Medizinische Fakultät
der
Universität Duisburg-Essen

Aus dem Institut für
Medizinische Informatik, Biometrie und Epidemiologie

Varianzunterschiede in klinischen Studien

I n a u g u r a l - D i s s e r t a t i o n
zur
Erlangung des Doktorgrades
der Naturwissenschaften in der Medizin
durch die Medizinische Fakultät
der Universität Duisburg-Essen

Vorgelegt von
Christian Lösch
aus Lörrach
2010

Dekan:	Herr Univ.-Prof. Dr. med. M. Forsting
1. Gutachter:	Herr Prof. Dr. rer. nat. M. Neuhäuser
2. Gutachter:	Herr Univ.-Prof. Dr. rer. nat. K.-H. Jöckel
Tag der mündlichen Prüfung:	17. November 2010

Publikationen

Zeitschriftenbeiträge

Lösch C, Neuhäuser M. The statistical analysis of a clinical trial when a protocol amendment changed the inclusion criteria. BMC Medical Research Methodology 2008 Apr 8;8.

Tagungsabstracts

Lösch C, Neuhäuser M. Varianzänderungen nach Modifikation der Einschlusskriterien in klinischen Studien und ein Kombinationstest zur Auswertung der Gesamtstudie, Vortrag, GMDS-Jahrestagung 2006

Lösch C, Neuhäuser M. Vergleich mehrerer Prozeduren zur wiederholten verblindeten Fallzahlrekalkulation in klinischen Studien, Vortrag, GMDS-Jahrestagung 2009

Inhaltsverzeichnis

1 Einleitung

1.1 Die generelle Fragestellung

In dieser Arbeit sollen drei Aspekte zur Varianz medizinischer Zielgrößen klinischer Studien untersucht werden: Erstens soll die mögliche Größenordnung von Varianzänderungen durch Analyse von real durchgeführten Studien abgeschätzt werden. Zweitens werden neue Möglichkeiten der Erkennung und der Reaktion auf eine Fehlspezifikation der Varianz zur Planungszeit beleuchtet und schließlich drittens wird ein alternatives Testverfahren im Falle von unterschiedlichen Varianzen über die Zeit analysiert.

Da bei der Planung einer klinischen Studie generell eine Schätzung der benötigten Fallzahl für den Test der primären Hypothese durchgeführt wird, ist die Frage nach der Varianz von erheblicher Bedeutung. Die Schätzung basiert dabei auf mehreren Informationen, wie sie z.B. in [1] zusammengefasst sind:

$$\text{Fallzahl} = \text{Varianz} \cdot f(\text{Fehlerraten})/(\text{minimaler relevanter Effekt})^2,$$

wobei f eine vom statistischen Verfahren abhängige Funktion der Fehler 1. und 2. Art ist, die man sich bei der Planung vorgibt. Im Folgenden wird häufig von der Gegenwahrscheinlichkeit des Fehlers 2. Art die Rede sein, die auch mit *Trennschärfe*, *Güte* oder *Power* des Tests bezeichnet wird und die für den medizinisch Forschenden von hohem Interesse ist.

Bei Überlegenheitsstudien besteht die Überzeugung, dass eine neue Behandlung hinsichtlich der untersuchten Zielgröße um mindestens um einen medizinisch relevanten Unterschied besser ist als die Standardtherapie. Dieser Unterschied ist ebenfalls in der obigen Fallzahlformel als *minimaler relevanter Effekt* enthalten. Durch seine Festlegung bestimmt man, wie klein ein medizinisch interessanter Unterschied sein kann, den der statistische Test noch mit hoher Güte aufdecken können soll. Die Fallzahlformel lässt auch sofort erkennen, wie die Varianz und der minimale relevante Effekt die Fallzahl beeinflussen: Bei ansonsten gleichen Parametern führt eine höhere Varianzannahme zu einer höheren Fallzahl, beim relevanten Effekt ist es genau umgekehrt: je größer der vermutete Effekt, desto geringer darf die benötigte Fallzahl sein.

An diesem Punkt wird eine Problematik bei der Planung klinischer Studien deutlich: Während die beiden Fehler per Konvention festgelegt sind und der minimale relevante Effekt durch medizinische Überlegungen bestimmt werden kann, ist die Größe der Varianz in der noch nicht durchgeführten Studie zuerst einmal unbekannt und es können nur Annahmen getroffen werden. Hierzu können verschiedene Quellen wie z.B. Angaben aus der Literatur oder Vorwissen aus früheren Studien dienen. Es besteht jedoch die Möglichkeit, dass die

Schätzungen aus diesen Quellen nicht auf die aktuelle Studie zutreffen, d.h. die Fallzahlplanung liefert in der Konsequenz ein inadäquates Ergebnis. Eine überschätzte Varianz bewirkt, dass mehr Patienten als nötig in die Studie eingeschlossen werden. Eine unterschätzte Varianz führt dagegen zu einer zu kleinen Fallzahl und somit zu einem Fehler 2. Art, der größer ist als bei der Planung festgesetzt.

Beides stellt nicht zuletzt ein ethisches Problem dar, denn im ersten Fall werden mehr Patienten als nötig einer experimentellen Behandlung ausgesetzt, im zweiten Fall werden zwar wenige Patienten eingeschlossen, jedoch werden ihre Daten vermutlich nicht die gewünschte Differenz aufdecken können. Die Studie würde somit ihren Zweck nicht erfüllen. Des Weiteren stellt eine zu groß angelegte Studie übermäßige Anforderungen an Budget und Zeit.

Außer der beschriebenen Fehleinschätzung der Varianz im Planungsstadium, kann es auch aus anderen Gründen zu einer Dynamik der Varianz während des Studienverlaufes kommen. Zum Teil liegen die Gründe in der Durchführung der Studie, aber auch externe Faktoren können die Ursache sein. Ein Beispiel für die erste Quelle ist das Verändern des Studienprotokolls während der Studie, was in der Regel ein Amendment nötig macht. Ein Beispiel für die zweite Quelle sind sich verbessernde Mess- und/oder Behandlungsmethoden bei Studien mit langer Laufzeit.

1.2 Reale Studien

Der erste Teil der vorliegenden Arbeit untersucht, ob und wie stark sich die Varianz von wichtigen Zielgrößen bereits durchgeführter realer klinischer Studien während der Studiendauer geändert hat und welche Faktoren hierbei eine Rolle gespielt haben könnten. Dies ist notwendig gewesen, da in der Literatur eine derartige Untersuchung nicht gefunden wurde.

Darüber hinaus wird in dieser Arbeit die Rekrutierung bei der Frage nach Ursachen von Varianzungleichheit als nahe liegender Faktor angesehen, da Ungleichheiten oder Probleme bei der Rekrutierung leicht zu unterschiedlicher Variabilität oder sogar Behandlungseffekten führen können.

Zur Rekrutierung in klinischen Studien gibt es bereits viele Veröffentlichungen, darunter z.B. die Übersichtsarbeit von Lovato [2], welche von über 4000 Publikationen zwischen 1987 und 1995 berichtet. Lovato spricht auch das Problem an, die benötigte Anzahl von Studienteilnehmern in der geplanten Zeit einzuschließen. Korrekturen einer schleppenden Rekrutierung bergen die Gefahr, dass die Variabilität der primären Endpunkte erhöht wird. Dieses ist besonders dann zu erwarten, wenn die Ein- und/ oder Ausschlusskriterien gelockert werden, wie es z.B. in [3] und [4] genannt wird.

Im letzteren Artikel wird auch ein Grund für die Notwendigkeit der Änderung der Kriterien genannt: Bei der Planung der Studie wird normalerweise bei den teilnehmenden Studienzentren eine voraussichtliche Anzahl an Teilnehmern erfragt. Während diese Angabe aus der Erfahrung meist hinreichend genau beantwortet werden kann, wird dagegen die Anzahl der Patienten oft stark unterschätzt, die gegen die Einschlusskriterien verstoßen oder die eine Teilnahme ablehnen.

In [5] geht es darum, dass bereits die Art der Rekrutierung Einfluss auf die Variabilität in

der Gruppe der Studienteilnehmer haben kann. In der betreffenden Studie wurden Patienten mit chronischen Schmerzen einerseits durch „Laien-Medien“ und andererseits in speziellen Schmerzkliniken rekrutiert. Es zeigte sich, dass die beiden Gruppen sich hinsichtlich wichtiger prognostischer Faktoren unterschieden. Diese Veröffentlichung legt damit ebenso nahe, dass die Chronizität der Krankheit über den Weg der Rekrutierung eine Rolle für die Variabilität spielen kann. Ebenso können sich die Studienteilnehmer je nach Rekrutierungsweg unterscheiden [6]. Des Weiteren kann eine fluktuierende Rekrutierungsgeschwindigkeit u.a. dazu führen, dass Patientencharakteristika nicht mehr einheitlich sind [7], was wiederum zu einem Varianzanstieg bei der primären Zielgröße führen kann.

1.3 Verblindete Fallzahladaption

Im zweiten Teil dieser Arbeit werden mehrere neue Möglichkeiten dargestellt, mit denen die Varianz während des Studienverlaufes gemessen und/oder bewertet werden kann. Ziel dieser Vorgehensweise ist, Fehlspezifikationen bei der initialen Fallzahlplanung oder zeitliche Verläufe der Varianz, wie sie im ersten Teil bei realen Studien vorgefunden wurden, zu identifizieren, um entsprechend darauf reagieren zu können.

In der Literatur der letzten 20 Jahre gibt es einige Arbeiten, die sich mit dem Thema der Flexibilisierung von klinischen Studien befassen. Eine Übersicht und Kategorisierung dieser findet sich in [1], der zufolge zwischen Fallzahlrekalkulation und Zwischenanalysen zur Wirksamkeit unterschieden werden kann: Während bei der ersten Methodik die Variabilität der Zielgröße die entscheidende Rolle spielt, um die Fallzahl der Trennschärfe gemäß anzupassen, wird bei der zweiten die Effektstärke (d.h. z.B. ein Maß des Gruppenunterschieds) zur Veränderung von Studienparametern verwendet. Solche Studienparameter können die Fallzahl, aber auch allgemeinere wie der zu verwendende primäre Endpunkt bei der Endauswertung selbst sein.

Weiterhin besteht die grundlegende Idee bei der Fallzahlrekalkulation darin, die Studie *weiterzuführen*, denn eine klassische Anwendung birgt die Erhöhung der Fallzahl bei zu großer Variabilität im Verlauf der Studie. Im Gegensatz dazu zielt der Einsatz von Interimanalysen mit Entblindung und Bestimmung des Effektes eher auf ein vorzeitiges *Beenden* der Studie ab, wenn schon zeitig genügend „Beweiskraft“ für die Alternativhypothese vorhanden ist. Da beide Methodiken prinzipiell unabhängig voneinander sind, können sie sogar gleichzeitig in einer Studie Anwendung finden.

Die im zweiten Teil untersuchten Prozeduren sind der Fallzahlrekalkulation zuzuordnen. Diese unterteilen sich in drei Untergruppen, die sich im Grad der Entblindung der Daten wie folgt unterscheiden [1]: vollständige Entblindung, partielle Entblindung und keine Entblindung. Partielle Entblindung bedeutet, dass die Gruppenzuordnung zwar bekannt ist, welche Behandlung in der jeweiligen Gruppe durchgeführt wird, aber unbekannt ist.

Zu den Methoden, die mit partieller Entblindung auskommen, gehören Ansätze, die unter den Stichworten *conditional power* oder *stochastic curtailment* zusammengefasst werden können. Dabei reicht der beobachtete Unterschied zum Zeitpunkt der Interimanalyse, um die Fallzahl anzupassen. Da hierfür Informationen über die Teststatistik der Endauswertung ver-

wendet werden, kommt es zu einer Erhöhung des Fehlers 1. Art, auf die mit einer Veränderung der kritischen Grenzen für die Teststatistik reagiert wird.

Die Verfahren, die mit vollständiger Entblindung arbeiten, basieren zumeist auf einer frühen Arbeit von Stein [8], bei der die Varianz aus einem ersten Teil von normalverteilten Daten entblindet geschätzt wurde, um die benötigte Fallzahl zu bestimmen. Nachdem die so bestimmte Fallzahl erreicht wurde, wurde ein modifizierter t-Test durchgeführt, dessen Varianzterm die Daten aus der Zwischenstichprobe nicht verwendete. In [9] wurde diese Methode unter dem Namen der *internen Pilotstudie* u.a. dahingehend modifiziert, dass die *gesamte* Datenbasis in die Teststatistik einfloss. Außerdem wies diese Methode noch die Eigenschaft der *Restriktion* auf, bei der die endgültige Fallzahl nie unter der initial geplanten liegen durfte. Eine Optimierung des Verfahrens von Wittes und Brittain wurde von Denne und Jennison [10] vorgeschlagen.

