierungen umkehren kann. Deshalb wird die höchste Dosierung identifiziert, deren Mittelwert bezüglich der Zielgröße zumindest so groß wie der Mittelwert jeder anderen Gruppe mit einer höheren Dosierung ist. Falls keine Umkehrung vorliegt, wird die höchste Dosierung ausgewählt. Formal kann diese Regel zur Bestimmung des Schätzers $\hat{w}_1$ wie folgt beschrieben werden:

$$\hat{w}_1 = \begin{cases} \min\limits_{j=0,\ldots,d-1} \{ j : \overline{X}_i \geq \overline{X}_{i+1},\ j \leq i \leq k-1 \}, & \text{falls } \overline{X}_{k-1} \geq \overline{X}_k \\ k & \text{sonst.} \end{cases}$$

Dabei bezeichnet $\overline{X}_i$ den Gruppen-Mittelwert in Gruppe $i = 0,...,k,$ der normalverteilt ist mit Erwartungswert μ_i und Varianz $\sigma^2_{MW} = \frac{\sigma^2}{m}$.

2. Schritt:

In diesem Schritt werden standardisierte Helmert-Kontraste verwendet, um die niedrigste Dosis auf dem Plateau der Dosis-Wirkungs-Kurve zu bestimmen:

$$\hat{w}_2 = \max_{j=1,\ldots,\hat{w}_1-1} \{ j : \sum_{i=j}^{\hat{w}_1} h_{ji} \overline{X}_i > 0.3 (\max_{i=1,\ldots,k} \overline{X}_i - \min_{i=0,\ldots,k} \overline{X}_i) \} + 1$$

mit

$$h_{ji} = \begin{cases} \dfrac{1}{\sqrt{(\hat{w}_1 - j)(\hat{w}_1 - j + 1)}}, & \text{falls } i > j \\ -\dfrac{\hat{w}_1 - j}{\sqrt{(\hat{w}_1 - j)(\hat{w}_1 - j + 1)}}, & \text{falls } i = j \,. \end{cases}$$

Wenn kein $\hat{w}_2$ mit diesen Eigenschaften existiert, wird die Dosis-Gruppe mit der maximalen Response bezüglich der Zielgröße gewählt, d.h., $\hat{w}_2 = \arg\max_i (\overline{X}_i)$.

3. Schritt:

Der dritte Schritt besteht aus zwei Teilen. Zunächst wird überprüft, ob es eine niedrigere Dosis als $\hat{w}_2$ gibt, deren Mittelwert bezüglich der Zielgröße um mindestens 15% der Differenz zwischen der maximalen Response der Dosis-Gruppen und der minimalen Response aller Behandlungsgruppen unter dem Mittelwert für die Dosis $\hat{w}_2$ liegt:

$$\hat{w}_3 = \min_{j=1,\ldots,\hat{w}_2} \left\{ \; j : \overline{X}_j \geq \overline{X}_{\hat{w}_2} - 0.15(\max_{i=1,\ldots,k} \overline{X}_i - \min_{i=0,\ldots,k} \overline{X}_i) \; \right\}.$$

Danach wird die kleinste Dosis identifiziert, deren Mittelwert größer ist als die Summe des kleinsten aufgetretenen Mittelwerts und 75% der Differenz zwischen der maximalen Response der Dosis-Gruppen und der minimalen Response aller Behandlungsgruppen:

$$\hat{w}_4 = \min_{j=1,\ldots,k} \left\{ \; j : \overline{X}_j \geq \min_{i=0,\ldots,k} \overline{X}_i + 0.75(\max_{i=1,\ldots,k} \overline{X}_i - \min_{i=0,\ldots,k} \overline{X}_i) \; \right\}.$$

Der Helmert-Schätzer ist dann das Maximum der im zweiten und im dritten Schritt identifizierten Indices der Dosis-Gruppen:

$$\hat{w}_H = \max(\hat{w}_3, \hat{w}_4) \, .$$

3.3.2 Schwellenwert-Schätzer

Einfacher Schwellenwert-Schätzer

Der einfache Schwellenwert-Schätzer (ES) verwendet lediglich den zweiten Teil des dritten Schrittes der Bestimmung des Helmert-Schätzers, wobei der dort verwendete Wert 0.75 durch einen allgemeineren Wert $b \in (0,1)$ ersetzt wird:

$$\hat{w}_{ES} = \min_{j=0,\ldots,k} \left\{ \; j : \overline{X}_j \geq \min_{i=0,\ldots,k} \overline{X}_i + b(\max_{i=1,\ldots,k} \overline{X}_i - \min_{i=0,\ldots,k} \overline{X}_i) \; \right\} \text{ mit } b \in (0,1) \, .$$

Im übernächsten Abschnitt wird angegeben, wie der Schwellenwert b für dieses Verfahren zu wählen ist. Zuvor wird noch ein alternativer Schwellenwert-Schätzer eingeführt.

Schwellenwert-Schätzer mit gleitendem Mittelwert

Die Bestimmung des Schwellenwert-Schätzers mit gleitendem Mittelwert (GMS) besteht aus zwei Schritten. Im ersten Schritt werden die folgenden gleitenden Mittelwerte $\tilde{\overline{X}}_i, i = 0,\ldots,k,$ berechnet:

$$\tilde{\bar{X}}_i = \begin{cases} \frac{5}{6}\bar{X}_0 + \frac{1}{6}\bar{X}_1, & \text{für } i = 0 \\ \frac{1}{6}\bar{X}_{i-1} + \frac{2}{3}\bar{X}_i + \frac{1}{6}\bar{X}_{i+1}, & \text{für } 1 \le i \le k-1 \\ \frac{1}{6}\bar{X}_{k-1} + \frac{5}{6}\bar{X}_k, & \text{für } i = k. \end{cases}$$

Im zweiten Schritt wird dann der im vorangehenden Kapitel beschriebene einfache Schwellenwert-Schätzer statt auf die ursprünglichen Gruppen-Mittelwerte $\bar{X}_i$ auf die $\tilde{\bar{X}}_i$ angewendet, um den Schwellenwert-Schätzer mit gleitenden Mittelwerten $\hat{w}_{GMS}$ zu erhalten.

Wahl des Schwellenwertes

Bei der Bestimmung des Helmert-Schätzers wurde im dritten Schritt der Daten unabhängige Schwellenwert $b = 0.75$ verwendet. Man sollte jedoch annehmen, dass eine vernünftige Wahl des Schwellenwertes die maximale Differenz zwischen den Erwartungswerten bezüglich der Response $\Delta = \max_{i,j \in \{0,\ldots,k\}} \{\mu_i - \mu_j\} = \max_{i \in \{0,\ldots,k\}} \mu_i - \min_{i \in \{0,\ldots,k\}} \mu_i$ und die Varianz der Mittelwerte σ^2_{MW} berücksichtigen sollte. Insbesondere sollte man erwarten, dass ein optimaler Schwellenwert b vom Verhältnis der maximalen Differenz und der Standardabweichung der Gruppen-Mittelwerte abhängt, d.h., von $\theta = \frac{\Delta}{\sigma_{MW}} = \sqrt{m}\frac{\Delta}{\sigma}$. Darüber hinaus sollte die Wahl von b abhängen von der Anzahl der Dosis-Gruppen, weil die Verteilung der maximalen Differenz der Gruppen-Mittelwerte $\max_{i \in \{0,\ldots,k\}} \bar{X}_i - \min_{i \in \{0,\ldots,k\}} \bar{X}_i$ ebenfalls von k abhängt. Um einen daten-abhängigen optimalen Wert für b zu bestimmen, legen wir das in Kapitel 3.2 eingeführte Gütekriterium des Bayes-Risikos zugrunde. Für unsere Untersuchungen verwenden wir die quadratische Verlustfunktion und die Gleichverteilung als *a priori* Verteilung für den Index der niedrigsten Plateau-Dosis w. Es wurden umfangreiche Monte-Carlo-Simulationen durchgeführt, um die Wahl von b bezüglich dieses Kriteriums zu optimieren. Um den gesamten Bereich klinisch relevanter Situationen zu erfassen, wurden Simulationen für die Werte $\theta \in 1.0, 2.65$ und 5.0 durchgeführt. Details hierzu finden sich in FRIEDE, MILLER, BISCHOFF und KIESER (2000). Die Ergebnisse können wie folgt zusammengefasst werden. Der tatsächliche (unbekannte!) Wert von θ hat für moderate und große θ ($\theta \ge 2.65$) keinen wesentlichen Einfluss auf den optimalen Schwellenwert b. Um einen von θ unabhängigen Schwellenwert zu erhalten wurde der Mittelwert der optimalen Schwellenwerte für

$\theta \in 1.0, 2.65$ und 5.0 berechnet. Für die quadratische Verlustfunktion und Anzahlen von Dosis-Gruppen $k = 3, ..., 6$ sind die resultierenden Werte in Tabelle 3 angegeben. Die Ergebnisse für den einfachen Schwellenwert-Schätzer, für den die Schwellenwerte für alle $k = 3, ..., 6$ zwischen 0.72 und 0.76 liegen, rechtfertigen retrospektiv die Verwendung des Wertes 0.75 für den Helmert-Schätzer, wie von BAUER, BAUER und BUDDE (1998) empirisch festgelegt.

Tabelle 3: Mittelwert der optimalen Schwellenwerte für $\theta \in 1.0, 2.65$ und 5.0, $\theta = \sqrt{m} \cdot (\Delta/\sigma)$, für den einfachen Schwellenwert-Schätzer (ES) und den Schwellenwert-Schätzer mit gleitendem Mittelwert (GMS) bei *k* Dosis-Gruppen und für quadratische Verlustfunktion.

		Mittelwert der optimalen Schwellenwerte			
Schwellenwert-Schätzer	*k*	3	4	5	6
ES		0.72	0.74	0.76	0.76
GMS		0.66	0.68	0.70	0.72

Die entsprechenden Schwellenwerte sind für die LINEX-Verlustfunktion kleiner als für l_2, da diese eine Überschätzung mit einem größeren Verlust „bestraft“ und die entsprechend optimierte Change point Schätzung damit eher zu niedrigeren Dosierungen tendiert als bei quadratischer Verlustfunktion. Beispielsweise beträgt der Mittelwert der optimalen Schwellenwerte für $\theta \in 1.0, 2.65$ und 5.0 und $k = 6$ für die Verlustfunktion $l_{LINEX}(w, \hat{w} \mid \lambda = 0.5, s = 1)$ 0.60 für ES und 0.57 für GMS.

3.3.3. Kleinste-Quadrate-Schätzer

Das Regressions-Modell (3.1) wird durch drei Parameter festgelegt: Regression-Abschnitt β_1, Regressions-Steigung β_2 und Change point w. Der Kleinste-Quadrate-Schätzer $\hat{\beta}_{KQ}$ des Parametervektors $\beta = (\beta_1, \beta_2, w)^T$ minimiert den euklidischen Abstand zwischen den beobachteten und den durch das Modell vorhergesagten Werten. Die Kleinste-Quadrate-Schätzung $\hat{w}_{KQ}$ (KQ) ist gegeben durch die dritte Komponente der Schätzung dieses Vektors. Eine Beschreibung der Kleinste-Quadrate-Schätzung in Change point Regressions-Modellen vom Typ (3.1) findet sich beispielsweise in der Arbeit von HINKLEY (1969).

3.3.4 Anwendungsbeispiel

Beispiel 3:

Die Anwendung der in den vorangehenden Abschnitten beschriebenen Schätzmethoden soll nun an einer konkreten Dosis-Wirkungs-Studie mit dem blutdrucksenkenden Medikament Cilazapril illustriert werden (NATHOFF, ATTWOOD, EICHLER, KOGLER, KLEINBLOESEM und VAN BRUMMELEN, 1990; BAUER und RÖHMEL, 1995). In dieser Studie wurden folgende Ergebnisse für die Reduktion des im Sitzen gemessenen diastolischen Blutdrucks [mmHg] zwischen Therapiebeginn und nach 4 Wochen Behandlung mit Placebo und drei Dosierungen von 1 H täglich verabreichtem Cilazapril erhalten (Mittelwert ± SEM, Stichprobenumfang): Placebo 0.72 ± 1.90, $n = 24$; 1 mg Cilazapril 2.97 ± 1.78, $n = 26$; 2.5 mg Cilazapril 5.33 ± 2.26, $n = 24$; 5 mg Cilazapril 6.31 ± 1.76, $n = 26$). Das beobachtete Dosis-Wirkungs-Profil ist in Abbildung 5 dargestellt.

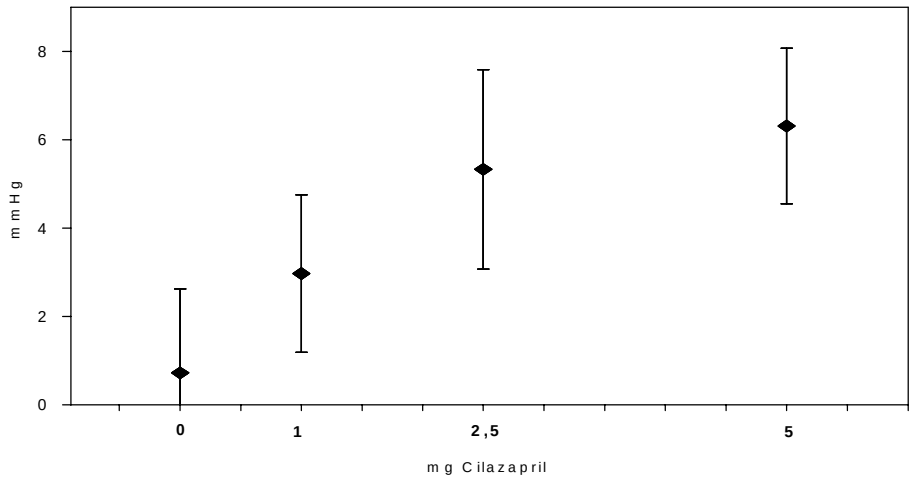

Abbildung 5: Reduktion des Blutdrucks im Sitzen in mmHg (Mittelwert ± SEM) unter Placebo und 1 mg, 2.5 mg und 5 mg Cilazapril (nach NATHOFF *et al.*, 1990)

Bei Inaugenscheinnahme der Dosis-Wirkungs-Kurve ist nicht evident, ob man die höchste oder die zweithöchste Dosis-Gruppe als kleinste Plateau-Dosis auswählen sollte. In einer solchen Situation ist eine Schätzmethode mit klar definierten Eigenschaften hilfreich für die Entscheidungsfindung. Im vorliegenden Fall würden alle vier Change point Schätzer die zweithöchste Dosis-Gruppe (2.5 mg Cilazapril) auswählen. Die integrierte Zusammenfassung aller mit Cilazapril durchgeführter Studien führte zu der Empfehlung, Cilazapril in einer Dosierung von 2.5 mg bis 5 mg täglich anzuwenden (siehe NATHOFF *et al.*, 1990).

3.4 Vergleich der Auswahlregeln

Die oben beschriebenen Auswahlregeln sollen nun anhand der in Kapitel 3.2 eingeführten Bewertungskriterien verglichen werden. Hierzu betrachten wir für $k = 6$ die in Abbildung 3 dargestellten Dosis-Wirkungs-Beziehungen mit den Werten 1.0, 2.65 und 5 für $\theta = \sqrt{m} \cdot (\Delta/\sigma)$; für einen standardisierten Therapieeffekt von beispielsweise $\Delta/\sigma = 0.5$ entsprechen diesen Werten Fallzahlen pro Gruppe m von 4, 28 und 100. In Monte-Carlo-Simulationen wurden für jede Situation 20 000 zufällige Datensätze erzeugt. Für die Selektionsraten (Tabellen 4 und 5) beträgt die Breite des zweiseitigen 95%-Konfidenzintervalls maximal 0.014. Aus Beispiel 7.2 in BISCHOFF und MILLER (2000) folgt, dass der maximale Standardfehler für das geschätzte Risiko dann auftritt, wenn der wahre Change point bei $w = 1$ oder $w = 6$ liegt und die Selektionswahrscheinlichkeiten für die beiden Dosierungen 0.5 beträgt. Die maximale Breite des zweiseitigen 95%-Konfidenzintervalls für das Risiko (Tabelle 6) beträgt damit 0.34. Die Simulationen wurden mit SAS/IML durchgeführt.

Tabelle 4 zeigt die *Raten korrekter Selektion* des wahren Change points. Für $2 \le w \le 5$ liefert der GMS-Schätzer (mit Ausnahme von $w = 5$ und $\theta = 5$) entweder die besten oder zweitbesten Ergebnisse. Der wesentlich kompliziertere Helmert-Schätzer übertrifft ES und GMS nur in wenigen Konstellationen und dann nur unwesentlich. Der KQ-Schätzer zeigt die besten Resultate, wenn der Change point entweder für die niedrigste oder die höchste untersuchte Dosis vorliegt ($w = 1, 6$). Dies sind allerdings Situationen, die man in der Praxis zu vermeiden sucht: Um sich ein umfassendes Bild von der Dosis-Wirkungs-Beziehung zu verschaffen, ist man bestrebt, die Dosierungen so zu wählen, dass die niedrigste Plateau-Dosis nicht am Rand des untersuchten Dosis-Bereichs liegt, da sich ansonsten nicht ausschließen lässt, dass die tatsächliche Change point Dosis kleiner oder größer als die ausgewählte Dosierung ist.

Tabelle 4: Rate korrekter Selektion der niedrigsten Plateau-Dosis für die in Abbildung 3 dargestellten Szenarien und für den Helmert-Schätzer (H), den einfachen Schwellenwert-Schätzer (ES), den Schwellenwert-Schätzer mit gleitenden Mittelwerten (GMS) und den Kleinste-Quadrate-Schätzer (KQ). (Die fettgedruckten Zahlen bezeichnen das jeweils beste Ergebnis für das entsprechende Szenario; $\theta = \sqrt{m} \cdot (\Delta/\sigma)$, 20 000 Replikationen).

		Simulierte Rate korrekter Selektion			
θ	Szenario	H	ES	GMS	KQ
1	1	.31	.33	.31	**.47**
	2	.27	**.30**	.28	.22
	3	.24	.25	**.26**	.17
	4	**.22**	.20	**.22**	.14
	5	**.19**	.16	.17	.14
	6	.15	.13	.16	**.30**
2.65	1	.40	.43	.31	**.71**
	2	.40	.42	**.46**	.37
	3	.37	.38	**.42**	.29
	4	.34	.33	**.35**	.24
	5	**.29**	.28	.28	.22
	6	.23	.21	.21	**.44**
5	1	.61	.61	.39	**.92**
	2	.62	.62	**.69**	.60
	3	.57	.57	**.64**	.50
	4	.48	**.49**	.48	.41
	5	.38	**.39**	.33	.36
	6	.26	.27	.16	**.56**

Von einem guten Change point-Schätzer würde man erwarten, dass er im Falle einer Fehl-Selektion zumindest eine Dosis-Gruppe, die nahe am wahren Change point liegt, auswählt. In Tabelle 5 sind für die Szenarien 2, 4 und 6 aus Abbildung 3 die *Selektionsraten für alle Dosis-Gruppen* angegeben. Man erkennt, dass die Ergebnisse für den einfachen Schwellenwert-Schätzer weitgehend identisch mit denen für den wesentlich komplizierteren Helmert-Schätzer sind. Die zusätzliche Anwendung von Helmert-Kontrasten führt damit nicht zu einer Verbesserung der Charakteristika des ES-Schätzers. Auf der anderen Seite zeigen die Ergebnisse für GMS, dass der einfache Schwellenwert-Schätzer verbessert werden kann, indem bei seiner Berechnung gleitende Mittelwerte verwendet werden. Weiterhin kann man Tabelle 5 entnehmen, dass der KQ-Schätzer zur Auswahl von Dosis-Gruppen tendiert, die am Rand des untersuchten Dosis-Bereichs liegen, und dieser Schätzer deshalb schlechte Ergebnisse für $w \neq 1$ und $w \neq 6$ erzielt.

Tabelle 6 zeigt für die in Abbildung 3 dargestellten Szenarien die *Risiken*, *maximalen Risiken* und *Bayes-Risiken* bei quadratischer Verlustfunktion und Gleichverteilung als *a priori* Verteilung. Für die Schätzer H, ES und GMS zeigen die Werte für das Risiko in Abhängigkeit vom Change point w einen U-förmigen Verlauf. Das gleiche gilt für den KQ-Schätzer und $\theta = 1$, während für $\theta = 2.65$ und $\theta = 5$ das Risiko mit wachsendem w monoton zunimmt. Der GMS-Schätzer liefert bezüglich des maximalen und des Bayes-Risikos die besten Resultate für $\theta = 1$ und $\theta = 2.65$ sowie das zweitbeste Ergebnis bezüglich des Bayes-Risikos für $\theta = 5$. Mit Ausnahme der Konstellationen $\theta = 2.65$, $w = 1$ und $\theta = 5, w = 1, 6$ wird von GMS auch für die Risiken zu den 18 einzelnen Change point-Situationen das beste oder zweitbeste Ergebnis erzielt.

Um die Unterschiede zur quadratischen Verlustfunktion zu illustrieren, werden im folgenden exemplarisch einige Ergebnisse zur LINEX-Verlustfunktion mit $\lambda = 0.5$ und $s = 1$ sowie für $\theta = 2.65$ angegeben. Wie bei der quadratischen Verlustfunktion wird als *a priori* Verteilung die Gleichverteilung verwendet. Die optimalen Schwellenwerte von 0.60 und 0.57 für ES und GMS sind aufgrund der größeren „Bestrafung“ einer Überschätzung kleiner als bei quadratischer Verlustfunktion (0.76 bzw. 0.72). Dies führt zu höheren Selektionsraten für die niedrigeren Dosis-Gruppen mit entsprechenden Konsequenzen für die Trefferraten: Für Szenario 2 beträgt die Rate korrekter Selektion 0.52 bzw. 0.57 für ES bzw. GMS (gegenüber 0.42 bzw. 0.46 bei quadratischer Verlustfunktion), für Szenario 6 dagegen nur 0.08 bzw. 0.07 (gegenüber jeweils 0.21). Wie zu erwarten schneiden ES und GMS bezüglich des Bayes-Risikos besser ab als ihre Konkurrenten, denn die Wahl der Schwellenwerte wurde unter diesem Kriterium optimiert. Die Bayes-Risiken betragen 0.30 und 0.33 für GMS und ES und 0.42 bzw. 0.46 für KQ und H. Dass der Kleinste-Quadrate-Schätzer bezüglich der LINEX-Verlustfunktion bessere Ergebnisse liefert als der Helmert-Schätzer (die beiden Schätzer weisen bei quadratischer Verlustfunktion das gleiche Bayes-Risiko auf) erklärt sich aus der Tatsache, dass die Selektionsraten von KQ für niedrigere Dosierungen größer sind als für höhere (siehe hierzu Tabelle 5).

Tabelle 5: Rate, mit der die Dosis-Gruppen i = 1, ..., 6, als vermeintlich niedrigste Plateau-Dosis für die in Abbildung 3 dargestellten Szenarien 2, 4 und 6 selektiert werden und für den Helmert-Schätzer (H), den einfachen Schwellenwert-Schätzer (ES), den Schwellenwert-Schätzer mit gleitenden Mittelwerten (GMS) und den Kleinste-Quadrate-Schätzer (KQ). (Die grau unterlegten Felder bezeichnen die wahre niedrigste Plateau-Dosis; μ_i: Erwartungswerte in Gruppe i; $\theta = \sqrt{m} \cdot (\Delta/\sigma)$, 20 000 Replikationen).

	Simulierte Selektionsrate														
	Szenario 2					Szenario 4					Szenario 6				
θ	μ_i	H	ES	GMS	KQ	μ_i	H	ES	GMS	KQ	μ_i	H	ES	GMS	KQ
1	0	.00	.00	.00	.00	0	.00	.00	.00	.00	0	.00	.00	.00	.00
	1/2	.17	.19	.18	.31	1/4	.14	.16	.14	.23	1/6	.16	.18	.16	.22
	1	.27	.30	.28	.22	1/2	.18	.20	.17	.15	2/6	.17	.19	.16	.13
	1	.20	.20	.21	.14	3/4	.21	.21	.22	.14	3/6	.18	.18	.18	.11
	1	.15	.14	.14	.10	1	.22	.20	.22	.14	4/6	.18	.17	.19	.11
	1	.11	.10	.10	.08	1	.15	.13	.14	.12	5/6	.17	.15	.16	.13
	1	.09	.07	.09	.15	1	.11	.09	.11	.22	1	.15	.13	.16	.30
2.65	0	.00	.00	.00	.00	0	.00	.00	.00	.00	0	.00	.00	.00	.00
	1/2	.08	.09	.05	.29	1/4	.03	.04	.02	.10	1/6	.04	.05	.02	.09
	1	.40	.42	.46	.37	2/4	.11	.12	.11	.12	2/6	.09	.10	.08	.08
	1	.23	.23	.24	.15	3/4	.24	.24	.28	.18	3/6	.15	.16	.16	.10
	1	.14	.13	.12	.07	1	.34	.33	.35	.24	4/6	.22	.23	.25	.12
	1	.09	.08	.07	.05	1	.18	.17	.16	.16	5/6	.26	.25	.28	.17
	1	.06	.05	.04	.07	1	.10	.09	.08	.20	1	.23	.21	.21	.44
5	0	.00	.00	.00	.00	0	.00	.00	.00	.00	0	.00	.00	.00	.00
	1/2	.03	.03	.01	.21	1/4	.00	.00	.00	.02	1/6	.00	.00	.00	.02
	1	.62	.62	.69	.60	2/4	.05	.05	.04	.07	2/6	.03	.03	.02	.03
	1	.20	.21	.22	.12	3/4	.28	.27	.37	.21	3/6	.11	.10	.10	.06
	1	.08	.08	.06	.03	1	.48	.49	.48	.41	4/6	.25	.25	.31	.12
	1	.04	.03	.03	.02	1	.14	.15	.10	.17	5/6	.36	.35	.41	.22
	1	.02	.01	.01	.01	1	.05	.05	.02	.12	1	.26	.27	.16	.56

Tabelle 6: Risiko bei quadratischer Verlustfunktion, maximales Risiko und Bayes-Risiko (bei Gleichverteilung als *a priori* Verteilung) für die in Abbildung 3 dargestellten Szenarien und für den Helmert-Schätzer (H), den einfachen Schwellenwert-Schätzer (ES), den Schwellenwert-Schätzer mit gleitenden Mittelwerten (GMS) und den Kleinste-Quadrate-Schätzer (KQ). (Die fettgedruckten Zahlen bezeichnen das jeweils beste Ergebnis für das entsprechende Szenario; $\theta = \sqrt{m} \cdot (\Delta/\sigma)$, 20 000 Replikationen).

		Simuliertes Risiko			
θ	Szenario	H	ES	GMS	KQ
1	1	5.64	**4.83**	5.44	5.60
	2	3.50	**3.02**	3.32	4.04
	3	2.37	**2.23**	2.33	3.40
	4	2.75	2.96	**2.73**	3.76
	5	4.96	5.37	**4.94**	5.56
	6	9.20	9.99	**9.05**	9.18
	Maximales Risiko	9.20	9.99	**9.05**	9.18
	Bayes Risiko	4.74	4.73	**4.63**	5.26
2.65	1	4.16	3.49	3.86	**2.27**
	2	2.66	2.32	**2.18**	2.34
	3	1.75	1.59	**1.43**	2.41
	4	1.57	1.61	**1.34**	2.54
	5	2.50	2.72	**2.21**	3.12
	6	5.04	5.46	**4.58**	5.03
	Maximales Risiko	5.04	5.46	**4.58**	5.03
	Bayes Risiko	2.95	2.87	**2.60**	2.95
5	1	1.77	1.50	1.74	**0.36**
	2	1.23	1.13	**0.80**	0.83
	3	0.87	0.86	**0.56**	1.14
	4	0.83	0.81	**0.70**	1.31
	5	1.31	**1.28**	**1.28**	1.41
	6	2.82	2.74	2.84	**2.11**
	Maximales Risiko	2.82	2.74	2.84	**2.11**
	Bayes Risiko	1.47	1.39	1.32	**1.19**

Zusammenfassend kann man schlussfolgern, dass der GMS-Schätzer einfacher zu berechnen ist als die konkurrierenden Helmert- und Kleinste-Quadrate-Schätzer und in den meisten der betrachteten Situationen die besseren Ergebnisse erzielt. Im folgenden Kapitel werden wir die Güte des adaptiven Zwei-Stufen-Designs mit einem nicht-adaptiven Design in der Dose-Response-Situation vergleichen und dabei den GMS-Schätzer zur Auswahl der niedrigsten Plateau-Dosis für den zweiten Studienteil verwenden.

3.5 Vergleich zwischen adaptivem und nicht-adaptivem Design

In diesem Kapitel soll gezeigt werden, wie Change point-Schätzer nutzbringend in Dose-Response-Studien mit adaptivem Design eingesetzt werden können. Dabei betrachten wir exemplarisch die folgenden Design-Situationen:

Ein-Stufen-Design:

- Die Studie wird mit jeweils n Beobachtungen in den $k+1$ Behandlungsgruppen durchgeführt. Wir nehmen an, dass in der Planungsphase von einem linearen Dosis-Wirkungs-Zusammenhang ausgegangen wird und deshalb festgelegt wird, dass die Null-Hypothese $H_{01} = \bigcap_{i=1}^{k} H_0^i,\ H_0^i : \mu_0 \geq \mu_i, i = 1,...,k,$ mit dem linearen Trend-Test (s.u.) zum einseitigen Niveau $\alpha = 0.025$ getestet wird.

Adaptives Zwei-Stufen-Design:

1. Studienteil:

- Der erste Teil der Studie wird mit jeweils n_1 Beobachtungen in den $k+1$ Behandlungsgruppen durchgeführt. Wie beim Ein-Stufen-Design nehmen wir an, dass in der Planungsphase ein linearer Dosis-Wirkungs-Zusammenhang angenommen wird und deshalb die Null-Hypothese H_{01} mit dem linearen Trend-Test getestet wird. Der Einfachheit halber betrachten wir die Situation ohne vorzeitiges Studienende mit Beibehaltung der Null-Hypothese ($\alpha_0 = 1.0$); der lineare Trend-Test wird damit zum einseitigen Niveau $\alpha_1 = c_\alpha = 0.00380$ ($\alpha = 0.025$) durchgeführt. Falls H_{01} nicht abgelehnt werden kann (d.h., falls für den *p*-Wert zum linearen Kontrast-Test gilt $p_1 > \alpha_1$), wird mit dem GMS-Schätzer die Dosis-Gruppe mit dem Index $\hat{w} \in \{1,...,k\}$ zur weiteren Untersuchung im zweiten Studienteil ausgewählt.

2. Studienteil:

- Der zweite Teil der Studie wird mit jeweils n_2 Beobachtungen in den Dosis-Gruppen 0 und $\hat{w}$ durchgeführt. Die Null-Hypothese $H_0^{\hat{w}}$ wird mit dem einseitigen Zwei-Stichproben *t*-Test getestet, der den *p*-Wert p_2 liefert. H_{01} kann abgelehnt werden falls $p_1 \cdot p_2 \leq c_\alpha$.

Für den Vergleich von adaptivem und nicht-adaptivem Design wurden für $k = 6$ die Szenarien 1-6 aus Abbildung 3 mit $\theta = \Delta/\sigma = 1.0$ betrachtet. Um die Betrachtung nicht unnötig zu verkomplizieren, wurden die Fallzahlen so gewählt, dass der Gesamtstichprobenumfang N für das Ein- und das Zwei-Stufen-Design gleich ist: $n = 13, n_1 = 7, n_2 = 21,$ und damit $N = (k+1) \cdot n = (k+1) \cdot n_1 + 2 \cdot n_2 = 91.$ Die zusätzliche Option adaptiver Designs, die Fallzahl im Verlauf der Studie entsprechend der jeweiligen Notwendigkeit anpassen zu können (siehe auch Kapitel 4), wurde damit nicht ausgenutzt.

Die Power des Zwei-Stufen-Designs mit der Auswahlregel GMS wurde mittels Monte Carlo Simulationen geschätzt. Es wurden jeweils 10 000 Replikationen generiert; die maximale Breite des zweiseitigen 95% Konfidenzintervalls für die Power beträgt dann 0.02 (bei einer tatsächlichen Power von 0.50). Die Simulationen wurden mit SAS/IML durchgeführt.

Die Power des Ein-Stufen-Designs lässt sich analytisch berechnen. Das gleiche gilt für das Zwei-Stufen-Design mit der „idealen“ Selektionsregel, die für den zweiten Studienteil stets die wahre niedrigste Plateau-Dosis w auswählt. Um einen Eindruck davon zu vermitteln, welche Power unter der oben beschriebenen Strategie maximal erzielt werden kann, wurde die Power für das Zwei-Stufen-Design mit dieser Auswahlregel (im folgenden mit TRUE bezeichnet) ebenfalls bestimmt. Die entsprechenden Berechnungen wurden mit Mathematica 3.0 durchgeführt.