In einer Reihe von Arbeiten von Gould und Shih [11, 12, 13, 14] wurde versucht, der Forderung von regulatorischer Seite gerecht zu werden, eine zwischenzeitliche Entblindung im Verlauf einer Studie zu vermeiden, da Verblindung in Zusammenhang mit Randomisierung zu den wichtigsten Maßnahmen gegen Bias gehört [15]. Eine zwischenzeitliche Entblindung erfordert des Weiteren, dass ein unabhängiges Data Safety Monitoring Committee (DSMC) eingerichtet werden muss, was den organisatorischen Aufwand der Studie erhöht.

Im ersten der o.g. Artikel [11] wurde die Prozedur von Stein [8] auf binäre Endpunkte übertragen. Die vorgeschlagene Verfahrensweise kommt ohne das Verwerfen der Daten aus der Pilotstichprobe aus, was durch die Verwendung des Gesamtanteils an Ereignissen (*pooled event rate, PER*) erreicht wird. Das Verfahren wurde zudem noch mit Hilfe von Methoden der Bayesschen Statistik sublimiert.

In [12] haben die Autoren eine Methode vorgeschlagen, mit der sich im Falle normalverteilter Zielgrößen die Fallzahl unter Aufrechterhaltung der Verblindung während des Studienverlaufes korrigieren lässt. Auch sie basiert letztlich auf der Idee von Stein und verwendet zur verblindeten Schätzung der Intragruppen-Varianz einerseits eine einfache Korrektur, andererseits wird ein EM-Algorithmus verwendet, dessen Motivation aus der Beobachtung kommt, dass beim Vereinigen der beiden Mengen an Daten der Behandlungsgruppen eine Mischung aus Normalverteilungen entsteht. Bei der Parameterschätzung für Mischverteilungen wiederum wird häufig der EM-Algorithmus verwendet. In der Folgezeit wurde der Algorithmus auf andere Statistiken und Studiendesigns übertragen.

Friede und Kieser zeigten [16] drei Gründe auf, warum der EM-Algorithmus zur Fallzahlrekalkulation nicht verwendet werden sollte. Erstens hängen die Schätzwerte die der Algorithmus liefert, von seinen Initialwerten ab. Zweitens ist das Konvergenzkriterium des Verfahrens nicht geeignet, um eine Stabilisierung zu erkennen. Drittens werden keine Informationen aus der Randomisierungsprozedur verwendet, so dass implizit von einfacher Randomisierung ausgegangen wird, während heutige klinische Studien aber meist eine Blockrandomisierung verwenden. In [17] wurde dies jedoch relativiert, denn Waksman konnte eine kleine Ungenauigkeit in der Formulierung des Algorithmus in [11] finden und korrigieren. Nach der Korrektur wurde durch den Algorithmus zwar der Maximum-Likelihood-Schätzer der Intragruppen-Varianz gefunden, aber er ist verzerrt und hat eine große Streubreite. Der Autor bewertet den ge-

poolten Schätzer als den besseren, obwohl dieser aufgrund der möglichen Mittelwertdifferenz einen systematisch zu großen Wert für die Intragruppen-Varianz liefert.

Der Ansatz aus [9] wurde in [18] hinsichtlich der Inflation des Fehlers 1. Art untersucht und eine obere Abschätzung für diese hergeleitet. Mit dieser Abschätzung ist es möglich, eine Studie mit interner Pilotphase durchzuführen, ohne dass es zu einem erhöhten Fehler 1. Art kommt, was durch eine entsprechend niedrigere Vorgabe für den Fehler erreicht wird.

Der Artikel [19] von Kieser und Friede bietet eine Alternative zum Verfahren von Gould und Shih. Die Grundidee ist hier die Verwendung des gepoolten Schätzers, den man optional mit einem Korrekturterm versehen kann, der der o.g. Überschätzung durch eine Mittelwertdifferenz ungleich Null begegnen soll. Man erhält dadurch den *adjustierten* Schätzer. Dieser ist in approximativer Form bereits in [12] zu finden und wird dort *simple adjustment* genannt. In exakter Form ist er auch im Appendix zu [20] zu finden. Die Autoren zeigten eine Rekalkulationsprozedur für die t-Test-Situation, die sowohl die Erhaltung der Verblindung garantiert, als auch den Fehler 1. Art zu berechnen und/oder zu kontrollieren erlaubt.

Friede und Kieser [21] kommen zum Schluss, dass der unadjustierte Varianzschätzer in den meisten Fällen eine gute Wahl ist, weil er zu einem Test führt, der das Niveau einhält, aber auch bei der Fallzahlplanung gute Ergebnisse liefert. Weiterhin ist die in [19] vorgeschlagene Prozedur durch die Verwendung von Permutationstests auch auf andere Testsituationen im Rahmen randomisierter Experimente übertragbar.

Zu der genannten Übertragung mit Hilfe von Permutationstests sind nach dem Artikel in [19] zahlreiche andere Möglichkeiten zur Anwendung oder Erweiterung dieser Grundidee gezeigt worden, so z.B. die Anwendbarkeit bei Nichtunterlegenheits- bzw. Äquivalenzstudien, [21], bei binären Endpunkten [22] und für Nichtunterlegenheitsstudien bei binären Endpunkten [23]. Eine Vorgehensweise für mehrarmige Studien [24], die auf dem Ansatz von Wittes und Brittain basiert, wurde dahingehend modifiziert, dass nun ein verblindeter und ggf. adjustierter Varianzschätzer verwendet werden kann [25]. Weiterhin wurde eine interne Pilotstudie bei ordinalen Endpunkten unter der Annahme von *proportional odds* angewendet [26]. Bei der genannten Studie wären auch Methoden der verblindeten Rekalkulation anwendbar gewesen [21]. Prinzipiell ist verblindete Fallzahladaption auch in gruppensequentiellen Designs für Überlebenszeitstudien möglich [27], genauso wie die Anwendung von Methoden aus [18] auf Cluster randomisierte Studien [28]. Bei Studien mit Messwiederholungen besteht die Möglichkeit der Verwendung von verblindeten Interimdaten nach Gould und Shih [29], wobei die Problematik des EM-Algorithmus berücksichtigt und auf das Permutationsargument in [20] hingewiesen wurde. In [30] wurde bei der verblindeten Rekalkulation ebenfalls auf die Verfahren von Gould und Shih verwiesen, allerdings wird eine Abschätzung des Fehlers 1. Art wie in [18] empfohlen. Des Weiteren können Daten von Patienten, die zwar schon in die Studie eingeschlossen sind, aber noch nicht das Ende erreicht haben, bei der Fallzahlrekalkulation genutzt werden [21]. Dies ist für normalverteilte [31] wie auch für binäre Endpunkte [32] untersucht worden.

Drei wichtige Ansätze zur Fallzahlrekalkulation bei binären Endpunkten wurden von Gould [11], Herson [33] und von Shih [14] vorgeschlagen. Der erste kann verblindet angewendet werden, seine Eigenschaften wurden durch Friede [22] genauer untersucht, während der zweite

nur durchgeführt werden kann, wenn die Erfolgsquote in der Kontrollgruppe bestimmt wird. Die Prozedur von Shih dagegen verwendet bis zum Zeitpunkt der Interimstichprobe eine künstliche dummy-Stratifizierung, die die Verblindung formal aufrecht erhält, aber den Behandlungseffekt in den Gruppen schätzen lässt.

In Erweiterung der genannten Literatur soll im zweiten Teil dieser Arbeit die Auswirkung der *mehrfachen* Anpassung der Fallzahl durch die Prozeduren von Kieser und Friede [19, 22] auf den Fehler 1. Art, sowie die benötigten Fallzahlen inklusive ihrer Streuung unter der Nullhypothese untersucht werden. Die mehrfache Anpassung hat den Vorteil, dass das Problem der richtigen Wahl des Zeitpunktes entfällt und auch Varianzveränderungen über die Zeit entgegnet werden kann.

1.4 Ein Kombinationstest bei geänderter Variabilität

Der dritte Teil der Arbeit beschäftigt sich mit dem Vergleich von drei Testverfahren in verschiedenen Situationen, in denen die Varianz der primären Zielgröße im zeitlichen Verlauf der Studie nicht konstant war [34]. Das erste und einfachste Verfahren – und das wohl am häufigsten in der Praxis verwendete – ignoriert diese Unterschiedlichkeit, das zweite wertet die Phasen getrennt aus und basiert auf Fishers Kombinationstest. Das dritte Verfahren ist dem Prinzip nach ein Abschlusstest, der wie bei der separaten Auswertung die Varianzunterschiede berücksichtigt.

Bei der Untersuchung wird modellhaft die Einführung eines Amendments während der Studie als Ursprung für die unterschiedlichen Varianzen angenommen. Amendments zu Protokollen klinischer Studien sind nicht ungewöhnlich und aus verschiedenen Gründen nötig, von denen einige in [15] und [35] genannt werden: In Studien mit langer Laufzeit kann sich der Stand der Wissenschaft zwischenzeitlich weiterentwickelt haben, so dass Planungsannahmen geändert oder optimale Behandlungsmethoden angepasst werden müssen. Außerdem können unerwartet kleine Rekrutierungsraten die Ausweitung der Ein- und Ausschlusskriterien erfordern. Überlegungen zur Sicherheit können jedoch ebenfalls Änderungen an den Zulassungskriterien nötig machen. Bei einem Amendment muss dargelegt werden, ob durch die geänderten Umstände die statistische Analyse modifiziert werden muss und wie das zu geschehen hat („amendment should also cover any statistical consequences [...] and alterations to the planned statistical analysis“) [35].

Wenn Ein-/Ausschlusskriterien während einer Studie geändert werden, unterscheiden sich höchstwahrscheinlich die Patientenkollektive vor und nach diesem Zeitpunkt. Auswirkungen von Amendments auf das Patientenkollektiv, die statistischen Konsequenzen und die Fallzahlanpassung zur Erhaltung der Trennschärfe wurden in [36] untersucht.

Weiterhin wird in [36] auf eine Vorgabe der FDA [37] verwiesen, der zufolge die genauen Bedingungen der zu behandelnden Erkrankung und damit die potentiell zukünftige Gruppe von Patienten durch die Ein-/Ausschlusskriterien definiert werden. Dabei ist es wichtig, dass diese so gewählt werden, dass der Schluss von der Studienpopulation auf die Gesamtpopulation valide und unverzerrt ist. Es besteht jedoch die Gefahr, dass statistische Standardmethoden nach der Änderung dieser Kriterien nicht mehr anwendbar sind [36], denn es entstehen jeweils

neuartige Zielpopulationen innerhalb der Studie, deren Maße zur Wirksamkeit schwer bis gar nicht mehr zu bestimmen sind. Zur Messung der Auswirkung von Amendments schlagen die Autoren vor, Größen zur Verschiebung der Mittelwerte, der Änderung der Variabilität und der Effektgröße in den Subpopulationen zu bestimmen.

Chow und Shao [38] zeigen ein Amendment, welches für die hier zu untersuchende Situation beispielhaft ist: Es handelt sich hierbei um eine placebokontrollierte Studie zu Asthma. Da die Rekrutierung dieser Studie zu langsam voran schritt, wurde die ursprüngliche Bedingung, nur Patienten mit einem FEV_1-Wert (*Forced Expiratory Volume in 1 second*) von 1.5 bis 2 Liter aufzunehmen, dahingehend geändert, dass die Obergrenze zunächst auf 2.5 Liter, dann in einem zweiten Amendment auf 3 Liter angehoben wurde. In dieser Studie betraf das Amendment die Differenz des FEV_1-Wertes am Ende der Studie zum Wert am Anfang und damit direkt die primäre Zielgröße.

In [39] berichten die Autoren von einer zweiarmigen Studie zu Rezidiven bei Hautkrebs. Durch Anheben der Obergrenze für den zulässigen Cholesterinspiegel in einem Amendment sollte die Rekrutierung beschleunigt werden.