Die Teststatistik des Kontrast-Tests mit Scores $c_i, i = 0, ..., k,$ bzw. des Zwei-Stichproben *t*-Tests zwischen den Gruppen 0 und $\hat{w}$ lauten für eine Fallzahl m pro Gruppe

$$T_{Kontrast} = \frac{\sum_{i=0}^{k} c_i \overline{X}_i}{\sqrt{\left(\sum_{i=0}^{k} \frac{c_i^2}{m}\right) S^2}} \text{ bzw. } T_t = \sqrt{\frac{m}{2}} \frac{\overline{X}_{\hat{w}} - \overline{X}_0}{S}.$$

S^2 bezeichnet dabei den “gepoolten” Varianzschätzer aus allen Dosis-Gruppen bzw. aus den Dosis-Gruppen 0 and $\hat{w}$. Für den linearen Trend-Test, der unter den Kontrast-Tests optimale Power für einen durchgehend linearen Anstieg besitzt, lauten die Scores $-k/2$, $-k/2+1$, ..., $k/2-1$, $k/2$; der *t*-Test ist damit ein Spezialfall des linearen Trend-Tests für $k = 1$. Unter der Null- (Alternativ-) Hypothese sind $T_{Kontrast}$ und T_t zentral (nicht-zentral) *t*-verteilt mit $(k+1)(m-1)$ bzw. $2(m-1)$ Freiheitsgraden. Im Ein-Stufen-Design mit Stichprobenumfang *n* pro Gruppe ist die Power des linearen Kontrast-Tests gegeben durch

$$\text{Power (Ein-Stufen-Design)} = \Pr(T_{Kontrast} \geq t_{(k+1)(n-1),1-\alpha}) = 1 - G_{t_{(k+1)(n-1),\vartheta}}(t_{(k+1)(n-1),1-\alpha}).$$

Dabei bezeichnet $t_{df,1-\alpha}$ das $(1-\alpha)$-Quantil der zentralen *t*-Verteilung mit *df* Freiheitsgraden, $G_{t_{df,\vartheta}}$ ist die Verteilungsfunktion der nicht-zentralen *t*-Verteilung mit *df* Freiheitsgraden und Nicht-Zentralitätsparameter

$$\vartheta = \frac{1}{\sqrt{\sum_{i=0}^{k} \frac{c_i^2}{n}}} \frac{\sum_{i=0}^{k} c_i \mu_i}{\sigma}.$$

Die Power für die Selektionsregel TRUE im adaptiven Zwei-Stufen-Design mit Fallzahlen n_1 und n_2 im ersten und zweiten Studienteil ist gegeben durch

$$\text{Power (TRUE)} = 1 - \int_{-\infty}^{u_1} \int_{-\infty}^{u_2(t_1)} g_{t_{(k+1)(n_1-1),\vartheta_1}}(t_1)\, g_{t_{2(n_2-1),\vartheta}}(t_2)\, dt_2\, dt_1\,, \tag{3.2}$$

wobei $g_{t_{df,nc}}$ die Dichtefunktion der nicht-zentralen *t*-Verteilung mit *df* Freiheitsgraden und Nicht-Zentralitätsparameter *nc* ist. Weiterhin bezeichnen u_1 und $u_2(t_1)$ die Grenzen des Ablehnungsbereichs für die Interimanalyse bzw. die Endauswertung $u_1 = t_{(k+1)(n_1-1),1-c_\alpha}$ $u_2(t_1) = t_{2(n_2-1),1-c_\alpha/(1-G_{t_{(k+1)(n_1-1)}}(t_1))}$ sowie $\vartheta_1 = \sqrt{\frac{n_1}{n}}\vartheta$ und $\vartheta_2 = \sqrt{\frac{n_2}{2}}\,\frac{\mu_{\hat{w}} - \mu_0}{\sigma}$. Da die Fallzahl pro Gruppe n_2 für den zweiten Studienteil fest vorgegeben ist und insbesondere nicht von den Ergebnissen des ersten Studienteils abhängt, lässt sich (3.2) wie folgt vereinfachen:

$$\begin{aligned}\text{Power (TRUE)} &= 1 - \int_{-\infty}^{u_1} g_{t_{(k+1)(n_1-1),\vartheta_1}}(t_1) \left[\int_{-\infty}^{u_2(t_1)} g_{t_{2(n_2-1),\vartheta_2}}(t_2)\, dt_2 \right] dt_1 \\ &= 1 - \int_{-\infty}^{u_1} g_{t_{(k+1)(n_1-1),\vartheta_1}}(t_1)\, G_{t_{2(n_2-1),\vartheta_2}}(u_2(t_1))\, dt_1.\end{aligned}$$

In Tabelle 7 sind die Ergebnisse des Powervergleichs angegeben. Betrachtet man die Power für den GMS-Schätzer und die idealisierte Selektionsregel TRUE im Zwei-Stufen-Design, so erkennt man, dass der Powerverlust durch Auswahl einer falschen Dosis-Gruppe für den zweiten Studienteil mit steigender Change point-Dosis anwächst. Dies entspricht der Erwartung, da eine Unterschätzung der niedrigsten Plateau-Dosis wahrscheinlicher ist, wenn der Change point im oberen Dosis-Bereich liegt.

Der Vergleich zwischen dem adaptiven und dem nicht-adaptiven Design zeigt, dass in den Situationen, in denen die tatsächliche Dosis-Wirkungs-Beziehung von dem angenommenen linearen Zusammenhang erkennbar abweicht, durch das adaptive Zwei-Stufen-Design in Verbindung mit dem GMS-Schätzer ein teilweise drastischer Gewinn an Power erzielt wird

(für die Szenarien 1 und 2 um 0.39 bzw. 0.16). Gleichzeitig ist auch für die Situation, in der die Planungsannahme perfekt erfüllt ist (Szenario 6), die Power des dann optimalen Ein-Stufen-Designs um lediglich 0.04 höher als für das adaptive Design.

In der Arbeit von BAUER, BAUER und BUDDE (1998) sind Simulationergebnisse für das gleiche Vorgehen und ebenfalls bei Verwendung des linearen Trend-Tests angegeben; die Auswahl der Dosis-Gruppe wird dort aber mit dem Helmert-Schätzer statt dem GMS-Schätzer durchgeführt. Beispielsweise wurde für die Szenarien 2, 4 und 6 eine Power von 0.92, 0.94 und 0.83 (gegenüber 0.93, 0.92 und 0.84 für den GMS-Schätzer) erzielt. Dies zeigt erneut, dass der GMS-Schätzer, der wesentlich leichter zu berechnen ist als der Helmert-Schätzer und im Gegensatz zu diesem intuitiv verständlich ist, hinsichtlich seiner Charakteristika zumindest als gleichwertig anzusehen ist.

Tabelle 7: Powervergleich zwischen nicht-adaptivem Design (linearer Trend-Test mit allen Dosierungs-Gruppen) und adaptivem Zwei-Stufen-Design nach BAUER und KÖHNE (1994) (linearer Trend-Test mit allen Dosierungs-Gruppen nach erstem Studienteil, *t*-Test mit Kontrollgruppe und selektierter Dosis-Gruppe nach zweitem Studienteil) für die in Abbildung 3 dargestellten Szenarien. GMS: Verwendung des Schwellenwert-Schätzers mit gleitenden Mittelwerten; TRUE: Verwendung des „idealen" Schätzers, der stets die wahren Change point Dosierung selektiert. (Details siehe Text).

	Power		
	Ein-Stufen-Design	Adaptives Zwei-Stufen Design	
Szenario		GMS	TRUE
1	0.52	0.91	0.92
2	0.77	0.93	0.95
3	0.88	0.93	0.97
4	0.92	0.92	0.97
5	0.92	0.90	0.97
6	0.88	0.84	0.97

Für einen systematischen Vergleich zwischen adaptivem und nicht-adaptivem Design muss die Komplexität der in der Praxis angewendeten Entscheidungsregeln naturgemäß simplifiziert werden. Ebenso ist es notwendig, sich auf einen oder zumindest wenige der zahlreichen Adaptionsmöglichkeiten zu beschränken. Dennoch zeigen die oben dargestellten Ergebnisse eindrucksvoll, wie in adaptiven Designs die akkumulierten Daten dazu verwendet werden können, falsche Planungsannahmen zu identifizieren und zu korrigieren. Weitere Beispiele dafür, wie dieses Potential in Dosis-Wirkungs-Studien genutzt werden kann, finden sich in den Arbeiten von BAUER und RÖHMEL (1995), BAUER, BAUER und BUDDE (1998), BAUER und KIESER (1999) und FRIEDE und KIESER (1999). In den nachfolgenden Kapiteln werden wir andere Anwendungsbereiche für dieses Prinzip kennenlernen.

4. Adaptive Fallzahlplanung

Entsprechend ihrer grundlegenden Bedeutung für eine fundierte Planung klinischer Studien und Experimente ist die Entwicklung und Verfeinerung von Methoden zur Berechnung des notwendigen Stichprobenumfanges von jeher eines der intensiv beforschten Gebiete der medizinischen Biometrie. Der Therapieforschung steht gegenwärtig ein breites und feingefächertes Spektrum an Verfahren für unterschiedlichste Studiendesigns, klinische Fragestellungen und Skalierungen der Zielvariablen zur Verfügung (siehe z.B. das Buch von BOCK, 1998, und die Übersichten von ROEBRUCK, ELZE, HAUSCHKE, LEVERKUS und KIESER, 1997; OELLRICH, FREISCHLÄGER, BENNER und KIESER, 1997; ORTSEIFEN, BRUCKNER, BURKE und KIESER, 1997). Jedes dieser Verfahren ist aber im besten Fall so gut wie die Annahmen, die der Fallzahlberechnung zugrunde liegen.

In der Situation normalverteilter Daten hängt der für eine klinische Studie notwendige Stichprobenumfang ab vom minimalen klinisch relevanten Behandlungsgruppen-Unterschied, vom Signifikanzniveau, von der vorgegebenen Power und von der Varianz der Zielgröße. Der minimale klinisch relevante Behandlungseffekt wird von medizinischen Experten spezifiziert und hängt unter anderem von der Wirksamkeit verfügbarer Therapien in der entsprechenden Indikation ab. Auf statistischer Ebene besteht ein weitgehender Konsens darüber, welche Werte für die Wahrscheinlichkeit eines Fehlers 1. Art (üblicherweise $\alpha = 0.025$ einseitig oder 0.05 zweiseitig) und 2. Art ($\beta = 0.10$ oder 0.20) zu wählen sind. Im Gegensatz dazu ist die Variabilität der Zielgröße häufig mit einer erheblichen Unsicherheit behaftet, da sie durch Faktoren beeinflusst wird, die spezifisch für die jeweilige klinische Studie sind. Beispiele hierfür sind die Ein- und Ausschlusskritierien, der Grad an Standardisierung der verwendeten Messmethoden, sowie die Anzahl und Ausstattung der beteiligten Zentren. Dennoch ist es eine gängige Praxis, bei einer zu planenden Studie die Annahme über die Populationsvarianz auf der Stichprobenvarianz aus einer abgeschlossenen Studie mit der gleichen Zielgröße zu basieren. Eine andere Möglichkeit besteht darin, vor Durchführung der eigentlichen klinischen Prüfung im gleichen Umfeld eine „kleine“ Pilotstudie durchzuführen, um damit Informationen über die Streuung zu erhalten. Es ist jedoch offensichtlich, dass die Durchführung einer solchen Pilotstudie mit dem alleinigen Ziel der Varianzschätzung im Sinne eines möglichst effektiven Weges zur Beantwortung einer klinischen Fragestellung eine Verschwendung von Zeit und Resourcen darstellt.

Zur Lösung des dargestellten Problems schlugen WITTES und BRITTAIN (1990) das sogenannte Internal Pilot Study Design vor. Dabei wird im Verlauf der Studie auf der Basis

der bislang vorliegenden Beobachtungen die Varianz der Zielgröße geschätzt und, falls notwendig, die ursprünglich vorgesehene Fallzahl modifiziert. In die Auswertung gehen dann die Daten aller rekrutierter Patienten ein, also auch die der internen Pilotphase.

Ein wichtiges Charakteristikum des Wittes-Brittain Designs besteht darin, dass nach der Beendigung der Pilotphase kein Hypothesentest durchgeführt wird und deshalb keine Option zu einer vorzeitigen Studienbeendigung wegen eines deutlichen oder fehlenden Therapiegruppen-Unterschiedes besteht. Die Daten der internen Pilotstudie werden lediglich zur Überprüfung der Annahmen über die Varianz der Zielgröße verwendet. Demgegenüber erlaubt die Klasse der adaptiven Designs, die wir in den vorangehenden Kapiteln anhand des Zwei-Stufen-Designs nach BAUER und KÖHNE (1994) prototypisch dargestellt haben, sowohl eine Modifikation der ursprünglich geplanten Fallzahl als auch einen vorzeitigen Studien-Stop unter Kontrolle der spezifizierten Wahrscheinlichkeit für einen Fehler 1. Art. Im Unterschied zum Design mit interner Pilotstudie wird hier im Rahmen einer Zwischenauswertung der relative Behandlungseffekt geschätzt und ein Hypothesentest durchgeführt. In der Nomenklatur einschlägiger Guidelines (CPMP, 1995; ICH, 1999) werden diese beiden Vorgehensweisen durch die Bezeichnungen „sample size adjustment“ bzw. „interim analysis“ („*Any analysis intended to compare treatment arms with respect to efficacy or safety at any time prior to the formal completion of a trial*“; ICH, 1999) unterschieden.

In diesem Kapitel werden Methoden zur adaptiven Fallzahlplanung für die Designs mit interner Pilotstudie und adaptive Designs mit Zwischenauswertung vorgestellt, und es werden die Charakteristika der resultierenden Stichprobenumfänge für beide Designtypen verglichen. Die Darstellung basiert auf den eigenen Vorarbeiten KIESER und FRIEDE (2000a, 2000b) und FRIEDE und KIESER (2000a, 2000b) zum Thema adaptiver Fallzahlplanung. Weiterhin gingen die eigenen Arbeiten KIESER und WASSMER (1996) und KIESER und HAUSCHKE (1999, 2000), über die Bestimmung des Stichprobenumfanges im nicht-adaptiven Design in die Betrachtungen ein. Die Verfahren werden anhand von klinischen Studien illustriert, in denen die entwickelte Methodik praktisch eingesetzt wurde (KIESER, 1999; MITFESSEL, ERXLEBEN, SCHULZE und KIESER, 1999).

4.1 Design mit interner Pilotstudie

Aus den in der Einleitung genannten Gründen sind Designs wünschenswert, die es erlauben, die Planungsannahmen über die Varianz durch Inspektion der bereits akkumulierten Daten

während des Studienverlaufs zu überprüfen und die Fallzahl eventuell anzupassen. Dieser Notwendigkeit wurde in der kürzlich verabschiedeten ICH-Guideline E9 „Statistical Principles for Clinical Trials“ (ICH, 1999) durch ein eigenes Kapitel 4.4 „Sample Size Adjustment“ Rechnung getragen. Dort steht: „*In long term trials there will usually be an opportunity to check the assumptions which underlay the original design and sample size calculations. This may be particularly important if the trial specifications have been made on preliminary and/or uncertain information. An interim check conducted on the blinded data may reveal that the overall response variances, event rates or survival experience are not as anticipated. A revised sample size may then be calculated using suitably modified assumptions, and should be justified and documented in a protocol amendment and in the clinical study report. The steps taken to preserve blindness and the consequences, if any for the type I error and the width of confidence intervals should be explained. The potential need for re-estimation of the sample size should be envisaged in the protocol whenever possible.*“ (Unterstreichungen nicht im Original). Die markierten Passagen bezeichnen Aspekte, die auch für die folgenden Betrachtungen eine zentrale Rolle spielen: Eine Fallzahlanpassung, sofern nicht im Rahmen einer formalen Zwischenauswertung durchgeführt, sollte erfolgen

- unter Beibehaltung der Verblindung sowie
- unter Kontrolle der spezifizierten Wahrscheinlichkeit für einen Fehler 1. Art.

Für die nachfolgenden Überlegungen nehmen wir an, dass die Zielgröße einer Normalverteilung folgt. Wir betrachten dabei ein balanciertes Design mit m Patienten in jeder der $k \geq 2$ Behandlungsgruppen und bezeichnen die j-te Beobachtung in Behandlungsgruppe i mit $X_{ij}, i = 1,...,k,\ j = 1,...,m$. Wir nehmen an, dass die X_{ij} unabhängig und normalverteilt mit Erwartungswert μ_i und gemeinsamer (unbekannter) Varianz σ^2 sind.

Das Konzept der internen Pilotstudie wurde von WITTES und BRITTAIN (1990) zur Anwendung in klinischen Studien vorgeschlagen. Das Vorgehen lässt sich durch die folgenden drei Schritte beschreiben:

(i) Vor Beginn der Studie wird eine vorläufige Fallzahl pro Gruppe $\hat{N}_{est}$ berechnet, die auf einer *a priori* Schätzung S_0^2 der Varianz σ^2 beruht. Der Stichprobenumfang pro Gruppe für die interne Pilotstudie $n_1,\ n_1 \leq \hat{N}_{est}$, wird festgelegt.

(ii) Wenn $2n_1$ Patienten die Studie abgeschlossen haben, wird die Populationsvarianz auf der Basis dieser Daten erneut geschätzt, und durch Einsetzen dieser Schätzung S_1^2

anstelle von σ^2 in die Fallzahlformel erhält man einen aktualisierten Wert für die notwendige Fallzahl pro Gruppe $\hat{N}_{re-est}$.

(iii) Weitere $n_2 = \hat{N} - n_1$ Patienten pro Gruppe werden rekrutiert. In die Auswertung der Studie gehen die Daten aller $2\hat{N}$ Studienteilnehmer ein.

Bemerkungen:

1. Die Idee eines Zwei-Stufen-Designs mit einer Überprüfung der Planungsannahme bezüglich der Varianz nach der ersten Stufe geht auf STEIN (1945) zurück. Die Prozedur von STEIN hat zum Ziel, dass die Breite des Konfidenzintervalls für den Erwartungswert einer normalverteilten Zufallsgröße eine vorgegebene Marge nicht überschreitet. Auf einen weiteren Unterschied zur Methode von WITTES und BRITTAIN werden wir in Kapitel 4.1.2.1 eingehen.
2. Der ursprüngliche Vorschlag von WITTES und BRITTAIN (1990) bestand darin, $\hat{N}$ zu wählen als das Maximum des ursprünglich vorgesehenen und des re-kalkulierten Stichprobenumfanges, d.h., $\hat{N} = \max(\hat{N}_{est}, \hat{N}_{re-est})$. Damit ist die endgültige Fallzahl nie kleiner als der in der Planungsphase vorgesehene Stichprobenumfang. Ein gravierender Nachteil dieser Strategie besteht darin, dass bei einer ursprünglich zu groß kalkulierten Fallzahl auch der endgültige Stichprobenumfang zu hoch ausfallen wird. Deshalb schlugen BIRKETT und DAY (1994) vor, auf diese Restriktion zu verzichten und $\hat{N}$ zu wählen als $\hat{N} = \max(n_1, \hat{N}_{re-est})$.
3. Zur Wahl der Fallzahl n_1 der internen Pilotstudien wurden verschiedene Vorschläge unterbreitet. Für das Zwei-Stufen-Verfahren von STEIN (1945) schlug SEELBINDER (1953) eine Regel zur Bestimmung von n_1 vor, mit der das Optimalitätskriterium „Minimiere das Maximum von $E(\hat{N}_{re-est}) - N$ über einen Bereich von σ“ erfüllt wird. MOSHMAN (1958) gab eine Verfeinerung dieser Strategie an, bei der zusätzlich der Aspekt, dass die Wahrscheinlichkeit eines sehr großen Stichprobenumfanges durch einen vorgegebenen Wert beschränkt sein soll, in das Minimierungskriterium eingeht. Für das oben beschriebene Internal Pilot Study Design wählten WITTES und BRITTAIN (1990) in ihren Simulationsuntersuchungen den Stichprobenumfang der internen Pilotstudie als die Hälfte der in der Planungsphase berechneten Fallzahl, d.h., $n_1 = 0.5 \cdot \hat{N}_{est}$. SANDVIK, ERIKSSEN, MOWINCKEL und RODLAND (1996) schlugen eine Methode vor, die das Ziel hat, den Stichprobenumfang der internen Pilotstudie so groß wie möglich zu wählen, und die

gleichzeitig die Wahrscheinlichkeit, mehr Patienten als für die Gesamtstudie notwendig sind, durch einen vorgegebenen Maximalwert begrenzt. Bei diesem Ansatz wird vorausgesetzt, dass die Daten, die vor Studienbeginn verfügbar sind und die zur initialen Varianzschätzung verwendet werden, eine Zufallsstichprobe der Studienpopulation darstellen. Diese Annahme ist aber kritisch zu bewerten: Falls sie erfüllt ist, kann die Fallzahlplanung ohne interne Pilotstudie so bewerkstelligt werden, dass die tatsächliche Power mit einer vorgegebenen Wahrscheinlichkeit mindestens so groß wie die gewünschte Power $1-\beta$ ist (siehe Kapitel 4.3.1 sowie BROWNE, 1995, und KIESER und WASSMER, 1996). Andererseits gelten bei Durchführung einer internen Pilotstudie die entsprechenden Aussagen in der gleichen Weise gelten ohne die obige Voraussetzung (siehe Kapitel 4.3.1 sowie KIESER und FRIEDE, 2000a). Deshalb bringt bei Gültigkeit der Annahme die Durchführung einer internen Pilotstudie keine Vorteile. SINGER (1999) wies darauf hin, dass bei der Festlegung des Stichprobenumfanges der internen Pilotstudie zusätzlich zu den Überlegungen von SANDVIK *et al.* (1996) die Rekrutierungsrate und die Beobachtungszeit der Studie zu berücksichtigen sind. Dies ist notwendig, um sicherzustellen, dass der tatsächlich notwendige Stichprobenumfang auch dann nicht überschritten wird, wenn (wie in der Praxis aus logistischen Gründen üblich) die Rekrutierung über die für die interne Pilotstudie notwendige Zahl von Patienten hinaus fortgesetzt wird, bis das Ergebnis der Varianzschätzung verfügbar ist. DENNE und JENNISON (1999) wählten als Optimalitätskriterium zur Wahl von n_1 die Minimierung des Verhältnisses $E(\hat{N}_{re-est})/N$ für den wahren Wert von σ. Sie schlugen eine Strategie vor, die einen Wert für diesen Quotienten liefert, der auch bei Fehl-Spezifikation nahe bei dem Minimum liegt.

Eine zentrale Rolle für die Effektivität des Designs mit interner Pilotstudie spielt die Varianzschätzung, die zur Re-Kalkulation des Stichprobenumfanges verwendet wird. Im folgenden Kapitel werden verschiedene Schätzmethoden vorgestellt und ihre Eigenschaften untersucht. Zuvor wird noch die Testsituation eingeführt, für die an verschiedenen Stellen die Anwendung der Verfahren beispielhaft dargestellt werden wird. Getestet wird die globale Null-Hypothese $H_0 : \mu_1 = \mu_2 = ... = \mu_k$, die die Gleichheit aller k Erwartungswerte der Behandlungsgruppen behauptet. Beispiele für Tests von H_0 sind der F-Test gegen allgemeine Alternativen oder Kontrast-Tests, die eine höhere Power beim Vorliegen spezifischer Erwartungswertprofile besitzen (siehe Kapitel 3.5). Die Methoden zur adaptiven

Fallzahlplanung, die im folgenden vorgestellt werden, können für beliebige Teststatistiken, die für dieses Testproblem geeignet sind, verwendet werden. Wir werden speziell den F-Test betrachten sowie seinen Spezialfall für $k = 2$, den *t*-Test. Die Teststatistik des F-Tests ist gegeben durch

$$F = \frac{m \cdot \sum_{i=1}^{k} (\bar{X}_{i.} - \bar{X}_{..})^2 / (k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{m} (\bar{X}_{ij} - X_{i.})^2 / (k \cdot (m-1))}$$

$$= \frac{m \cdot \sum_{i=1}^{k} (\bar{X}_{i.} - \bar{X}_{..})^2}{(k-1) \cdot S^2},$$

wobei $\bar{X}_{i.}$ und $\bar{X}_{..}$ die Mittelwerte in Behandlungsgruppe i bzw. der gesamten Stichprobe bezeichnet und S^2 den üblichen „gepoolten“ Schätzer der Varianz innerhalb der Gruppen. Unter H_0 ist F zentral F-verteilt mit $(k-1)$ und $k(m-1)$ Freiheitsgraden. Unter einer vorgegebenen Alternativ-Hypothese ist F nicht-zentral F-verteilt mit der selben Anzahl an Freiheitsgraden und Nicht-Zentralitätsparameter

$$\vartheta = m \cdot \frac{\sum_{i=1}^{k} (\mu_i - \bar{\mu})^2}{\sigma^2}.$$

Dabei bezeichnet $\bar{\mu} = \frac{1}{k} \sum_{i=1}^{k} \mu_i$ den Mittelwert der k Erwartungswerte der Behandlungsgruppen. Daraus folgt, dass die Fallzahl pro Gruppe N, die notwendig ist, um bei Vorliegen des klinisch relevanten Effektes die Null-Hypothese zum Niveau α mit einer Wahrscheinlichkeit $1-\beta$ zu verwerfen, gegeben ist durch die kleinste ganze Zahl, die die folgende Bedingung erfüllt:

$$1 - G_{F_{k-1,k(N-1),\vartheta}}(f_{k-1,k(N-1),1-\alpha}) \geq 1-\beta.$$

In der obenstehenden Formel bezeichnet $G_{F_{k-1,k(m-1),\vartheta}}$ die Verteilungsfunktion der F-Verteilung mit $(k-1)$ und $k(m-1)$ Freiheitsgraden und Nicht-Zentralitätsparameter ϑ. $f_{k-1,k(m-1),1-\alpha}$ bezeichnet das $(1-\alpha)$-Perzentil der zentralen F-Verteilung mit der gleichen Anzahl von Freiheitsgraden.

4.1.1 Varianzschätzer und ihre Eigenschaften

4.1.1.1 Einfache Varianzschätzer

Wir nehmen im folgenden an, dass die Stichprobe der internen Pilotstudie in jeder Behandlungsgruppe n_1 Beobachtungen enthält. Bei ungleicher Randomisierungswahrscheinlichkeit für die k Gruppen können die nachfolgenden Methoden entsprechend modifiziert werden.

Gepoolter k-Stichproben-Varianzschätzer

Die übliche gepoolte Schätzung der Varianz innerhalb der Gruppen ist gegeben durch

$$S^2_{k-sample} = \frac{1}{k(n_1 - 1)} \cdot \left[\sum_{i=1}^{k} \sum_{j=1}^{n_1} \left(X_{ij} - \overline{X}_{i.} \right)^2 \right].$$

Die Berechnung von $S^2_{k-sample}$ erfordert die Entblindung der Behandlungsgruppen-Zugehörigkeit für die Patienten der internen Pilotstudie, womit unweigerlich während der laufenden Studie die relativen Behandlungseffekte bekannt werden. Dies widerspricht der Charakterisierung von Fallzahladjustierungen als Verfahren des verblindeten Data Reviews in internationalen Guidelines (s.o.) und macht zudem die Etablierung eines unabhängigen Data and Safety Monitoring Committees (DSMC) notwendig (CPMP, 1995; ICH, 1999). Ein solches Board stellt sicher, dass die Ergebnisse über die Therapiegruppen-Unterschiede nicht an Personen, die unmittelbar an der Studie beteiligt sind, weitergegeben werden, was ansonsten eine Quelle für eine Verzerrung der Studienergebnisse wäre. Diese Vorsichtsmaßnahme bedeutet einen zusätzlichen logistischen Aufwand für die Studie. Eine weitere Unannehmlichkeit der vorzeitigen Entblindung innerhalb dieses Designs liegt darin, dass keine Option zu einer vorzeitigen Beendigung der Studie mit einer Ablehnung der Null-Hypothese besteht, selbst wenn für die interne Pilotstudie ein deutlicher Behandlungsgruppen-Unterschied mit einem hochsignifikanten Ergebnis vorliegen würde. Aber ist es wirklich glaubwürdig und sinnvoll, dass sich in einer solchen Situation das IDMB lediglich auf eine Überprüfung der notwendigen Fallzahl beschränkt? Es gibt also triftige Gründe, für das Design mit interner Pilotstudie verblindete Varianzschätzer vorzuziehen, sofern diese nicht erhebliche Nachteile bzgl. der Effizienz der Fallzahlschätzung mit sich bringen. Zur Untersuchung dieser Fragestellung werden wir deshalb später in diesem Kapitel die k-Stichproben-Varianzschätzung mit den nachfolgenden beschriebenen verblindeten Varianz-Schätzverfahren vergleichen.

Ein-Stichproben-Varianzschätzer

Die einfachste Varianzschätzung, die ohne Entblindung des Randomisierungscodes berechnet werden kann, erhält man durch Ignorieren der Tatsache, dass die Beobachtungen aus k Gruppen herrühren. Der Ein-Stichproben-Varianzschätzer ist gegeben durch

$$S^2_{1-sample} = \frac{1}{kn_1 - 1} \cdot \left[\sum_{i=1}^{k} \sum_{j=1}^{n_1} (X_{ij} - \overline{X}_{..})^2 \right].$$

Der Erwartungswert von $S^2_{1-sample}$ in der k-Stichproben-Situation lässt sich einfach aus der folgenden Varianzzerlegung ableiten:

$$\left[\sum_{i=1}^{k} \sum_{j=1}^{n_1} (X_{ij} - \overline{X}_{..})^2 \right] = \left[\sum_{i=1}^{k} \sum_{j=1}^{n_1} (X_{ij} - \overline{X}_{i.})^2 \right] + \left[n_1 \cdot \sum_{i=1}^{k} (\overline{X}_{i.} - \overline{X}_{..})^2 \right]. \qquad (4.1)$$

Daraus folgt:

$$\begin{aligned} E(S^2_{1-sample}) &= E\left(\frac{1}{kn_1 - 1} \cdot \left[\sum_{i=1}^{k} \sum_{j=1}^{n_1} (X_{ij} - \overline{X}_{..})^2 \right] \right) \\ &= \frac{1}{kn_1 - 1} \cdot \left[\left(k(n_1 - 1) \cdot \sigma^2 \right) + \left((k-1) \cdot \sigma^2 + n_1 \cdot \sum_{i=1}^{k} (\mu_i - \overline{\mu})^2 \right) \right] \\ &= \sigma^2 + \frac{n_1}{kn_1 - 1} \cdot \sum_{i=1}^{k} (\mu_i - \overline{\mu})^2 . \end{aligned}$$

Adjustierter Ein-Stichproben-Varianzschätzer

Man erhält den adjustierten Ein-Stichproben-Varianzschätzer S^2_{adj}, indem man die Ein-Stichproben-Varianz um ihre Verzerrung unter dem unter der Alternativ-Hypothese H_1 angenommenen Erwartungswertprofil $\mu_1^*, ..., \mu_k^*$ korrigiert. Damit ist S^2_{adj} definiert als

$$S^2_{adj} = S^2_{1-sample} - \frac{n_1}{kn_1 - 1} \cdot \sum_{i=1}^{k} (\mu_1^* - \overline{\mu}^*)^2 .$$

Nach Konstruktion ist S^2_{adj} ein unter H_1 unverzerrter Varianzschätzer, der ohne Entblindung der Behandlungsgruppen-Zugehörigkeit berechnet werden kann. Dieser Schätzer ist eine Verallgemeinerung der Vorschläge von GOULD und SHIH (1992) und ZUCKER, WITTES, SCHABENBERGER und BRITTAIN (1999) für zwei Behandlungsgruppen auf die allgemeine Situation $k \geq 2$.

Eigenschaften der einfachen Varianzschätzer

Im Zusammenhang mit den oben angegebenen einfachen Varianzschätzern stellen sich folgende miteinander verwandte Fragen:

1. Welchen Vorteil bzgl. der resultierenden Fallzahl bringt eine Entblindung der Behandlungsgruppen-Zugehörigkeit zum Zwecke der Varianzschätzung, d.h., welche Unterschiede ergeben sich bei der Fallzahl-Rekalkulation unter Verwendung von $S^2_{k-sample}$ im Vergleich zu $S^2_{1-sample}$ oder S^2_{adj}?
2. Wie groß ist der Unterschied bzgl. der resultierenden Fallzahl, wenn man statt der adjustierten Ein-Stichproben-Varianz S^2_{adj} die nicht-adjustierte Variante, d.h., die einfache Ein-Stichproben-Varianz $S^2_{1-sample}$ der gepoolten Stichprobe zur Fallzahl-Rekalkulation verwendet?
3. Welche Konsequenzen ergeben sich für die resultierende Fallzahl, wenn bei Verwendung von S^2_{adj} die für die Adjustierung angenommenen Erwartungswerte μ_i^* nicht mit den tatsächlichen Erwartungswerten μ_i übereinstimmen?