Als ein weiteres Beispiel kann die placebokontrollierte Phase-III-Studie zur Behandlung von Schuppenflechte mit Efalizumab [40] dienen. Durch ein Amendment wurden die so genannte „high-need“-Gruppe hinzugenommen. Die Protokollmodifikation wurde durchgeführt, um zusätzlich Sicherheits- und Wirksamkeitsdaten für diese Gruppe zu erhalten.

An diesen Beispielen sieht man, dass Änderungen an den Ein-/Ausschlusskriterien in Studien verschiedenster Indikation und zu unterschiedlichen Zwecken vorgenommen werden und dass die Anzahl dieser Modifikationen üblicherweise gering ist.

Oft wird auf die durch Amendments geänderte Populationen in der statistischen Analyse nicht reagiert, d.h. es werden dieselben Methoden angewendet, die im ursprünglichen Studienprotokoll spezifiziert worden sind, solange eine Vergrößerung des Fehlers 1. Art nicht abzusehen ist. Da die Veränderungen jedoch sehr grundlegend gewesen sein können, besteht die Gefahr, dass die Studie mit den geplanten statistischen Methoden die eigentliche Frage nicht mehr valide beantworten kann.

Durch das Poolen der Daten kann zudem ein Bias entstehen und/oder die Studie an Power verlieren – möglicherweise unter den bei der Planung angesetzten Wert. Adaptive Designs, wie sie durch Bauer [41] eingeführt wurden, können dabei helfen, dass die statistischen Schlussweisen korrekt bleiben. Chow et al. [36] weisen jedoch dabei auf die Gefahr für die wissenschaftliche Validität und für die Trennschärfe bei kleineren Fallzahlen und mehreren Amendments hin.

In dieser Arbeit soll der Fall untersucht werden, in dem ein Amendment das Patientenkollektiv verändert hat, ohne dass vorher ein flexibles Design gewählt wurde. Dazu wird eine alternative auf dem Kombinationstest von Fisher basierende Auswertungsmethode vorgeschlagen und untersucht. Bei ihr wird die Zielgröße wahlweise unabhängig von jeglichen Kovariablen in jeder Phase getrennt getestet und die so gewonnenen p-Werte am Ende zu einem Gesamtergebnis kombiniert.

2 Methodik

2.1 Reale Studien

2.1.1 Quellen

Die Daten zu realen Studien wurden aus mehreren Quellen gewonnen. Die wichtigsten Quellen waren dabei eine institutseigene Projektdatenbank und ein Verzeichnis von Publikationen von Mitarbeitern des IMIBE. Des Weiteren wurde ebenfalls nach öffentlich zugänglichen Studiendaten im Internet recherchiert und auch Beispieldaten aus der Literatur gesucht. Nachdem Kandidatenstudien identifiziert waren, wurden verschiedene Zusatzinformationen über sie ermittelt, anhand derer die nachstehenden Kriterien überprüft wurden.

2.1.1.1 Kriterien

Folgende Kriterien mussten erfüllt sein, damit die Rohdaten einer Studie verwendet werden konnten:

1. Verfügbarkeit ausreichender Informationen über die Studie
   Nach Auffinden eines möglichen Kandidaten mussten genug Informationen vorhanden sein, um die restlichen Fragen dieses Katalogs beantworten zu können.

2. Studie am Menschen
   Die Studie musste mit Hilfe von Patienten oder gesunden Probanden durchgeführt worden sein. Laborexperimente (auch solche mit menschlichen Zellen), Tierversuche oder auch technische Untersuchungen (z.B. Testreihen zu Röntgenmaterial und einem künstlichen Torso) wurden ausgeschlossen.

3. ausreichende Fallzahl
   Die Größe einer Untersuchungsgruppe musste mindestens 30 Probanden / Patienten betragen, damit bei einer Teilung in zwei Phasen mindestens 15 Werte pro Phase und Behandlungsgruppe für die Variabilitätsbestimmung zur Verfügung standen.

4. *echter* Rekrutierungsprozess
   Die Daten sollten so erhoben worden sein, dass der Einschluss der Patienten ein Prozess mit zeitlicher Ausdehnung war, denn im Normalfall müssen Patienten sukzessive in eine Studie eingeschlossen werden. Durch diese Forderung fallen offensichtlich zwei Studientypen heraus: Querschnittstudien und rein retrospektive Studien. Ebenso wurden Registerdaten ausgeschlossen, denn diese stellen oft Datenbanken dar, die über viele Jahrzehnte und bundes(land)weit gesammelt wurden und sind daher eher untypisch für den Prozess während einer zeitlich und räumlich begrenzten klinischen Studie.

5. Verfügbarkeit der Studiendaten
   Fehlende Datenhoheit, Nichtverfügbarkeit in elektronischer Form und Fehlen aufgrund abgelaufener Aufbewahrungsfrist konnten hier zum Ausschluss einer Studie führen.

6. Information zum Rekrutierungszeitpunkt
   Idealerweise sollte diese Information aus dem Rekrutierungsdatum selbst bestehen. In der Praxis zeigte sich aber oft, dass dieses nicht in der Datenbank erfasst war. In diesen Fällen wurde ersatzweise auf den Zeitpunkt der Randomisierung, der Einverständniserklärung oder den der Erstuntersuchung ausgewichen. Es wurde *nicht* auf Probanden-Identifikationsnummern ausgewichen, denn diese spiegeln erstens nicht den Verlauf der Rekrutierung in der Kalenderzeit wieder, sondern bestenfalls die Reihenfolge und zweitens kann in multizentrischen Studien nicht einmal diese rekonstruiert werden.

7. Skalenniveau der Zielgröße
   Die Betrachtung wurde auf stetige Zielgrößen eingeschränkt. Bei Studien, die zensierte oder binäre primäre Endpunkte hatten, wurde nach sekundären Größen gesucht, über die eine Auswertung als stetige Variable möglich war und die mit dem Endpunkt verbunden waren, d.h. dass z.B. bei onkologischen Studien von der Variable *Gesamtüberleben* auf *Lebensqualität* ausgewichen wurde.

Eine einheitliche Aufstellung über die Gründe des Ausschlusses der nicht verwendeten Studien kann es nicht geben, denn es reichte ein einzelnes nicht erfülltes Kriterium, um die Studie zu verwerfen, auch wenn zu diesem Zeitpunkt andere Kriterien noch unbekannt waren.

2.1.1.2 Metadaten

Wenn eine Studie die oben genannten Kriterien erfüllte, wurden eine Reihe von Daten erhoben, die wichtige Eigenschaften zur Unterscheidung der Studien widerspiegelten. Aufgrund dieser Daten sollte eine etwaige Abhängigkeit von Schwankungen in der Variabilität oder der Rekrutierungsgeschwindigkeit gefunden werden.

1. Art der Studie
   Es wurde festgehalten, ob es sich um eine Interventionsstudie (IS) oder eine Nichtinterventionsstudie (NIS) handelte. Eine Studie war dann eine Interventionsstudie, wenn ein Einwirken auf die *Behandlung* eines Probanden vorlag – z.B. Zuteilung zu einer Behandlungsgruppe durch Randomisierung. Bei einer Studie handelte es sich daher nicht um eine Interventionsstudie, wenn nur Einwirkung auf den menschlichen Organismus im Sinne von Medikamentengabe vorlag. Entsprechend wurde eine Studie ohne Einfluss auf einen Behandlungsverlauf als Nichtinterventionsstudie eingeordnet. Daraus folgt, dass Anwendungsbeobachtungen unter Nichtinterventionsstudie fallen, denn obwohl zwar u.U. Medikamente eingenommen werden, wird doch kein Einfluss auf die individuelle Behandlung ausgeübt.
   Da auf diese Weise aber sehr unterschiedliche Studientypen unter eine Bezeichnung

fallen, nämlich ein Teil der klinischen Phase IV-Studien und Anwendungsbeobachtungen einerseits und epidemiologische Studien andererseits, wurden die ersteren noch zusätzlich als *patientenbasiert* und die letzteren als *bevölkerungsbasiert* klassifiziert.

2. Phase in der Medikamentenentwicklung
   Die Anforderungen an Studien in verschiedenen Phasen der Medikamentenentwicklung unterscheiden sich z.T. stark. Die Patientenkollektive einer Phase III-Studie und einer Phase IV-Studie können sehr unterschiedlich zusammengesetzt und auch unterschiedlich groß sein. Daher wurde vermutet, dass sich diese Studien auch hinsichtlich ihres Rekrutierungsverhaltens wie auch ihrer Variabilität unterscheiden. Es wurde eine Einteilung der klinischer Studien in die Phasen I bis IV vorgenommen, wobei die Zuordnung ausschließlich über Informationen aus der Dokumentation zur jeweiligen Studie stattfand.

3. Chronizität Erkrankung
   Die Chronizität der Erkrankung könnte Einfluss auf die Rekrutierung haben, wobei dieser in beiden Richtungen denkbar war. Möglich ist eine anfänglich „zu schnelle“ Rekrutierung, bei der der Studienarzt schon einige chronische Patienten kennt und sie schnell zu Anfang einschließen kann. Als Beispiel kann eine Studie zur Behandlung von Migräne und Kopfschmerz mit Akupunktur [42] dienen, bei der Patientendateien nach möglichen Studienteilnehmern durchsucht worden sind, statt diese zu rekrutieren, wenn sie vorstellig wurden. Es gibt auch die gegenläufige Argumentation [43]: Patienten mit chronischer Krankheit kommen bei laufender Behandlung weniger oft zu ihrem Arzt und können daher nicht so einfach in eine Studie eingeschlossen werden.
   Für diese Variable wurden die Ausprägungen *chronisch*, *akut* und *beides möglich* verwendet.

4. Anzahl Zentren
   Die Anzahl der Zentren in einer Studie könnte offensichtlich auch die Variabilität der Zielgröße beeinflussen. Es gibt Untersuchungen zum Thema der Varianzheterogenität bei multizentrischen Studien [44]. Diese übersetzt sich in eine zeitliche Heterogenität, wenn die Zentren nicht gleichzeitig mit der Rekrutierung beginnen oder gar, wenn eine schleppende Rekrutierung das Eröffnen neuer Zentren nötig macht. Abgesehen davon kann eine unterschiedliche Rekrutierungsdynamik der einzelnen Zentren zu einer Veränderung der Variabilität führen.

5. Randomisierung
   Ein Zusammenhang von Randomisierung und Rekrutierung bzw. Variabilität wäre insofern denkbar, dass bei nichtrandomisierten Studien eine geringere Variabilität vorliegen könnte, die aus einer Selektion homogener Behandlungsgruppen resultiert. Andererseits könnte eine Randomisierung zu einer Selektion von einer Subgruppe „randomisierungswilliger“ Patienten führen, die wiederum zu schleppender Rekrutierung und Amendment mit Variabilitätsveränderung führen könnte. In die gleiche Richtung wirkt das Problem, dass auch Ärzte (und damit Zentren) zu klinischen Studien rekrutiert werden

müssen. Bei Randomisierung ist denkbar, dass Ärzte eine Teilnahme ablehnen, weil sie nicht mehr unabhängig über die Therapie ihrer Patienten entscheiden können [45].

6. Verblindung
   So wie die Randomisierung der Vermeidung von Strukturungleichheit dienen soll, so soll die Verblindung Beobachtungsungleichheit verhindern, die ihrerseits indirekt auf die Variabilität wirken kann. Die möglichen Ausprägungen dieser Variablen waren *unverblindet*, *einfach blind* und *doppelblind*.

7. Jahr der Studie
   In den letzten 30 Jahren gab es mehrere Veränderungen bei den regulatorischen Bedingungen für klinischen Studien. Es ist durch diese Veränderungen ein Einfluss auf die Dynamik der Variabilität der primären Zielgrößen zwischen Studien aus verschiedenen Zeiträumen denkbar. Es wurden dazu nur Interventionsstudien untersucht, da z.B. für epidemiologische Studien völlig andere Richtlinien existieren. Wichtige Veränderungen der gesetzlichen bzw. wissenschaftlichen Richtlinien fanden zu folgenden Zeitpunkten statt [46, 47]:
   - 1. Juli 1991: EG-GCP-Leitlinie tritt in Kraft
   - 1. Januar 1997: ICH-GCP-Leitlinie in der EU in Kraft
   - 6. August 2004: 12. AMG-Novelle: EU-GCP verpflichtend

8. Art der Kontrollgruppe
   Anhand dieses Metadatums wurden die Studien eingeteilt in solche ohne Kontrollgruppe (wie z.B. bei Anwendungsbeobachtungen), solche mit Placebo- oder mit aktiver Kontrolle. Die unterschiedliche Akzeptanz der einen oder anderen Option bei den potentiellen Studienteilnehmern könnte zu unterschiedlicher Dynamik bei der Rekrutierung und Varianz geführt haben. So wurde z.B. bereits beim Design einer Studie zur Behandlung von Migräne und Kopfschmerz mit Akupunktur [42] darauf hingewiesen, dass sowohl die *Compliance* des Patienten, d.h. der Grad, in dem er dem Studienprotokoll folgt, und auch die Rekrutierungsrate der Studie bei Placebo-Kontrollen geringer sein können.