Bevor wir diesen Fragen für den Spezialfall des F-Tests nachgehen, werden die hierzu notwendigen Eigenschaften der einfachen Varianzschätzer im folgenden Satz zusammengefasst.

Satz 7:

Seien $X_{ij} \sim N(\mu_i, \sigma^2)$, $i = 1,...,k$, $j = 1,...,m$, unabhängig. Dann gilt für die Erwartungswerte von $S^2_{k-sample}$, $S^2_{1-sample}$ und S^2_{adj}:

$$E(S^2_{k-sample}) = \sigma^2,$$

$$E(S^2_{1-sample}) = \sigma^2 + \frac{m}{km-1}\Delta^2,$$

$$E(S^2_{adj}) = \sigma^2 + \frac{m}{km-1}(\Delta^2 - \delta^2).$$

Für die Varianzen von $S^2_{k-sample}$, $S^2_{1-sample}$ und S^2_{adj} gilt:

$$Var(S^2_{k-sample}) = \frac{2\sigma^4}{k(m-1)},$$

$$Var(S^2_{1-sample}) = Var(S^2_{adj}) = \frac{2\sigma^4}{km-1}\left(1 + \frac{2m}{km-1}\frac{\Delta^2}{\sigma^2}\right).$$

Dabei bezeichnen $\Delta^2 = \sum_{i=1}^{k} (\mu_i - \bar{\mu})^2$ und $\delta^2 = \sum_{i=1}^{k} (\mu_i^* - \bar{\mu}^*)^2$ mit den unter der Alternativ-Hypothese angenommenen Erwartungswerten μ_i^*, $i = 1, ..., k$.

Beweis:

Nach Definition der nicht-zentralen Chi-Quadrat-Verteilung gilt für unabhängige Zufallsvariablen $Y_j \sim N(\eta_j, \sigma^2)$, $j = 1, ..., q$,

$$\frac{\sum_{j=1}^{q} (Y_j - \bar{Y}_\cdot)^2}{\sigma^2} \sim \chi^2_{q-1}(\vartheta),$$

wobei $\chi^2_{q-1}(\vartheta)$ die nicht-zentrale Chi-Quadrat-Verteilung mit $q-1$ Freiheitsgraden und Nicht-Zentralitätsparameter

$$\vartheta = \frac{\sum_{j=1}^{q} (\eta_j - \bar{\eta})^2}{\sigma^2}$$

bezeichnet. Weiterhin gilt für Erwartungswert und Varianz von $\chi^2_{q-1}(\vartheta)$:

$$E(\chi^2_{q-1}(\vartheta)) = (q-1) + \vartheta \tag{4.2}$$

$$Var(\chi^2_{q-1}(\vartheta)) = 2((q-1) + 2\vartheta) \tag{4.3}$$

(siehe z.B. JOHNSON und KOTZ, 1970). Daraus folgt $\frac{k(m-1)}{\sigma^2} S^2_{k-sample} \sim \chi^2_{k(m-1)}(\vartheta_{k-sample})$ und $\frac{km-1}{\sigma^2} S^2_{1-sample} \sim \chi^2_{km-1}(\vartheta_{1-sample})$ mit

$$\vartheta_{k-sample} = \frac{\sum_{i=1}^{k} m(\mu_i - \mu_i)^2}{\sigma^2} = 0, \tag{4.4}$$

$$\vartheta_{1-sample} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{m} (\mu_i - \bar{\mu})^2}{\sigma^2} = \frac{m}{\sigma^2} \sum_{i=1}^{k} (\mu_i - \bar{\mu})^2 . \tag{4.5}$$

Durch Einsetzen von (4.4) und (4.5) in (4.2) und (4.3) und Berücksichtigung der Definition $S^2_{adj} = S^2_{1-sample} - \frac{m}{km-1} \delta^2$ folgen die Behauptungen. ∎

Im folgenden Satz 8 sind die Konsequenzen für den resultierenden Stichprobenumfang bei Verwendung der oben genannten Varianzschätzer zur Fallzahladjustierung zusammengefasst.

Dort wird die für eine Vielzahl von Testproblemen gültige Annahme getroffen, dass sich der notwendige Stichprobenumfang pro Gruppe N durch eine Formel der folgenden Struktur approximieren lässt:

$$N = \lambda(k, \alpha, \beta) \cdot \frac{\sigma^2}{\delta^2}. \tag{4.6}$$

Dabei bezeichnet α das Niveau, $1-\beta$ die angestrebte Power und δ der minimale klinisch relevante Therapieeffekt mit $\delta^2 = \sum_{i=1}^{k} \left(\mu_i^* - \bar{\mu}^*\right)^2$ und den unter der Alternativ-Hypothese angenommenen Erwartungswerten μ_i^*, $i = 1, ..., k$.

Für die Null-Hypothese $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ und den F-Test gilt nach FRIEDE (2000a) die Darstellung (4.6) mit

$$\lambda(k, \alpha, \beta) = (\sqrt{\chi^2_{k-1,1-\alpha} - (k-2)} + z_{1-\beta})^2. \tag{4.7}$$

In Formel (4.7) bezeichnet $\chi^2_{k-1,1-\alpha}$ das $(1-\alpha)$ - Quantil der zentralen Chi-Quadrat-Verteilung mit $(k-1)$ Freiheitsgraden und $z_{1-\beta}$ das $(1-\beta)$ - Quantil der Standardnormalverteilung. Eine Fallzahl-Approximationsformel vom Typ (4.6) gilt darüber hinaus beispielsweise für Kontrast-Tests (siehe Kapitel 3.5; Herleitung der Approximationsformel analog zu den in KIESER und HAUSCHKE, 1999, verwendeten Methoden), für Äquivalenz-, Nicht-Unterlegenheits- und Überlegenheits-Studien in der Zwei-Stichproben-Situation und dem Parallelgruppen-Design (KIESER und HAUSCHKE, 1999) sowie in Crossover-Studien mit zwei Perioden und zwei Behandlungen (LIU und CHOW, 1992; KIESER und HAUSCHKE, 2000).

Satz 8:

Für $X_{ij} \sim N(\mu_i, \sigma^2)$, $i = 1,...,k$, $j = 1,...,m$, unabhängig, soll eine Null-Hypothese H_0 zum Niveau α getestet werden. Falls die durch die Erwartungswerte μ_i^*, $i = 1, ..., k$, spezifizierte Alternativ-Hypothese gilt, soll die Null-Hypothese mit der Wahrscheinlichkeit $1-\beta$ abgelehnt werden können. Die hierzu notwendige Fallzahl pro Gruppe N lasse sich approximativ mit einer Formel von der Struktur (4.6) bestimmen, wobei $\delta^2 = \sum_{i=1}^{k} \left(\mu_i^* - \bar{\mu}^*\right)^2$.

Im Rahmen einer internen Pilotstudie erfolge eine Fallzahl-Rekalkulation, bei der die mit $n_1 \leq N$ Beobachtungen pro Gruppe berechnete Varianzschätzung S^2 in die Approximationsformel (4.6) eingesetzt wird. Die resultierende Fallzahl pro Gruppe sei mit $\hat{N}(S^2)$ bezeichnet. Dann gilt:

$$\hat{N}(S^2_{1-sample}) = \hat{N}(S^2_{adj}) + \frac{m}{km-1} \cdot \lambda(k, \alpha, \beta) \tag{4.8a}$$

$$\approx \hat{N}(S^2_{adj}) + \frac{1}{k} \cdot \lambda(k, \alpha, \beta), \tag{4.8b}$$

$$E(\hat{N}(S^2_{adj})) = N + \frac{m}{km-1} \cdot \lambda(k, \alpha, \beta) \cdot \left(\left(\frac{\Delta}{\delta} \right)^2 - 1 \right) \tag{4.9a}$$

$$\approx N + \frac{1}{k} \cdot \lambda(k, \alpha, \beta) \cdot \left(\left(\frac{\Delta}{\delta} \right)^2 - 1 \right), \tag{4.9b}$$

$$Var(\hat{N}(S^2_{1-sample})) = Var(S^2_{adj}) = \frac{2\sigma^4}{km-1} \left(1 + \frac{2m}{km-1} \left(\frac{\Delta}{\sigma} \right)^2 \right)$$

$$= \frac{km-1}{k(m-1)} \cdot \left(1 + \frac{2m}{km-1} \left(\frac{\Delta}{\sigma} \right)^2 \right) \cdot Var(S^2_{k-sample}) \tag{4.10a}$$

$$\approx \left(1 + \frac{2}{k} \left(\frac{\Delta}{\sigma} \right)^2 \right) \cdot Var(S^2_{k-sample}). \tag{4.10b}$$

Beweis:

Der Beweis von (4.8a) folgt direkt durch Einsetzen von $S^2_{1-sample}$ und S^2_{adj} in (4.6). Die Beziehung (4.9a) folgt aus der Linearität von $\hat{N}(S^2)$ in S^2 und den Ergebnissen für die Erwartungswerte von $S^2_{1-sample}$ und S^2_{adj} in Satz 7. (4.10a) folgt aus den entsprechenden Ergebnissen für $Var(S^2_{1-sample})$ und $Var(S^2_{adj})$. Die Approximationen (4.8b), (4.9b) und (4.10b) ergeben sich direkt aus den für die üblichen Stichprobenumfänge gültigen Beziehungen $\frac{lm}{km-1} \approx \frac{l}{k}$, $l = 1, 2,$ und $\frac{km-1}{k(m-1)} \approx 1$. ■

In der folgenden Tabelle 8 ist für den Test von $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ mit dem F-Test der approximative Bias von $\hat{N}_{adj}$ nach (4.9a) angegeben, der aus einer Fehl-Spezifikation von Δ, d.h., für $\Delta = \tau \cdot \delta$, $\tau \neq 1$, bei der Berechnung von S^2_{adj} resultiert. Für festen Wert von τ ist der Bias monoton fallend mit steigender Anzahl k der Behandlungsgruppen, so dass der maximale Bias für $k = 2$ erreicht wird. Für $\tau = \sqrt{2}$ entspricht der Wert für den Bias dem Unterschied zwischen $\hat{N}_{1-sample}$ und $\hat{N}_{adj}$; diese Differenz ist nach (4.8b) unabhängig von den

Werten für Δ^2 und σ^2. Die Abweichung des Erwartungswertes von $\hat{N}_{adj}$ vom tatsächlich benötigten Stichprobenumfang pro Gruppe N aufgrund einer Fehl-Spezifikation von Δ ist selbst für den extremen Fall, dass der wahre Therapiegruppen-Unterschied 50% größer als die klinisch relevante Differenz ist, moderat; für die Mehrzahl von Studien, die im Design mit interner Pilotstudie durchgeführt werden, ist dieser Unterschied irrelevant. Das gleiche gilt für den Unterschied zwischen der Fallzahl bei Verwendung der unadjustierten Ein-Stichproben-Varianz $S^2_{1-sample}$ bzw. der adjustierten Variante S^2_{adj}.

Tabelle 8: Approximativer Bias von $\hat{N}_{adj}$ nach (4.8a), der aus einer Fehl-Spezifikation von Δ bei Berechnung von S^2_{adj} resultiert, für den Test von $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ mit dem *F*-Test. Zweiseitiges Signifikanzniveau $\alpha = 0.05$, Power $1-\beta = 0.80\,(0.90)$, Anzahl Behandlungsgruppen $k = 2, ..., 5$, wahrer Behandlungsgruppen-Unterschied Δ, unter Alternativ-Hypothese angenommener Behandlungsgruppen-Unterschied δ und $\Delta = \tau \cdot \delta$. (Die Werte für $\tau = \sqrt{2}$ entsprechen der Differenz $\hat{N}_{1-sample} - \hat{N}_{adj}$ der Fallzahlen pro Gruppe bei Verwendung der nicht-adjustierten und der adjustierten Ein-Stichproben-Varianz).

	Approximativer Bias von $\hat{N}_{adj}$			
k	2	3	4	5
τ				
0.5	-3	-2	-2	-2
	(-4)	(-3)	(-3)	(-2)
$\sqrt{2}$	4	3	3	2
	(5)	(4)	(3)	(3)
1.5	5	4	3	3
	(7)	(5)	(4)	(4)
2.0	12	9	8	7
	(16)	(12)	(10)	(9)

Formel (4.10b) gibt Aufschluss über die Variabilität der resultierenden Fallzahlschätzungen. Demnach ist die Varianz der Fallzahl pro Gruppe bei Verwendung der verblindeten Schätzer $S^2_{1-sample}$ oder S^2_{adj} um den Faktor $1+\frac{2}{k}\left(\frac{\Delta}{\sigma}\right)^2$ größer ist als für den Schätzer $S^2_{k-sample}$, der Entblindung voraussetzt. Unter der Annahme $\delta = \Delta$ ist dieser Faktor beispielsweise stets ≤ 1.25 für $\left(\frac{\delta}{\sigma}\right)^2 \leq \frac{k}{8}$. Für die Hypothese $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ und den F-Test entspricht dies für $\alpha = 0.05$ und $1-\beta = 0.80$ (0.90) notwendigen Fallzahlen pro Gruppe von 31 (42) für $k = 2$ und 18 (23) für $k = 5$. Da in der Praxis derartige Fallzahlen eher eine untere Grenze darstellen, um die Implementierung einer internen Pilotstudie überhaupt in Erwägung zu

ziehen, ist der „Preis“, der für die Nicht-Entblindung bezüglich der Varianzerhöhung der resultierenden Fallzahl zu bezahlen ist, in der Regel moderat.

Zusammenfassend wird aus diesen Betrachtungen deutlich, dass es sowohl hinsichtlich des Bias’ als auch der Variabilität der resultierenden Fallzahl in aller Regel keinen für die Praxis relevanten Unterschied bedeutet, ob man eine sehr einfache verblindete Varianzschätzung wie die adjustierte oder nicht-adjustierte Ein-Stichproben-Varianz der gepoolten Stichprobe oder den entblindeten Varianzschätzer verwendet. Dieses Ergebnis ist nicht zuletzt vor dem Hintergrund der im nächsten Kapitel beschriebenen EM-Algorithmus-basierten Methode zur verblindeten Varianzschätzung im Auge zu behalten.

4.1.1.2 EM-Algorithmus-basierter Varianzschätzer

Die Anwendung des EM-Algorithmus’ auf das Problem der verblindeten Varianzschätzung für eine Mischung von Normalverteilungen wurde von GOULD und SHIH (1992) für $k = 2$ Behandlungsgruppen vorgeschlagen. Im folgenden wird die Verallgemeinerung dieser Prozedur für beliebiges $k \geq 2$ beschrieben (siehe hierzu auch KIESER und FRIEDE, 2000b). In der Situation verblindeter Behandlungsgruppen-Zugehörigkeiten können dieselbigen als fehlende Werte angesehen werden. Entsprechend des Vorschlags von GOULD und SHIH (1992) geht die EM-Algorithmus-basierte Prozedur dann wie folgt vor. Nach einer adäquaten Initialisierung der Gruppenmittelwerte und der Varianz wird in einem sogenannten E-Schritt für jede Beobachtung die bedingte Wahrscheinlichkeit berechnet, einer bestimmten Behandlungsgruppe anzugehören, gegeben die aktuellen Parameterschätzungen. Die Berechnung erfolgt mit dem Bayes’schen Theorem unter der Annahme einer *a priori* Wahrscheinlichkeit von $1/k$ für jede Beobachtung, einer bestimmten Behandlungsgruppe anzugehören. Im folgenden M-Schritt werden diese bedingten Wahrscheinlichkeiten verwendet, um Maximum-Likelihood-Schätzer der Erwartungswerte und der Varianz für die vollständigen Daten zu berechnen. Die E- und M-Schritte werden wiederholt, bis sich zwei aufeinander folgende Schätzungen für die Varianz um weniger als ein vorgegebenes ε unterscheiden (Vorschlag von GOULD und SHIH, 1992, mit $\varepsilon = 0.01$) bzw. bis sich die aufeinander folgenden Schätzungen für alle Parameter um nicht mehr als eine vorgegebene Genauigkeit unterscheiden (Vorschlag von SHIH, 1992, mit $\varepsilon = 0.001$).

Um geeignete Werte für die Initialisierung zu finden, wurde von GOULD und SHIH (1992) vorgeschlagen, wie folgt vorzugehen: Aus dem Q-Q-Plot der beobachteten Daten gegen die

Perzentile der Normalverteilung wird mit der Kleinste-Quadrate-Methode die zugehörige Regressionsgerade geschätzt. Die Steigung dieser Geraden wird als Startwert für die Standardabweichung verwendet. Einen Startwert für den Wertebereich der Behandlungsgruppen-Mittelwerte erhält man, indem man den Startwert für σ mit einer Schätzung für den maximalen standardisierten Behandlungseffekt $\theta = (\mu_{\max} - \mu_{\min})/\sigma$ multipliziert, wobei $\mu_{\max} = \max_{i \in \{1,\ldots,k\}} \mu_i$ und $\mu_{\min} = \min_{i \in \{1,\ldots,k\}} \mu_i$. Von GOULD und SHIH (1992) wurde als „typischer" standardisierter Behandlungseffekt in klinischen Studien der Wert 0.35 vorgeschlagen. Die Startwerte für die $\mu_i, i = 1, \ldots, k$, erhält man aus diesem Wertebereich durch symmetrische und äquidistante Wahl um die Abszisse der geschätzten Regressionslinie. Beachtet man, dass die Steigung der Regressionsgeraden ungefähr identisch mit der geschätzten Ein-Stichproben-Standardabweichung $S_{1-sample}$ ist und die Abszisse mit dem Mittelwert aller Beobachtungen der gepoolten Daten übereinstimmt, so lassen sich sehr ähnliche Startwerte wesentlich einfacher und ohne Berechnung der Regressionsgeraden bestimmen. Eine detaillierte Beschreibung der EM-Schritte findet sich in GOULD und SHIH (1992) sowie in FRIEDE (2000b) und FRIEDE und KIESER (2000b). In den letztgenannten Arbeiten ist auch das zugrundeliegende statistische Modell und sein Zusammenhang mit dem verwendeten Randomisierungsverfahren beschrieben.

Review der in der Literatur berichteten Eigenschaften des EM-Algorithmus-basierten Varianzschätzers

Nach der Publikation von 1992 wurden in einer Vielzahl von Folgearbeiten von GOULD und SHIH zahlreiche Eigenschaften der EM-Algorithmus-basierten Prozedur untersucht. Sämtliche Ergebnisse basieren allerdings auf Monte-Carlo-Simulationen, aus denen die folgenden Schlussfolgerungen gezogen wurden. Die Prozedur liefert demnach Varianzschätzungen, die sehr gut mit den wahren Werten von σ^2 übereinstimmen (GOULD und SHIH, 1992; GOULD, 1995; GOULD, 1997), und die nicht von der Wahl der Startwerte abhängen (GOULD und SHIH, 1992; SHIH, 1992). Das eigentliche Ziel der Varianzschätzung ist deren Verwendung für die Fallzahladjustierung. Ergebnisse verschiedener Simulationsstudien legen nahe, dass die EM-Algorithmus-basierte Prozedur die gewünschte Power erreicht, auch wenn in der Planungsphase σ^2 falsch spezifiziert wurde (SHIH und GOULD, 1995; GOULD, 1997; GOULD und SHIH, 1998; SHIH und LONG, 1998). Falls die EM-Algorthmus-basierte Prozedur

tatsächlich gute Schätzer für die Varianz innerhalb der Behandlungsgruppen liefert, scheint daraus zu folgen, dass man über die Varianzzerlegung (4.1) Informationen über den tatsächlichen Behandlungsgruppen-Unterschied in Erfahrung bringen kann, ohne den Randomisierungs-Code zu brechen. GOULD (1995, 1997) schloss aus den Ergebnissen seiner Simulationen, dass der vorgeschlagene Algorithmus die Varianz zwar gut schätzt, aber mit Hilfe der Relation (4.1) keine Schlussfolgerungen über die Gruppenunterschiede gezogen werden können. Aufgrund des algebraischen Zusammenhangs zwischen den beiden Größen ist dies kein intuitiv einleuchtender Befund. Darüber hinaus wird berichtet, dass auch aus den Schätzungen, die der Algorithmus für die Erwartungswerte liefert, keine Information über den tatsächlichen Therapieeffekt abgeleitet werden kann. Auch diese Eigenschaft ist in hohem Maße unplausibel, da der Algorithmus keinen der zu schätzenden Parameter vor dem anderen auszeichnet und deshalb zu erwartet ist, dass entweder alle oder keine der Schätzungen brauchbar sind. Weiterhin wurde in Simulationen gezeigt, dass die Prozedur robust gegen das Vorliegen von Heterogenität und Block-Effekten ist, wie sie z.B. in Multicenter-Studien auftreten, und damit das Verfahren auch für diese Anwendungssituation geeignet erscheint (SHIH und LONG, 1998). Außerdem wurde die Verwendung der Prozedur in Longitudinal-Studien (SHIH und GOULD, 1995) sowie in Studien mit gruppensequentiellem Design (GOULD und SHIH, 1998) vorgeschlagen.

Neuere Ergebnisse über Eigenschaften des EM-Algorithmus-basierten Varianzschätzers

Es ist naheliegend, einen Vergleich zwischen den einfachen Varianzschätzern und dem komplizierten EM-Algorithmus-basierten Verfahren durchzuführen, zumal in der Arbeit, in der die Prozedur eingeführt wurde (GOULD und SHIH, 1992), auch eine Variante der adjustierten Ein-Stichproben-Varianz beschrieben wurde. Dennoch beruhen sämtliche im vorangehenden Abschnitt genannten Ergebnisse zum EM-Algorithmus-basierten Verfahren auf Simulationsuntersuchungen ohne Berücksichtigung eines alternativen Verfahrens als Kontrolle. In der Arbeit von KIESER und FRIEDE (2000b) wurden deshalb in einer Monte-Carlo-Simulationsstudie die Eigenschaften des EM-Algorithmus-basierten Verfahrens mit denen der einfachen Varianzschätzer verglichen. Betrachtet wurden die Situationen $k = 3$ und $k = 5$ bei einer Populationsvarianz von $\sigma^2 = 1$ und verschiedenen Szenarien bezüglich der Fallzahlen pro Gruppe der internen Pilotstudie und der Erwartungswerte der Behandlungsgruppen, mit maximalen Behandlungsgruppen-Unterschieden zwischen 0.2 und

0.7. Für Werte von $\alpha = 0.05, 1-\beta = 0.80$ und den F-Test entsprechen diesen Situationen Fallzahlen pro Gruppe zwischen 26 und 483. In der folgenden Tabelle 9 sind die Mittelwerte und Standardabweichungen für die Schätzwerte von σ^2 angegeben, die in 10 000 Replikationen für $n_1 = 30$ ermittelt wurden. Für die EM-Algorithmus-basierte Prozedur wurden die Iterationen durchgeführt bis zwei aufeinanderfolgende Schätzungen der Varianz um weniger als $\varepsilon = 0.001$ differierten. Wie wir von Satz 7, Kapitel 4.1.1, wissen, lassen sich für die einfachen Varianzschätzer die Mittelwerte und Standardabweichungen auch analytisch berechnen. Aus Gründen der Vergleichbarkeit sind im folgenden auch für diese Schätzer die Simulationsergebnisse angegeben.

Tabelle 9: Simulierte Mittelwerte (Standardabweichung) der Varianzschätzer bei einem wahren Wert $\sigma^2 = 1$ und Fallzahl pro Gruppe n_1 für die Varianzschätzung (10 000 Replikationen).

	Simulierter Mittelwert (SD) der geschätzten Varianz			
Szenario ($\mu_1, \mu_2, ..., \mu_k$)	EM-Algorithmus-basierter Varianzschätzer	Adj. Ein-Stichproben-Varianzschätzer	Ein-Stichproben-Varianzschätzer	*k*-Stichproben-Varianzschätzer
(0.0, 0.1, 0.2)	0.978 (.146)	1.002 (.151)	1.009 (.151)	1.002 (.151)
(0.0, 0.175, 0.35)	0.991 (.148)	1.001 (.153)	1.022 (.153)	1.002 (.152)
(0.0, 0.25, 0.5)	1.009 (.151)	0.999 (.156)	1.041 (.156)	0.999 (.151)
0.0, 0.35, 0.7)	1.051 (.157)	1.002 (.162)	1.084 (.162)	1.001 (.152)
(0.0, 0.05, 0.1, 0.15, 0.2)	0.984 (.114)	1.001 (.114)	1.006 (.117)	1.001 (.118)
(0.0, 0.875, 0.175, 0.265, 0.35)	0.993 (.115)	.0999 (.118)	1.015 (.118)	0.999 (.117)
(0.0, 0.125, 0.25, 0.375, 0.5)	1.011 (.118)	1.002 (.121)	1.033 (.121)	1.001 (.119)
(0.0, 0.175, 0.35, 0.525, 0.7)	1.039 (.121)	1.000 (.123)	1.062 (.123)	1.000 (.119)
(0.0, 0.0, 0.1, 0.2, 0.2)	0.986 (.115)	0.999 (.118)	1.007 (.118)	1.000 (.119)
0.0, 0.0, 0.175, 0.35, 0.35)	1.002 (.116)	1.000 (.119)	1.024 (.119)	1.000 (.117)
(0.0, 0.0, 0.25, 0.5, 0.5)	1.029 (.119)	1.001 (.121)	1.052 (.121)	1.002 (.117)
(0.0, 0.0, 0.35, 0.7, 0.7)	1.075 (.124)	1.000 (.126)	1.098 (.126)	1.000 (.117)

Die Ergebnisse zeigen, dass zwischen den verschiedenen Verfahren lediglich geringfügige Unterschiede bestehen. Der Mittelwert der geschätzten Varianz hängt bei der EM-Algorithmus-basierten Prozedur vom (unbekannten) maximalen Behandlungsgruppen-Unterschied ab, wobei für kleinere Effekte die tatsächliche Varianz unterschätzt und für

größere überschätzt wird. Die Standardabweichung der geschätzten Varianzen ist für die EM-Algorithmus-basierte Prozedur weitgehend identisch mit denen der anderen Varianzschätzer.

Auch Vergleiche zwischen den Varianzschätzverfahren bezüglich anderer Charakteristika (z.B. Betrachtung der Wahrscheinlichkeit, vorgegebene Werte für die Power zu erreichen bzw. zu unterschreiten) führten zu dem Ergebnis, dass sich die Methoden nur marginal unterscheiden.

Diese Befunde lassen Zweifel aufkommen, ob die in der Literatur und der Anwendungspraxis bislang zu beobachtende Bevorzugung des komplexen EM-Algorithmus-basierten Verfahrens gegenüber den einfachen Schätzverfahren gerechtfertigt ist. Eine tiefer gehende Untersuchung zeigt, dass die Prozedur von GOULD und SHIH grundsätzliche Defizite aufweist, die zur Konsequenz haben, dass sie zur verblindeten Varianz- und Fallzahlschätzung nicht brauchbar ist. Eine detaillierte Darstellung der Methoden und Ergebnisse sowie eine Diskussion dieses Problemkreises findet sich in FRIEDE (2000b) und FRIEDE und KIESER (2000b).

Dem EM-Algorithmus-basierten Verfahren sind folgende Defizite inhärent:

1. Die Information über das verwendete Randomisierungsverfahren wird beim Algorithmus nicht berücksichtigt:

Wie in FRIEDE (2000b) und FRIEDE und KIESER (2000b) gezeigt wird, hängt das statistische Modell und damit die Likelihood-Funktion, die dem Problem zugrunde liegt, von dem in der konkreten Anwendungssituation benutzten Randomisierungsverfahren ab. Die *a priori* Wahrscheinlichkeit $1/k$ für die Zugehörigkeit zu einer speziellen der k Behandlungsgruppen, die von GOULD und SHIH für den Algorithmus vorgeschlagen wurde, nimmt implizit an, dass die Randomisierung entsprechend dem Verfahren durchgeführt wird, das für $k = 2$ der Münzrandomisierung entspricht. Falls ein anderes Randomisierungsverfahren, beispielsweise die Blockrandomisierung, zum Einsatz kommt, muss die Likelihood-Funktion entsprechend modifiziert werden.

2. Das Stop-Kriterium sichert nicht die Konvergenz des Algorithmus':

Dieser Sachverhalt liegt darin begründet, dass die Differenz zwischen den Schätzern zweier aufeinander folgender EM-Schritte nicht monoton abnimmt, sondern zwar typischerweise bereits nach wenigen Iterationen „klein" ist, danach aber deutlich anwächst und sich erst anschließend nach Überschreiten eines Maximums streng monoton dem Wert Null annähert. Iteriert man die Prozedur so lange, bis Konvergenz erreicht ist, so können hierzu, unter anderem abhängig von der Initialisierung, teilweise mehrere tausend EM-Schritte notwendig

sein. In den Abbildungen 6 und 7 des folgenden Kapitels 4.1.1.3 ist dieser Sachverhalt für zwei Anwendungsbeispiele anhand der Darstellung der Schätzwerte und der Differenzen aufeinander folgender Schätzungen in Abhängigkeit vom Iterationsschritt illustriert.

3. Die Schätzwerte hängen von der Initialisierung ab:
Dieser Befund steht im Widerspruch zu den oben zitierten anderslautenden Behauptungen in der Literatur und ist eine direkte Konsequenz des vorgenannten Defizits: Für verschiedene Initialisierungen wird das Stop-Kriterium nach unterschiedlicher Zahl von Iterationsschritten mit jeweils unterschiedlichen aktuellen Schätzungen erreicht bevor es zu einer Konvergenz des Algorithmus‘ gekommen ist. Daraus resultieren unterschiedliche Schätzwerte.

Paradoxerweise erklärt das Konvergenzverhalten des Algorithmus‘ zusammen mit der angewendeten Stop-Regel, warum die zahlreichen Simulationsuntersuchungen zur EM-Algorithmus-basierten Prozedur vernünftige Ergebnisse lieferten: Das vorgegebene Stop-Kriterium führt in den meisten Fällen dazu, dass nur sehr wenige Iterationen durchgeführt werden. Der resultierende Varianzschätzer liegt damit sehr nahe bei dem Startwert (aber konsistent darunter), der, wie wir oben gesehen haben, selbst nicht weit von der Ein-Stichproben-Varianz entfernt ist.

Die Defizite 1 und 2 (und damit auch 3) ließen sich durch entsprechende Modifikationen des Verfahrens eliminieren, um damit ein adäquates Verfahren zur Maximum-Likelihood-Schätzung der Varianz bei unbekannter Gruppenzugehörigkeit zu erhalten. Das folgende Ergebnis aus FRIEDE (2000b) und FRIEDE und KIESER (2000b) zeigt, dass dieser Weg nicht erfolgversprechend ist: In der Situation verblindeter Behandlungsgruppen-Zugehörigkeit und Fallzahlen der internen Pilotstudie, die kleiner als der tatsächlich notwendige Stichprobenumfang sind, liefert die Maximum-Likelihood-Schätzung Werte für die Varianz, die bei großer Streuung deutlich unter dem wahren Wert liegen. Damit scheidet das EM-Algorithmus-basierte Verfahren endgültig als ernsthafter Konkurrent zu den einfachen Methoden aus.

Im folgenden Kapitel wird die praktische Anwendung der verschiedenen Verfahren zur verblindeten Fallzahladjustierung an zwei klinischen Studien, die im Design mit interner Pilotstudie durchgeführt wurden, vorgestellt. Insbesondere werden an diesen Datensätzen die Defizite 2 und 3 der EM-Algorithmus-basierten Prozedur illustriert.