9. Amendment
   Wie bereits erwähnt, kann ein Amendment die Folge von Rekrutierungsproblemen sein, die wiederum in den anderen Faktoren begründet sein können. Aber auch unabhängig davon kann ein Amendment zu Rekrutierungs- und Varianzschwankungen führen, besonders, wenn es Ein-/Ausschlusskriterien betrifft. Ein solches Amendment kann z.B. auch bei wiederholten (schweren) Protokollverletzungen nötig werden. Es wurden allerdings nur Amendments berücksichtigt, die Einfluss auf die Rekrutierung haben sollten oder konnten und die *während* der Rekrutierungsphase der Studie gültig geworden sind. Amendments, die z.B. nur Änderungen der sprachlichen Formulierungen im Prüfplan korrigiert haben oder solche, die *vor* dem Einschluss des ersten Patienten in Kraft getreten sind, sind nicht als Amendments in diese Analyse eingegangen.

10. Art der Intervention
    Die Art der Intervention spielt eine Rolle dafür, welche Regulatorien gültig werden, denn es fallen beispielsweise Arzneimittel und Lifestyleinterventionen oder auch Operationsmethoden unter verschiedene gesetzliche Regelungen. Die damit verbundenen unterschiedlichen Qualitätsanforderungen schlagen sich in der grundsätzlichen Durchführung einer Studie nieder und könnten sich daher in der Dynamik der Variabilität der Zielgrößen widerspiegeln. Hinsichtlich der Rekrutierung gibt es eine aktuelle Diskussion, ob und wie z.B. Studien, die Operationsmethoden beinhalten, größere Probleme bei der Patientenrekrutierung haben könnten als andere [48, 49].

11. Dauer der Studie
    Die Dauer der Studie kann eine Rolle beim Rekrutierungsverlauf, aber auch bei der Variabilität spielen. Bei kurzen Studien wurden weniger Schwankungen bei diesen beiden Parametern als bei länger laufenden Studien erwartet. Je länger eine Studie läuft, desto höher ist die Gefahr, dass dadurch Änderungen in der Rekrutierung und auch in der Varianz geschehen oder Amendments nötig werden. Die Dauer wurde als die Zeitspanne zwischen dem Eintreten des ersten Patienten und des letzten Patienten anhand der Variable zum Rekrutierungszeitpunkt (oder Ersatzvariable) definiert.

12. Anzahl der Untersuchungsgruppen
    Es ist ein Zusammenhang zwischen der Anzahl der Untersuchungsgruppen und der Rekrutierungsgeschwindigkeit denkbar und zwar besonders dann, wenn mehrere Gruppen gleich groß sein sollen und ähnliche Struktur aufweisen sollen. Möglicherweise sind Studien mit mehr Behandlungsgruppen mit größeren Problemen bei der Rekrutierung konfrontiert.

2.1.2 Maße zur Rekrutierung

Es wurde ein Maß für die Ungleichmäßigkeit des Rekrutierungsverlaufes berechnet. Dazu wurde jede Studie bei einem der drei folgenden Zeitpunkte aufgeteilt: Datum eines Amendments, Datum einer Interimanalyse oder bei der Hälfte, wenn keine der zwei erstgenannten Möglichkeiten zutraf. Dann wurde bestimmt, wann sich dieser Zeitpunkt t relativ zur Gesamtzeit T der Studie befand. Anschließend wurde der Anteil n der bis zum Zeitpunkt t rekrutierten Patienten an der Gesamtzahl N bestimmt und von diesem Anteil der vorher bestimmte abgezogen. Diese Differenz $\delta_{\text{rek}} = n/N - t/T$ sollte als Maß für die Gleichmäßigkeit der Rekrutierung dienen.

Beispiel: Eine fiktive Studie startete am 01.06.2001 und endete am 20.06.2001, dann war die Rekrutierungsdauer T genau 20 Tage. Wenn es ein Amendment am 10.06.2001 gegeben hätte ($t = 10$), hatte dort die Teilung stattgefunden und diese wäre bei 1/2 der Gesamtzeit aufgetreten. Wenn weiterhin bei dieser Studie insgesamt $N = 50$ Patienten und bis zum 10.01.2001 schon $n = 30$ rekrutiert worden sind, sind also bis dahin 3/5 der Patienten in die Studie aufgenommen worden. Damit wäre $\delta_{\text{rek}} = n/N - t/T = 80\% - 50\% = 30\%$.

Das Maß δ_{rek} spiegelt damit die Unterschiede in der Rekrutierung in zwei Phasen einer kli-

nischen Studie wieder, wenn diese zwei Phasen durch ein besonderes Ereignis wie ein Amendment oder eine Interimanalyse bestimmt wurden. Wie man sofort sieht, ist $\delta_{\text{rek}} > 0$, wenn mehr Patienten rekrutiert worden sind, als es bei gleichmäßiger Rekrutierung der Fall gewesen wäre und < 0, wenn eine anfänglich schleppende Rekrutierung vorlag.

Zu δ_{rek} wurden dann passende Lage und Streumaße zuerst für alle, dann für die nach den aufgeführten Metadaten gruppierten Studien getrennt berechnet, um Anhaltspunkte für mögliche Abhängigkeiten zu bekommen.

Außerdem wurde auch die Rekrutierungsgeschwindigkeit v_{rek} bestimmt. Dazu wurde die erreichte Anzahl an Patienten bzw. Probanden zur Dauer der Studie ins Verhältnis gesetzt.

Je nach Fragestellung und Verteilungseigenschaften wurde die Rekrutierungsgeschwindigkeit mit Hilfe des $t-$Tests oder des Fisher-Pitman-Permutationstests verglichen oder mit den zugehörigen Verallgemeinerungen für mehrere Gruppen (ANOVA bzw. Permutation One-Way ANOVA).

Der Fisher-Pitman-Permutationstest [50] soll an dieser Stelle kurz beschrieben werden. Die Darstellung basiert auf [51], wobei die dortigen Bezeichnungen übernommen werden. Der Test wird angewendet, wenn zwei unabhängige Gruppen auf Lokationsunterschiede verglichen werden sollen. Dabei seien mit $u_{11}, \ldots, u_{n_1 1}$ die n_1 Werte der ersten Gruppe und entsprechend mit $u_{12}, \ldots, u_{n_2 2}$ die n_2 Werte der zweiten Gruppe bezeichnet, womit sich die Gesamtgröße $N = n_1 + n_2$ der Stichprobe ergibt. Wenn F_1 und F_2 die Verteilungsfunktionen der beiden Verteilungen sind, kann die Nullhypothese des Tests als $H_0 : F_1 = F_2$ formuliert werden.

Allgemein wird bei Tests dieses Typs jeder Beobachtung u_{ij} ein *Score* w_{ij} zugewiesen, die Scores vereinigt und wieder auf die Gruppen verteilt, so dass zwar die Gruppengrößen wieder die ursprünglichen sind, dass aber die Beobachtungen nun völlig unterschiedlich auf die Gruppen verteilt werden können. Um alle möglichen Neuverteilungen aufzählen, wird die Reihenfolge der Beobachtungen permutiert. Wenn W nun die Menge aller Permutationen $\tilde{w}_{ij}$ ist, kann die beobachtete Reihenfolge w_{ij} als eine Realisation der $\tilde{w}_{ij}$ betrachtet und aus den Scores eine Teststatistik konstruiert werden. Die Teststatistik $T \equiv T(\tilde{w}_{ij})$ des Fisher-Pitman-Test ergibt sich aus der Summe

$$T = \sum_{i=1}^{n_1} \tilde{w}_{i1}$$

der Scores der ersten Gruppe. Während bei Wahl z.B. der Ränge als Scores der bekannte Wilcoxon-Mann-Whitney-Test resultiert, werden beim Fisher-Pitman-Test die ursprünglichen Werte selbst als Scores verwendet. Nach Adjustierung für Bindungen kann der p-Wert des Tests wie bei anderen exakten Tests bestimmt werden: Er ergibt sich als der Anteil der Permutationen, die zu einer Teststatistik führen, die mindestens so „extrem“ ist wie die beobachtete. Wenn also t der beobachtete Wert der Teststatistik T ist, ergibt sich der zweiseitige p-Wert als

$$p_2 = \Pr(|T - E(T)| \geq |t - E(T)|),$$

wobei der Erwartungswert $E(T)$

$$E(T) = \sum_{i=1}^{N} \frac{n_1 w_i}{N}$$

beträgt.

Ein Permutationstest, der mehrere Gruppen vergleicht, ist die Permutation One-Way ANOVA [51]: Werden K unabhängige Gruppen mit jeweils $n_1, n_2, \ldots, n_K$ Beobachtungen u_{ij} mit $j = 1, \ldots, K$ und $i = 1, \ldots, n_j$ betrachtet und bezeichnet F_j die Verteilungsfunktion der Werte in der j-ten Gruppe, dann ist die Nullhypothese des Tests:

$$H_0 : F_1 = F_2 = \ldots = F_K.$$

Analog zum Fisher-Pitman-Permutationstest werden aus den Daten u_{ij} die Scores w_{ij} wieder als die Originalwerte bestimmt, d.h. $w_{ij} := u_{ij}$. Von diesen werden alle Permutationen $\tilde{w}$ bestimmt unter der Bedingung, dass in die Gruppe j wieder genau n_j Beobachtungen fallen – diese Menge an Permutationen sei wieder mit W, die Teststatistik einer Permutation mit $T(\tilde{w}) \equiv T$ und die konkrete Realisation der vorliegenden Daten mit $t(w)$ bezeichnet. Die Teststatistik wird nun im Gegensatz zum Fisher-Pitman-Test folgendermaßen berechnet: Zuerst wird die Summe der Scores w_j in jeder Gruppe berechnet und dann der Mittelwert $\bar{w}$ dieser Scoresummen bestimmt, außerdem wird noch mit S^2 eine Streuung der Scores zum mittleren Score berechnet:

$$\tilde{w}_j = \sum_{i=1}^{n_j} \tilde{w}_{ij} \quad , \qquad \bar{w} = \frac{1}{N} \sum_{j=1}^{K} w_j \quad , \qquad S^2 = \frac{1}{N-1} \sum_{j=1}^{K} \sum_{i=1}^{n_j} (w_{ij} - \bar{w})^2$$

Die Teststatistik T ergibt sich dann als

$$T(\tilde{w}) = \frac{1}{S^2} \sum_{j=1}^{K} \frac{(\tilde{w}_j - n_j \bar{w})^2}{n_j}.$$

Durch Aufzählung aller Permutationen aus W kommt man zur exakten Verteilung von T unter der Nullhypothese. Dabei wird ausgenutzt, dass unter der Nullhypothese alle Permutationen gleich wahrscheinlich sind — die Wahrscheinlichkeit einer Permutation sei mit $h(\tilde{w})$ bezeichnet und man bestimmt wieder diejenigen Permutationen, die eine Teststatistik $T(\tilde{w})$ aufweisen, die mindestens so extrem ist, wie die beobachtete $t(w)$ und bestimmt deren Wahrscheinlichkeit. Bei der Teststatistik der Permutation One-Way ANOVA bedeutet „extrem“, dass der Wert der größer oder gleich dem der beobachteten Statistik sein muss. Die Alternativhypothese dieses Tests ist ungerichtet, d.h. es wird keine Anordnung der Gruppen bei Ablehnung der Nullhypothese angenommen. Der zweiseitige p-Wert beträgt

$$p_2 = \Pr(T \geq t) = \sum_{\{\tilde{w} \,:\, T \geq t\}} h(\tilde{w}).$$

2.1.3 Untersuchung der Variabilität

Die Unterschiede bei der Variabilität der untersuchten Studien wurden ermittelt, indem die Varianz der zweiten Phase zu der der ersten Phase ins Verhältnis gesetzt wurde. Zur besseren Anschaulichkeit bei der Deskription wird hier das Verhältnis der Standardabweichungen

berichtet, welches die Wurzel aus dem Verhältnis der Varianzen ist.

Da bei Studien, zu denen mehrere Zielgrößen vorlagen, die beschriebene Bewertung für jede Einzelvariable durchgeführt wurde, werden hier zur Untersuchung drei Ansätze verfolgt:

1. einfache Deskription mit Ignorieren der Korreliertheit von Variablen, die aus der gleichen Studie stammen,
2. Verwendung aller Verhältnisse und Vergleich mit der theoretischen Mischverteilung, die die Korreliertheit einbezieht,
3. Berücksichtigung der Clusterstruktur durch Verwendung eines hierarchischen Modells.

2.1.3.1 Deskription der Varianzunterschiede

Für eine erste Deskription wurden alle Varianzverhältnisse als unabhängig betrachtet, so dass die Tatsache, dass sie teilweise aus der gleichen Studie stammen und daher eine Clusterstruktur aufweisen, außer Acht gelassen wurde. Es wurden folgende beschreibende Statistiken berechnet, wie sie z.B. in [52] beschrieben werden: Lagemaße (Mittelwert, Median), Streumaße (Standardabweichung, Interquartilsabstand), und weitere Maße zur Beschreibung der Verteilung wie Perzentile, Maximum, Minimum, Schiefe und Kurtosis.

2.1.3.2 Vergleich mit der Mischverteilung

Ein Weg, die Verteilung der Varianzverhältnisse zu beurteilen und dabei die Clusterstruktur zu berücksichtigen ist, sie zu vereinigen, während man ihre jeweiligen Verteilungen beachtet. Das kann durch die Verwendung von Mischverteilungen erreicht werden. Dazu werden zuerst die theoretischen Verteilungen der einzelnen beobachteten Varianzverhältnisse benötigt.

Es ist ein bekanntes Resultat der Statistik, dass die Zufallsgröße $\frac{S_1^2}{S_2^2}$ eine $F_{m-1,n-1}$-Verteilung aufweist, wenn diese Bedingungen erfüllt sind: S_1^2, S_2^2 sind die erwartungstreuen Schätzer der Varianz zweier Stichproben vom Umfängen n und m von jeweils unabhängig identisch normalverteilten Zufallsvariablen und weiterhin sind die zugrunde liegenden wahren Varianzen gleich (Nullhypothese beim F-Test).

Im Anwendungsfall dieser Arbeit können die Beobachtungen der ersten und zweiten Phase als stochastisch unabhängig betrachtet werden (denn sie werden von verschiedenen Probanden erhoben), so dass für die beschriebenen Verhältnisse der Varianzen eine F-Verteilung unter der Nullhypothese gleicher Varianz über die Phasen hinweg angenommen werden kann.

Die Varianzverhältnisse innerhalb einer Studie entspringen immer *derselben* F-Verteilung, denn diese ist nur abhängig von den Stichprobenumfängen, d.h. hier den Umfängen der beiden Phasen. Zwischen den Studien jedoch ist im Allgemeinen mit unterschiedlichen Freiheitsgraden für die F-Verteilung zu rechnen. Bildet man also die Vereinigung aller Varianzverhältnisse, dann entspringen diese einer Mischung aus verschiedenen F-Verteilungen. Für diese Mischung kann unter der Nullhypothese eine theoretische Dichtefunktion hergeleitet werden.

Die Dichte einer Mischverteilung entspricht dem gewichteten Mittel der Dichten der einzelnen Verteilungen, d.h. sind die Dichtefunktionen f_i, $i = 1, \ldots, n$ und die Gewichte $a_1, \ldots, a_n$

mit $\sum a_i = 1$ gegeben, dann ist die Dichte g der Mischverteilung bestimmt durch

$$g(x) = \sum_{i=1}^{n} a_i f_i(x).$$

Aus dieser Darstellung können die Perzentile der Mischverteilung auf einfache Weise berechnet werden.

Durch Vergleich des Histogramms der beobachteten Varianzverhältnisse mit der theoretischen Dichte der Mischverteilung wurde optisch beurteilt, ob grobe Verstöße gegen die Nullhypothese gleicher Varianzen vorliegen.

Die gewählten Gewichte für jede Studie entsprachen dem Anteil an Variablen, den diese Studie zur Gesamtzahl aller beobachteten Variablen beigesteuert hatte.

Eine Alternative zur Dichteschätzung mit Hilfe eines Histogramms ist die Verwendung eines *Kernschätzers* für die Dichte. Dieser nimmt an einer Stelle x den folgenden Wert an [53]

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

wobei K eine *Kernfunktion* ist, n die Anzahl der vorliegenden Daten und h die *Bandbreite*. Eine Kernfunktion ist eine Funktion, die nur nichtnegative Funktionswerte aufweist und deren Integral Eins ergibt. Für die Untersuchungen in der vorliegenden Arbeit wurde zur Schätzung die `SAS`-Prozedur `PROC CAPABILITY` verwendet, bei der standardmäßig die Gaußsche Glockenkurve als Kernfunktion zum Einsatz kommt. Von ihr konnte gezeigt werden, dass der sich ergebende Schätzer asymptotisch erwartungstreu ist und die Dichtefunktion sogar schwach konsistent annähert. Die Bandbreite h bestimmt wie „glatt“ der Dichteschätzer sein wird. Die oben genannte `SAS`-Prozedur kann einen Wert für h bestimmen, wobei die Approximationsgüte in Form des IMSE (integrated mean square error) optimiert wird.

Die realen Varianzverhältnisse wurden zusätzlich mit dem Ein-Stichproben-Kolmogorov-Smirnov-Test auf Unterschied zur erwarteten Mischverteilung getestet. Bei diesem Test kann die empirische Verteilungsfunktion einer Stichprobe mit einer theoretischen Vorgabe verglichen werden. Nach [54] haben Tests, die auf der empirischen Verteilungsfunktion basieren eine höhere Trennschärfe als z.B. der χ^2-Test. Die Voraussetzungen, dass die Vergleichsverteilung stetig ist und dass diese ohne Schätzung der Verteilungsparameter angegeben werden kann, sind in diesem Falle offensichtlicherweise erfüllt. Der Vergleich beruht auf dem maximalen vertikalen Abstand dieser beiden Kurven, welcher mit der Teststatistik D bezeichnet wird. Der Wert von D wurde dem Artikel von Stephens [54] gemäß modifiziert zu

$$T = D \cdot (\sqrt{n} + 0.12 + 0.11/\sqrt{n}),$$

der kritische Wert der modifizierten Teststatistik zum 1%-Level beträgt 1.628.

Schließlich kann zur Visualisierung der Verteilung im Vergleich zur theoretischen Mischverteilung auch ein P-P-Plot (*percentage plot*) gezeichnet werden. Bei P-P-Plots wird die empirische Summenhäufigkeitsfunktion der beobachteten Daten gegen die theoretischen Wer-

te der Testverteilung aufgetragen. Formal wird dabei folgendermaßen vorgegangen [55]: Zum Vergleich zweier Verteilungen F und G werden zu festen Quantilen q_i $(i = 1, \ldots, k)$ die Werte $F(q_i)$ und $G(q_i)$ bestimmt. Der P-P-Plot ist dann der Scatterplot der Punkte $(F(q_i), G(q_i))$.

Im betrachteten Fall kann man für G die empirische kumulative Verteilungsfunktion der beobachteten Varianzquotienten und für F die Mischverteilung der theoretischen $F_{m-1,n-1}$-Verteilungen wählen.

Entstammen die beobachteten Daten der theoretischen Verteilung, werden die Punkte $(F(q_i), G(q_i))$ des Scatterplots nahe der Winkelhalbierenden liegen.

2.1.3.3 Untersuchung mit hierarchischem Modell

Da pro Studie z.T. mehrere Variablen untersucht wurden, ergab sich eine hierarchische Struktur der Daten. Die 182 Messwerte, die aus den Verhältnissen der Standardabweichung entstanden, konnten nicht als unabhängige Beobachtungen betrachtet werden. Insbesondere nicht, weil die zu untersuchenden Metavariablen auf der Ebene der *Studien* gegeben waren. Die Daten wurden daher mit einem gemischten linearen Modell untersucht, wie es z.B. in [56] beschrieben ist. Gemischte Modelle dieser Art können so motiviert werden, dass Parameter einer Regression selbst wieder durch eine Regression geschätzt werden. Dadurch können Kovarianzstrukturen modelliert werden, die die Berücksichtigung der besonderen Clusterstruktur der hierarchischen Daten erlauben. Mit Hilfe der gemischten Modelle können dann Varianzkomponenten, aus denen sich die Gesamtvarianz zusammensetzt, bestimmt und verglichen werden.

Es wurde zuerst auf ein *unconditional means model* zurückgegriffen. Dieses geschah wie in [56] beschrieben: Die zu modellierende Größe (in der vorliegenden Arbeit also das Verhältnis der Standardabweichung) wird mit Y_{ij} bezeichnet, wobei i der Index für das Level 1 (die einzelne Variable in einer Studie) und j der Index für das Level 2 (die Studien) verwendet wird. Das Modell lautete nun:

$$Y_{ij} = \beta_{0j} + r_{ij}, \text{ mit } r_{ij} \sim \mathcal{N}(0, \sigma^2),$$

d.h. die Zielvariable wird durch einen Mittelwert β_{0j} pro Studie und einen individuellen Fehler r_{ij} zusammengesetzt. Die β_{0j} werden werden nun ihrerseits mit Hilfe linearer Regression dargestellt:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \text{ mit } u_{0j} \sim \mathcal{N}(0, \tau_{00}).$$

Setzt man nun die zweite in die erste Gleichung ein, erhält man das hierarchische Modell:

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}, \text{ mit } r_{ij} \sim \mathcal{N}(0, \sigma^2), \text{ und } u_{0j} \sim \mathcal{N}(0, \tau_{00}).$$

Dieses Modell, welches nur die Studien als Einflussparameter enthält, kann Auskunft über die Varianzkomponenten geben, die aus der Variabilität innerhalb der Studien und der Restvarianz bestehen. Zu der Varianzkomponente τ_{00} und der Reststreuung σ^2 lassen sich Tests berechnen, die beide Werte gegen Null vergleichen. Ist τ_{00} groß, wird viel Varianz durch die Gruppierung oder Clusterung in Studien erklärt. Der Anteil der Streuung der auf diese

Intragruppen-Varianz entfällt wird auch Intraklassen-Korrelation ρ genannt:

$$\hat{\rho} = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}^2},$$

wobei das Dach über den Variablen zeigt, dass es sich um geschätzte Werte handelt.

Der Schätzer für den Modellparameter γ_{00} steht für den Mittelwert der Varianzverhältnisse, der über das hierarchische Modell, d.h. unter der Berücksichtigung der Clusterstruktur ermittelt werden kann.

Sollen nun weitere erklärende Variablen auf dem Level der Studien zusätzlich in das Modell aufgenommen werden, wird die zweite der obigen Modellgleichungen modifiziert zu [56]:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \cdot x_j + u_{0j}, \text{ mit } u_{0j} \sim \mathcal{N}(0, \tau_{00}),$$

wobei nun x_j ein weiterer Prädiktor auf Studienlevel ist – z.B. einer, der die Studien in prä- und post GCP einteilt. Damit erweitert sich das resultierende Modell zum *conditional model*

$$Y_{ij} = \gamma_{00} + \gamma_{01} \cdot x_j + u_{0j} + r_{ij}, \text{ mit } r_{ij} \sim \mathcal{N}(0, \sigma^2), \text{ und } u_{0j} \sim \mathcal{N}(0, \tau_{00}),$$

wobei der Teil, der mit griechischen Buchstaben bezeichnet wurde, derjenige der *fixed effects* ist, während der Teil mit den lateinischen Buchstaben die zufälligen Effekte (*random effects*) kennzeichnet.

Nach diesen Vorüberlegungen ist es möglich, nach weiteren Einflussfaktoren auf Studienebene zu suchen, die auf Varianzunterschiede deuten, während man die Clusterstruktur der Messwerte weiterhin berücksichtigt. In [56] werden zwei Möglichkeiten dazu beschrieben: Erstens kann bestimmt werden, wie stark sich die Varianzkomponente τ_{00} beim Übergang vom allgemeineren Modell ohne den zusätzlichen Prädiktor auf Studienebene zu dem Modell mit diesem Prädiktor relativ geändert hat. Ergab sich im allgemeineren Modell τ^a_{00} und Modell mit Prädiktor τ^b_{00}, dann wird der Wert

$$\tau^d_{00} := \frac{\tau^b_{00} - \tau^a_{00}}{\tau^a_{00}}$$

bestimmt.