4.1.1.3 Anwendungsbeispiele

Beispiel 4:

In einer randomisierten, doppelblinden Multicenter-Studie, die zum Zeitpunkt der Erstellung dieser Schrift noch nicht abgeschlossen war, wird die Wirksamkeit und Verträglichkeit zweier Dosierungen eines Antidementivums mit Placebo bei Patienten mit Alzheimer Krankheit verglichen. Der 26-wöchigen doppelblinde Therapiephase geht eine 4-wöchige medikationsfreie Run-in-Phase voraus, in der die Untersuchungen zur Verifizierung der Einschlussdiagnose durchgeführt werden. Entsprechend den Vorgaben in einschlägigen Guidelines (LEBER, 1990; CPMP, 1997) ist die Wirksamkeit von Antidementiva in den beiden Ebenen „kognitive Leistungsfähigkeit“ und „globales klinisches Zustandsbild“ nachzuweisen. Als Messinstrument zur Beurteilung von Therapieeffekten bezüglich des erstgenannten Bereiches wurde in dieser Studie die kognitive Subskala der Alzheimer’s Disease Assessment Scale (ADAS-cog; Zielgröße: Differenz zwischen Therapiebeginn und Therapieende (Woche 26)) verwendet (ROSEN, MOHS und DAVIS, 1984), die gewissermaßen das Standardverfahren für diese Fragestellung darstellt. Zur Erfassung der Wirksamkeit in der zweiten Ebene wurde der Fragebogen Alzheimer’s Disease Cooperative Study - Clinical Global Impression of Change (ADCS-CGIC) eingesetzt (Zielgröße: ADCS-CGIC bei Therapieende (Woche 26)). Dieses Verfahren wurde durch die Alzheimer’s Disease Cooperative Study Group (ADCS) im Hinblick auf eine Empfehlung der US-amerikanischen Food and Drug Administration (FDA) entwickelt. Entsprechen dieses Vorschlages der FDA soll in klinischen Studien mit antidementiven Medikamenten eine globale Beurteilung verwendet werden, die auf einem Interview des Patienten durch einen Kliniker basiert (SCHNEIDER *et al*., 1997). Im Rahmen der konfirmatorischen Auswertung der vorliegenden Studie ist die Wirksamkeit simultan für beide Zielgrößen nachzuweisen, und zwar möglichst für beide aktiven Behandlungsgruppen gegen Placebo und, falls ein deutlicher Dosis-Trend erkennbar ist, unter Umständen auch für die höhere Dosierung gegen die niedrigere. Damit liegt ein komplexes multiples Testproblem vor, dessen Umsetzung in eine adäquate Analysestrategie (Festlegung der multiplen Testprozedur und der Tests für die entsprechenden Hypothesen) erheblichen Einfluss auf den notwendigen Stichprobenumfang hat. Um die Betrachtungen nicht unnötig zu verkomplizieren, nehmen wir im folgenden an, dass die Veränderung des ADAS-cog die einzige Zielgröße ist, und dass die Auswertung mit dem *F*-Test zum zweiseitigen Niveau $\alpha = 0.05$ erfolgen soll. Die Planungsannahmen für die Fallzahlberechnung $(1-\beta = 0.90)$ waren $\delta = 2.04$ $(\mu_1 = 0.0, \mu_2 = \mu_3 = 2.5)$ und $\sigma = 6.0$. Im Prüfplan war festgelegt, dass die Standardabweichung nach der Rekrutierung von 80% der Patienten (jedoch auf der Basis von

mindestens 60 abgeschlossenen Patienten) im Rahmen einer internen Pilotstudie geschätzt werden und der Stichprobenumfang falls notwendig angepasst werden sollte. Im ursprünglichen Protokoll war zur verblindeten Varianzschätzung die von GOULD und SHIH vorgeschlagene Methode, erweitert auf $k = 3$ Behandlungsgruppen (siehe Kapitel 4.1.1.2 und KIESER und FRIEDE, 2000b) vorgesehen. Nachdem die oben genannten Defizite der EM-Algorithmus-basierten Prozedur evident wurden, wurde in einer Prüfplanänderung vor Durchführung der Varianzschätzung festgelegt, dass die adjustierte Ein-Stichprobenvarianz-Methode verwendet werden sollte.

Die interne Pilotstudie umfasste $n = 96$ Patienten und lieferte Werte für die nicht-adjustierte bzw. adjustierte Ein-Stichproben-Standardabweichungen von $s_{1-sample} = 5.83$ bzw. $s_{adj} = 5.71$ (KIESER, 1999). In Abbildung 6 sind die Schätzwerte für die Standardabweichung sowie die Differenzen aufeinander folgender Schätzungen der EM-Algorithmus-basierten Prozedur für die Initialisierungen $\theta = 0.35$ (schwarze Linien) und $\theta = 0.50$ (graue Linien) in Abhängigkeit vom Iterationsschritt dargestellt.

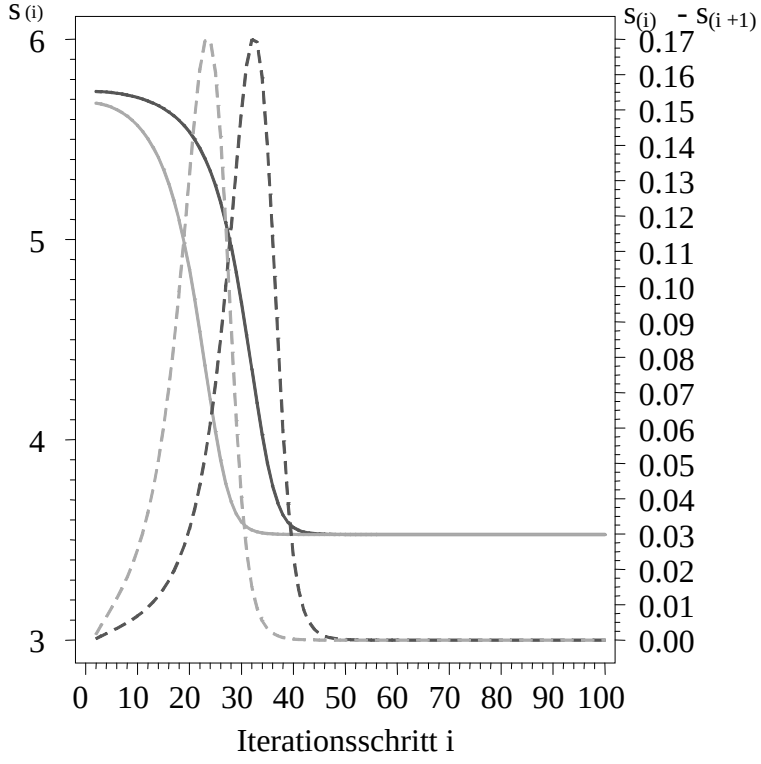

Abbildung 6: Schätzwerte für die Standardabweichung $s_{(i)}$ (durchgezogene Linien) sowie Differenzen aufeinander folgender Schätzungen $s_{(i)} - s_{(i+1)}$ (gestrichelte Linien) der EM-Algorithmus-basierten Prozedur für die Initialisierungen $\theta = 0.35$ (schwarze Linien) und $\theta = 0.50$ (graue Linien) in Abhängigkeit vom Iterationsschritt. Dreiarmige Studie in der Indikation Alzheimer Krankheit, Zielgröße Differenz des ADAS-cog Scores zwischen Therapiebeginn und -ende.

Man erkennt den oben beschriebenen typischen Verlauf der EM-Schritte mit zunächst kleinen Differenzen zwischen den sukzessiven Schätzwerten, die dann deutlich anwachsen um nach Überschreiten eines Maximalwertes gegen Null zu konvergieren. Wählt man, wie von GOULD und SHIH (1992) vorgeschlagen, als Kriterium für den Stop des Algorithmus‘ das Unterschreiten einer Schranke ε für die Differenz der geschätzten Varianzen, so können völlig unterschiedliche Ergebnisse resultieren. Beispielsweise stoppt für $\theta = 0.35$ und $\varepsilon = 0.01$ die Iteration bereits nach zwei Schritten ($s_{(1)} - s_{(2)} = 0.008$) mit einer geschätzten Standardabweichung von $s_{(2)} = 5.72$. Der Algorithmus ist aber zu diesem Zeitpunkt noch weit von der Konvergenz entfernt, denn der Unterschied zwischen zwei Schätzwerten nimmt danach deutlich zu bis zu dem maximalen Wert $s_{(33)} - s_{(34)} = 0.17$ nach 33 Iterationsschritten, und unterschreitet erst im 46. Zyklus die Genauigkeit $\varepsilon = 0.001$. Der Schätzwert für die Standardabweichung von $s_{(46)} = 3.53$ verändert sich danach auch bei fortgesetzter Iteration nicht mehr. Der Wert, gegen den der Algorithmus konvergiert, ist für beide Initialisierungen gleich, er wird für die kleinere Initialisierungskonstante aber später erreicht.

Unter den ursprünglichen Planungsannahmen wäre ein Stichprobenumfang pro Gruppe von 92 notwendig. Der geschätzten nicht-adjustierten bzw. adjustierten Ein-Stichproben-Standardabweichung entsprechen Fallzahlen von 87 bzw. 83, so dass auf der Basis dieser Schätzungen nicht unbedingt die Notwendigkeit zu einer Änderung der ursprünglich vorgesehenen Anzahl zu rekrutierender Patienten besteht. Hätte man jedoch die EM-Algorithmus-basierte Prozedur mit dem Schätzwert, der sich nach Konvergenz des Algorithmus‘ ergibt, verwendet, so hätte man geschlussfolgert, dass es ausreicht, lediglich 33 Patienten pro Behandlungsgruppe in die Studie einzuschließen. Nimmt man die Planungsannahmen als korrekt an, so beträgt die Power mit der durch die Prozedur von GOULD und SHIH nahegelegten Fallzahl nicht wie gewünscht 90% sondern lediglich 52%. Zwar ist die Studie derzeit noch nicht abgeschlossen und deshalb die gepoolte entblindete k-Stichproben-Standardabweichung noch nicht verfügbar, doch kann aufgrund der oben angegebenen Defizite des Verfahrens nach GOULD und SHIH und der Ergebnisse für die einfachen verblindeten Schätzmethoden davon ausgegangen werden, dass die aus der EM-Algorithmus-basierte Prozedur resultierende Schätzung deutlich unter der gepoolten Standardabweichung liegt und die resultierende Fallzahl die tatsächlich notwendige erheblich unterschätzt.

Dass dieses Verhalten des EM-Algorithmus’ nicht darauf zurückzuführen ist, dass der ursprüngliche Vorschlag von GOULD und SHIH (1992) für zwei Behandlungsgruppen auf die

Situation $k > 2$ verallgemeinert wurde oder der ADAS-cog-Score eine ordinal-skalierte Variable ist (allerdings mit einem breiten Wertebereich zwischen 0 und 70 Punkten), zeigt das folgende Beispiel für $k = 2$ und normalverteilte Zielgröße.

Beispiel 5:

In einer randomisierten, doppelblinden referenz-kontrollierten Multicenter-Studie mit 12 Wochen Behandlungsdauer wurde die Anwendung von Budesonid mit zwei unterschiedlichen Inhalationssystemen bei Patienten mit leichtem bis mittelschwerem Brochialasthma untersucht (MITFESSEL, ERXLEBEN, SCHULZE und KIESER, 1999). Ziel dieser Studie war es, die Äquivalenz eines neuen Flourchlorkohlenwasserstoff (FCKW)-freien Inhalators, des MAGhalers®, und eines konventionellen FCKW-haltigen Inhalators nachzuweisen. Die Zielgröße war der Mittelwert des forcierten exspiratorischen Volumens in 1 Sekunde (FEV_1) der Messungen nach 4, 8 und 12 Wochen Therapie. In der Planungsphase konnte über eine Literaturrecherche keine Studie mit dieser Zielgröße identifiziert werden. Als Grundlage der Fallzahlplanung diente deshalb eine Schätzung der Varianz, die über Extrapolation der Ergebnisse einer abgeschlossenen Studie mit einer wesentlich kürzerer Behandlungszeit gewonnen wurde. Aufgrund dieser unsicheren Vorinformation wurde im Prüfplan vorgesehen, die Planungsannahme $\sigma = 0.70$ mit der EM-Algorithmus-basierten Prozedur nach GOULD und SHIH zu überprüfen, nachdem 60 prüfplan-konforme Patienten die Studie abgeschlossen haben.

Für die Varianzschätzung waren die Daten von 61 Per-protocol-Patienten verfügbar. Für die Initialisierung $\theta = 0.35$ lag die Differenz zweier aufeinander folgender Schätzungen der Standardabweichung nach lediglich zwei EM-Schritten unter der spezifizierten Genauigkeit von $\varepsilon = 0.01$. Der Algorithmus stoppte mit einem geschätzten Wert von $s_{(2)} = 0.705$, der praktisch identisch mit dem für σ angenommenen Wert ist. Allerdings zeigt die Differenz zweier aufeinander folgender Schätzungen wiederum den bereits aus Kapitel 4.1.1.2 und aus Beispiel 4 bekannten Verlauf (Abbildung 7). Das Maximum der Differenz wird für $\theta = 0.35$ hier erst nach 130 Iterationsschritten erreicht, und führt man mehr als 170 Iterationen durch, so erhält man als Schätzwert $s_{(170)} = 0.52$, was in krassem Widerspruch zur Planungsannahme steht. Falls man dem Schätzwert, gegen den das EM-Algorithmus-basierte Verfahren konvergiert, Glauben schenkt, so würde dies bedeuten, dass man nur 55% der ursprünglich geplanten Patientenzahl benötigen würde, um die vorgesehene Power zu erreichen (Fallzahlplanung nach KIESER und HAUSCHKE, 1999).

Die Ein-Stichproben-Schätzung der Standardabweichung ist unter der Alternativ-Hypothese der gleichen Wirksamkeit der beiden Inhalationssysteme unverzerrt und in dieser Situation identisch mit der adjustierten Schätzung. Der erhaltene Schätzwert ist $s_{1-sample} = s_{adj} = 0.72$ und liegt damit so nahe bei der Planungserwartung, dass keine Fallzahladjustierung notwendig ist.

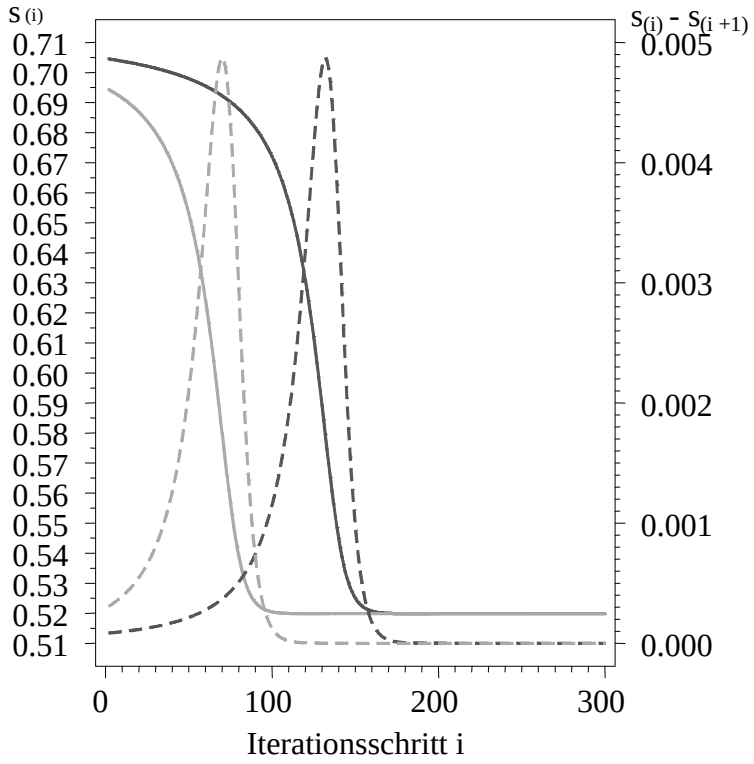

Abbildung 7: Schätzwerte für die Standardabweichung $s_{(i)}$ (durchgezogene Linien) sowie Differenzen aufeinander folgender Schätzungen $s_{(i)} - s_{(i+1)}$ (gestrichelte Linien) der EM-Algorithmus-basierten Prozedur für die Initialisierungen $\theta = 0.35$ (schwarze Linien) und $\theta = 0.50$ (graue Linien) in Abhängigkeit vom Iterationsschritt. Zweiarmige Äquivalenzstudie in der Indikation leichtes bis mittelschweres Brochialasthma, Zielgröße Mittelwert des forcierten exspiratorischen Volumens in 1 Sekunde (FEV_1) der Messungen nach 4, 8 und 12 Wochen Therapie.

4.1.2 Kontrolle der Wahrscheinlichkeit eines Fehlers 1. Art

4.1.2.1 Fallzahladaption mit entblindetem Varianzschätzer

Simulationsuntersuchungen von WITTES und BRITTAIN (1990) und BIRKETT und DAY (1994) zeigten, dass für das Design mit interner Pilotstudie bei Verwendung des entblindeten gepoolten Varianzschätzers zur Fallzahladjustierung und der üblichen *t*-Teststatistik zur Auswertung eine Überschreitung der nominalen Wahrscheinlichkeit eines Fehlers 1. Art auftreten kann. Die tatsächliche Fehlerrate lag zwar in den meisten praktisch relevanten Situationen nicht wesentlich über dem vorgegebenen Niveau α, doch führten die mit dieser Frage verbundenen Unwägbarkeiten offensichtlich dazu, den Passus „ ... *the consequences* (of a sample size adjustment)*, if any for the type I error ... should be explained*" in die ICH E9-Guideline (ICH, 1999) aufzunehmen. Erstaunlicherweise dauerte es nach Erscheinen der Arbeit von WITTES und BRITTAIN fast 10 Jahre, bis für das Problem der Fehlerkontrolle Lösungen publiziert wurden. Unabhängig voneinander aber nahezu gleichzeitig fanden mehrere Forschungsgruppen unterschiedliche Zugänge zur Berechnung bzw. Kontrolle der Wahrscheinlichkeit eines Fehlers 1. Art.

DENNE und JENNISON (1999) schlugen vor, zur Kontrolle des α-Niveaus die Freiheitsgrade des *t*-Quantils, das den Ablehnungsbereich des *t*-Tests festlegt, zu adjustieren, um so der Tatsache Rechnung zu tragen, dass die Gesamtfallzahl aufgrund einer Varianzschätzung aus den Studiendaten erfolgt. Obwohl die Idee intuitiv einleuchtend ist, bleibt dieser Ansatz unbefriedigend, weil der Wert zur Adjustierung der Freiheitsgrade durch Simulationen ermittelt wurde und die Einhaltung des nominalen Niveaus ebenfalls nur durch Simulationen überprüft wurde. Die analytische Berechnung der tatsächlichen Wahrscheinlichkeit eines Fehlers 1. Art gelang COFFEY und MULLER (1999) für das lineare Modell und WITTES, SCHABENBERGER, ZUCKER, BRITTAIN und PROSCHAN (1999) und KIESER und FRIEDE (2000a) für die *t*-Test-Situation. Durch entsprechende Kalkulationen unter der Alternativ-Hypothese konnten die Power-Charakteristika des Internal Pilot Study Designs für verschiedene Fallzahladjustierungs-Strategien ebenfalls analytisch untersucht werden (COFFEY und MULLER, 1999; ZUCKER, WITTES, SCHABENBERGER und BRITTAIN, 1999; KIESER und FRIEDE, 2000a).

In der Arbeit von WITTES *et al.* (1999) findet sich eine schöne Begründung dafür, warum bei Fallzahladjustierung mit dem entblindeten gepoolten Varianzschätzer und Auswertung mit der üblichen *t*-Teststatistik das α-Niveau verzerrt ist. Dort wird gezeigt, dass die gepoolte Stichprobenvarianz, die im Design mit festem Stichprobenumfang σ^2 erwartungstreu schätzt,

in der vorliegenden Situation in der Erwartung kleiner als die Populationsvarianz ist. Als Konsequenz überschreitet die *t*-Teststatistik die kritische Schranke zu häufig. Bei dem ursprünglichen Vorschlag von STEIN (1945) gingen in die Varianzschätzung, die für die Teststatistik verwendet wird, nur die Daten der internen Pilotstudie ein. Damit wurde zwar die Kontrolle des α-Niveaus garantiert, doch ist dieses Verfahren insbesondere bei vergleichsweise großem Anteil des zweiten Studienabschnitts am Gesamtstichprobenumfang für die Praxis nicht zufriedenstellend. In einer Arbeit von PROSCHAN und WITTES (2000) wurde ein Varianzschätzer angegeben, der wie die übliche gepoolte Stichprobenvarianz die Daten aller Studienteilnehmer berücksichtigt, aber auch in Designs mit der Option zu einer Fallzahladjustierung σ^2 unverzerrt schätzt. Bei Verwendung der zugehörigen Teststatistik im Internal Pilot Study Design wird deshalb das nominale Niveau α nicht überschritten.

Die Techniken, die zur analytischen Berechnung des tatsächlichen Fehlers 1. Art verwendet wurden, weisen bei allen Unterschieden charakteristische Ähnlichkeiten auf. Im folgenden sollen die wesentlichen Ideen am Beispiel des Ansatzes von KIESER und FRIEDE (2000a) für die *t*-Test-Situation illustriert werden. Dabei betrachten wir das einseitige Testproblem $H_0 : \mu_1 = \mu_2$ gegen $H_1 : \mu_1 > \mu_2$. Bei vorgegebenem Niveau α soll beim Vorliegen eines klinisch relevanten Therapiegruppen-Unterschiedes $\delta = \mu_1^* - \mu_2^* > 0$ die Null-Hypothese mit einer Wahrscheinlichkeit $1-\beta$ abgelehnt werden. Für bekanntes σ^2 kann der notwendige Stichprobenumfang pro Gruppe N sehr gut approximiert werden durch

$$N = 2 \cdot \frac{(z_{1-\alpha} + z_{1-\beta})^2}{\delta^2} \cdot \sigma^2 . \qquad (4.11)$$

Es sei angemerkt, dass sich die Näherungsformel (4.11) als Spezialfall für $k = 2$ aus (4.6) und (4.7) ergibt.

Der erste Schritt zur Berechnung des tatsächlichen α-Niveaus besteht darin, die bei der Auswertung zu verwendende Teststatistik so zu zerlegen, dass die Bestandteile unabhängig oder bedingt auf die Varianzschätzung der internen Pilotstudie (die die Gesamtfallzahl determiniert) unabhängig sind. Dies ermöglicht die Bestimmung der gemeinsamen Dichte der Teststatistik. Die Integration dieser Dichte über den Ablehnungsbereich liefert dann die tatsächliche Rate für einen Fehler 1. Art α_{act}. In KIESER und FRIEDE (2000a) wurde dieser Weg für die Teststatistik T^* beschritten. Der einzige Unterschied zwischen der üblichen *t*-Teststatistik und T^* besteht darin, dass letztere eine Varianzschätzung verwendet, die sich aus den beiden gepoolten Varianzen S_i^2, $i = 1, 2$, der Stichproben vor ($i = 1$) bzw. nach der Fallzahladjustierung ($i = 2$) zusammensetzt:

$$T^* = \sqrt{\frac{n_1 + n_2}{2}} \cdot \frac{\overline{X}_1 - \overline{X}_2}{S^*},$$

wobei

$$S^* = \sqrt{\frac{2(n_1 - 1)}{2(n_1 + n_2 - 2)} S_1^2 + \frac{2(n_2 - 1)}{2(n_1 + n_2 - 2)} S_2^2}\ . \tag{4.12}$$

Bei festem Stichprobenumfang $2(n_1 + n_2)$ folgt T^* unter H_0 der zentralen *t*-Verteilung mit $2(n_1 + n_2 - 2)$ Freiheitsgraden. Die Struktur von S^* ermöglicht die Bestimmung der Dichte der Teststatistik T^* auch für das Design mit interner Pilotstudie, bei dem die Fallzahl eine Zufallsvariable ist. Mit den Transformationen $D = \sqrt{\frac{n_1 + n_2}{2}} \cdot \frac{\overline{X}_1 - \overline{X}_2}{\sigma^2}$, $V_i = \frac{2(n_i - 1)S_i^2}{\sigma^2}$, $i = 1, 2$, lässt sich die Teststatistik darstellen in der Form

$$T^* = \frac{D}{\sqrt{\frac{v_1 + v_2}{2(n_1 + n_2 - 2)}}},$$

und die Fallzahl-Adjustierungsregel lautet

$$\begin{aligned} \hat{N}_{re-est}(S_1^2) &= 2 \cdot \frac{(z_{1-\alpha} + z_{1-\beta})^2}{\delta^2} \cdot S_1^2 \\ &= \frac{N}{2(n_1 - 1)} \cdot V_1\,. \end{aligned} \tag{4.13}$$

Die Zufallsgröße V_1 ist chi-quadrat-verteilt mit $2(n_1 - 1)$ Freiheitsgraden. Bedingt nach V_1 ist V_2 (= „transformierte“ Varianz der nach der Fallzahladjustierung rekrutierten Stichprobe) chi-quadrat-verteilt mit $2(n_2 - 1)$ Freiheitsgraden. Weiterhin sind D und V_2 bedingt nach V_1 unabhängig, und D ist normalverteilt mit Erwartungswert $\sqrt{\frac{n_1 + n_2}{2}} \frac{\mu_1 - \mu_2}{\sigma}$ und Varianz 1. Damit ist die gemeinsame Dichte von D, V_1 und V_2 gegeben durch das Produkt der entsprechenden Dichten, und unter der Null-Hypothese H_0 gilt:

$$f(d, v_1, v_2 | H_0) = g_{N(0,1)}(d) \cdot g_{\chi^2_{2(n_1-1)}}(v_1) \cdot g_{\chi^2_{2(n_2-1)}}(v_2).$$

Dabei bezeichnet $g_{N(\mu,\sigma^2)}$ die Dichte der Normalverteilung mit Erwartungswert μ und Varianz σ^2 und $g_{\chi^2_{df}}$ die Dichte der zentralen Chi-Quadrat-Verteilung mit *df* Freiheitsgraden.

Die tatsächliche Wahrscheinlichkeit für einen Fehler 1. Art α_{act} für den einseitigen t^*-Test erhält man nun durch Integration der Dichte $f(d, v_1, v_2 | H_0)$ über den Ablehnungsbereich:

$$\alpha_{act} = \int_0^\infty \int_0^\infty \int_{(t_{2(n_1+n_2-2),1-\alpha}) \cdot \sqrt{\frac{v_1+v_2}{2(n_1+n_2-2)}}}^\infty f(d, v_1, v_2 | H_0)\ dd\ dv_2\ dv_1 . \qquad (4.14)$$

Hier ist zu bemerken, dass n_2 und damit auch die Dichte $f(d, v_1, v_2 | H_0)$, der kritische Wert $t_{2(n_1+n_2-2),1-\alpha}$ und α_{act} vom nominalen Niveau α, der Power $1-\beta$, dem Stichprobenumfang der internen Pilotstudie $2n_1$, der Populationsvarianz σ^2, dem klinisch relevanten Unterschied Δ und von der angewendeten Fallzahladaptionsregel abhängen. Mit der Fallzahlformel (4.11) können die Werte für $1-\beta$, Δ und σ^2 zusammengefasst werden zu dem notwendigen Stichprobenumfang pro Gruppe N bei gegebenem Niveau α. Im Gegensatz zu den Arbeiten von COFFEY und MULLER (1999) und WITTES *et al.* (1999), wo die tatsächliche Fehlerrate in Abhängigkeit von mehreren unbekannten Parametern angegeben wurde, kann man sich nun bei der Darstellung auf einen einzigen unbekannten Parameter (N) und die bekannten Größen α und n_1 beschränken. Diese Eigenschaft werden wir uns weiter unten zu Nutzen machen, wenn wir ein Verfahren zur Fehlerkontrolle herleiten. Abbildung 8 zeigt das tatsächliche Niveau α_{act} für $\alpha = 0.025$ in Abhängigkeit von n_1 und N für die Fallzahladaptionsregel von BIRKETT und DAY. Die Berechnungen wurden mit der Software Mathematica 3.0 durchgeführt.

Für feste Werte von N nimmt die tatsächliche Rate eines Fehlers 1. Art mit zunehmendem Stichprobenumfang der internen Pilotstudie ab und nähert sich dem nominalen Niveau α. Umgekehrt liegt bei festem Wert von n_1 das tatsächliche Niveau α_{act} nahe beim nominalen Niveau α, wenn die Werte von N im Vergleich zu n_1 hinreichend „groß“ oder „klein“ sind. Heuristisch erklärt sich dieses Verhalten aus der Tatsache, dass in diesen Situationen die Zahl der Patienten, die nach der internen Pilotstudie zusätzlich aufgenommen werden, sehr groß bzw. sehr klein ist im Vergleich zu der Zahl bereits rekrutierter Studienteilnehmer. Die Prozedur mit Fallzahladaption verhält sich deshalb in diesen Situationen sehr ähnlich zum Design mit festem Stichprobenumfang.

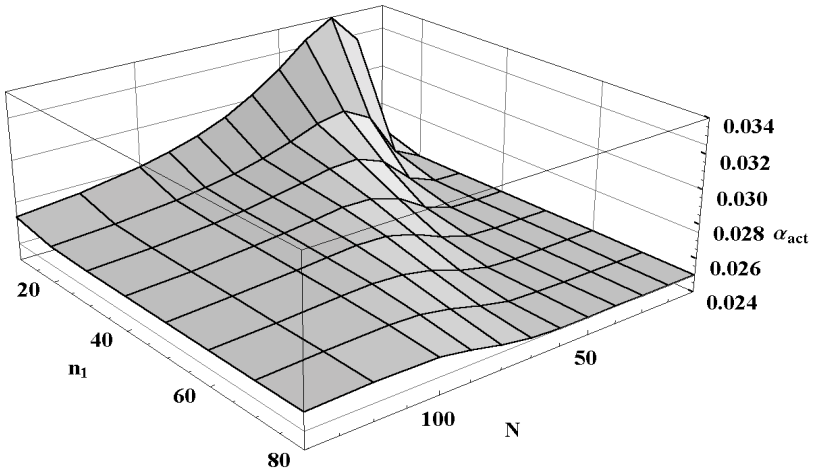

Abbildung 8: Tatsächliche Wahrscheinlichkeit für einen Fehler 1. Art α_{act} für ein nominales einseitiges Niveau $\alpha = 0.025$ und den t^*-Test im Design mit interner Pilotstudie und der Fallzahladaptionsregel nach BIRKETT und DAY (1994). Fallzahl pro Gruppe der internen Pilotstudie n_1, notwendige Fallzahl pro Gruppe N.

Aufgrund der Abhängigkeit von α_{act} vom unbekannten tatsächlich notwendigen Stichprobenumfang pro Gruppe N ist das tatsächliche Niveau in einer konkreten praktischen Situation unbekannt. Da α_{act} über (4.14) stetig von N abhängt und diese Funktion den oben beschriebenen Verlauf aufweist, existiert jedoch für jeden festen Wert von n_1 ein Wert N^{max} für den das tatsächliche Fehlerniveau maximiert wird. Dies ermöglicht die Bestimmung einer oberen Schranke α_{act}^{max} für die Wahrscheinlichkeit eines Fehlers 1. Art, die lediglich von den bekannten Design-Parametern α, n_1 und der Fallzahladaptionsstrategie abhängt. Somit kann das maximal erreichte Niveau für ein beliebiges Design mit interner Pilotstudie bestimmt und damit der in der ICH E9-Guideline (ICH, 1999) erhobenen Forderung Rechnung getragen werden.

In Tabelle 10 sind für beispielhafte Situationen das maximal erreichte Fehlerniveau α_{act}^{max} und die Fallzahl pro Gruppe N^{max}, für die dieses Maximum erreicht wird, für die Adaptionsregel von BIRKETT und DAY angegeben. Man sieht, dass sich die Werte für N^{max} für verschiedenes α nur marginal unterscheiden und im wesentlichen vom Stichprobenumfang der internen Pilotstudie abhängen. Die durch lineare Regression ermittelte Näherung $N^{max} \approx 1.7 \cdot n_1 + 17$ approximiert den wahren Wert sehr gut und kann als Startwert für die exakte Bestimmung von N^{max} verwendet werden.

Tabelle 10: Maximale tatsächliche Wahrscheinlichkeit eines Fehlers 1. Art α_{act}^{max} und zugehörige Fallzahl pro Gruppe N^{max} für den t^*-Test im Design mit interner Pilotstudie zum einseitigen nominalen Niveau α und Fallzahladaptionsregel nach BIRKETT und DAY (1994).