Für die Varianzkomponente τ_{00} kann außerdem ein Test gegen den Wert Null durchgeführt werden, so dass also entschieden werden kann, ob sich die Variabilität der Achsenabschnitte (*intercepts*) der Cluster signifikant von Null unterscheidet. Praktisch würde das bedeuten, dass selbst nach Einschluss des fixen Effekts auf dem Level der Studien, immer noch ein signifikanter Anteil an Streuung zwischen den Studien unerklärt wäre. Denn die Größe von τ^d_{00} stellt nur eine relative Veränderung dar und es ist möglich, dass diese absolut gesehen nur einen kleinen Teil darstellt [56].

Die zweite Möglichkeit besteht darin, eine residuale Intraklassen-Korrelation zu berechnen. In [56] wird dies sinngemäß so formuliert: Es wird die Intraklassen-Korrelation der Studien mit vergleichbarer Ausprägung der jeweiligen Meta-Eigenschaft berechnet (also z.B. alle Studien

mit gleicher Ausprägung der Amendment-Variable). Man bestimmt dazu ρ genau wie oben und interpretiert es nun wie einen partiellen Intraklassen-Korrelationskoeffizienten, der um den vermittelnden Einfluss des jeweiligen Einflussfaktors bereinigt ist.

2.2 Verblindete Fallzahladaption

Bei der verblindeten Fallzahladaption ging es hauptsächlich um die Untersuchung verschiedener Adaptionsprozeduren hinsichtlich der Einhaltung des Niveaus des resultierenden Tests. Diese Frage sollte im Rahmen einer Simulationsstudie geklärt werden, bei der fiktive klinische Studien mit verschiedenen Eigenschaften erzeugt wurden. Die Studien unterschieden sich hinsichtlich des Skalenniveaus der Zielgröße (stetig, binär), in der initialen Fallzahl (96-400) und einer möglichen Fehlspezifikation der Varianz zur Planungszeit. Bei den stetigen Zielgrößen ergaben sich acht unterschiedliche Prozeduren aus der Kombination von drei verschiedenen Eigenschaften der Adaption. Im Falle der binären Größen waren es vier Prozeduren. Es wurde untersucht, wie sich die Prozeduren in den verschiedenen Rahmenbedingungen verhielten, d.h. wie oft der Test am Ende zur Ablehnung unter der Nullhypothese führte und wie groß die jeweils erreichte Fallzahl und ihre Variabilität war.

Allen Simulationen war gemeinsam, dass es drei Zeitpunkte gab, an denen die Varianz verblindet geschätzt und die Fallzahl auf ihrer Grundlage neu bestimmt wurde. Zu jedem dieser Zeitpunkte wurde die Varianz der Endpunkte *aller* bis jeweils dahin rekrutierten Patienten bestimmt. Mit Hilfe dieser Varianzschätzung ergab sich dann unter Umständen eine neue Fallzahl, die dann statt der initialen weiterverwendet wurde. Die erste Anpassung fand immer nach 1/4 der ursprünglich geplanten Fallzahl statt, die nächste nach 1/3 der *noch zu rekrutierenden* Fallzahl, die letzte dann bei 1/2 der *verbleibenden* Fallzahl.

Ein Beispiel für den Idealfall sieht so aus: Bei der Planung einer Studie wurde aufgrund von Vorinformationen die Variabilität genau richtig eingeschätzt und die initiale Fallzahlplanung hat 400 ergeben. Die Variabilität bleibt während der Studie immer gleich dem vorher richtig angenommenen Wert und die erste Schätzung findet dann nach dem Einschluss des (400/4=)100. Patienten statt. Die Varianzschätzung ergibt den Wert, der bei der Planung verwendet wurde und resultiert folglich in der gleichen Fallzahl wie ursprünglich geplant, d.h. 400. Es sind damit noch 300 weitere Patienten zu rekrutieren. Nach Einschluss des (300/3=)100. Patienten nach dem ersten Varianzmonitoring (d.h. nach Einschluss des 200. Patienten insgesamt) wird die Varianz aller bis dahin rekrutierten 200 Patienten bestimmt und darauf aufbauend eine neue Fallzahl bestimmt. Bei dieser Idealstudie wird erneut 400 als alte und neue Fallzahl festgesetzt und es bleiben weitere 200 Patienten, die in die Studie eingeschlossen werden müssen. Bei (200/2=)100 dieser Patienten findet das letzte Varianzmonitoring statt, bei dem dann die Varianz bei den inzwischen 300 Patienten ermittelt wird. Es wird sich bei der folgenden Fallzahlplanung wieder 400 ergeben und die Studie wird dann nach weiteren 100 Patienten enden.

Vereinfacht sieht der Algorithmus so aus:

1. Sei N_i, $i = 0, 1, 2, 3$, die aktuelle Gesamtfallzahl nach der i-ten Rekalkulation
   Sei M_i, $i = 0, 1, 2, 3$, die Anzahl zusätzlicher Patienten, die für das i-te Varianzmonitoring benötigt werden

2. Sei $M_0 = 0$,
   bestimme initiale Fallzahl N_0 durch konventionelle Fallzahlplanung

3. für $i = 1, 2, 3$:
   a) bestimme neues $M_i := (N_{i-1} - \sum_{j=0}^{i-1} M_j)/(4 - i + 1)$
   b) bestimme verblindet die Varianz für die ersten $\sum_{j=0}^{i} M_j$ Patienten
   c) berechne die neue Fallzahl N_i auf Basis der Varianzschätzung durch konventionelle Fallzahlplanung
   d) ist $N_i < \sum_{j=0}^{i} M_j$, dann setze $N_{\text{final}} = \sum_{j=0}^{i} M_j$ und breche Schleife ab, andernfalls setze Schleife mit dem nächsten i fort

4. setze die gesuchte Fallzahl N_{final} auf N_3

Nachdem die simulierte Rekrutierung der Patienten beendet war, wurde ein Test auf Gruppenunterschiede durchgeführt. Es gab immer zwei Behandlungsgruppen, deren Werte unter der Nullhypothese mit Hilfe der Zufallszahlenfunktionen `RANNOR` (normalverteilte Endpunkte) bzw. `RANBIN` (binomiale Endpunkte) der Statistiksoftware `SAS 9.1` simuliert wurden. Die Anzahl simulierter Studien betrug jeweils 10^5 und ein Test wurde als signifikant betrachtet, wenn sein p-Wert kleiner als 5% war. Dann wurde untersucht, ob die 95%-Konfidenzintervalle um die erreichten Anteile an signifikanten Tests die angestrebten 5% überdeckten.

2.2.1 Die untersuchten Prozeduren

Die untersuchten Prozeduren ergaben sich aus der Kombination von drei (bzw. zwei) Prinzipien, die entweder befolgt wurden, oder nicht – was zu acht (bzw. vier) Prozeduren führte. Diese Prinzipien waren:

- Restriktion (*u/r* für *unrestringiert / restringiert*)
- Korrektur für Verblinden (*os/kf* für *one-sample / Kieser-Friede-Korrektur*)
- Einsatz eines Control-Charts (*nc/cc* für *no control chart / control chart*)

Da die Prozedur uosnc bzw. uon keine der möglichen Sondereigenschaften hat, kann sie als Ausgangspunkt bei der Konstruktion der anderen sieben verstanden werden. Daher wird sie im Folgenden auch *Basisprozedur* genannt.

2.2.1.1 Restriktion

Wies eine Prozedur die Eigenschaft der Restriktion auf, musste die neue Fallzahl nach jeder Rekalkulation mindestens so groß sein wie die ursprünglich geplante [9]. In der Sprechweise

des obigen Algorithmus musste also immer gelten: $N_i \geq N_0$ für $i = 1, 2, 3$. Wurde demnach eine neue Fallzahl N_i' bestimmt, dann war die zu verwendende Fallzahl

$$N_i = \max(N_i', N_0).$$

Diese Bedingung war bei stetigen wie bei dichotomen Prozeduren möglich. Die Bedingung c) im Algorithmus wird bei einer Prozedur mit Restriktion nie erfüllt sein. Wurde bei einer Prozedur die Bedingung der Restriktion beachtet, dann begann ihr Name mit *r* (restringiert), andernfalls mit *u* (unrestringiert).

Es gibt noch ein zweites Konzept der Restriktion, bei dem die Fallzahl nie unter die Größe der Stichprobe fallen darf, die zur Varianzbestimmung verwendet wurde [57]. Diese zweite Variante wird hier stillschweigend *immer* angewendet, denn es ist nicht sinnvoll, weniger Patienten in die Analyse eingehen zu lassen, als bereits rekrutiert worden sind.

2.2.1.2 Korrektur für Verblinden

Wie bereits erwähnt, kann eine verblindete Varianzschätzung dazu führen, dass der wahre Wert der Varianz überschätzt wird. Der Bias kann mit Hilfe der Varianzzerlegung quantifiziert und ein entsprechender Korrekturterm angegeben werden [19]. Der Korrekturterm kann dann von der gepoolten Varianz abgezogen werden. Die Korrektur beruht auf der bei der Planung der Studie angesetzten Mittelwertdifferenz und ist jedoch nur für normalverteilte Endpunkte formuliert worden, weswegen dieses Prinzip bei den Prozeduren für dichotome Endpunkte entfiel.

Die unkorrigierte Variante ist der *one-sample*-Schätzer oder auch die *gepoolte* Varianz, welche mit den Notationen aus [19] so aufgeschrieben werden kann:

$$S^2_{\text{os}} = \frac{1}{2n_1 - 1} \left(\sum_{i=1,2} \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 \right),$$

wobei die Beobachtungen aus den beiden Gruppen $X_{11}, X_{12}, \ldots, X_{1n_1}$ und $X_{21}, X_{22}, \ldots, X_{2n_2}$ bestehen und die Varianz nach n_1 Beobachtungen pro Gruppe verblindet bestimmt wird. Nach Varianzzerlegung kann der Bias dieses Schätzers in Abhängigkeit von der Differenz Δ der Mittelwerte beider Gruppen bestimmt werden. Da diese Differenz vorher unbekannt ist, muss man den erwarteten Unterschied δ unter der Alternativhypothese verwenden und kommt somit zum adjustierten Schätzer

$$S^2_{\text{adj}} = S^2_{\text{os}} - \frac{n_1}{2(2n_1 - 1)} \delta^2.$$

Wie Kieser und Friede [19] darlegen ist dies der Schätzer, der von Zucker et al. [20] vorgeschlagen wurde. Die Verwendung dieser Korrektur spiegelte sich im Namen der Prozedur durch *os* für den gepoolten (d.h. one-sample-) Schätzer oder durch *kf* für den adjustierten Schätzer wieder.

2.2.1.3 Einsatz eines Control-Charts

Control-Charts werden in verschiedenen Bereichen zur statistischen Qualitätssicherung verwendet (s. z.B. [58]). Dabei können ein Zielwert sowie Warn- und Kontrollgrenzen für eine statistische Größe vorgeben werden. Bei den Grenzen handelt es sich um Schranken, die durch „unterschiedlich scharfe Tests“ [59] bestimmt werden. Das Verlassen des Bereiches, der durch die *Warn*grenzen markiert ist, führt nur dazu, dass der Prozess genauer beobachtet werden sollte. Gegenmaßnahmen sind jedoch angezeigt, wenn die *Kontroll*grenzen überschritten werden.

Die Idee des Control-Charts wurde dahingehend vereinfacht, dass erstens nur ein Paar an Grenzen verwendet wurde – nämlich Kontrollgrenzen – und zweitens, dass diese Grenzen nicht durch den Ablehnbereich eines statistischen Tests definiert wurden, sondern durch maximal erlaubte Abweichungen einer zu überwachenden Größe von einem Idealwert.

Die zu überwachende Größe war die Trennschärfe (Güte, Power) des Tests, wie er nach Ende der Rekrutierung ausgeführt werden sollte. Dazu wurde zu jedem Monitoring-Zeitpunkt die Varianz verblindet (korrigiert / unkorrigiert) geschätzt und die Power des Tests unter der vorher festgelegten Alternativhypothese bestimmt, wobei als Gesamtfallzahl die zu diesem Zeitpunkt aktuelle verwendet wurde.