Fallzahl pro Gruppe der internen Pilotstudie n_1	Maximales tatsächliches Niveau α_{act}^{max}		Zugehörige Fallzahl pro Gruppe N^{max}	
	$\alpha = 0.025$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.05$
10	0.0339	0.0629	30	26
20	0.0294	0.0564	40	38
30	0.0279	0.0543	52	50
50	0.0268	0.0526	76	76
100	0.0259	0.0514	136	134
150	0.0256	0.0509	192	192
200	0.0255	0.0507	248	248

Die Ergebnisse für $\alpha_{act}^{\max}$ können nun dazu benutzt werden, ein adjustiertes nominales Niveau α_{adj} so zu bestimmen, dass die zugehörige maximal erreichte Wahrscheinlichkeit eines Fehlers 1. Art α nicht überschritten wird. Für feste Werte von $1-\beta$ und n_1 und eine spezifizierte Fallzahladaptionsregel ist das tatsächliche Niveau α_{act}^{max} eine stetige und monotone Funktion der nominalen Fehlerrate. Damit kann das adjustierte Niveau α_{adj} durch Lösen der folgenden Gleichung gefunden werden:

$$\alpha_{act}^{max}(\alpha_{adj}) = \alpha . \tag{4.15}$$

Gleichung (4.15) kann wie folgt iterativ gelöst werden:

- Startwert für α_{adj}: $\alpha_{adj}^{(0)} = \alpha \cdot \dfrac{\alpha}{\alpha_{act}^{max}(\alpha)}$
- Schritt i des Algorithmus' ($i \geq 1$): $\alpha_{adj}^{(i)} = \alpha_{adj}^{(i-1)} \cdot \dfrac{\alpha}{\alpha_{act}^{max}(\alpha_{adj}^{(i-1)})}$
- Stop des Algorithmus', falls $\alpha_{act}^{max}(\alpha_{adj}^{(s)}) - \alpha \leq \varepsilon$ für eine vorgegebene Genauigkeit $\varepsilon > 0$; definiere $\alpha_{adj} = \alpha_{adj}^{(s)}$.

In Tabelle 11 sind die adjustierten Niveaus α_{adj} für den einseitigen t^*-Test und die nominalen Wahrscheinlichkeiten eines Fehlers 1. Art $\alpha = 0.025$ und 0.05 und verschiedene Stichprobenumfänge der internen Pilotstudie n_1 angegeben. Für die Berechnungen wurde eine Genauigkeit von $\varepsilon = 0.0001$ verwendet.

Tabelle 11: Adjustiertes Niveau α_{adj}, für das die tatsächliche Wahrscheinlichkeit eines Fehlers 1. Art für den einseitigen t^*-Test im Design mit interner Pilotstudie und Fallzahladaptionsregel nach BIRKETT und DAY (1994) durch α kontrolliert wird.

Fallzahl pro Gruppe der internen Pilotstudie n_1	Adjustiertes Niveau α_{adj}	
	$\alpha = 0.025$	$\alpha = 0.05$
10	0.0178	0.0387
20	0.0210	0.0440
30	0.0223	0.0459
50	0.0233	0.0475
100	0.0241	0.0487
150	0.0244	0.0491
200	0.0245	0.0493

Bemerkungen:

1. Im Zusammenhang mit der Verwendung verblindeter Varianzschätzer zur Fallzahladjustierung wurde die Teststatistik t^* von GOULD und SHIH (1992) benutzt, um für hypothetische Beispiele die Auswirkung einer Fallzahlanpassung auf das Fehlerniveau zu illustrieren. Aufgrund der Tatsache, dass in dieser Arbeit die Überlegungen stets in Abhängigkeit vom unbekannten Wert der Populationsvarianz σ^2 angestellt wurden, blieb durch diese Betrachtungen die für die Praxis relevante Frage des tatsächlichen Fehlerniveaus in einer gegebenen Design-Situation sowie die Kontrolle derselben unbeantwortet.
2. Mit den Ergebnissen für den t^*-Test lassen sich sofort obere Schranken für die tatsächliche Fehlerrate (und damit auch adjustierte Niveaus) für den *t*-Test herleiten. Für die übliche gepoolte Varianzschätzung S^2 folgt aus (4.12) die Abschätzung

$$S^2 \geq \frac{n_1 + n_2 - 4}{n_1 + n_2 - 2}(S^*)^2,$$

und damit

$$\Pr_{H_0}\left(t \geq t_{n_1+n_2-2,1-\alpha}\right) \leq \Pr_{H_0}\left(t^* \geq \sqrt{\frac{n_1 + n_2 - 4}{n_1 + n_2 - 2}} t_{n_1+n_2-2,1-\alpha}\right).$$

Nach Publikation der Arbeit von KIESER und FRIEDE (2000a) wurde evident, dass es nicht notwendig ist, diesen Umweg zu beschreiten. Vielmehr lassen sich die Berechnungen mit ähnlichen Argumenten, wie sie für den t^*-Test benutzt wurden, auch direkt für die t-Teststatistik durchführen. In Kapitel 4.1.2.2 wird dieses Vorgehen am Beispiel der Fallzahladjustierung mit der verblindeten adjustierten Ein-Stichproben-Varianz dargestellt.

3. In Analogie zu den oben angestellten Überlegungen oder unter Verwendung der Techniken von COFFEY und MULLER (1999) lassen sich α_{act}^{max} und α_{adj} auch für den Vergleich von k Behandlungsgruppen mit dem *F*-Test berechnen. Wir haben uns bei der Darstellung auf den Zwei-Stichproben-Fall beschränkt, weil sich für diesen, anders als für die Situation $k > 2$, die Verfahren zur Fehlerkontrolle bei Verwendung der entblindeten Stichprobenvarianz direkt auf entsprechende Methoden bei Verwendung verblindeter Varianzschätzungen übertragen lassen (siehe Kapitel 4.1.2.2).

4.1.2.2 Fallzahladaption mit verblindetem Varianzschätzer

Verwendet man wie durch die ICH E9-Guideline nahegelegt einen verblindeten Varianzschätzer zur Fallzahladjustierung, so gibt es zwei Möglichkeiten, das vorgegebene Signifikanzniveau zu kontrollieren. Zum einen kann bei der Auswertung ein Randomisierungstest verwendet werden. Satz 9 im folgenden Kapitel zeigt, dass dann die Wahrscheinlichkeit eines Fehlers 1. Art auch im Design mit interner Pilotstudie durch α kontrolliert wird, wenn der Test zum Niveau α durchgeführt wird. Falls bei der Auswertung die Original-Teststatistik verwendet werden soll, lassen sich durch Verallgemeinerung der Ergebnisse des vorangehenden Kapitels Aussagen über die tatsächliche Fehlerrate gewinnen. Wir werden aber sehen, dass dies, im Gegensatz zur erstgenannten Option, für jede Teststatistik separat bewerkstelligt werden muss, und schon im Falle des *F*-Tests eine einfache Übertragung der für den *t*-Test erfolgreichen Methodik nicht zielführend ist.

Auswertung mit Randomisierungstest

Randomisierungstests basieren auf dem Modell, dass die in einer randomisierten Studie beobachteten Werte der Zielgröße feste Zahlen sind und die Null-Hypothese behauptet, dass die Behandlung keinen Einfluss auf diese Zahlenwerte hat. Die Behandlungsgruppen-Zugehörigkeiten sind dann unter der Null-Hypothese austauschbar. Die Null-Hypothese kann getestet werden, indem alle Permutationen der Gruppen-Zuordnung, die unter dem verwendeten Randomisierungsschema möglich sind, gebildet werden und die entsprechende Verteilung einer geeigneten Teststatistik T berechnet wird. Wir gehen auf die Theorie dieser Tests im folgenden nur insoweit ein, wie wir sie für unsere Zwecke benötigen. Für weitergehende Einführungen sei auf die Lehrbücher von GOOD (1994) und EDGINGTON (1995) verwiesen. BERGER (2000) hat kürzlich in einem lesenswerten Beitrag häufig genannte Argumente für und wider Randomisierungstests einer kritischen Prüfung unterzogen und deren tatsächliche Vor- und Nachteile bei der Auswertung klinischer Studien gegenübergestellt.

Mit T sei im folgenden die für den Randomisierungstest gewählte Teststatistik bezeichnet, mit $\underline{X}_n = (X_1, ..., X_n)$ der Vektor der unabhängig und identisch verteilten Zielgrößen bei einem festen Stichprobenumfang n und mit $\underline{Z}_n = (Z_1, ..., Z_n)$ die randomisierten Gruppen-Zugehörig-keiten. Bezeichnet man Zufallsvariablen mit Großbuchstaben und ihre Realisierungen mit Kleinbuchstaben, dann benutzt der zu T gehörige Randomisierungstest die Verteilung von $T(n, \underline{x}_n, \underline{Z}_n)$. Die Verteilung hängt ab von der Teststatistik selbst, dem Stichprobenumfang n, den beobachteten Daten $\underline{x}_n$ und der Menge Ω_n aller unter dem Studiendesign möglichen Randomisierungs-Zuordnungen für $\underline{x}_n$. Einen Niveau-α-Test erhält man, indem man das $(1-\alpha)$-Perzentil $t_{1-\alpha}(n, \underline{x}_n)$ der Verteilung von $T(n, \underline{x}_n, \underline{Z}_n)$ unter der zu testenden Null-Hypothese H_0 bestimmt und $t_{1-\alpha}(n, \underline{x}_n)$ als Grenze des Ablehnungsbereichs definiert. Ohne Einschränkung der Allgemeinheit sei angenommen, dass große Werte von T gegen die Gültigkeit von H_0 sprechen. Dann gilt konstruktionsgemäß

$$\Pr_{H_0}\left(T(n, \underline{x}_n, \underline{Z}_n) > t_{1-\alpha}(n, \underline{x}_n)\right) \leq \alpha\,. \tag{4.16}$$

Ungleichung (4.16) lässt sich auch schreiben als

$$\sum_{\underline{z}_n \in \Omega_n} \Pr_{H_0}\left(\underline{Z}_n = \underline{z}_n\right) \cdot I\left(T(n, \underline{x}_n, \underline{z}_n) > t_{1-\alpha}(n, \underline{x}_n)\right) \leq \alpha\,, \tag{4.17}$$

wobei I die Indikatorfunktion bezeichnet:

$$I(A) = \begin{cases} 1 & \text{falls A wahr ist} \\ 0 & \text{falls A falsch ist.} \end{cases}$$

In der Situation eines Designs mit interner Pilotstudie und der Option zur Fallzahladjustierung ist der Stichprobenumfang nicht fest sondern zufällig. Die entsprechende Zufallsvariable sei mit $\hat{N}$ bezeichnet, die Menge aller möglichen Realisierungen von $\hat{N}$ mit $\Psi \subseteq \{n : n \geq n_1\}$. Die Wahrscheinlichkeit eines Fehlers 1. Art berechnet sich dann analog zu (4.17) als

$$\Pr_{H_0}\left(T(\hat{N}, \underline{x}_{\hat{N}}, \underline{Z}_{\hat{N}}) > t_{1-\alpha}(\hat{N}, \underline{x}_{\hat{N}})\right)$$
$$= \sum_{n \in \Psi} \sum_{\underline{z}_n \in \Omega_n} \Pr_{H_0}\left(\hat{N} = n, \underline{Z}_n = \underline{z}_n\right) \cdot I\left(T(n, \underline{x}_n, \underline{z}_n) > t_{1-\alpha}(n, \underline{x}_n)\right).$$

Der Stichprobenumfang wird bestimmt über eine Regel, die ausschließlich die Daten $\underline{x}_{n_1}$ verwenden, wobei n_1 fest ist und $n_1 \leq n$ für alle $n \in \Psi$ gilt. Weiterhin sind unter der Null-Hypothese $\underline{X}_{n_1}$ und $\underline{Z}_n$ für alle $n \in \Psi$ unabhängig, woraus folgt, dass $\hat{N}$ und $\underline{Z}_n$ für alle $n \in \Psi$ unabhängig sind. Daraus folgt

$$\Pr_{H_0}\left(T(\hat{N}, \underline{x}_{\hat{N}}, \underline{Z}_{\hat{N}}) > t_{1-\alpha}(\hat{N}, \underline{x}_{\hat{N}})\right)$$
$$= \sum_{n \in \Psi} \left[\sum_{\underline{z}_n \in \Omega_n} \Pr_{H_0}(\underline{Z}_n = \underline{z}_n) \cdot I\left(T(n, \underline{x}_n, \underline{z}_n) > t_{1-\alpha}(n, \underline{x}_n)\right) \right] \cdot \Pr_{H_0}(\hat{N} = n)$$
$$\leq \alpha \cdot \sum_{n \in \Psi} \Pr_{H_0}(\hat{N} = n) = \alpha .$$

Damit ist der folgende Satz bewiesen.

Satz 9:

Im Design mit interner Pilotstudie werde die Fallzahladjustierung mit einer verblindeten Varianzschätzung durchgeführt. Bei der Auswertung werde ein Randomisierungstest zum Niveau α verwendet. Dann wird die Wahrscheinlichkeit für einen Fehler 1. Art durch α kontrolliert.

Beweis:

Siehe oben. Der Beweis folgt der Idee von EDWARDS (1999) für die Situation, dass im Rahmen eines Blind Data Review das Auswertungsmodell spezifiziert wird. Der wesentliche Unterschied besteht darin, dass dort die Modellwahl auf den gleichen Daten wie die Auswertung basiert, womit keine Zufälligkeit bezüglich der Fallzahl besteht. ■

Auswertung mit Original-Test

Im vorangehenden Kapitel haben wir gesehen, dass bei Verwendung eines Randomisierungstests das Niveau α für klinische Studie mit interner Pilotstudie unabhängig von der Wahl der speziellen Teststatistik und der Fallzahladaptionsregel kontrolliert wird. Derartig weitreichende Aussagen sind (derzeit) nicht möglich, wenn die Analyse nicht mit der Randomisierungs-Version sondern mit dem Original-Test durchgeführt werden soll. Vielmehr muss dann die tatsächliche Fehlerrate für jede konkret vorliegende Situation berechnet werden. Das gleiche gilt analog für den Fall, dass Überschreitungen des nominalen Niveaus beobachtet werden und die strikte Kontrolle des α-Fehlers notwendig ist. Für die Berechnung bietet sich eine Übertragung der Techniken an, mit denen in Kapitel 4.1.2.1 diese Probleme für die Fallzahladjustierung mit der entblindeten Varianzschätzung gelöst wurden. Diesen Weg hat FRIEDE (2000b) für den *t*-Test beschritten. Nachfolgend wird das Vorgehen sowie seine Analogien und Unterschiede zu dem bei Verwendung der gepoolten Zwei-Stichproben-Varianz skizziert. Wir betrachten dabei die Situation, dass die adjustierte Ein-Stichproben-Varianz zur Re-Kalkulation des Stichprobenumfanges benutzt wird; die Resultate für die nicht-adjustierte Version folgen als Spezialfall für $\delta = 0$.

Die Fallzahladaptionsregel bei Verwendung des adjustierten Ein-Stichproben-Varianzschätzers

$$S_{adj}^2 = \frac{2(n_1 - 1)}{2n_1 - 1} \cdot S_1^2 + \frac{n_1}{2(2n_1 - 1)} \cdot \left(\overline{X}_{11} - \overline{X}_{12}\right)^2 - \frac{n_1}{2(2n_1 - 1)} \cdot \delta^2$$

und der Approximationsformel (4.11) lautet

$$\begin{aligned} \hat{N}_{re-est}(S_{adj}^2) &= 2 \cdot \frac{(z_{1-\alpha} + z_{1-\beta})^2}{\delta^2} \cdot S_{adj}^2 \\ &= \frac{N}{2n_1 - 1} \cdot \left(\frac{2(n_1 - 1)S_1^2}{\sigma^2} + \frac{n_1}{2} \cdot \frac{(\overline{X}_{11} - \overline{X}_{12})^2}{\sigma^2} - \frac{n_1}{2} \cdot \left(\frac{\delta}{\sigma}\right)^2 \right), \end{aligned} \qquad (4.18)$$

wobei $\overline{X}_{ij}$, $i, j = 1, 2,$ die Mittelwerte in Gruppe j und Studienabschnitt i bezeichnen. Mit den folgenden Transformationen, die in gleicher bzw. ähnlicher Form bereits in Kapitel 4.1.2.1 verwendet wurden, lässt sich die Gleichung (4.18) so umformen, dass die Formel zur Fallzahladaption nur noch Konstanten und Zufallsgrößen enthält, aus denen die t-Teststatistik zusammengesetzt ist (siehe (4.19) unten): $D_i = \sqrt{\frac{n_i}{2}} \cdot \frac{\overline{X}_{i1} - \overline{X}_{i2}}{\sigma^2}$, $V_i = \frac{2(n_i - 1)S_i^2}{\sigma^2}$, $i = 1, 2$.

Dann gilt

$$\hat{N}_{re-est}(S^2_{adj}) = \frac{N}{2n_1 - 1} \cdot \left(V_1 + D_1^2\right) - \frac{n_1}{2n_1 - 1} \cdot (z_{1-\alpha} + z_{1-\beta})^2 .$$

Es sei angemerkt, dass im Unterschied zur Verwendung der entblindeten Varianzschätzung (Formel (4.13) in Kapitel 4.1.2.1) die Fallzahladaptionsregel nun zusätzlich von der zufälligen Größe D_1^2 abhängt. Die *t*-Teststatistik lässt sich mit den transformierten Größen wie folgt darstellen:

$$T = \frac{\sqrt{\frac{n_1}{n_1 + n_2}} D_1 + \sqrt{\frac{n_2}{n_1 + n_2}} D_2}{\sqrt{\frac{V_1 + V_2^*}{2(n_1 + n_2 - 1)}}} \tag{4.19}$$

mit $V_2^* = V_2 + \frac{2n_1 n_2}{n_1 + n_2} \cdot \left((\overline{X}_{11} - \overline{X}_{21})^2 + (\overline{X}_{12} - \overline{X}_{22})^2\right)$. Für die (bedingten) Verteilungen der Komponenten der Teststatistik gilt $V_1 \sim \chi^2_{2(n_1-1)}$, $V_2^* \sim \chi^2_{2n_2}$, $D_i \sim N(\sqrt{\frac{n_i}{2}} \frac{\mu_1 - \mu_2}{\sigma}, 1)$, $i = 1, 2$. Weiterhin sind diese Zufallsgrößen (bedingt) unabhängig, so dass sich die gemeinsame Dichte als Produkt der einzelnen Dichten darstellen lässt. Unter H_0 gilt damit

$$f(d_1, v_1, d_2, v_2^* | H_0) = g_{N(0,1)}(d_1) \cdot g_{\chi^2_{2(n_1-1)}}(v_1) \cdot g_{N(0,1)}(d_2) \cdot g_{\chi^2_{2n_2}}(v_2^*) .$$

Durch Integration über den Ablehnungsbereich erhält man die tatsächliche Wahrscheinlichkeit eines Fehlers 1. Art. Aufgrund der gewählten Parametrisierung der Fallzahladaptionsregel hängt α_{act} bei Verwendung von S^2_{adj} lediglich von den bekannten Werten für n_1, α und $1-\beta$, der Adaptionsregel sowie dem unbekannten tatsächlich notwendigen Stichprobenumfang pro Gruppe N ab. Für den nicht-adjustierten Varianzschätzer ist α_{act} wie in Kapitel 4.1.2.1 bereits durch n_1, α, die Adaptionsregel und N determiniert. Für eine konkrete Design-Situation lässt sich deshalb die maximale Wahrscheinlichkeit für einen Fehler 1. Art bestimmen, indem das tatsächliche Niveaus α_{act} für alle möglichen Werte, die das unbekannte N annehmen kann, berechnet wird.

Bemerkungen:

1. Wegen des in Kapitel 4.1.2.1 beschriebenen Verhaltens von α_{act} in Abhängigkeit vom relativen Verhältnis zwischen N und n_1 muss zur Bestimmung der maximalen Wahrscheinlichkeit eines Fehlers 1. Art die tatsächliche Fehlerrate nicht für alle Werte, die N annehmen kann, berechnet werden, sondern nur für einen kompakten Wertebereich.

2. Mit den Darstellungen (4.19) der *t*-Teststatistik und (4.18) der Fallzahladaptionsregel kann das tatsächliche Niveau für den *t*-Test auch bei Verwendung der entblindeten Varianzschätzung S_1^2 berechnet werden.

Die tatsächliche Wahrscheinlichkeit eines Fehlers 1. Art α_{act} wurde von FRIEDE (2000b) für den *t*-Test und $\alpha = 0.025,\ 0.05$ (einseitig), $1-\beta = 0.80, 0.90$, $n_1 = 20, 30, ..., 200,\ N = 20, 40, ..., 200$ und die Regel nach BIRKETT und DAY nach dem oben beschriebenen Verfahren berechnet. Es wurde weder bei Fallzahladaption mit der adjustierten noch mit der nicht-adjustierten Ein-Stichproben-Varianz eine Überschreitung des nominalen Niveaus um mehr als 0.0001 beobachtet (FRIEDE, 2000b). Damit sind für die Auswertung von Studien mit interner Pilotstudie in der *t*-Test-Situation keine zusätzlichen Maßnahmen zur Kontrolle der Wahrscheinlichkeit eines Fehlers 1. Art notwendig, sondern es kann der übliche Ablehnungsbereich zugrunde gelegt werden.

Dieses erfreuliche Resultat wird getrübt durch die Einschränkung, dass es für eine Übertragung dieses Ergebnisses auf andere Tests jeweils von neuem notwendig ist, den oben für den *t*-Test beschriebenen Weg zu beschreiten. Man sieht sofort ein, dass dies für Kontrast-Tests (siehe Kapitel 3.5) ganz analog zum t-Test möglich ist. An die Grenzen dieser Technik stößt man jedoch bereits beim F-Test: Dort steht im Zähler der Teststatistik die Summe der quadratischen Differenzen der Mittelwerte der Behandlungsgruppen, die sich nicht in der oben beschriebenen Weise in entsprechende Anteile aus der Stichprobe vor und nach der Fallzahladjustierung splitten lässt. Die Ergebnisse für den *t*-Test und das allgemeine Resultat für Randomisierungstests mit beliebiger Teststatistik geben Anlass zu Mutmaßungen darüber, ob bei anderen als den oben untersuchten Situationen Überschreitungen des nominalen Niveaus überhaupt zu erwarten sind. Eine Methodik, die es erlaubt, die Fehlerrate für eine beliebig vorgegebene Teststatistik zu berechnen oder gar einen Satz mit ähnlicher Allgemeinheit wie Satz 9 für Original-Tests zu beweisen, ist derzeit nicht in Aussicht. Der einzige allerdings unbefriedigende Ausweg besteht deshalb darin, Simulationsuntersuchungen durchzuführen und, falls notwendig, zur Fehlerkontrolle simulations-basierte Adjustierungen vorzunehmen (siehe DENNE und JENNISON, 1999, und Kapitel 4.1.2.1).

4.2 Adaptives Zwei-Stufen-Design

In Analogie zum Design mit interner Pilotstudie lässt sich das Vorgehen bei der Planung und Adaption der Fallzahl im Zwei-Stufen-Design nach BAUER und KÖHNE (1994), bei dem die Studie nach der Zwischenauswertung mit Ablehnung oder Beibehaltung der Null-Hypothese H_0 beendet werden kann, wie folgt beschreiben:

(i) Vor Beginn der Studie wird eine vorläufige Fallzahl pro Gruppe $\hat{N}_{est}$ berechnet, die auf einer *a priori*-Schätzung S_0^2 der Varianz σ^2 beruht. Der Stichprobenumfang pro Gruppe für den ersten Studienteil n_1, $n_1 \leq \hat{N}_{est}$, und die Stop-Grenzen α_0 und α_1 werden festgelegt.

(ii) Wenn $2n_1$ Patienten die Studie abgeschlossen haben, wird die Zwischenauswertung durchgeführt, die einen *p*-Wert p_1 liefert. Für $p_1 \leq \alpha_1$ oder $p_1 \geq \alpha_0$ wird die Studie beendet. Für $\alpha_1 < p_1 < \alpha_0$ kann H_0 bei der Endauswertung abgelehnt werden, falls für den *p*-Wert p_2 des zweiten Studienabschnitts gilt $p_2 \leq c_\alpha / p_1$. Es ist deshalb sinnvoll, die Fallzahl pro Gruppe n_2 des zweiten Studienteils zum Niveau c_α / p_1 und zur Power $1-\beta$ zu bestimmen; die Power ist bedingt auf das Ereignis, dass ein zweiter Studienteil durchgeführt wird.

(iii) Für den zweiten Studienteil werden n_2 Patienten pro Gruppe rekrutiert, die Auswertung liefert einen *p*-Wert p_2. Die Null-Hypothese kann abgelehnt werden, falls $p_1 \cdot p_2 \leq c_\alpha$, andernfalls wird sie beibehalten.

Im Vergleich zum Design mit interner Pilotstudie bestehen charakteristische Unterschiede. Zum einen stellen die beiden kritischen Punkte des Internal Pilot Study Designs „Notwendigkeit zur verblindeten Varianzschätzung“ und „Kontrolle der Wahrscheinlichkeit eines Fehlers 1. Art“ für Designs mit adaptiver Zwischenauswertung kein Problem dar. Aufgrund der Tatsache, dass im Rahmen der Zwischenauswertung ein Hypothesentest durchgeführt wird, der eine Entblindung der Studie notwendig macht, steht der k-Stichproben-Varianzschätzer zur Fallzahladjustierung zur Verfügung. Die Einhaltung des nominalen Niveaus ist per Konstruktion auch unter (Fallzahl-) Adaption gewährleistet (siehe Kapitel 2.1).

Ein weiterer Unterschied zum Design mit interner Pilotstudie besteht darin, dass das Niveau für den zweiten Studienteil, das in die Berechnung von n_2 eingeht, vom *p*-Wert des

ersten Studienteils p_1 abhängt. Die resultierende Fallzahl wird damit von einer weiteren Zufallsgröße beeinflusst. Die Konsequenzen für deren Verteilung werden wir in Kapitel 4.3.3 untersuchen.

Aufgrund der bei der Zwischenauswertung notwendigen Entblindung kann für die Fallzahl-Adaption mehr Informationen aus den Studiendaten verwendet werden als beim Design mit interner Pilotstudie. Zum einen kann wie dort die Fallzahl auf der Basis der Varianzschätzung aus dem ersten Studienteil re-kalkuliert werden, wobei nun die gepoolte Stichprobenvarianz zur Verfügung steht. Weiterhin wurde vorgeschlagen, den beobachteten Therapiegruppen-Unterschied zur Fallzahladaption zu nutzen. Im folgenden wird eine zusammenfassende Übersicht über den derzeitigen Forschungsstand zu diesem Themenkomplex gegeben, wobei die letztgenannte Strategie einer kritischen Überprüfung unterzogen wird.

4.2.1 Fallzahladaption unter Verwendung der geschätzten Varianz

POSCH und BAUER (2000) untersuchten mittels analytischer Betrachtungen die Fallzahlplanung und -adaption für das Design von BAUER und KÖHNE (1994) in der Zwei-Stichproben *t*-Test-Situation. Die Fallzahlberechnung für den zweiten Studienteil erfolgt nach der oben beschriebenen Strategie unter Verwendung der gepoolten Varianz aus der Zwischenauswertung zu einer festen, gegenüber der ursprünglichen Planung unveränderten Alternativ-Hypothese $H_1 : \Delta = \delta$. Hierbei sind grundsätzlich die Fälle $\alpha_1 < 1$ und $\alpha_0 = 1$ zu unterscheiden.

Für Designs, die eine vorzeitige Beendigung der Studie mit Beibehaltung der Null-Hypothese vorsehen ($\alpha_0 < 1$), lässt sich unter der Annahme $\sigma^2 = s_0^2$ der Stichprobenumfang des ersten Studienteils so bestimmen, dass die Gesamt-Power der Studie $1-\beta$ beträgt. Für diese Wahl von n_1 lassen sich kritische Grenzen α_0 und α_1 herleiten, für die die mittlere Gesamt-Fallzahl unter H_0 bzw. H_1 minimiert wird; die resultierenden Schranken sind bei Minimierung unter der Null- und der Alternativ-Hypothese sehr ähnlich. Falls die Varianzschätzung s_0^2 vor Studienbeginn kleiner als σ^2 ist, werden n_1 und α_0 zu niedrig gewählt und die Gesamt-Power liegt je nach Grad der Fehlspezifikation mehr oder weniger deutlich unter $1-\beta$. Als Ausweg schlagen POSCH und BAUER vor, unter Verwendung der Varianzschätzung s_1^2 aus der Zwischenauswertung die vor Studienbeginn festgelegte Grenze α_0 so abzuändern,

dass für die aktuelle Wahl von n_1 unter der Annahme $\sigma^2 = s_1^2$ die Gesamt-Power $1-\beta$ beträgt. Bei Adaption von α_0 muss dann entweder α_1 oder α_2 so abgeändert werden, dass die Niveau-Bedingung (2.1.b) des Kombinationstests erfüllt ist. Wie POSCH und BAUER zeigen, wird durch diese daten-abhängige Adaption der Entscheidungsgrenzen erreicht, dass die Gesamt-Power auch bei einer *a priori*-Unterschätzung der Varianz nur unwesentlich unter dem angestrebten Wert $1-\beta$ liegt. Gleichzeitig sind die berechneten Niveau-Überschreitungen nur minimal. Bei einer Überschätzung von σ^2 in der Planungsphase ist (mit und ohne Adaption von α_0) die erwartete Fallzahl unter H_0 und unter H_1 größer als bei Planung einer Studie mit festem Stichprobenumfang und der gleichen Annahme über die Varianz; wenn die Annahme $\sigma^2 = s_0^2$ stimmt oder die Varianz vor Studienbeginn unterschätzt wird, ist der mittlere Stichprobenumfang des Zwei-Stufen Designs kleiner. Der entscheidende Vorteil des adaptiven Designs besteht aber darin, dass im Gegensatz zum Design mit festem Stichprobenumfang im Mittel die Power $1-\beta$ erreicht wird, unabhängig von der Korrektheit der Annahme über σ^2 in der Planungsphase.

Für $\alpha_0 = 1$ ist die Gesamt-Power der Studie stets größer als $1-\beta$. Um die Überschreitung der gewünschten Power möglichst niedrig zu halten, ist es vorteilhaft, die Stop-Grenze möglichst niedrig zu wählen, also $\alpha_1 = c_\alpha$. Der optimale Stichprobenumfang für den ersten Studienteil kann nun wie für $\alpha_0 < 1$ unter der Annahme $\sigma^2 = s_0^2$ bestimmt werden. Wenn vor Studienbeginn die Varianz nicht überschätzt wird, ist unter H_1 (bei einer höheren erwarteten Power) der mittlere Gesamt-Stichprobenumfang für diese Wahl von n_1 ungefähr so groß wie die unter den gleichen Annahmen über die Varianz berechnete Fallzahl des Designs mit festem Stichprobenumfang; für $s_0^2 > \sigma^2$ ist wie für $\alpha_0 < 1$ die mittlere Fallzahl des adaptiven Designs größer. Unter H_0 ist die mittlere Fallzahl des Zwei-Stufen Designs stets größer als für das Design mit festem Stichprobenumfang, denn die Wahl $\alpha_0 = 1$ hat zur Konsequenz, dass auch für große *p*-Werte p_1 ein zweiter Studienteil durchgeführt wird mit niedrigem Signifikanzniveau c_α / α und dementsprechend hoher Fallzahl n_2.

4.2.2 Fallzahladaption unter Verwendung des geschätzten Behandlungsgruppen-Unterschieds

Von mehreren Autoren wurde vorgeschlagen, in adaptiven Designs den Stichprobenumfang für den nachfolgenden Studienteil zu berechnen, indem unter der Annahme bekannter Varianz σ^2 der in der Zwischenauswertung beobachtete Behandlungsgruppen-Unterschied anstelle der klinisch relevanten Differenz δ in die Fallzahlformel eingesetzt wird (FISHER, 1998; SHEN und FISHER, 1999; CUI, HUNG und WANG, 1999; WASSMER, 1999b; Funke, 2000; FUNKE und WASSMER, 2000).

WASSMER (1999b) führte für diese Adaptionsregel exakte Berechnungen der Gesamt-Power und des mittleren Stichprobenumfanges durch. Dabei zeigte sich, dass die Ergebnisse für die verschiedenen adaptiven Zwei-Stufen-Designs (untersucht wurden die Verfahren von BAUER und KÖHNE, 1994, PROSCHAN und HUNSBERGER, 1995, und LEHMACHER und WASSMER, 1999) nur geringfügige Unterschiede aufweisen. Selbst wenn man sich einer Bewertung der Relevanz der Unterschiede enthält, kann nicht von einem besten Design gesprochen werden, da eine höhere Power auch stets mit einem höheren mittleren Stichprobenumfang verbunden ist. In Vergleichen zwischen dem adaptiven Design nach PROSCHAN und HUNSBERGER (1995) und dem einstufigen Design mit festem Stichprobenumfang sowie dem nicht-adaptiven gruppensequentiellen Zwei-Stufen-Design nach DEMETS und WARE (1980) weist das adaptive Verfahren häufig niedrigere mittlere Stichprobenumfänge bei gleichzeitig höherer Power auf.

FUNKE (2000) und FUNKE und WASSMER (2000) untersuchten die Charakteristika der Adaptionsstrategie für die kanonischen Verallgemeinerungen der oben genannten adaptiven Designs auf mehr als zwei Stufen. Analytische Berechnungen für verschiedenste Parameterkonstellationen zeigten, dass die vierte Stufe nur noch für sehr kleine Werte des tatsächlichen Gruppen-Unterschiedes Δ erreicht wird, und die fünfte Stufe praktisch überhaupt nicht mehr. Die Autoren schlossen daraus, dass man sich bei der praktischen Planung in aller Regel auf ein zwei- oder maximal dreistufiges Design beschränken kann.

Die Verwendung des beobachteten Therapieeffektes anstelle der minimalen klinisch relevanten Differenz bei der Fallzahlberechnung ist aus konzeptioneller Sicht fragwürdig: Die Fallzahl wird für die aus den Daten geschätzten Alternativ-Hypothese, dass der tatsächliche Behandlungsgruppen-Unterschied gleich dem beobachteten ist, bestimmt. Überlegungen zur klinischen Relevanz des Therapieeffektes gehen hier in keiner Weise ein; die Adaptions-Strategie wählt auch bei völlig unbedeutenden Unterschieden die Fallzahl so, dass unter der

oben genannten Annahme für den zu planenden Studienabschnitt eine bedingte Power $1-\beta$ zur Ablehnung von H_0 besteht. Die grundsätzliche Philosophie der Fallzahlplanung wird durch dieses Vorgehen *ad absurdum* geführt. Für Therapieeffekte, die kleiner als die minimale klinisch relevante Differenz sind ($\Delta < \delta$), resultiert ein „hunting for significance".

Diese Eigenschaften der Adaptionsstrategie spiegeln sich in einigen der Ergebnisse der Untersuchungen von WASSMER (1999b) wider. Das Ersetzen von Δ durch den beobachteten Unterschied führt auch dann zu einer Erhöhung der Power gegenüber den nicht-adaptiven Designs, wenn der tatsächliche Therapieeffekt kleiner ist als die minimale Differenz, die es wert ist, entdeckt zu werden. Insbesondere für Stop-Grenzen $\alpha_0 \geq 0.30$ ist dies mit einer teilweise erheblichen Erhöhung des Stichprobenumfanges verbunden: Im Vergleich zum nicht-adaptiven Zwei-Stufen-Design nach DEMETS und WARE (1980) werden (für uninteressante Alternativen!) bis zu achtfache mittlere Fallzahlen beobachtet (WASSMER, 1999b, S. 99, Tabelle 3.6). Eine Adaptionsstrategie, die die beobachtete Behandlungsgruppen-Differenz verwendet und die unerwünschten Eigenschaften für klinisch nicht relevante Therapieeffekte nicht aufweist, ist ein lohnendes Feld für zukünftige Forschungsaktivitäten.

4.3 Vergleich zwischen Design mit interner Pilotstudie und adaptivem Zwei-Stufen-Design

4.3.1 Erreichen der gewünschten Power mit vorgegebener Wahrscheinlichkeit

Häufig wird bei der Planung einer Studie mit festem Stichprobenumfang eine Varianzschätzung verwendet, die aus einer vorangehenden Studie in der gleichen Indikation mit der gleichen Zielgröße stammt. Bei dieser Vorgehensweise ist auch im Ein-Stufen Design ohne Fallzahlanpassung die Fallzahl eine Zufallsvariable. KIESER und WASSMER (1996) bewiesen für den Spezialfall des Zwei-Stichproben *t*-Tests das folgende allgemeine Resultat für die resultierende Power, das durch Simulationsuntersuchungen von BROWNE (1995) nahegelegt wurde.

Satz 10:

Für unabhängig normalverteilte Zufallsvariablen mit Populationsvarianz σ^2 lasse sich zu einem gegebenen Testproblem die Fallzahl pro Behandlungsgruppe, die notwendig ist, um beim Vorliegen eines klinisch relevanten Unterschieds δ die Null-Hypothese zum Niveau α mit einer Wahrscheinlichkeit $1-\beta$ ablehnen zu können, mit einer Approximationsformel der Struktur (4.6) berechnen:

$$N = \lambda(k, \alpha, \beta) \cdot \frac{\sigma^2}{\delta^2}$$

Dabei bezeichnet k die Anzahl der Behandlungsgruppen. Für ein Design mit festem Stichprobenumfang werde die Fallzahl bestimmt, indem man in obige Formel statt σ^2 die auf einer Stichprobe aus einer Population mit Varianz σ^2 beruhende (a) gepoolte Varianzschätzung bzw. (b) die obere Schranke des einseitigen $(1-\gamma)$-Konfidenzintervalls für σ^2 einsetzt. Dann beträgt die Wahrscheinlichkeit, dass mindestens die angestrebte Power $1-\beta$ erreicht wird, (a) weniger als 0.50 bzw. (b) $1-\gamma$.

Beweis:

Wir nehmen an, dass die gepoolte Varianzschätzung $S^2_{l-sample}$ bzw. die obere Schranke des einseitigen $(1-\gamma)$-Konfidenzintervalls für σ^2 (UCL_γ) auf der Basis einer Stichprobe von l Gruppen mit insgesamt n Beobachtungen erfolgt, $n > l$. Falls für die Bestimmung der Fallzahl ein Vielfaches $c_n \cdot S^2_{l-sample}$, $c_n > 0$, der gepoolten Varianz verwendet wird, resultiert aus der Formel (4.6) die Schätzung

$$\hat{N} = N \cdot \frac{c_n}{(n-l)} \cdot \frac{(n-l)S^2_{l-sample}}{\sigma^2}.$$

Für die zugehörige Power $\hat{\pi}$ gilt dann

$$\begin{aligned} \Pr(\hat{\pi} \geq 1-\beta) &= \Pr(\hat{N} \geq N) \\ &= \Pr\left(\frac{(n-l)S^2_{l-sample}}{\sigma^2} \geq \frac{(n-l)}{c_n} \right) \\ &= 1 - G_{\chi^2_{n-l}}\left(\frac{(n-l)}{c_n} \right), \end{aligned} \tag{4.20}$$

wobei $G_{\chi^2_{n-l}}$ die Verteilungsfunktion der zentralen Chi-Quadrat-Verteilung mit $n-l$ Freiheitsgraden bezeichnet. Wegen $G_{\chi^2_{n-l}}(n-l) > 0.50$ für $n > l$ folgt Aussage (a). Aussage

(b) folgt wegen $UCL_\gamma = \frac{(n-l)}{\chi^2_{\gamma,n-l}} \cdot S^2_{n-l}$ direkt aus (4.20), indem man in (4.20) $c_n = \frac{(n-l)}{\chi^2_{\gamma,n-l}}$ setzt.

■

Falls man also in einer Studie mit festem Stichprobenumfang mit hoher Wahrscheinlichkeit sicherstellen möchte, dass mindestens die vorgegebene Power erreicht wird, empfiehlt es sich, bei der Fallzahlplanung nicht die in der Planungsphase verfügbare Varianzschätzung selbst, sondern die zugehörige obere Schranke des einseitigen $(1-\gamma)$-Konfidenzintervalls für σ^2 zu verwenden. Dies führt zu einer erwarteten Power $>1-\beta$ mit dem Preis einer erwarteten Fallzahl, die größer als die notwendige Fallzahl N ist (KIESER und WASSMER, 1996).

Die Aussagen von Satz 10 sind direkt auf die Planung des ersten bzw. zweiten Studienteils einer Studie mit adaptivem Zwei-Stufen Design anzuwenden, vorausgesetzt, die Stichproben, die zur *a priori* Varianzschätzung herangezogen werden, stammen aus einer Population mit der gleichen Varianz wie die des jeweils zu planenden Studienteils. Diese Annahme ist für die Planung des zweiten Studienteils in den meisten Fällen unkritisch: Falls eine Studie nach der Zwischenauswertung fortgesetzt wird, erfolgt dies in aller Regel mit den gleichen Zentren und unter den gleichen Studienbedingungen. Gerade bei Durchführung eines zweiten Studienteils möchte man sicherstellen, dass die vorgegebene (bedingte) Power für die finale Auswertung auch tatsächlich erreicht oder überschritten wird.

Für das Design mit interner Pilotstudie ist das Resultat von Satz 10 nicht anwendbar, denn aufgrund der Option zur Fallzahlanpassung sind die Verteilungen des Stichprobenumfanges und der Power verschieden von denen bei festem Stichprobenumfang. Zur Untersuchung der Frage, ob analoge Resultate auch für dieses Design gelten, wurde in der Arbeit von KIESER und FRIEDE (2000a) die Verteilungsfunktion der Power berechnet. Dabei wurde ausgegangen vom Vorschlag von WITTES und BRITTAIN (Verwendung der gepoolten Varianzschätzung zur Fallzahladaption), für den in Kapitel 4.1.2.1 die entsprechenden Berechnungen unter H_0 durchgeführt wurden. Im folgenden werden die gleichen Bezeichnungen wie dort benutzt.

Für einen gegebenen Wert von v_1 (d.h. für gegebenes n_1 und s_1^2) ist die bedingte Power des einseitigen t^*-Tests unter der Alternativ-Hypothese H_1 im Design mit interner Pilotstudie gegeben durch

$$cp(v_1) = \int_0^\infty \int_{t(2(n_1+n_2-2),1-\alpha_{adj})\cdot\sqrt{\frac{v_1+v_2}{2(n_1+n_2-2)}}}^\infty f(d, v_1, v_2 | H_1)\, dd\, dv_2 ,$$

wobei α_{adj} das nominale Niveau bezeichnet, für das die tatsächliche Wahrscheinlichkeit eines Fehlers 1. Art durch α kontrolliert wird (siehe KIESER und FRIEDE, 2000a). Die Dichte f unter der Alternative H_1 kann analog zu $f(d, v_1, v_2 | H_0)$ (siehe Kapitel 4.1.2.1) hergeleitet werden und ist gegeben durch

$$f(d, v_1, v_2 | H_1) = g_{N(\sqrt{\frac{(n_1+n_2)}{2}} \cdot \frac{\mu_1-\mu_2}{\sigma}, 1)}(d) \cdot g_{\chi^2_{2(n_1-1)}}(v_1) \cdot g_{\chi^2_{2(n_2-1)}}(v_2).$$

Die Verteilungsfunktion der Power $\hat{\pi}_{WB}$ des Designs mit interner Pilotstudie erhält man damit als

$$\Pr(\hat{\pi}_{WB} \le z) = \int_{v_1 \in D_z} g_{\chi^2_{2(n_1-1)}}(v_1)\, dv_1,\ 0 \le z \le 1,$$

wobei der Integrationsbereich gegeben ist durch $D_z = \{v_1 : cp(v_1) \le z\}$.

In Tabelle 12 ist die Wahrscheinlichkeit, mindestens die vorgegebene Power $1-\beta = 0.80$ zu erreichen, für ein tatsächliches einseitiges Niveau $\alpha_{act} = 0.025$ des t^*-Tests, die Adaptionsregel von BIRKETT und DAY mit verschiedenen Varianzschätzern, $\sigma^2 = 1$ und verschiedene Alternativen $\Delta = \delta$ angegeben; die Berechnungen wurden mit Mathematica 3.0 durchgeführt. Unabhängig von der Fallzahl der internen Pilotstudie ist diese Wahrscheinlichkeit kleiner als 0.50, wenn die gepoolte Stichprobenvarianz der internen Pilotstudie zur Re-Kalkulation der Fallzahl verwendet wird. Die Resultate von Satz 10 bzgl. der oberen Schranke des oberen $(1-\gamma)$-Konfidenzintervalls gelten für Stichprobenumfänge der internen Pilotstudie, die deutlich unter der tatsächlich notwendigen Fallzahl liegen; für größeres n_1 liegt die Wahrscheinlichkeit, $1-\beta$ zu erreichen oder zu überschreiten, über $1-\gamma$.

Wie beim Design mit festem Stichprobenumfang und beim Zwei-Stufen-Design mit adaptiver Zwischenauswertung ist damit auch beim Design mit interner Pilotstudie die gepoolte Stichprobenvarianz mit einem Exzess-Faktor zu korrigieren, wenn das Erreichen der geplanten Power mit hoher Wahrscheinlichkeit sichergestellt werden soll. Auch unter diesem Aspekt ist die Verwendung der Ein-Stichproben-Varianz zur Fallzahladaption eine attraktive Alternative: Da sie unter H_1 die Populationsvarianz σ^2 überschätzt, wird die Wahrscheinlichkeit, mindestens die Power $1-\beta$ zu erreichen, größer sein als für S_1^2. Die exakten Eigenschaften dieser Strategie lassen sich analytisch bestimmen, indem die in Kapitel 4.1.2.2.2 angewandten Methoden auf die Situation der Alternativ-Hypothese übertragen werden.

Tabelle 12: Wahrscheinlichkeit, im Design mit interner Pilotstudie mindestens die geplante Power $1-\beta$ zu erreichen. Die Fallzahl für die Re-Kalkulation wird bestimmt nach der Regel von BIRKETT und DAY (1994) unter Verwendung der angegebene Schätzmethode für σ^2. t^*-Test, tatsächliches einseitiges Niveau $\alpha_{act} = 0.025$, $1-\beta = 0.80$, $\sigma^2 = 1$.

			Wahrscheinlichkeit, die geplante Power $1-\beta$ zu erreichen				
$\Delta = \delta$	Notwendige Fallzahl N	Fallzahl der internen Pilotstudie n_1	Stichproben-varianz der Pilotstudie s_1^2	0.60 UCL* für σ^2	0.70 UCL* für σ^2	0.80 UCL* für σ^2	0.90 UCL* für σ^2
0.30	350	10	0.43	0.60	0.70	0.80	0.90
		20	0.45	0.60	0.70	0.80	0.90
		30	0.46	0.60	0.70	0.81	0.91
		50	0.46	0.60	0.71	0.82	0.92
		100	0.46	0.61	0.73	0.84	0.94
		200	0.45	0.64	0.79	0.91	0.98
		300	0.43	0.71	0.91	0.99	1.00
0.50	126	10	0.42	0.59	0.70	0.80	0.90
		20	0.44	0.59	0.70	0.81	0.91
		30	0.44	0.60	0.71	0.82	0.92
		50	0.43	0.61	0.74	0.86	0.95
		100	0.40	0.67	0.87	0.99	1.00
0.70	64	10	0.41	0.59	0.70	0.80	0.91
		20	0.41	0.59	0.71	0.83	0.93
		30	0.40	0.60	0.74	0.86	0.96
		50	0.39	0.65	0.87	0.98	1.00

*$1-\gamma$ UCL = obere Schranke des einseitigen $(1-\gamma)$ - Konfidenzintervalls

4.3.2 Anwendung einer quasi-sequentiellen Prozedur

Um die Vorteile einer strikt-sequentiellen Fallzahlplanung auf ein zweistufiges Design zu übertragen, schlugen BETENSKY und TIERNEY (1997) das folgende quasi-sequentielle Verfahren vor. Die grundlegende Idee besteht darin, zukünftige Daten durch Resampling mit

Zurücklegen aus den bereits eingebrachten Daten der internen Pilotstudie zu simulieren. Bevor eine neue Beobachtung gezogen wird, wird die Varianzschätzung $S^2_{[r]}$ berechnet auf der Basis der kumulierten Stichprobe aus den beobachteten und den gezogenen Daten mit insgesamt r Beobachtungen. Durch Einsetzen von $S^2_{[r]}$ anstelle von σ^2 in die Fallzahlformel wird der zugehörige Stichprobenumfang $\hat{N}_{[r]}$ berechnet, und das Resampling wird gestoppt, sobald $r \geq \hat{N}_{[r]}$. Diese Schritte werden w mal wiederholt, und der resultierende Stichprobenumfang berechnet sich als der Mittelwert der w Stichprobenumfänge $\hat{N}_{[r]}$ beim Stop der Prozedur. Wie BETENSKY und TIERNEY (1997) wählen wir für unsere Simulationsuntersuchungen $w = 25$.

Diese von BETENSKY und TIERNEY für die Ein-Stichproben-Situation konzipierte Prozedur kann leicht auf $k \geq 2$ und die in Abschnitt 4.1.1 vorgestellten Varianzschätzer übertragen werden. In jedem Schritt werden nun k Beobachtungen gezogen, für den adjustierten und nicht-adjustierten Ein-Stichproben-Varianzschätzer aus den gepoolten Daten, für den k-Stichproben-Varianzschätzer jeweils eine Beobachtung aus jeder der Behandlungsgruppen. Das quasi-sequentielle Verfahren mit den verblindeten Varianzschätzern eignet sich für die Fallzahladaption im Design mit interner Pilotstudie. Mit der gepoolten Varianz der Stichprobe des ersten Studienteils kann die quasi-sequentielle Prozedur für die Planung des zweiten Studienteils eines adaptiven Designs mit Zwischenauswertung angewendet werden.

In FRIEDE und KIESER (2000a) wurde für $k = 2$ das quasi-sequentielle Verfahren mit gepoolter Varianzschätzung mit dem Standardvorgehen mit gepoolter Varianzschätzung, adjustierter Ein-Stichprobenvarianz und mit der EM-Algorithmus-basierten Prozedur verglichen. Die quasi-sequentielle Prozedur hatte dabei (wie bei den Untersuchungen von BETENSKY und TIERNEY, 1997, in der Ein-Stichproben-Situation) zu einer Reduktion der Mean Squared Errors (MSE) der resultierenden Fallzahl geführt. Der MSE ist die Summe des quadrierten Bias und der Varianz des Schätzers und vereinigt damit sowohl den Aspekt der Verzerrung als auch der Variabilität der Schätzung. Weiterhin kann er im entscheidungstheoretischen Rahmen als Risiko bei quadratischer Verlustfunktion interpretiert werden (siehe Kapitel 3.2). Dies ist für viele Anwendungssituationen ein realistisches Szenario, da bei der Fallzahlschätzung sowohl zu kleine als auch zu große Stichprobenumfänge (mit der Konsequenz ungenügender Power bzw. unnötig hoher Patientenzahlen) unerwünscht sind.

Im folgenden werden die Ergebnisse einer Monte-Carlo-Simulationsstudie vorgestellt, bei der für den gepoolten Varianzschätzer und den adjustierten und nicht-adjustierten Ein-Stichproben-Varianzschätzer die quasi-sequentielle Prozedur mit dem Standardverfahren zur

Fallzahladaption verglichen wurde. Für $k = 2$, einseitiges Niveau $\alpha = 0.025$, $1-\beta = 0.80$, $\sigma^2 = 1$ und $\Delta = \delta = 0.5$ wurden verschiedene Fallzahlen n_1 der internen Pilotstudie betrachtet. Für jede Situation wurden 10 000 Replikationen erzeugt. Die Simulationen wurden mit SAS/IML durchgeführt. In Tabelle 13 sind Mittelwert und Standardabweichung sowie MSE der resultierenden Fallzahlen pro Gruppe angegeben.

Tabelle 13: Simulierter Erwartungswert (MW) der resultierenden Gesamtfallzahl pro Gruppe sowie Standardabweichung (SD) und Mean Squared Error (MSE). Fallzahl pro Gruppe zur Varianzschätzung n_1; $\alpha = 0.025$ (einseitig), geplante Power $1-\beta = 0.80$, $\Delta = \delta = 0.5$, $\sigma^2 = 1$, d.h., $N = 64$. 10 000 Replikationen, 25 Wiederholungen für Resampling-Strategie.

	Simulierte Fallzahl pro Gruppe											
	Gepoolte Zwei-Stichproben-Varianz						Ein-Stichproben-Varianz					
	ohne Resampling			mit Resampling			ohne Resampling			mit Resampling		
n_1	MW	(SD)	MSE	MW	(SD)	MSE	MW	(SD)	MSE	MW	(SD)	MSE
10	64.0	(20.7)	428.8	58.0	(18.5)	378.8	68.0	(21.6)	483.7	64.8	(20.5)	420.0
20	64.3	(14.3)	204.6	61.7	(13.5)	186.7	68.3	(15.0)	243.8	67.0	(14.6)	220.8
30	64.2	(11.6)	134.0	62.9	(11.1)	123.5	68.2	(12.2)	165.0	67.5	(11.9)	153.3
50	64.5	(8.7)	76.0	64.5	(8.4)	71.4	68.3	(9.3)	105.8	68.5	(9.2)	104.5

Man erkennt, dass die Anwendung der Resampling-Prozedur zu einer Reduktion der Streuung der resultierenden Fallzahl führt. Insbesondere für kleine Fallzahlen n_1 ist der Erwartungswert der resultierenden Fallzahl für die quasi-sequentielle Prozedur kleiner als be einfachem Einsetzen des entsprechenden Schätzwertes der Varianz in die Fallzahlformel. Für die Ein-Stichprobenvarianz führt dies zu einer Reduktion, für die gepoolte Zwei-Stichproben-Varianz hingegen zu einer Erhöhung des Bias im Vergleich zur Prozedur ohne Resampling. Hinsichtlich des Mean square errors ist die jeweilige quasi-sequentielle Prozedur für alle Situationen überlegen, der Vorteil ist umso größer, je kleiner die Fallzahl n_1 der Stichprobe ist, auf der die Varianzschätzung beruht. Insgesamt kann damit festgehalten werden, dass durch Anwendung des Resampling-Verfahrens sowohl im adaptiven Zwei-Stufen-Design mit Zwischenauswertung (Verwendung der gepoolten Varianz) als auch im Design mit interner Pilotstudie (Verwendung der adjustierten oder nicht-adjustierten Varianz) die Präzision der re-kalkulierten Fallzahl erhöht werden kann.

4.3.3 Vergleich der resultierenden Fallzahl und Power

Wir hatten bereits in Kapitel 4.2 darauf hingewiesen, dass für das adaptive Design mit Zwischenauswertung neben der Option zu einem vorzeitigen Studien-Stop ein weiterer Unterschied zum Design mit interner Pilotstudie darin besteht, dass die resultierende Fallzahl vom *p*-Wert des ersten Studienteils abhängt. Es ist zu erwarten, dass sich dies in den Charakteristika von Fallzahl und Power niederschlägt. Den Fragen, welcher Art die Unterschiede sind und ob man von einer Überlegenheit eines der Designs sprechen kann, wird in diesem Abschnitt nachgegangen. Um den Vergleich der beiden Ansätze nicht unnötig zu erschweren, betrachten wir die Zwei-Stichproben *t*-Test-Situation und nehmen an, dass für beide Designs die Adaption des Stichprobenumfanges auf der Basis der gepoolten Varianzschätzung erfolgt. Die Fallzahl wird für den *a priori* spezifizierten minimalen klinisch relevanten Effekt δ und zur (bedingten) Power $1-\beta$ bestimmt; für das Design mit interner Pilotstudie wird der endgültige Stichprobenumfang nach der Regel von BIRKETT und DAY festgelegt. Für das BAUER-KÖHNE-Design wurden die Stop-Grenzen $\alpha_0 = 1$ und $\alpha_1 = c_\alpha$ verwendet. Die Berechnungen wurden mit Mathematica 3.0 durchgeführt. Die Ergebnisse dieser Betrachtungen finden sich auch in FRIEDE und KIESER (2000a).

4.3.3.1 Erwartete Fallzahl und Power

Aus der Erwartungstreue der gepoolten Varianzschätzung und der Fallzahl-Approximationsformel (4.11) folgt, dass für den Erwartungswert der Fallzahl pro Gruppe $\hat{N}_{WB}$ im Design mit interner Pilotstudie approximativ gilt $E\left(\hat{N}_{WB}\right) = \max\{n_1, N\}$. Die erwartete Power ist damit gegeben durch die Power für den Stichprobenumfang $\max\{n_1, N\}$ im Design mit festem Stichprobenumfang (FRIEDE, 2000b).

Für das adaptive Zwei-Stufen Design nach BAUER und KÖHNE haben POSCH und BAUER (2000) die erwartete Fallzahl und Power für allgemeine Stop-Grenzen hergeleitet. Für den hier untersuchten Spezialfall $\alpha_0 = 1,\ \alpha_1 = c_\alpha$ lässt sich der erwartete Stichprobenumfang pro Gruppe wie folgt berechnen:

$$E(\hat{N}_{BK}) = n_1 + \int_0^\infty \int_{A(v_1)} n_2(v_1, p_1(d_1, v_1)) \cdot g_{\chi^2_{2(n_1-1)}}(v_1) \cdot g_{N(\sqrt{\frac{n_1}{2}}\frac{\mu_1-\mu_2}{\sigma},1)}(d_1)\, dd_1\, dv_1 .$$

Dabei bezeichnet $V_1 = \frac{2(n_1-1)S_1^2}{\sigma^2}$ wie in Abschnitt 4.1.2.1 die transformierte Varianz, und p_1 bezeichnet den *p*-Wert nach der ersten Stufe. Es gilt $p_1 = p_1(d_1, v_1) = 1 - G_{t_{2(n_1-1)}}(t_1)$ mit dem Wert der Teststatistik t_1 nach der ersten Stufe, d.h. $t_1(d_1, v_1) = \sqrt{2(n_1-1)}\,\frac{d_1}{\sqrt{v_1}}$, und der Verteilung der zentralen *t*-Verteilung mit *df* Freiheitsgraden $G_{t_{df}}$. Die innere Integration erfolgt über alle Werte d_1, für die bei gegebenem v_1 die Null-Hypothese nach dem ersten Studienteil nicht abgelehnt wird:

$$A(v_1) = \{d_1 \mid t_1 \le t_{2(n_1-1),1-c_\alpha}\} = \{d_1 \mid d_1 \le (t_{2(n_1-1),1-c_\alpha})\sqrt{\frac{v_1}{2(n_1-1)}}\ \}.$$

Die Fallzahl n_2 für den zweiten Studienteil wird zum Niveau c_α / p_1 und zur Power $1-\beta$ bestimmt und berechnet sich in Analogie zu (4.13) als

$$n_2(v_1,\ p_1) = 2 \cdot \frac{\left(z_{1-c_\alpha/p_1} + z_{1-\beta}\right)^2}{\delta^2} \cdot \frac{v_1 \cdot \sigma^2}{2(n_1-1)}.$$

Die erwartete Power ist gegeben durch

$$E(\hat{\pi}_{BK}) = 1 - \int_0^\infty \int_{A(v_1)} \int_{-\infty}^{t_{2(n_1-1),1-c_\alpha/p_1}} g_{\chi^2_{2(n_1-1)}}(v_1) \cdot g_{N(\sqrt{\frac{n_1}{2}}\frac{\mu_1-\mu_2}{\sigma},1)}(d_1) \cdot g_{t_{2(n_2-1),\sqrt{\frac{n_2}{2}}\frac{\mu_1-\mu_2}{\sigma}}}(t_2)\,dt_2\,d\,d_1\,dv_1\ .$$

Tabelle 14 zeigt die erwartete Fallzahl und Power des adaptiven Zwei-Stufen Designs mit Zwischenauswertung für die oben beschriebene Adaptionsstrategie und für das einseitige Niveau $\alpha = 0.025$, $1-\beta = 0.80$, $\sigma^2 = 1$, $\Delta = \delta = 0.5$ und verschiedene Fallzahlen des ersten Studienteils. Wie bereits in Kapitel 4.2.1 beschrieben beträgt die Gesamt-Power stets mindestens $1-\beta$, da kein vorzeitiges Studienende mit Beibehaltung der Null-Hypothese möglich ist. Aufgrund der streng monotonen Abhängigkeit der Power des ersten Studienteils von der Fallzahl n_1 wächst auch die erwartete Gesamt-Power mit zunehmendem Stichprobenumfang des ersten Abschnitts. Im Gegensatz dazu verläuft die erwartete Fallzahl $E(\hat{N}_{BK})$ als Funktion von n_1 U-förmig. Die minimale erwartete Fallzahl ist ungefähr so groß wie die Fallzahl N, die im Design mit festem Stichprobenumfang benötigt wird, um eine Power von $1-\beta$ zu erreichen. Das Minimum wird erreicht für eine Fallzahl des ersten Studienteils von etwa $0.4 \cdot N$. Diese Ergebnisse gelten in gleicher Weise auch für die Vielzahl anderer Parameterkonstellationen, die wir untersuchten, und decken sich mit dem Ergebnis von POSCH und BAUER (2000). Dort wird gezeigt, dass unter der asymptotisch gültigen Annahme bekannter Varianz stets eine eindeutig bestimmte Fallzahl n_1 existiert, die den

Gesamt-Stichprobenumfang unter der oben beschriebenen Strategie minimiert und dass dieses Minimum für eine Fallzahl nahe $0.4 \cdot N$ angenommen wird**.**

Tabelle 14: Erwartete Fallzahl pro Gruppe $E(\hat{N}_{BK})$ und erwartete Power $E(\hat{\pi}_{BK})$ für das adaptive Design mit Zwischenauswertung nach BAUER und KÖHNE (1994). *t*-Test, einseitiges Niveau $\alpha = 0.025$, $\sigma = 1$, $\Delta = \delta = 0.5$, Fallzahl pro Gruppe n_1 für den ersten Studienteil, geplante Power $1-\beta = 0.80$ für den zweiten Studienteil. $E(\hat{N}_{WB})$ ist die erwartete Fallzahl pro Gruppe, um im Design mit interner Pilotstudie nach WITTES und BRITTAIN (1990) unter den gleichen Design-Spezifikationen eine erwartete Power von $E(\hat{\pi}_{BK})$ zu erzielen.

	BAUER-KÖHNE Design		WITTES-BRITTAIN Design
Fallzahl pro Gruppe n_1	Erwartete Fallzahl $E(\hat{N}_{BK})$	Erwartete Power $E(\hat{\pi}_{BK})$	Erwartete Fallzahl $E(\hat{N}_{WB})$ zur Power $E(\hat{\pi}_{BK})$
10	67.0	0.80	60.8
20	63.6	0.82	66.3
30	63.4	0.84	70.7
40	65.4	0.87	75.4
60	74.4	0.91	86.1

Eine Möglichkeit, das Zwei-Stufen-Design mit Zwischenauswertung mit dem Design mit interner Pilotstudie zu vergleichen, besteht darin, für letzteres die erwartete Fallzahl zu berechnen, die notwendig ist, um die unter dem BAUER-KÖHNE-Design erwartete Power $E(\hat{\pi}_{BK}) \geq 1-\beta$ zu erreichen. Den Ergebnissen in Tabelle 14 kann man entnehmen, dass für Fallzahlen $n_1 \geq 20$ ein Vorteil für das Design mit Zwischenauswertung besteht.

4.3.3.2 Verteilung der Fallzahl

Nach den Erwartungswerten vergleichen wir nun die Verteilungsfunktionen der resultierenden Fallzahlen für die beiden adaptiven Designs mit und ohne Zwischenauswertung. Für das Design mit interner Pilotstudie ist die kumulative Verteilungsfunktion der Fallzahl $G_{WB}(n)$ für $n < n_1$ gleich 0. Für $n \geq n_1$ ist $G_{WB}(n)$ bestimmt durch den Wert der gepoolten Varianz s_1^2, die zu einer Fallzahl von n führt. Es folgt deshalb aus (4.13)

$$G_{WB}(n) = G_{\chi^2_{2(n_1-1)}}\left(\frac{2(n_1-1)}{N} n\right).$$

Im adaptiven Design mit Zwischenauswertung ist die kumulative Verteilungsfunktion der Fallzahl $G_{BK}(n)$ ebenfalls 0 für $n < n_1$. Für $n = n_1$ ist $G_{BK}(n)$ gleich der Power für den ersten Studienteil, und für $n > n_1$ ist $G_{BK}(n)$ gegeben durch

$$G_{BK}(n) = 1 - G_{t_{2(n_1-1),\sqrt{\frac{n_1}{2}}\frac{\mu_1-\mu_2}{\sigma}}}(t_{2(n_1-1),1-\alpha_1}) + \int_{-\infty}^{+\infty} \int_{B(d_1)} g_{\chi^2_{2(n_1-1)}}(v_1) \cdot g_{N(\sqrt{\frac{n_1}{2}}\frac{\mu_1-\mu_2}{\sigma},1)}(d_1)\, dv_1\, d\, d_1,$$

wobei $B(d_1) = \{v_1 \mid t_1 \le t_{2(n_1-1),1-\alpha_1}$ und $n_2 \le n - n_1\}$ die Menge der Werte $v_1 > 0$ bezeichnet, die für gegebene Differenz d_1 nicht zur Ablehnung der Null-Hypothese nach dem ersten Studienteil führt sowie zu einer Fallzahl n_2 mit $n_2 \le n - n_1$ für den zweiten Studienteil. $G_{t_{df,\vartheta}}$ bezeichnet die Verteilungsfunktion der nicht-zentralen *t*-Verteilung mit *df* Freiheitsgraden und Nicht-Zentralitätsparameter ϑ. Der Summand $1 - G_{t_{2(n_1-1),\sqrt{\frac{n_1}{2}}\frac{\mu_1-\mu_2}{\sigma}}}(t_{2(n_1-1),1-\alpha_1})$ ist die Power des ersten Studienabschnitts.

Abbildung 9 zeigt die Verteilungsfunktion für die oben beschriebenen Design-Spezifikationen und eine Fallzahl pro Gruppe für den ersten Studienteil bzw. die interne Pilotstudie $n_1 = 30$.