Es wurden 85% als Zielwert der Power für den Control-Chart verwendet, sowie 80% als untere und 90% als obere Kontrollgrenze. Wenn bei einem Varianzmonitoring die Power dieses Band verlassen hat, wurde eine Rekalkulation der Fallzahl durchgeführt, ansonsten nicht. Bei der Benennung der Prozedur wurde die Verwendung eines Control-Chart durch die Buchstaben *cc*, die Nichtverwendung durch die Buchstaben *nc* verdeutlicht.

2.2.1.4 Die Prozeduren

Aus der Kombination der drei oben genannten Prinzipien ergaben sich Prozeduren, deren Namen sich aus den einzelnen Buchstaben ergeben, die jeweils genannt wurden, wie es in Tabelle 2.1 aufgeschlüsselt ist.

Restriktion	Korrektur	Control-Chart	Prozedur, stetig	Prozedur, binär
nein	nein	nein	uosnc	unc
nein	nein	ja	uoscc	ucc
nein	ja	nein	ukfnc	-entfällt-
nein	ja	ja	ukfcc	-entfällt-
ja	nein	nein	rosnc	rnc
ja	nein	ja	roscc	rcc
ja	ja	nein	rkfnc	-entfällt-
ja	ja	ja	rkfcc	-entfällt-

Tabelle 2.1:
Schema der Namensgebung für die Rekalkulationsprozeduren im stetigen wie im binären Fall.

2.2.2 Details zu den stetigen Endpunkten

In diesem Kapitel werden Besonderheiten bei der Betrachtung der stetigen Endpunkte genannt, die zu den allgemeinen Beschreibungen der vorhergehenden Kapitel unter 2.2 hinzugefügt werden müssen.

Es wurde ein einseitiger t-Test für die stetigen Endpunkte berechnet, d.h. es wurde für die Mittelwertdifferenz $\Delta := \mu_1 - \mu_2$ der zwei Behandlungsgruppen die Hypothese $H_0 : \Delta = 0$ gegen die Alternative $H_1 : \Delta > 0$ getestet.

2.2.2.1 Alternativhypothese und Powerberechnung

Die Berechnung der Power beim Varianzmonitoring mit Control-Charts fand per Iteration auf der Basis der Formel (7) aus [60] statt, welche für den zweiseitigen t-Test in originaler Notation

$$1 - \beta = 1 - \text{Probt}\left(t_{1-\alpha/2, n_A(r+1)-2}, \quad n_A(r+1) - 2, \quad \sqrt{\frac{r n_A d^2}{(r+1)\sigma^2}}\right)$$

lautet, wobei gilt, dass $1 - \beta$ die Power ergibt, r das das Verhältnis der Größen der beiden Behandlungsgruppen (hier $r = 1$), n_A die Größe der ersten Gruppe, d die Differenz der Mittelwerte und Probt die Verteilungsfunktion einer nichtzentralen t-Verteilung ist. In die Formel direkt eingetragen ist der Nichtzentralitätsparameter $\sqrt{\left[r n_A d^2/((r+1)\sigma^2)\right]}$ sowie die Anzahl $n_A(r+1) - 2$ an Freiheitsgraden. Unter den genannten Bedingungen, wegen der Annahme von Homoskedastizität mit $\sigma^2 = 1$ und wegen der Tatsache eines einseitigen Tests vereinfacht sich diese Formel zu:

$$1 - \beta = 1 - \text{Probt}\left(t_{1-\alpha, n-2}, \quad n - 2, \quad d\sqrt{\frac{n}{2}}\right),$$

in welcher nun n für die Größe der *gesamten* Gruppe steht. Eine weitere Vereinfachung bei der Implementation stellte die Verwendung der zentralen t-Verteilung statt der nichtzentralen dar. Der Unterschied zwischen den beiden war sehr gering und die Laufzeit der Simulation mit der zentralen Variante war deutlich kürzer.

Der Grund für den auf dieser Formel basierenden iterativen Ansatz besteht darin, dass *dieselbe* Formel innerhalb des Rekalkulationsalgorithmus zur Powerberechnung verwendet und somit ein systematischer Fehler aufgrund von unterschiedlichen Formeln und deren unterschiedlichen Approximationen ausgeschlossen werden sollte.

Mit der Vorinformation über Δ und mit der Varianz, die bei jedem Monitoring bestimmt wurde, konnte somit jeweils die Power geschätzt werden, mit der die Studie enden würde, wenn sich die Variabilität ab dem betreffenden Zeitpunkt nicht mehr ändern würde. Bei Verwendung eines Control-Charts war dies der Wert, der zwischen den Kontrollgrenzen liegen musste. Wurden diese verlassen, fand eine Neuberechnung der Fallzahl statt. Wurde kein Control-Chart verwendet, wurde in *jedem* Fall eine Neuberechnung der Fallzahl durchgeführt.

2.2.2.2 Neuberechnung der Fallzahl

Aus Perfomance-Gründen wurde nicht die `SAS`-Prozedur `PROC POWER` verwendet, sondern es wurde eine eigene Routine zur Fallzahlschätzung programmiert, welche genau wie die Berechnung von Δ mit der Formel aus [60] die gesuchte Fallzahl n liefern sollte. Dieses Mal wurde aus der Mittelwertdifferenz Δ, der gewünschten Power 85% und dem Signifikanzniveau 5% die Fallzahl berechnet.

Bei der Neuberechnung wurde dem Vorschlag aus [60] gefolgt und Gleichung (4) aus diesem Paper zur ersten Näherung verwendet, welche eine Fallzahl liefert, die allgemein etwas zu niedrig ist:

$$n_A = \frac{(r+1)(Z_{1-\beta} + Z_{1-\alpha/2})^2\sigma^2}{rd^2} \rightsquigarrow n_{\text{init}} = \frac{2(Z_{1-\beta} + Z_{1-\alpha})^2\hat{\sigma}_i^2}{\Delta^2}.$$

Auf der linken Seite steht wie oben n_A für die Größe einer Gruppe, r für das Verhältnis der Größen den beiden Gruppen, Z_q für das q-Quantil der Standardnormalverteilung. Dies kann analog für den vorliegenden Fall vereinfacht werden. Mit n_{init} wird der Startwert für die Fallzahlkalkulation, mit σ_i^2 der beim i-ten Varianzmonitoring (korrigiert oder unkorrigiert) bestimmte Varianzschätzer und mit Δ der vorgegebene Gruppenunterschied bezeichnet. Vom Startwert ausgehend wurde dann die Fallzahl iterativ erhöht und jeweils mit Formel (7) aus [60] geprüft, ob die benötigte Power schon erreicht war. Dafür sind meist nur sehr wenige Iterationsschritte nötig, da n_{init} nah an der benötigten Fallzahl liegt.

2.2.2.3 Spezifikation der Varianz

Es wurden konstante Varianzen angenommen, wobei zwei Fälle unterschieden werden können: Einerseits der Idealfall, in dem die tatsächliche Varianz (=1) auch der entspricht, die bei der Planung der Studie angenommen wurde. Auf der anderen Seite wurden Fehlspezifikationen angenommen, bei denen die wirkliche Varianz 0.5 oder 2 betrug, aber wie vorher eine Varianz von 1 angenommen wurde. In diesen Fällen wurde die Varianz um den Faktor 1/2 bzw. 2 falsch angenommen, die wahren Standardabweichungen wurden damit um ca. 30% unter- bzw. 40% überschätzt.

2.2.2.4 Zusammenfassung der Szenarien für stetige Endpunkte

Tabelle 2.2 beschreibt zusammenfassend die einzelnen Szenarien der Simulationen zu den stetigen Endpunkten. Da die Untersuchung der Einhaltung des Niveaus der Prozeduren galt, waren die Mittelwerte der normalverteilten Endpunkte für beide Gruppen und zu allen Zeitpunkten immer gleich Null. Die Größen der beiden Behandlungsgruppen waren immer gleich, nur die Gesamtgröße der Studie ergab sich erst dynamisch durch die Rekalkulation. In der Simulation wurden Studien mit durch acht teilbaren initialen Gesamtgrößen von 96, 200, 304 und 400 Patienten verwendet.

Zur Notation: Die Verteilung der Zufallsgrößen in beiden Behandlungsgruppen waren Normalverteilungen

$$\mathcal{L}_{\text{G}} = \mathcal{N}(\mu_{\text{G}}, v_{\text{G}}),$$

wobei G für die BehandlungsGruppe steht und die Werte T für Treatment / Verum / Behandlungsgruppe bzw. C für Control / Placebo / Kontrollgruppe annehmen kann. Damit lassen sich die Bedingungen für die Simulation schreiben als $n_T = n_C$ (Behandlungsgruppen immer gleich groß), $\mu_T = \mu_C = \mu_G = 0$ (kein Gruppenunterschied) und $v_T = v_C = v_G$ (gleiche Varianzen in den Behandlungsgruppen).

Die Szenarien der Simulation ergaben sich aus den Kombinationen der verschiedenen initialen Fallzahlen und Varianzspezifikationen (s. Tabelle 2.2).

Szenario	N_{init}	v_{ann}	v_{real}
1	96	1	1
2	200	1	1
3	304	1	1
4	400	1	1
5	96	1	0.5
6	200	1	0.5
7	304	1	0.5
8	400	1	0.5
9	96	1	2
10	200	1	2
11	304	1	2
12	400	1	2

Tabelle 2.2:
Szenarien bei normalverteilten Endpunkten. N_{init} steht für die initial auf Basis der angenommenen Varianz $v_{ann} = 1$ geplante Gesamtfallzahl, v_{real} steht für die reale Fallzahl. Der erste Block fasst alle Szenarien mit Verwendung der korrekten Varianz für die verschiedenen Studiengrößen zusammen. Die anderen Blöcke stellen die Über- und Unterschätzung der Varianz für die verschiedenen Studiengrößen dar.

2.2.3 Details zu den dichotomen Endpunkten

In diesem Abschnitt werden zusätzliche Informationen zur Simulation der dichotomen Endpunkte gegeben. Es wurde wieder von zwei Gruppen ausgegangen, die jedoch hinsichtlich ihrer Anteile π_1, π_2 an positiven Ergebnissen der binären Zielvariable verglichen werden sollten. Die verwendete Teststatistik war die χ^2-Statistik nach Pearson, die zu einer zweiseitigen Alternative hin ausgewertet wurde, d.h. formal wurde die Nullhypothese $H_0 : \pi_1 = \pi_2$ gegen die Alternative $H_1 : \pi_1 \neq \pi_2$ getestet.

2.2.3.1 Die zentrale Formel zur Fallzahl- und Powerberechnung

Wie bei den stetigen Endpunkten wurde auch hier zur Bestimmung von δ, N und der Power immer die gleiche Formel verwendet. Als Grundlage diente dabei [22], in deren Notation die

Formel (1) (e.b.d.) für die Fallzahl folgendermaßen lautet:

$$n = \left(2 + \theta + \frac{1}{\theta}\right) \cdot \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^{*2}} \cdot \pi(1-\pi) \rightsquigarrow 4\frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\Delta^{*2}} \cdot \pi(1-\pi)$$

wobei nun n die Gesamtfallzahl der Studie bezeichnet, θ das Verhältnis der beiden Gruppengrößen, z_q wieder das q-Quantil der Standardnormalverteilung, Δ^* die nachzuweisende Differenz der beiden Proportionen und π die *Gesamtereigniswahrscheinlichkeit* (*pooled event rate*, PER) ist. Letztere kann aus den Einzelwahrscheinlichkeiten durch $\pi = (\pi_1 + \theta\pi_2)/(1+\theta)$ berechnet werden. Die vereinfachte Formel nach dem Pfeil resultiert aus der Verwendung gleicher Gruppengrößen ($\theta = 1$).

2.2.3.2 Alternativhypothese, Powerberechnung und Neuberechnung der Fallzahl

Auch für die binären Endpunkte wurden zuerst n, π, α und β vorgegeben, um Δ^* zu bestimmen, was durch einfaches Freistellen der obigen Gleichung erreicht wurde. Für n wurde die jeweils vorher festgelegte initiale Fallzahl verwendet (96, 200, 304, 400). Nach Bestimmung dieses Wertes wurde die Rekrutierung unter Gültigkeit der Nullhypothese simuliert. Das bedeutete, dass unter den verschiedenen PERs für die beiden Einzelwahrscheinlichkeiten galt: $\pi_1 = \pi_2 = \pi =$PER.