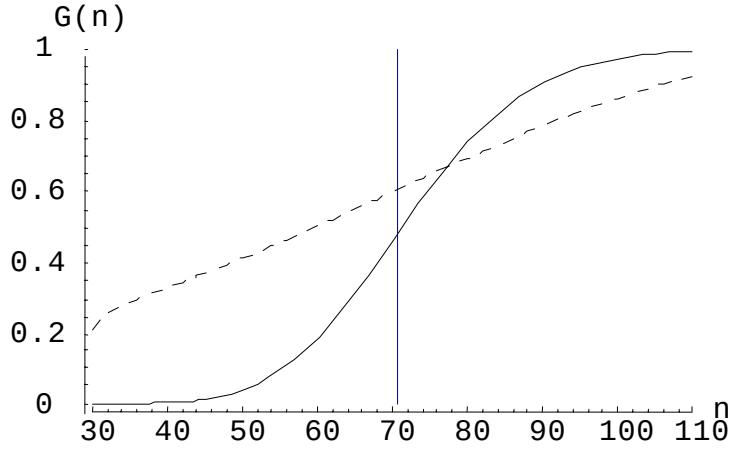

Abbildung 9: Verteilungsfunktion $G(n)$ der resultierenden Fallzahl pro Gruppe für das adaptive Design mit Zwischenauswertung nach BAUER und KÖHNE (1994) (gestrichelte Linie) und das Design mit interner Pilotstudie nach WITTES und BRITTAIN (1990) (durchgezogene Linie). *t*-Test, einseitiges Niveau $\alpha = 0.025$, $\sigma = 1$, $\Delta = \delta = 0.5$, Fallzahl pro Gruppe für den ersten Studienteil $n_1 = 30$. Geplante Power $1 - \beta = 0.80$ für den zweiten Studienteil des Designs nach BAUER und KÖHNE, die Fallzahl für das Design mit interner Pilotstudie nach WITTES und BRITTAIN ist so gewählt, dass die gleiche erwartete Power wie für das Design nach BAUER und KÖHNE erzielt wird ($E(\hat{\pi}_{BK}) = E(\hat{\pi}_{BK}) = 0.844$; die notwendige Fallzahl pro Gruppe im Design mit festem Stichprobenumfang beträgt für diese Power $N = 71$).

Man sieht, dass für das adaptive Design mit Zwischenauswertung das Auftreten extrem großer oder kleiner Gesamtfallzahlen eine wesentlich höhere Wahrscheinlichkeit besitzt als für das

Design mit interner Pilotstudie. Hohe Fallzahlen rühren von „großen“ *p*-Werten im ersten Studienabschnitt her, kleine Fallzahlen von Abbrüchen nach der Zwischenauswertung oder „kleinen“ *p*-Werten p_1. Die Variation der resultierenden Fallzahl ist somit für das Design mit der Option zum vorzeitigen Studienende größer und damit die Unsicherheit darüber, welchen Stichprobenumfang die Studie insgesamt haben wird.

4.3.3.3 Schlussfolgerungen

Die im vorangehenden Abschnitt dargestellten Vergleiche der Charakteristika von Power und Fallzahl für das adaptive Zwei-Stufen-Design und das Design mit interner Pilotstudie zeigen, dass das BAUER-KÖHNE-Design für moderate Fallzahlen für den ersten Studienteil eine höhere Power als das WITTES-BRITTAIN-Design besitzt. Auf der anderen Seite ist die Variabilität der Fallzahl ebenfalls größer mit einer größeren Chance eines kleineren, aber auch einem erheblichen Risiko eines größeren Stichprobenumfanges als im Internal Pilot Study Design. In der Praxis wird die Entscheidung darüber, ob eine Studie mit oder ohne Zwischenauswertung durchgeführt wird, in aller Regel nicht allein aufgrund der statistischen Eigenschaften der Designs getroffen werden. Eine wesentliche Rolle spielt die Frage, ob sich eine Zwischenauswertung sinnvoll in den Studienablauf integrieren lässt. Hierzu ist es unter anderem notwendig, dass die individuelle Beobachtungszeit wesentlich kürzer als die für die Studie veranschlagte Rekrutierungszeit ist. Ansonsten ist zum Zeitpunkt der Zwischenauswertung die Zahl der Patienten, die in die Studie eingeschlossen wurden und diese zumindest teilweise durchlaufen haben, erheblich größer als die der bei der Interimanalyse berücksichtigten. Der Nutzen eines vorzeitigen Abbruchs der Studie wäre dann erheblich eingeschränkt. Weiterhin hängt die Entscheidung davon ab, ob für die untersuchte Therapie bereits fundierte Kenntnisse vorliegen oder die Studie in einer frühen Phase der Arzneimittelentwicklung durchgeführt wird. Gerade in letzterem Fall ist es häufig wünschenswert, im Rahmen einer Zwischenauswertung Wirksamkeit und Verträglichkeit vor Erreichen der initial geplanten Fallzahl zu beurteilen und neben der Fallzahl auch andere Design-Merkmale verändern zu können. Bei einer Studie mit einer erprobten Behandlung ist es dagegen oftmals ausreichend, sicherzustellen, dass die statistische Power zum Erreichen des Studienziels bei der aktuellen Streuung der Zielgröße den gewünschten Wert erreicht. Wenn prinzipiell beide Studiendesigns in Frage kommen, so können die Ergebnisse dieses Kapitels die Entscheidung zwischen den beiden Alternativen unterstützen.

5. Adaptive Auswahl der Teststatistik

Für feste Werte für das Signifikanzniveau und die Fallzahl und bei gegebener Alternativ-Hypothese hängt die Power eines statistischen Tests von der Verteilung, die den Daten zugrunde liegt, ab. Es ist deshalb eine attraktive Option, das in der Planungsphase aufgrund von (in der Regel unsicheren) Annahmen festgelegte Testverfahren auf der Basis neu hinzugekommener Informationen im Studienverlauf verändern zu können. Dass in adaptiven Designs eine solche Änderung unter Einhaltung der Wahrscheinlichkeit eines Fehlers 1. Art möglich ist, haben wir in Satz 6, Kapitel 2.4 bewiesen. LANG, AUTERITH und BAUER (2000) machten sich dies in der Situation zunutze, dass die Gleichheit der Erwartungswerte von k Behandlungsgruppen unter einer geordneten monotonen Alternative getestet werden soll. Hier werden die Daten des ersten Studienteils dazu benutzt, um die optimalen Scores für das unbekannte Erwartungswertprofil zu schätzen; der zugehörige Kontrast-Test (siehe Kapitel 3.5) wird dann bei der Auswertung des zweiten Studienteils verwendet. Es zeigte sich, dass insbesondere bei ausreichend großer Fallzahl für die Lernstichprobe mit der adaptiven Wahl der Scores eine höhere Power als beim nicht-adaptiven Vorgehen erzielt werden kann.

Im folgenden Abschnitt 5.1 wird ein Bootstrap-Verfahren vorgestellt, mit dem unter Verwendung beobachteter Daten (z.B. Daten einer Pilotstudie oder einer Zwischenauswertung) die statistische Power für Shift-Alternativen approximiert werden kann, ohne dass Annahmen über die Form der Verteilung getroffen werden müssen. In der Situation zweistufiger adaptiver Designs kann dieses Verfahren dazu verwendet werden, für den zweiten Studienteil diejenige Teststatistik auszuwählen, für die mit den Daten der Zwischenauswertung die maximale Power erzielt wird. Die Charakteristika einer solchen Selektionsstrategie werden in Abschnitt 5.2 in einer Monte-Carlo-Simulationsstudie untersucht. Die Anwendung des Verfahrens wird an einer klinischen Studie in der Indikation Depression illustriert, die im adaptiven Zwei-Stufen-Design durchgeführt wurde und bei der die vorgeschlagene Prozedur tatsächlich zu einem Wechsel der Teststatistik nach der Zwischenauswertung geführt hat (DIENEL und KIESER, 1999).

5.1 Schätzung der Power für Zwei-Stichproben-Probleme nach COLLINGS und HAMILTON (1988)

Die folgende Methode wurde von COLLINGS und HAMILTON (1988) zur Schätzung der Power des Zwei-Stichproben Wilcoxon-Tests bei Lokations-Alternativen vorgeschlagen. TROENDLE

(1999) zeigte in einer Simulationsstudie, dass dieses Verfahren für eine breite Klasse von Verteilungen bessere Ergebnisse liefert als die Schätzung der Power unter Annahme normalverteilter Daten (siehe z.B. LEHMANN, 1975) und die Approximation via Edgeworth-Entwicklung (siehe z.B. BARNDORFF-NIELSEN und COX, 1979). MAHONEY und MAGEL (1996) verallgemeinerten das Verfahren auf das k-Stichprobenproblem und den Kruskal-Wallis-Test. Wir werden die Prozedur für $k = 2$, aber in ihrer allgemeinen Form für einseitige verteilungsfreie Zwei-Stichproben-Tests mit Teststatistik T darstellen.

Die beiden unabhängigen Stichproben werden mit $X_{ij}, i = 1, 2, j = 1,..., n_i,$ bezeichnet. Die $X_{11},..., X_{1n_1}$ sind eine Zufallsstichprobe aus der Verteilung mit stetiger kumulativer Verteilungsfunktion $F_{X_1} = F$, und die $X_{21},..., X_{2n_2}$ sind eine Zufallsstichprobe aus der Verteilung mit stetiger Verteilungsfunktion F_{X_2}. Wir setzen voraus, dass diese Verteilungen gleiche Form haben, benötigen aber keine Kenntnis über diese Form. Weiterhin nehmen wir an, dass die Verteilungen um Δ gegeneinander verschoben sind: $F_{X_2}(x_2) = F_{X_1}(x_2 - \Delta) = F(x_2 - \Delta)$ für alle x_2. Wir betrachten das einseitige Testproblem $H_0 : \Delta = 0$ gegen $H_1 : \Delta > 0$ und einen zugehörigen Niveau-α-Test mit Teststatistik T, für die große Werte für die Alternativ-Hypothese H_1 sprechen. Die kumulative Verteilungsfunktion von T, die durch F und Δ vollständig festgelegt ist, sei mit $F_T(\cdot, \Delta)$ bezeichnet, die kritische Schranke zum Niveau α mit t_α, und die statistische Power für die Verteilung F, die Lokations-Alternative Δ und das Signifikanzniveau α mit $\pi(F, T; \Delta, \alpha)$. Dann gilt: $\alpha = 1 - F_T(t_\alpha; 0)$ und $\pi(F, T; \Delta, \alpha) = 1 - F_T(t_\alpha; \Delta)$. Für verteilungsfreie Tests ist die Null-Verteilung von T, $F_T(\cdot; 0)$, unabhängig von der zugrundeliegenden stetigen Verteilung F. Demgegenüber hängt die statistische Power sehr wohl von F ab, weshalb diese in der Praxis häufig unter einer spezifisch angenommenen Verteilung F berechnet wird - in der Situation, dass ein verteilungsfreier Test deshalb ausgewählt wurde, weil die Form von F unbekannt ist, eine geradezu widersinnige Strategie.

Bei der Methode von COLLINGS und HAMILTON (1988) wird eine nicht-parametrische Schätzung $\hat{F}$ von F erzeugt und die tatsächliche Power $\pi(F, T; \Delta; \alpha)$ durch $\pi(\hat{F}, T; \Delta, \alpha)$ geschätzt. Zur approximativen Berechnung von $\pi(\hat{F}, T; \Delta, \alpha)$ wird ein Bootstrap-Verfahren angewandt. Sowohl für die Schätzung der Verteilungsfunktion als auch für die Approximation von $\pi(\hat{F}, T; \Delta, \alpha)$ gibt es mehrere Möglichkeiten. Im folgenden wird das Verfahren

angewendet, das in umfangreichen Vergleichsuntersuchungen zu den besten Ergebnissen geführt hat (COLLINGS und HAMILTON, 1986, 1988; HAMILTON und COLLINGS, 1991).

Unter Verzicht auf die technischen Einzelheiten kann das von COLLINGS und HAMILTON vorgeschlagene Verfahren zur Schätzung der Power für einen Stichprobenumfang $m_1 + m_2$ wie folgt beschrieben werden. Für jede der beiden Stichproben $X_{11},...,X_{1n_1}$ und $X_{21},...,X_{2n_2}$ wird folgendermaßen vorgegangen:

- Die zugrundeliegende Verteilungsfunktion wird durch die empirische Verteilungsfunktion geschätzt.
- Es wird eine Zufallsstichprobe vom Umfang $m_1 + m_2$ aus der geschätzten Verteilung gezogen; zu den m_2 Beobachtungen $m_1 + 1, ..., m_1 + m_2$ wird der Wert Δ addiert.
- Mit diesen Daten wird der entsprechende Test durchgeführt, und es wird gezählt, ob die Null-Hypothese abgelehnt wird oder nicht.
- Die letzten beiden Schritte werden w mal durchgeführt und die Power durch den Anteil der Ablehnung unter den w Wiederholungen geschätzt.

Die resultierende Power-Schätzung ist dann das gewichtete Mittel der Schätzungen, die aus den beiden Stichproben $X_{11},...,X_{1n_1}$ und $X_{21},...,X_{2n_2}$ resultieren.

Im einzelnen werden nach COLLINGS und HAMILTON (1988) folgende Schritte durchgeführt:

Schätzung der kumulativen Verteilungsfunktion F

Wir nehmen an, dass q Beobachtungen $z_1, ..., z_q$ vorliegen. (Wir werden weiter unten sehen, dass die z entweder die x_1- oder die x_2-Werte sind.) Die zugehörigen geordneten Werte seien mit $z_{(1)}, ..., z_{(q)}$ bezeichnet, und wir definieren $z_{(0)} = 2z_{(1)} - z_{(2)}$ und $z_{(q+1)} = 2z_{(q)} - z_{(q-1)}$; dann erhält man die stetige Verteilungsfunktion $\hat{F}$, indem man jedem Intervall $(z_{(i)}, z_{(i+1)}), i = 0, ..., q$ die Wahrscheinlichkeit $1/(q+1)$ zuweist.

Approximation von $\pi(\hat{F}, T; \Delta, \alpha)$ durch Resampling aus der Verteilung $\hat{F}$

Für eine Schätzung $\hat{F}$ von F kann die Power $\pi(\hat{F}, T; \Delta, \alpha)$ für einen Stichprobenumfang von $m_1 + m_2$ durch die folgende Resampling-Prozedur approximiert werden.

1. Es wird eine Zufallsstichprobe vom Umfang $m_1 + m_2$ aus $\hat{F}$ gezogen. Die ersten m_1 Beobachtungen bilden eine simulierte Stichprobe von X_1 und werden mit $X_{11}^0, ..., X_{m_1}^0$ bezeichnet. Dann wird zu jedem der verbleibenden m_2 Beobachtungen der Stichprobe der

Wert Δ addiert, um eine simulierte Stichprobe von X_2 zu erhalten, die mit $X_{21}^0, ..., X_{2m_2}^0$ bezeichnet wird. (Die X_1^0 und X_2^0 werden aus der gleichen Verteilung generiert, mit dem einzigen Unterschied, dass die Verteilung der X_2^0 um Δ Einheiten nach rechts verschoben ist.)

2. Die Teststatistik T^0 wird für die X_1^0 und X_2^0 berechnet. Falls $T^0 \geq t_\alpha$ dann wird eine Ablehnung von H_0 gezählt, im anderen Fall eine Beibehaltung von H_0.
3. Die Schritte 1 und 2 werden w mal wiederholt. (COLLINGS und HAMILTON, 1988, verwendeten $w = 2500$.) Die Power $\pi(\hat{F}, T; \Delta, \alpha)$ wird approximiert durch die Anzahl der Ablehnungen von H_0 unter den w Wiederholungen. Diese Schätzung wird mit $\hat{\pi}(\hat{F}, T; \Delta, \alpha)$ bezeichnet.

COLLINGS und HAMILTON (1986) untersuchten mehrere mögliche Varianten des Bootsstrap-Verfahrens zur Schätzung der Power. Die folgende Methode erwies sich als die mit Abstand effizienteste:

F wird zum einen ausschließlich auf der Basis der Werte X_1 (Verteilungsfunktion $\hat{F}_{X_1}$) und zum anderen ausschließlich auf der Basis der Werte X_2 (Verteilungsfunktion $\hat{F}_{X_2}$) geschätzt. Die Power wird dann approximiert durch das gewichtete Mittel

$$\hat{\pi}(\hat{F}, T; \Delta, x) = \frac{n_1 \hat{\pi}(\hat{F}_{X_1}, T; \Delta, \alpha) + n_2 \hat{\pi}(\hat{F}_{X_2}, T; \Delta, \alpha)}{n_1 + n_2}.$$

Es ist offensichtlich, dass das Verfahren mit naheliegenden Modifikationen bei vorgegebenem Signifikanzniveau α, Power $1 - \beta$ und Alternative Δ zur Berechnung des notwendigen Stichprobenumfanges bei der Anwendung nicht-parametrischer Tests verwendet werden kann (siehe HAMILTON und COLLINGS, 1991). Wir werden im folgenden Kapitel die Frage untersuchen, ob diese Methode dazu geeignet ist, basierend auf den Daten aus einer Zwischenauswertung eine effiziente Teststatistik für den zweiten Studienteil eines adaptiven Designs auszuwählen. Die Möglichkeit, dieses Verfahren für eine adaptive Strategie zu nutzen wurde von COLLINGS und HAMILTON (1988) und BÜNING und TRENKLER (1994, S. 316) erwähnt. Allerdings wurde dabei weniger an eine Strategie, wie sie im folgenden angegeben wird, gedacht, sondern an eine im Sinne sogenannter adaptiver Tests. Bei diesem Typ von Tests wird die Entscheidung darüber, welche Teststatistik für die Auswertung benutzt wird, auf der Basis des aktuellen Datensatzes getroffen. Bislang wurden aber auch zu

dieser Anwendungsmöglichkeit des Verfahrens keinerlei Untersuchungen angestellt. Bevor wir die Eigenschaften der oben beschriebenen Strategie im Rahmen eines adaptiven Zwei-Stufen-Designs in einer Simulationsstudie untersuchen, soll die praktische Anwendung der Methode anhand einer konkreten klinischen Studie illustriert werden.

Beispiel 6:

In einer doppelblinden randomisierten placebo-kontrollierten Multicenter-Studie wurde die Wirksamkeit und Verträglichkeit von *Hypericum perforatum* bei Patienten mit leichter bis mittelschwerer Depression untersucht (DIENEL und KIESER, 1999). Die Studie wurde im adaptiven Zwei-Stufen-Design nach BAUER und KÖHNE (1994) zu einem einseitigen Signifikanzniveau $\alpha = 0.025$ und den Design-Charakteristika $\alpha_0 = 0.20, \alpha_1 = 0.0152$ und $c_\alpha = 0.00380$ durchgeführt. Zielgröße war die Differenz des Gesamtscores der Hamilton-Depressions-Skala (HAMD, 17-Item-Version) zwischen Therapiebeginn (Tag 0) und Therapieende (Tag 42). Als Auswertungsverfahren für den ersten Studienteil wurde im Protokoll der *t*-Test für unverbundene Stichproben festgelegt. Aufgrund der asymptotischen Verteilungsfreiheit des *t*-Tests, der aufgrund des Randomisierungsverfahrens zu erwartenden Balanciertheit der Behandlungsgruppen und der analytischen Berechnungen von HEEREN und D´AGOSTINO (1987) kann man davon ausgehen, dass der *t*-Test auch für die ordinal-skalierte Zielgröße das vorgegebene Niveau α einhält. Welche Power der *t*-Test in der vorliegenden Situation im Vergleich zu anderen Verfahren besitzt, wird durch diese Betrachtungen allerdings nicht beantwortet (siehe z.B. die Untersuchungen zu Power-Unterschieden zwischen verschiedenen Zwei-Stichproben-Tests bei bekannten stetigen Verteilungen in BÜNING, 1991, S. 119ff). In die Zwischenauswertung gingen die Daten von $2 \times 84 = 168$ Patienten ein, die im Rahmen der Intention-to-treat-Auswertung analysiert wurden. Der einseitige *p*-Wert betrug $p_1 = 0.043$, womit zur Ablehnung der Null-Hypothese im zweiten Studienteil ein *p*-Wert erzielt werden muss, der unter $c_\alpha / p_1 = 0.088$ liegt. Mit den Daten der Zwischenauswertung wurde die Bootstrap-Methode nach COLLINGS und HAMILTON (1988) angewendet, um für den *t*- und den U-Test die Power für einen Stichprobenumfang von 2×100 Patienten zu schätzen. Es wurden $w = 10\ 000$ Replikationen durchgeführt, womit die Länge des 95%-Konfidenzintervalls für die tatsächliche Power maximal $\pm$ 0.01 beträgt (für eine tatsächliche Power von $1 - \beta = 0.50$). Es wurde das einseitige Signifikanzniveau 0.088 zugrunde gelegt, der U-Test wurde in der asymptotischen Version angewandt. Tabelle 15 zeigt die geschätzte Power für Shift-Alternativen $\Delta = 0.0$ bis 4.0 (0.5). Der U-Test liefert

konsistent höhere Werte für die geschätzte Power verglichen mit dem *t*-Test, weshalb im Rahmen eines Prüfplan-Amendments festgelegt wurde, die Auswertung des zweiten Studienteils mit dem U-Test durchzuführen.

Tabelle 15: Geschätzte Power für den *t*- und den U-Test für Shift-Alternativen Δ. Fallzahlen $n_1 = n_2 = 100$, $\alpha = 0.088$, 10 000 Bootstrap-Ziehungen.

Δ	Geschätzte Power	
	t-Test	U-Test
0.0	0.086	0.086
0.5	0.203	0.239
1.0	0.377	0.434
1.5	0.593	0.645
2.0	0.780	0.815
2.5	0.906	0.920
3.0	0.970	0.972
3.5	0.992	0.991
4.0	0.999	0.998

5.2 Vergleich zwischen adaptivem und nicht-adaptivem Design

Im folgenden sollen die Charakteristika einer adaptiven Auswahl der Teststatistik auf der Basis einer Schätzung der Power für einige Anwendungssituationen exemplarisch untersucht werden. Hierzu betrachten wir das adaptive Zwei-Stufen-Design nach BAUER und KÖHNE (1994) mit $\alpha = 0.025$ (einseitig), $\alpha_0 = 1.0$ (d.h. keine vorzeitige Beendigung mit Beibehaltung der Null-Hypothese) und $\alpha_1 = c_\alpha = 0.00380$. In der Simulationsstudie wurden die Normalverteilung, die Exponentialverteilung und die Log-Normalverteilung zugrunde gelegt. Für jede dieser Verteilungen wurde die Null-Hypothese und die Lagealternativen $\Delta = 0.25, 0.45$ betrachtet. Neben dem Zwei-Stichproben *t*-Test wurden drei Tests aus der Klasse der linearen Rangtests in die Betrachtung einbezogen. Für das oben beschriebene Shift-Modell hat diese Klasse von Tests eine große Bedeutung, weil sie unter der Null-Hypothese das nominale Niveau für beliebige stetige Verteilungen einhalten (sog. Verteilungsfreiheit). Der zweite Grund liegt in der Tatsache begründet, dass für einige Repräsentanten aus dieser Klasse die Power bei normalverteilten Daten nicht wesentlich unter der des gleichmäßig besten *t*-Tests liegt, während bei Nicht-Vorliegen der Normalverteilung

die Power des *t*-Tests teilweise deutlich übertroffen wird. Aufgrund dieser beiden Aspekte sind lineare Rangtests insbesondere auch für die vorliegende Fragestellung interessante Alternativen.

Bezeichnet man mit $\underline{Z} = (Z_1, ..., Z_{n_1+n_2})$ den Vektor mit Komponenten $Z_i = 1$, falls die *i*-te Variable in der kombinierten und geordneten Stichprobe aus der Stichprobe 1 stammt und $Z_i = 0$ falls aus Stichprobe 2, so können die linearen Rangstatistiken für das Zwei-Stichproben-Problem in der folgenden Form geschrieben werden:

$$L_{n_1+n_2} = \sum_{i=1}^{n_1+n_2} s(i) \cdot Z_i$$

mit geeigneten Scores $s(i)$. Wir betrachten speziell folgende Repräsentanten aus dieser Klasse von Tests:

- U-Test (WILCOXON, 1945): $s(i) = i$
- Van der Waerden-Test (VAN DER WAERDEN, 1952/1953): $s(i) = \Phi^{-1}\left(\frac{i}{n_1 + n_2 + 1}\right)$, wobei Φ^{-1} die Quantilfunktion der Standardnormalverteilung bezeichnet.
- Median-Test (WESTENBERG, 1948): $s(i) = 0$ für $i \leq \frac{n_1 + n_2 + 1}{2}$ und $s(i) = 1$ für $i > \frac{n_1 + n_2 + 1}{2}$.

Aussagen zur lokalen Optimalität und zur asymptotischen relativen Effizienz linearer Rangtests finden sich zum Beispiel in BÜNING und TRENKLER (1994) und RANDLES und WOLFE (1979).

Wir werden bei den Berechnungen die Tests nicht in der exakten sondern in der asymptotischen Version verwenden; dies ist aufgrund der verwendeten Stichprobenumfänge gerechtfertigt und löst zumindest eines der Probleme bei den rechenintensiven Simulationen. In RANDLES und WOLFE (1979) wird gezeigt, dass unter gewissen Nebenbedingungen, die für die hier verwendeten Tests erfüllt sind, die Teststatistik

$$T_{n_1+n_2} = \frac{L_{n_1+n_2} - E(L_{n_1+n_2})}{\sqrt{Var(L_{n_1+n_2})}}$$

asymptotisch (d.h. für $\min(n_1, n_2) \to \infty$) standardnormalverteilt ist. Dabei ist

$$E(L_{n_1+n_2}) = \frac{n_1}{n_1 + n_2} \cdot \sum_{i=1}^{n_1+n_2} s(i)$$

und

$$Var(L_{n_1+n_2}) = \frac{n_1 \cdot n_2}{(n_1+n_2)^2 \cdot (n_1+n_2-1)} \cdot \left((n_1+n_2) \cdot \sum_{i=1}^{n_1+n_2} s(i)^2 - \left(\sum_{i=1}^{n_1+n} s(i) \right)^2 \right).$$

Es wurden die Szenarien betrachtet, dass bei der Zwischenauswertung der *t*-Test bzw. der U-Test verwendet wird (Teststatistik T), und dass für die Auswertung des zweiten Studienteils mit der Methode von COLLINGS und HAMILTON (1988) entweder der *t*-Test, der U-Test, der van der Waerden-Test oder der Median-Test ausgewählt wird (Teststatistik T^*). Zur Schätzung der Power der Prozedur mit adaptiver Auswahl der Teststatistik wurden in der Simulationsstudie für jede der betrachteten Verteilungskonstellationen und die beiden Teststatistiken T folgende Schritte $r = 2\ 500$ mal durchgeführt:

1. Schritt: Generieren der Daten des ersten Studienteils

Es werden zwei unabhängige Stichproben aus der Verteilung F vom Umfang $n_1 + n_2 = 2 \times 50 = 100$ erzeugt, die um Δ gegeneinander verschoben sind.

2. Schritt: Zwischenauswertung

Es wird ein einseitiger Test zum Niveau $\alpha_1 = 0.00380$ mit der Teststatistik T durchgeführt:

- $p_1^T \le \alpha_1$: Ablehnung der Null-Hypothese, Fortsetzung der Simulation mit Schritt 1.
- $\alpha_1 < p_1^T \le 1$: Nach dem Verfahren von COLLINGS und HAMILTON (1988) wird mit $w = 2500$ die Power zum Niveau $\alpha_2^T = c_\alpha / p_1^T$ für die Verteilung F mit Shift-Alternative Δ und den Stichprobenumfang $m_1 + m_2 = 2 \times 50 = 100$ geschätzt. Für den zweiten Studienteil wird die Teststatistik T^* mit der größten geschätzten Power ausgewählt.

3. Schritt: Generieren der Daten des zweiten Studienteils und Endauswertung

Es werden zwei unabhängige Stichproben aus der Verteilung F vom Umfang $2 \times 50 = 100$ erzeugt, die um Δ gegeneinander verschoben sind. Es wird ein einseitiger Test zum Niveau c_α / p_1^T mit der Teststatistik T^* durchgeführt:

- $p_2^{T^*} \le c_\alpha / p_1^T$: Ablehnung der Null-Hypothese.
- $p_2^{T^*} > c_\alpha / p_1^T$: Beibehaltung der Null-Hypothese.

Fortsetzung der Simulation mit Schritt 1.

Die Power der adaptiven Strategie bei Verwendung der Teststatistik T im ersten Studienteil wird geschätzt durch

$$\hat{\pi}(F, T, T^*; \Delta, \alpha) = \frac{\#\{\text{Ablehnung der Null} - \text{Hypothese}\}}{\#\{\text{Replikationen}\}}.$$

Die Power für die adaptive Strategie wird verglichen mit der

- Power $\hat{\pi}(F, T, T; \Delta, \alpha)$ des adaptiven Zwei-Stufen-Designs bei fester Wahl der Teststatistik T für beide Studienabschnitte und Stichprobenumfänge von jeweils 2×50 zum gleichen Gesamt-Niveau α und gleichen Werten für α_0 und α_1 wie bei der adaptiven Strategie.
- Power $\hat{\pi}(F, T; \Delta, \alpha)$ des „klassischen" Ein-Stufen-Design mit Teststatistik T und Stichprobenumfang 2×100 zum Niveau α.

In Tabelle 16 sind die geschätzten Ablehnraten für das Ein-Stufen-Design bei Auswertung mit Test *T (T*), das Zwei-Stufen-Design bei Auswertung beider Studienteile mit Test *T (T-T*) und das Zwei-Stufen-Design bei Auswertung des ersten Studienteils mit Test *T* und des zweiten Studienteils mit dem ausgewählten Test T^* *(T-*T^*) angegeben. Bei der durchgeführten Anzahl von Replikationen ($r = 2\ 500$) beträgt die maximale Länge des 95%-Konfidenzintervalls für die Power $\pm$ 0.02 (bei einer tatsächlichen Power von $1 - \beta = 0.50$).

Für die Normalverteilung wurde von mehreren Autoren (BAUER und KÖHNE, 1994; BANIK, KÖHNE und BAUER, 1996; WASSMER, 1997, 1998) mittels analytischer Berechnungen die Power des gleichmäßig besten Vorgehens (= Anwendung des *t*-Tests auf die Gesamtstichprobe) mit der Power bei Anwendung des *t*-Tests auf zwei Teilstichproben und Kombination mit dem Fisher-Produkttest verglichen. Es zeigte sich jeweils, dass der Power-Verlust nur gering ist. Wie man Tabelle 16 entnehmen kann, gilt diese Aussage auch, wenn man für das Ein-Stufen-Design oder den ersten Studienteil eines adaptiven Designs einen optimalen oder zufriedenstellenden Test gewählt hat und für den zweiten Teil einen Wechsel zu einem anderen Testverfahren zulässt. Auf der anderen Seite besitzt die adaptive Strategie ein großes Potential zu einem Gewinn an Power, wenn für die Zwischenauswertung ein Test festgelegt wurde, der für die vorliegende Verteilung schlechte Power-Eigenschaften besitzt. Das Verfahren „lernt" aus den Daten des ersten Studienteils und wählt für den zweiten Teil einen Test aus, der für diese Datenstruktur eine höhere Power aufweist. Dies führt zu einer teilweise erheblichen Erhöhung der Ablehnrate gegenüber den nicht-adaptiven Strategien *T* und *T-T* (siehe Ergebnisse in Tabelle 16 für den *t*-Test und die Exponential- und die Log-Normalverteilung).

Tabelle 16: Geschätzte Power für verschiedene Verteilungen, Shift-Alternativen Δ und folgende Strategien: Ein-Stufen-Design, Auswertung mit Test *T (T)*; Zwei-Stufen-Design, Auswertung beider Studienteile mit Test *T (T-T)*; Zwei-Stufen-Design, Auswertung erster Studienteil mit Test *T* und zweiter Studienteil mit ausgewähltem Test T^* *(T-*T^**)*. $\alpha = 0.025$ (einseitig), Stop-Grenzen für Zwei-Stufen-Design nach BAUER und KÖHNE (1994): $\alpha_0 = 1, \quad \alpha_1 = c_\alpha = 0.00380$; Fallzahlen: Ein-Stufen-Design 2×100, Zwei-Stufen-Design $2\times(2\times50)$; 2 500 Replikationen, 2 500 Bootstrap-Ziehungen für Strategie *(T-*T^**)*.

		N (0, 1)	
		T	
Δ	Strategie	*t*-Test	U-Test
0.0	*T*	0.026	0.024
	T-T	0.025	0.025
	T-T^*	0.028	0.025
0.25	*T*	0.422	0.401
	T-T	0.406	0.382
	T-T^*	0.390	0.378
0.45	*T*	0.890	0.878
	T-T	0.860	0.842
	T-T^*	0.842	0.839
		Exp (1)	
		T	
Δ	Strategie	*t*-Test	U-Test
0.0	*T*	0.024	0.027
	T-T	0.022	0.023
	T-T^*	0.023	0.026
0.25	*T*	0.450	0.784
	T-T	0.386	0.734
	T-T^*	0.633	0.762
0.45	*T*	0.891	0.996
	T-T	0.871	0.990
	T-T^*	0.967	0.993

Tabelle 16 (Fortsetzung):

		LN (0,1)	
		T	
Δ	Strategie	*t*-Test	U-Test
0.0	*T*	0.020	0.024
	T-*T*	0.021	0.028
	T-*T**	0.021	0.028
0.25	*T*	0.140	0.578
	T-*T*	0.152	0.542
	T-*T**	0.357	0.550
0.45	*T*	0.366	0.956
	T-*T*	0.388	0.945
	T-*T**	0.783	0.950

6. Zusammenfassung und Ausblick

Bei herkömmlichen Studiendesigns ist es notwendig, die Rahmenbedingungen der Durchführung, wie Fallzahl, Fragestellungen und statistische Testverfahren für die konfirmatorische Auswertung vor Studienbeginn festzulegen. Eine datenabhängige Veränderung der Design-Elemente im Studienverlauf ist dann nicht mehr möglich, ohne Gefahr zu laufen, die vorgegebene Wahrscheinlichkeit eines falsch-positiven Ergebnisses zu überschreiten. Angesichts der Tatsache, dass es eher die Regel als die Ausnahme ist, dass in der Planungsphase einer klinischen Studie die Vorinformationen, die für eine optimale Design-Spezifikation notwendig sind, gänzlich fehlen oder mit großer Unsicherheit behaftet sind, ist das Unbehagen der Therapieforschung an diesen starren Rahmenbedingungen herkömmlicher Studiendesigns verständlich.

Die in jüngster Zeit entwickelten adaptiven Verfahren beheben dieses Defizit. Sie folgen dem intuitiven Konzept, die verfügbaren Daten im Studienverlauf mit den ursprünglichen Planungsannahmen abzugleichen und, falls notwendig, das Design entsprechend anzupassen. Auf das breite Spektrum möglicher Design-Änderungen wurde in der Literatur bereits vielfach hingewiesen. Bislang gibt es aber nur wenige Untersuchungen zu der Frage, welche Entscheidungsregeln hierbei verwendet werden sollen und inwieweit die Freiheit einer flexiblen Design-Gestaltung auch tatsächlich zu einer größeren Effektivität führt. Dieser Weg wurde in der vorliegenden Arbeit für einige wichtige Anwendungssituationen beschritten.

Die in Kapitel 2 vorgestellten multiplen Testprozeduren sind die Voraussetzung dafür, dass auch Studien, in denen mehrere Fragestellungen untersucht werden, im adaptiven Design durchgeführt und inferenzstatistisch bearbeitet werden können. Diese Option ist von besonderem Interesse, weil eine präzise Planung umso schwieriger ist, je komplexer die Zielsetzung der Studie ist. Ein wichtiges Beispiel hierfür sind Dosis-Findungs-Studien, bei deren Planung das Wirksamkeitsprofil der untersuchten Substanz nur vermutet werden kann. Für derartige Studien wurde eine Strategie vorgeschlagen, mit der im adaptiven Zwei-Stufen-Design sowohl die explorative Untersuchung der Dosis-Wirkungs-Beziehung als auch der Wirksamkeitsnachweis in eine einzige Studie integriert werden kann. Dies ist inbesondere deshalb bemerkenswert, weil diese Ziele im herkömmlichen Ansatz nicht nur in verschiedenen Studien, sondern in unterschiedlichen Phasen der Arzneimittelentwicklung verfolgt werden, und das alternative Vorgehen eine wesentliche Erhöhung der Effizienz verspricht. Die Strategie besteht darin, zunächst mit einer Menge von aussichtsreichen Dosierungen zu beginnen, und im Rahmen der Zwischenauswertung die Dosis-Gruppe mit

dem günstigsten Nutzen-Risiko-Verhältnis auszuwählen. Falls notwendig kann die Studie dann mit einem zweiten Studienteil fortgesetzt werden, um die Wirksamkeit der selektierten Dosis nachzuweisen. Die entwickelten multiplen Testprozeduren bilden die Basis für die konfirmatorische Analyse dieses Studientyps. In Kapitel 3 wurden Methoden vorgestellt, mit denen unter einem Plateau-Modell für den Dosis-Wirkungs-Verlauf die minimale Dosierung mit der maximalen Wirksamkeit identifiziert werden kann. Mittels Simulationsuntersuchungen wurde gezeigt, dass durch Anwendung dieser Strategie im Rahmen adaptiver Zwei-Stufen-Designs eine erheblicher Gewinn an statistischer Power im Vergleich zum herkömmlichen Ein-Stufen-Ansatz erzielt werden kann.

Falls die tatsächliche Streuung der Zielvariablen von dem bei der Fallzahlplanung angenommenen Wert abweicht, so ist bei herkömmlichen Studiendesigns entweder der berechnete Stichprobenumfang unnötig groß, oder die Chance, das Studienziel zu erreichen, ist kleiner als vorgesehen. In Kapitel 4 wurden für normalverteilte Daten Methoden zur adaptiven Bestimmung des Stichprobenumfanges hergeleitet, die dieses Manko beseitigen. Dabei wurde neben den adaptiven Designs mit Zwischenauswertung auch die Variante betrachtet, dass die Daten im Studienverlauf ohne Kenntnis der Therapiegruppen-Zugehörigkeit („verblindet“) im Hinblick auf die bei der Planung angenommene Streuung der Zielgröße inspiziert werden. Während für die erstgenannten Designs die Einhaltung der Wahrscheinlichkeit eines Fehlers 1. Art aufgrund ihrer Konstruktion auch unter einer Fallzahladaption garantiert ist, erforderte dies für Designs ohne Zwischenauswertung die Entwicklung spezieller Verfahren. Weiterhin wurden Methoden zur verblindeten Varianzschätzung hergeleitet, die so konzipiert sind, dass sie für Studien mit einer beliebigen Anzahl von Behandlungsgruppen eingesetzt werden können. Die Untersuchung der Charakteristika der resultierenden Fallzahlen zeigte, dass für beide Design-Varianten durch die Option zur Fallzahl-Adaption im Mittel der tatsächlich notwendige Stichprobenumfang erreicht wird, unabhängig davon, ob die Planungsannahme korrekt ist oder nicht.

Die statistische Power eines Testverfahrens hängt unter anderem von der Verteilung der zu analysierenden Variablen ab. Da die Festlegung des Auswertungsverfahrens in der Planungsphase in der Regel auf einer unsicheren Verteilungsannahme erfolgt, ist die Möglichkeit zu einem Wechsel des Testverfahrens wünschenswert, wenn sich diese Annahme im Studienverlauf als falsch erweist. In der Arbeit wurde gezeigt, dass adaptive Designs diese Option unter Einhaltung der Wahrscheinlichkeit eines Fehlers 1. Art bieten. Es wurde ein Resampling-Verfahren vorgestellt, das aus einer Klasse von Testverfahren denjenigen Test für die Auswertung des zweiten Studienteils auswählt, der für die bei der Zwischenauswertung

beobachteten Daten die maximale Power erzielt. Die Ergebnisse von Simulationsuntersuchungen zeigten das gleiche Bild wie für die anderen Anwendungssituationen: Wenn die Planungsannahmen korrekt sind, ist der Power-Verlust des adaptiven Designs gegenüber dem dann optimalen Ein-Stufen-Design nur gering. Andererseits besitzen adaptive Designs im Falle einer Fehl-Spezifikationen bei der Planung einer klinischen Studie ein erhebliches Potential zu einer effizienteren Nutzung der verfügbaren Information, was sich in einer Erhöhung der statistischen Power und damit in einer Reduktion der notwendigen Patientenzahlen niederschlägt.

Die zukünftigen Entwicklungsmöglichkeiten im Bereich adaptiver Designs sind vielfältig. In Kapitel 3 der vorliegenden Arbeit wurde gezeigt, dass die Durchführung von Dosis-Findungs-Studien im adaptiven Design vielversprechende Vorteile aufweist. Gerade in dieser Anwendungssituation sind die in der Praxis angewendeten Entscheidungskriterien komplex, denn neben den Daten bezüglich Wirksamkeit und Verträglichkeit der aktuellen Studie gehen auch die Ergebnisse bereits abgeschlossener Studien in eine integrierte Bewertung ein. Durch eine Modellierung des Prozesses der Arzneimittelentwicklung und durch entsprechende optimale Entscheidungsregeln, die das jeweilige Vorwissen über die untersuchte Therapie berücksichtigen, könnte die Flexibilität adaptiver Designs noch effizienter genutzt werden.

Ein Resultat von Kapitel 4 dieser Arbeit war die Aussage, dass im Design mit interner Pilotstudie bei normalverteilten Daten, zwei Stichproben und bei Verwendung des *t*-Tests in der Auswertung die Wahrscheinlichkeit eines Fehlers 1. Art nicht überschritten wird, wenn die verblindete Fallzahladaption mit der adjustierten oder nicht-adjustierten Ein-Stichproben-Varianz erfolgt. Entsprechende Resultate für mehr als zwei Stichproben sind derzeit ebensowenig verfügbar wie für nicht-normalverteilte Daten, wie z.B. Raten. Bei adaptiven Designs mit Zwischenauswertung kann neben der geschätzten Varianz auch der beobachtete Behandlungsgruppen-Unterschied zur Fallzahladjustierung verwendet werden. Die gegenwärtigen Vorschläge hierzu sind nicht befriedigend, so dass die Entwicklung entsprechender Strategien ebenfalls ein lohnender Bereich für weitere Forschungsaktivitäten ist.

Bei allen bisherigen Vergleichen zwischen adaptiven und nicht-adaptiven Designs beschränkte sich die Adaption auf einen einzelnen Design-Aspekt. Es spricht vieles dafür, dass erst durch eine simultane Anwendung effizienter Entscheidungsregeln, die unterschiedliche Design-Elemente betreffen, das tatsächliche Potential adaptiver Designs ausgeschöpft werden wird.

7. Literaturverzeichnis

ARMITAGE, P. (1975). *Sequential Medical Trials*. Blackwell, Oxford.

BANIK, N., KÖHNE, K., BAUER, P. (1996). On the power of Fisher's combination test for two stage sampling in the presence of nuisance parameters. *Biometrical Journal* **38**, 25-37.

BARNDORFF-NIELSEN, O., COX, D.R. (1979). Edgeworth and saddlepoint approximations with statistical applications. *Journal of the Royal Statistical Society, Series B* **41**, 279-312.

BAUER, P. (1989). Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie* **4**, 130-148.

BAUER, P. (1991). Multiple testing in clinical trials. *Statistics in Medicine* **10**, 871-890.

BAUER, M., BAUER, P., BUDDE, M. (1998). A simulation program for adaptive two stage designs. *Computational Statistics & Data Analysis* **26**, 351-371.

BAUER, P., BRANNATH, W., POSCH, M. (2000). Flexible two-stage designs. Erscheint in *Methods of Information in Medicine*.

BAUER, P., KIESER, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* **18**, 1833-1848.

BAUER, P., KÖHNE, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 1029-1041. Correction in *Biometrics* **52**, 380.

BAUER, P., RÖHMEL, J. (1995). An adaptive method for establishing a dose response relationship. *Statistics in Medicine* **14**, 1595-1607.

BERGER, V.W. (2000). Pros and cons of permutation tests in clinical trials. *Statistics in Medicine* **19**, 1319-1328.

BETENSKY, R.A., TIERNEY, C. (1997). An examination of methods for sample size recalculation during an experiment. *Statistics in Medicine* **16**, 2587-2598.

BIRKETT, M.A., DAY, S.J. (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine* **13**, 2455-2463.

BISCHOFF, W., FIEGER, W., WULFERT, S. (1995). Minimax- and Γ-minimax estimation of a bounded normal mean under LINEX loss. *Statistics and Decisions* **13**, 287-298.

BISCHOFF, W., MILLER, F. (2000). Asymptotically optimal tests and optimal designs for testing the mean in regression models with applications to change-point problems. Erscheint in *Annals of the Institute of Statistical Mathematics*.

BOCK, J. (1998). *Bestimmung des Stichprobenumfangs für biologische Experimente und kontrollierte klinische Studien*. Oldenbourg-Verlag, München.

BRANNATH, W., POSCH, M., BAUER, P. (1999). Recursive combination tests. Submitted.

BROWNE, R.H. (1995). On the use of a pilot sample for sample size determination. *Statistics in Medicine* **14**, 1933-1940.

BÜNING, H. (1991). *Robuste und adaptive Tests*. De Gruyter Verlag, Berlin.

BÜNING, H., TRENKLER, G. (1994). *Nichtparametrische Statistische Methoden*. 2. Auflage. De Gruyter Verlag, Berlin.

COFFEY, C.S., MULLER, K.E. (1999). Exact test size and power of a gaussian error linear model for an internal pilot study. *Statistics in Medicine* **18**, 1199-1214.

COLLINGS, B.J., HAMILTON, M.A. (1986). Technical Report 9-8-86. Department of Mathematical Sciences, Montana State University Statistical Center, Bozeman.

COLLINGS, B.J., HAMILTON, M.A. (1988). Estimating the power of the two-sample Wilcoxon test for location shift. *Biometrics* **44**, 847-860.

CPMP Working Party on Efficacy of Medicinal Products (1995). Biostatistical methodology in clinical trials in applications for marketing authorizations for medicinal products. *Statistics in Medicine* **14**, 1659-1682.

CPMP Working Party on Efficacy of Medicinal Products (1997). Note for guidance on medicinal products in the treatment of Alzheimer´s disease. The European Agency for the Evaluation of Medicinal Products, London .

CUI, L., HUNG, H.M.J., WANG S.-J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**, 321-324.

DEMETS, D.L., WARE, J.H. (1980). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* **67**, 651-660.

DEMETS, D.L., LAN, K.K.G. (1984). An overview of sequential methods and their application in clinical trials. *Communications in Statistics – Theory and Methods* **13**, 2315-2338.

DENNE, J.S., JENNISON, C. (1999). Estimating the sample size for a *t*-test using an internal pilot. *Statistics in Medicine* **18**, 1575-1585.

DIENEL, A., KIESER, M. (1999). A double-blind, randomized, placebo-controlled, parallel group study of the efficacy and safety of St John´s Wort (*Hypericum*) extract 300 mg three times a day in patients with mild to moderate depression. Internal Report of the Interim Analysis. Dr. Willmar Schwabe Pharmaceuticals, Karlsruhe.

EDGINGTON, E.S. (1995). *Randomization Tests*. 2. Auflage. Marcel Dekker, New York.

EDWARDS, D. (1999). On model prespecification in confirmatory randomised studies. *Statistics in Medicine* **18**, 771-785.

FISHER, L.D. (1998). Self-designing clinical trials. *Statistics in Medicine* **17**, 1551-1562.

FISHER, R.A. (1932). *Statistical Methods for Research Workers*. 4. Auflage. Oliver & Boyd, London.

FRIEDE, T. (2000a). Approximate sample size formulas for one-way analysis of variance. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* **31**, 22-31.

FRIEDE, T. (2000b). Methoden zur Bestimmung der Fallzahl in klinischen Studien mit interner Pilotstudie. Dissertation zur Erlangung des Doctor scientiarum humanarum. Medizinische Fakultät, Universität Heidelberg.

FRIEDE, T., KIESER, M. (1999). Decision making in dose-response trials with adaptive two-stage designs. Submitted.

FRIEDE, T., KIESER, M. (2000a). A comparison of methods for adaptive sample size adjustment. Submitted.

FRIEDE, T., KIESER, M. (2000b). Unveiling the mystery of an EM-algorithm based procedure for blinded variance estimation. Submitted.

FRIEDE, T., MILLER, F., BISCHOFF, W., KIESER, M. (2000). A note on change point estimation in dose-response trials. Erscheint in *Computational Statistics & Data Analysis*.

FUNKE, K. (2000). Powervergleich bei adaptiven Testverfahren. Vortrag im Rahmen des Workshops „Adaptive Studiendesigns“, 25.-26.5.2000, Heidelberg.

FUNKE, K., WASSMER, G. (2000). Powervergleich bei adaptiven Testverfahren. Vortrag im Rahmen des 46. Biometrischen Kolloquiums der Deutschen Region der Internationalen Biometrischen Gesellschaft, 20.-23.3.2000, Rostock.

GOOD, P. (1994). *Permutation Tests*. Springer-Verlag, New York.

GOULD, A.L. (1995). Planning and revising the sample size for a trial. *Statistics in Medicine* **14**, 1039-1051.

GOULD, A.L. (1997). Issues in blinded sample size re-estimation. *Communications in Statistics – Simulation and Computation* **26**, 1229-1239.

GOULD, A.L., SHIH, W.J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics – Theory and Methods* **21**, 2833-2853.

GOULD, A.L., SHIH, W.J. (1998). Modifying the design of ongoing trials without unblinding. *Statistics in Medicine* **17**, 89-100.

HAMILTON, M. (1976). 048 HAMA, Hamilton Anxiety Scale. In: *ECDEU Assessment Manual for Psychopharmacology*, W. Guy (Hrsg.), 193-198. Rockville.

HAMILTON, M. (1986). The Hamilton rating scale for depression. In: *Assessment of Depression*, N. Sartorius, T.A. Ban (Hrsg.), 143-152. Springer-Verlag, Berlin.

HAMILTON, M.A., COLLINGS, B.J. (1991). Determining the appropriate sample size for nonparametric tests for location shift. *Technometrics* **33**, 327-337.

HAYBITTLE, J.L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology* **44**, 793-797.

HEEREN, T., D´AGOSTINO, R. (1987). Robustness of the two independent samples *t*-test when applied to ordinal scaled data. *Statistics in Medicine* **6**, 79-90.

HINKLEY, D.V. (1969). Inference about the intersection in two-phase regression. *Biometrika* **56**, 495-504.

HOCHBERG, Y., TAMHANE, A. (1987). *Multiple Comparison Procedures*. Wiley, New York.

HOGG, R.V., FISHER, D.M., RANDLES, R.H. (1975). A two sample adaptive distribution-free test. *Journal of the American Statistical Association* **70**, 656-661.

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65-70.

HOMMEL, G. (2000). Hypothesenmodifikation nach Interimanalysen - theoretische und praktische Probleme. Vortrag im Rahmen des Workshops "Adaptive Studiendesigns", 25.-26.5.2000, Heidelberg.

HWANG, I.K., SHIH, W.J., DECANI, J.S. (1990). Group sequential designs using a family of Type I error probability spending functions. *Statistics in Medicine* **9**, 1439-1445.

ICH (1994). ICH Harmonized Tripartite Guideline E4: Dose-Response Information to Support Drug Registration. ICH Technical Coordination, London.

ICH (1999). ICH Harmonized Tripartite Guideline E9: Statistical Principles for Clinical Trials. *Statistics in Medicine* **18**, 1905-1942.

JENNISON, C., TURNBULL, B. (1999). *Group Sequential Methods with Applications to Clinical Trials.* Chapman & Hall, London.

JOHNSON, N.L., KOTZ, S. (1970). *Continuous Univariate Distributions – 2*. Houghton Mifflin, Boston.

JONCKHEERE, A.R. (1954). A distribution-free *k*-sample test against ordered alternatives. *Biometrika* **41**, 133-145.

KIESER, M. (1999). Report of the planned blinded data review for sample size re-estimation. Randomized, double-blind, placebo-controlled multicenter trial to demonstrate the clinical efficacy and safety of two different doses of an antidementia drug in patients suffering from

Dementia of the Alzheimer´s Type according to DSM-IV and NINCDS/ADRDA criteria. Interner Bericht Dr. Willmar Schwabe GmbH & Co., Karlsruhe.

KIESER, M., BAUER, P., LEHMACHER, W. (1999). Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal* **41**, 261-277.

KIESER, M., FRIEDE, T. (2000a). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine* **19**, 901-911.

KIESER, M., FRIEDE, T. (2000b). Blinded sample size reestimation in multiarmed clinical trials. *Drug Information Journal* **34**, 455-460.

KIESER, M., HAUSCHKE, D. (1999). Approximate sample sizes for testing hypotheses about the ratio and difference of two means. *Journal of Biopharmaceutical Statistics* **9**, 641-650.

KIESER, M., HAUSCHKE, D. (2000). Statistical methods for demonstrating equivalence in crossover trials based on the ratio of two location parameters. *Drug Information Journal* **34**, 563-568.

KIESER, M., KÖPCKE, W. (1998). Gruppensequentielle Verfahren. In: *Verfahrensbibliothek - Versuchsplanung und -auswertung, Band II*, D. Rasch, G. Herrendörfer, J. Bock, N. Victor, V. Guiard (Hrsg.), 724-738. Oldenbourg Verlag, München.

KIESER, M., LEHMACHER, W. (1995). Multiples Testen bei klinischen Prüfungen mit Zwischenauswertungen und a-priori geordneten Hypothesen. In: *Proceedings der 40. Jahrestagung der GMDS*, 162-165. MMV Medizin-Verlag, München.

KIESER, M., REITMEIR, P., WASSMER, G. (1995). Test procedures for clinical trials with multiple endpoints. In: *Biometrie in der chemisch-pharmazeutischen Industrie*, Volume 6, J. Vollmar (Hrsg.), 40-59. Fischer Verlag, Stuttgart.

KIESER, M., WASSMER, G. (1996). On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. *Biometrical Journal* **38**, 941-949.

KROPF, S., HOMMEL, G., SCHMIDT, U., BRICKWEDEL, J., JEPSEN, M.S. (2000). Multiple comparisons of treatments with stable multivariate tests in a two-stage adaptive design, including a test for non-inferiority. Submitted

LAAKMANN, G., SCHÜLE, C., BAGHAI, T., KIESER, M. (1998). St. John's wort in mild to moderate depression: The relevance of hyperforin for clinical efficacy. *Pharmacopsychiatry* **31** (S1), 54-59.

LAN, K.K.G., DEMETS, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659-663.

LANG, T., AUTERITH, A., BAUER, P. (2000). Trendtests with adaptive scoring. Submitted.

LEBER, P. (1990). Guidelines for the clinical evaluation of antidementia drugs. First Draft. US Food and Drug Administration, Rockville.

LEHMACHER, W., KIESER, M., HOTHORN, L.A. (2000). Sequential and multiple testing for dose-response analysis. *Drug Information Journal* **34**, 591-597.

LEHMACHER, W., WASSMER, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics* **55**, 1286-1290.

LEHMACHER, W., WASSMER, G., REITMEIR, P. (1991). Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* **47**, 511-521.

LEHMANN, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

LIU, J.-P., CHOW, S.-C. (1992). Sample size determination for the two one-sided tests procedure in bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics* **20**, 101-104.

MAHONEY, M., MAGEL, R. (1996). Estimation of the power of the Kruskal-Wallis test. *Biometrical Journal* **38**, 613-630.

MALSCH, U., KIESER, M. (2000). Efficacy of Kava-Kava special extract WS 1490 in the treatment of patients with nervous anxiety, tension and restlessness states of non-psychotic origin after pre-treatment with benzodiazepines. Submitted.

MARCUS, R., PERITZ, E., GABRIEL, K.R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655-660.

MAURER, W., HOTHORN, L.A., LEHMACHER, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In: *Biometrie in der chemisch-pharmazeutischen Industrie*, Volume 6, J. Vollmar (Hrsg.), 3-18. Fischer, Stuttgart.

MITFESSEL, A., ERXLEBEN, M., SCHULZE, B., KIESER, M. (1999). Efficacy, safety and acceptance of Budesonide admininstered via a new dry powder inhaler or CFC mitered dose inhaler in patients with mild to moderate asthma. *European Respiratory Journal*, **14**, 105s (Abstract).

MOSHMAN, J. (1958). A method for selecting the size of the initial sample in Stein's two sample procedure. *Annals of Mathematical Statistics* **29**, 1271-1275.

MÜLLER, H.-H., SCHÄFER, H. (1999a). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. Submitted.

MÜLLER, H.-H., SCHÄFER, H. (1999b). Changing a design during the course of an experiment. Submitted.

NATHOFF, I.L., ATTWOOD, M.A., EICHLER, D.A., KOGLER, P., KLEINBLOESEM, C.H., VAN BRUMMELEN, P. (1990). Cilazapril. *Cardiovascular Drug Reviews* **8**, 1-24.

O'BRIEN, P.C., FLEMING, T.R. (1979). A multiple test procedure for clinical trials. *Biometrics* **35**, 549-555.

OELLRICH, S., FREISCHLÄGER, F., BENNER, A., KIESER, M. (1997). Sample size determination on survival time data - a review. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* **28**, 64-85.

ORTSEIFEN, C., BRUCKNER, T., BURKE, M., KIESER, M. (1997). An overview of software tools for sample size determination. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* **28**, 91-118.

PAMPALLONA, S., TSIATIS, A.A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favour of the null-hypothesis. *Journal of Statistical Planning and Inference* **42**, 19-35.

PETO, R., PIKE, P., ARMITAGE, P., BRESLOW, N.E., COX, D.R., HOWARD, S.V., MANTEL, N., MCPHERSON, K., PETO, J., SMITH, P.G. (1976). Design and analysis of randomised clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer* **35**, 585-611.

POCOCK, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191-199.

POSCH, M., BAUER, P. (1999). Adaptive two stage designs and the conditional error function. *Biometrical Journal* **41**, 689-696.

POSCH, M., BAUER, P. (2000). Interim analysis and sample size reassessment. Erscheint in *Biometrics*.

PROSCHAN, M.A., HUNSBERGER, S.A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315-1324.

PROSCHAN, M.A., WITTES, J. (2000). An improved double sampling procedure based on the variance. Erscheint in *Biometrics*.

RANDLES, R.H., WOLFE, D.A. (1979). *Introduction to the Theory of Nonparametric Statistics*. Wiley, New York.

REITMEIR, P., WASSMER, G. (1996). One-sided multiple endpoint testing in two-sample comparisons. *Communications in Statistics - Simulation and Computation* **25**, 99-119.

ROEBRUCK, P., ELZE, M., HAUSCHKE, D., LEVERKUS, F., KIESER, M. (1997). Literaturübersicht zur Fallzahlplanung für Äquivalenzprobleme. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie* **28**, 51-63.

ROSEN, W.G, MOHS, R.C., DAVIS, K.L. (1984). A new rating scale for Alzheimer´s disease. *American Journal of Psychiatry* **141**, 1345-1364.

RUBERG, S.J. (1995a). Dose response studies. I. Some design considerations. *Journal of Biopharmaceutical Statistics* **5**, 1-14.

RUBERG, S.J. (1995b). Dose response studies. II. Analysis and interpretation. *Journal of Biopharmaceutical Statistics* **5**, 15-42.

SANDVIK, L., ERIKSSEN, J., MOWINCKEL, P., RODLAND, E.A. (1996). A method for determining the size of internal pilot studies. *Statistics in Medicine* **15**, 1587-1590.

SCHNEIDER, L.S., OLIN, J.T., DOODY, R.S., CLARK, C.M., MORRIS, J.C., REISBERG, B., SCHMITT, F.A., GRUNDMAN, M., THOMAS, R.G., FERRIS, S.H., AND THE ALZHEIMER´S DISEASE COOPERATIVE STUDY (1997). Validity and reliability of the Alzheimer´s Disease Cooperative Study-Clinical Global Impression of Change. *Alzheimer´s Disease and Associated Disorders* **11** (Suppl. 2), S22-S32.

SEELBINDER, B.M. (1953). On Stein´s two-stage sampling scheme. *Annals of Mathematical Statistics* **24**, 640-649.

SHEINER, L.B. (1997). Learning versus confirming in clinical drug development. *Clinical Pharmacology and Therapeutics* **61**, 275-291.

SHEN, Y., FISHER, L.D. (1999). Statistical inference for self-designing clinical trials with a one sided hypothesis. *Biometrics* **55**, 190-197.

SHIH, W.J. (1992). Sample size reestimation in clinical trials. In: *Biopharmaceutical Sequential Statistical Applications*, K. Peace (Hrsg.), 285-301, Marcel Dekker, New York.

SHIH, W.J., GOULD A.L (1995). Re-evaluating design specifications of longitudinal clinical trials without unblinding when the key response is rate of change. *Statistics in Medicine* **14**, 2239-2248.

SHIH, W.J., LONG, J. (1998). Blinded sample size re-estimation with unequal variances and center effects in clinical trials. *Communications in Statistics - Theory and Methods* **27**, 395-408.

SINGER, J. (1999). Letter to the editor: A method for determining the size of internal pilot studies. *Statistics in Medicine* **18**, 1151-1153.

STEIN, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* **16**, 243-258.

TANG, D.-I., GELLER, N. (1999). Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics* **55**, 1188-1192.

TROENDLE, J.F. (1999). Approximating the power of Wilcoxon´s rank-sum test against shift alternatives. *Statistics in Medicine* **18**, 2763-2773.

VARIAN, H.R. (1975). A Bayesian approach to real estate assessment. In: *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage,* S.E. Fienberg, A. Zellner (Hrsg.), 195-208. North-Holland, Amsterdam.

WAERDEN, B.L. VAN DER (1952/1953). Order tests for the two-sample problem and their power. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen* (Indagationes Mathematicae 14), 453-458 und 56 (Indagationes Mathematicae 15), 303-316. Errata: Ibid. (1953), 80.

WALD, A. (1947). *Sequential Analysis*. Wiley, New York.

WASSMER, G. (1997). A technical note on the power determination for Fisher´s combination test. *Biometrical Journal* **39**, 831-838.

WASSMER, G. (1998). A comparison of two methods for adaptive interim analyses in clinical trials. *Biometrics* **54**, 696-705.

WASSMER, G. (1999a). Multistage adaptive test procedures based on Fisher´s product criterion. *Biometrical Journal* **41**, 279-293.

WASSMER, G. (1999b). *Statistische Testverfahren für gruppensequentielle und adaptive Pläne in klinischen Studien.* Alexander Mönch-Verlag, Köln.

WASSMER, G., REITMEIR, P., KIESER, M., LEHMACHER, W. (1999). Procedures for testing multiple endpoints in clinical trials: an overview. *Journal of Statistical Planning and Inference* **82**, 69-81.

WESTENBERG, J. (1948). Significance test for median and interquartile range in samples from continuous populations of any form. *Proceedings Koninklijke Nederlandse Akademie van Wetenshappen* **51,** 252-261.

WILCOXON, F. (1945). Individual comparisons by ranking methods. *Biometrics* **1**, 80-83.

WITTES, J., BRITTAIN, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* **9**, 65-72.

WITTES, J., SCHABENBERGER, O., ZUCKER, D., BRITTAIN, E., PROSCHAN, M. (1999). Internal pilot studies I: type I error rate of the naive *t*-test. *Statistics in Medicine* **18**, 3481-3491.

ZERSSEN, D. VON (unter Mitarbeit von D.-M. KOELLER) (1976). *Klinische Selbstbeurteilungs-Skalen (KSb-S) aus dem Münchener Psychiatrischen Informationssystem (PSYCHIS München).*

Allgemeiner Teil. Beltz Test, Weinheim.

ZUCKER, D.M., WITTES, J.T., SCHABENBERGER, O., BRITTAIN, E. (1999). Internal pilot studies II: comparison of various procedures. *Statistics in Medicine* **18**, 3493-3509.

Danksagung

Mein Dank gilt Herrn Prof. Victor, der mich bei der Erstellung dieser Arbeit sehr unterstützt hat und zu ihrem Gelingen beigetragen hat. Zu großem Dank bin ich Herrn Dipl.-Math. Tim Friede verpflichtet. Die gemeinsame Arbeit im Rahmen des DFG-Projektes „Methoden zur adaptiven Fallzahlplanung in klinischen Studien“ war eine wesentliche Quelle der Motivation und hat zu vielen fruchtbaren Diskussionen und Ideen geführt. Dies war der kreative und produktive Nährboden für große Teile dieser Arbeit. Sehr viel zu verdanken habe ich auch Herrn Prof. Bauer und Herrn Prof. Lehmacher, die meinen wissenschaftlichen Werdegang in den letzten Jahren nachhaltig geprägt haben, und die mich stets ermunterten, Neues zu denken und Angedachtes aufzuschreiben.

Vor allem aber danke ich meiner Frau Yvonne für ihre Liebe und Geduld in der Zeit der Entstehung dieser Schrift. Ohne ihr Verständnis und ihre Unterstützung wäre die Arbeit nicht zustande gekommen.