Während eines Monitoringzeitpunktes wurde die PER durch p geschätzt, was offensichtlich unter Aufrechterhaltung der Verblindung möglich war. Die zentrale Fallzahlformel wurde nach $z_{1-\beta}$ aufgelöst:

$$z_{1-\beta} = \sqrt{\frac{n\Delta^{*2}}{4p(1-p)}} - z_{1-\alpha/2},$$

woraus unter Anwendung der Verteilungsfunktion der Standardnormalverteilung die Power $1-\beta$ berechnet wurde. Bei Verwendung eines Control-Charts konnte diese wiederum mit den Kontrollgrenzen verglichen werden. Bei Verlassen der Kontrollgrenzen oder bei unbedingter Neuberechnung der Fallzahl wurde π in der zentralen Formel durch die aktuelle Schätzung p der PER ersetzt.

2.2.3.3 Zusammenfassung der Szenarien für binäre Endpunkte

Wie für die stetigen Endpunkte folgt eine Aufstellung der einzelnen Simulationsparameter bei den binären Endpunkten. Die Nullhypothese gleicher Anteile in beiden Gruppen sollte auch bei diesen immer gelten. Die Variabilität wird durch die *pooled event rate*, PER, quantifiziert, welche über die Studie hinweg konstant blieb. Die Gruppengrößen waren wieder identisch zu denen bei den stetigen Endpunkten. Der Index G soll auch hier die Ausprägungen T und C annehmen können. Die Eintrittswahrscheinlichkeit beim Bernoulli-Experiment wird mit π_{G} und die Größe der Stichprobe mit n_{G} bezeichnet. Es wurden gleich große Stichproben für die Behandlungs- und Kontrollgruppe gezogen, d.h. $n_{\mathrm{T}} = n_{\mathrm{C}}$.

Zu jeder der Studiengrößen 96, 200, 304 und 400 und PERs von 0.5, 0.6 und 0.7 wurde die korrespondierende Differenz δ der Proportionen bestimmt und damit die Simulation durchgeführt (Tabelle 2.3, Szenarien 1 bis 12).

Um eine Fehleinschätzung der PER zu simulieren, wurde eine Studienplanung angenommen, die auf der Annahme der mittleren PER (und dem zugehörigen δ) basierte, bei der aber in der Simulation eine der beiden anderen PERs verwendet wurde. Bei Verwendung der PER zum kleineren δ wäre die Variabilität überschätzt worden, bei der zum größeren δ wäre sie unterschätzt worden.

Beispiel: Bei der Planung einer Studie sei $\delta = 0.3$ bei einer PER von 0.6 wie in Szenario 2 angenommen worden, dagegen wurden Daten wie in Szenario 1 simuliert, bei dem eine höhere PER, d.h. eine niedrigere Varianz vorherrschte. Die Planung wäre demnach von einer zu hohen Variabilität und damit einer zu hohen Fallzahl ausgegangen.

Die Fälle mit den Fehleinschätzungen werden in Tabelle 2.3 durch die Szenarien 13 bis 20 repräsentiert.

Szenario	N_{init}	PER	δ
1	96	0.7	0.280
2	96	0.6	0.300
3	96	0.5	0.306
4	200	0.7	0.194
5	200	0.6	0.208
6	200	0.5	0.212
7	304	0.7	0.158
8	304	0.6	0.168
9	304	0.5	0.172
10	400	0.7	0.137
11	400	0.6	0.147
12	400	0.5	0.150
13	96	0.7	0.300
14	96	0.5	0.300
15	200	0.7	0.208
16	200	0.5	0.208
17	304	0.7	0.168
18	304	0.5	0.168
19	400	0.7	0.147
20	400	0.5	0.147

Tabelle 2.3:
Szenarien bei binomial verteilen Endpunkten

Anders als bei den stetigen Endpunkten kann bei den binären nicht sofort die tatsächlich benötigte Fallzahl bei Fehlspezifikation der PER abgeleitet werden. Diese Angaben folgen in der Tabelle 2.4.

Parameter	angenommen	bei Überschätzung	bei Unterschätzung
PER	0.6	0.7	0.5
Varianz	0.24	0.21	0.25
N	96	83.8	99.8
N	200	174.3	207.5
N	304	267.2	318.1
N	400	349.0	415.5

Tabelle 2.4:
Angenommene und tatsächlich benötigte Fallzahlen bei Fehlspezifikation im Falle binärer Endpunkte. Über- bzw. Unterschätzung bezieht sich auf die Varianz.

2.3 Ein Kombinationstest bei geänderter Variabilität

Dieser Abschnitt stellt die Methodik aus dem Artikel von Lösch und Neuhäuser [34] dar, in welchem untersucht wurde, wie sich verschiedene Testprozeduren hinsichtlich des α-Fehlers und der Power verhalten, wenn während einer klinischen Studie ein Amendment zum Prüfplan notwendig wurde. Dazu wurde wieder eine Simulationsstudie verwendet, in der verschiedene Situationen untersucht wurden. Diese Situationen ergaben sich aus zwei Szenarien bei der Rekrutierung, verschiedenen Anstiegen der Varianz und mehreren Möglichkeiten des Verlaufes der Mittelwerte über die Zeit.

Jede simulierte Studie sollte die Situation eines Zweigruppenvergleiches auf Überlegenheit bei einer normalverteilten Zielgröße nachbilden. Dabei sollte die Hypothese einseitig getestet werden, d.h.

$$H_0 : \mu_\mathrm{T} = \mu_\mathrm{C} \qquad \text{gegen} \qquad H_1 : \mu_\mathrm{T} > \mu_\mathrm{C},$$

wobei T für *Treatment* (Verum, Novum, Behandlung), C für *Control* (Standard, Placebo, Kontrolle) und μ_G für den Erwartungswert der jeweiligen Gruppe G steht.

In allen Fällen wurde zu einem bestimmten Zeitpunkt der Einfluss eines Amendments simuliert - es ergibt sich also im Gegensatz zu den Untersuchungen der Adaptionsprozeduren eine Änderung der Verteilung *im Lauf* der Studie. Das Amendment sollte Einfluss auf die Rekrutierung nehmen und zwar über die Änderung der Ein-/Ausschlusskriterien. Nach dem Amendment konnte sich sowohl die Varianz als auch der Mittelwert in den Behandlungsgruppen ändern, wobei auch der Fall ohne Änderung betrachtet wurde. Durch den Zeitpunkt des Amendments wurde die Studie in zwei Hälften geteilt, die nachfolgend als *Phasen* bezeichnet werden. Somit sind nur die Werte jeder Gruppe innerhalb einer Phase normalverteilt und die Werte einer Gruppe über die gesamte Studie hinweg folgen einer Mischung zweier Normalverteilungen.

Dann wurden die verschiedenen zu untersuchenden statistischen Tests auf die so erzeugten Daten angewendet und der Anteil an Signifikanzen bestimmt. Im Falle gleicher Mittelwerte erhielt man so ein empirisches α-Level, im Falle unterschiedlicher Mittelwerte die empirische Trennschärfe des Tests.

2.3.1 Details zur Simulation

Für die Simulation wurde der Zufallzahlengenerator `RANNOR` der Statistiksoftware `SAS 9.1` verwendet. Die Anzahl der Replikationen war beim Ermitteln des empirischen Fehlers 1. Art immer 10^5, während bei den Untersuchungen zur Trennschärfe immer 10^4 verwendet wurde.

Es gab zwei Szenarien, die untersucht wurden: Entweder wurde das Amendment in der Mitte der Studie (Szenario 1), d.h. nach der Hälfte der Gesamtfallzahl wirksam oder schon nach 1/3 der Gesamtfallzahl (Szenario 2). Das zweite Szenario soll den Fall einer anfänglich schleppenden Rekrutierung darstellen, in dem durch das Amendment eine Beschleunigung erfolgreich bewirkt wird.

Die Fallzahlen in den simulierten Studie waren wie folgt:

$n_{\mathrm{b}}^{\mathrm{C}}$: Anzahl Patienten in der Kontrollgruppe vor dem Amendment (**b**efore)
$n_{\mathrm{b}}^{\mathrm{T}}$: Anzahl Patienten in der Behandlungsgruppe vor dem Amendment (**b**efore)
$n_{\mathrm{a}}^{\mathrm{C}}$: Anzahl Patienten in der Kontrollgruppe nach dem Amendment (**a**fter)
$n_{\mathrm{a}}^{\mathrm{T}}$: Anzahl Patienten in der Behandlungsgruppe nach dem Amendment (**a**fter)

Damit sind die Szenarien festgelegt durch:

Szenario 1: $n_{\mathrm{b}}^{\mathrm{C}} = n_{\mathrm{b}}^{\mathrm{T}} = n_{\mathrm{a}}^{\mathrm{C}} = n_{\mathrm{a}}^{\mathrm{T}} = 50$
Szenario 2: $n_{\mathrm{b}}^{\mathrm{C}} = n_{\mathrm{b}}^{\mathrm{T}} = 25$ und $n_{\mathrm{a}}^{\mathrm{C}} = n_{\mathrm{a}}^{\mathrm{T}} = 50$

Der Varianzanstieg in der zweiten Phase wurde als Vielfaches zur ersten Phase formuliert und es wurden Varianzinflationsfaktoren von 1 (keine Veränderung), 1.5, 2, 2.5 bzw. 3 verwendet, was zu einer Steigerung der Standardabweichung um 0, 22.5, 41.4, 58.1 bzw. 73.2% führte.

Die angesetzten Konfigurationen $(\mu_{\mathrm{b}}^{\mathrm{T}}, \mu_{\mathrm{a}}^{\mathrm{T}}, \mu_{\mathrm{b}}^{\mathrm{C}}, \mu_{\mathrm{a}}^{\mathrm{C}})$ der Mittelwerte können in vier Gruppen eingeteilt werden:

1. zur Untersuchung des Fehlers 1. Art: $(\mu_{\mathrm{b}}^{\mathrm{T}}, \mu_{\mathrm{a}}^{\mathrm{T}}, \mu_{\mathrm{b}}^{\mathrm{C}}, \mu_{\mathrm{a}}^{\mathrm{C}}) = (0, 0, 0, 0)$

2. zur Untersuchung der Trennschärfe im Falle, dass alle Mittelwerte konstant sind: $(\mu_{\mathrm{b}}^{\mathrm{T}}, \mu_{\mathrm{a}}^{\mathrm{T}}, \mu_{\mathrm{b}}^{\mathrm{C}}, \mu_{\mathrm{a}}^{\mathrm{C}}) = (s, s, 0, 0)$, d.h. bei einer konstanten Mittelwertdifferenz s=0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.40, 0.50, 0.60, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, und 1.00 zwischen der Behandlungs- und der Kontrollgruppe

3. zur Untersuchung der Trennschärfe im Falle, dass die Mittelwerte zwar nicht konstant über die Phasen hinweg sind ($d = \mu_{\mathrm{a}}^{\mathrm{T}} - \mu_{\mathrm{b}}^{\mathrm{T}} = \mu_{\mathrm{a}}^{\mathrm{C}} - \mu_{\mathrm{b}}^{\mathrm{C}} \neq 0$), dass es aber die Gruppendifferenz ist ($s = \mu_{\mathrm{b}}^{\mathrm{T}} - \mu_{\mathrm{b}}^{\mathrm{C}} = \mu_{\mathrm{a}}^{\mathrm{T}} - \mu_{\mathrm{a}}^{\mathrm{C}} \neq 0$). Dies führt also zu einer Konfiguration $(\mu_{\mathrm{b}}^{\mathrm{T}}, \mu_{\mathrm{a}}^{\mathrm{T}}, \mu_{\mathrm{b}}^{\mathrm{C}}, \mu_{\mathrm{a}}^{\mathrm{C}}) = (s, s+d, 0, d)$, wobei für s und d die Werte 0.1, 0.5 und 1.0 verwendet wurden.

4. zur Untersuchung der Trennschärfe, wenn im Vergleich zum vorhergehenden Punkt auch die Mittelwertdifferenz sich über die Phasen hinweg verändern darf. Dabei wurden die Spezialfälle
   $(\mu_{\mathrm{b}}^{\mathrm{T}}, \mu_{\mathrm{a}}^{\mathrm{T}}, \mu_{\mathrm{b}}^{\mathrm{C}}, \mu_{\mathrm{a}}^{\mathrm{C}}) = (0.5, 0.2, 0, 0)$ im Szenario 1 und
   $(\mu_{\mathrm{b}}^{\mathrm{T}}, \mu_{\mathrm{a}}^{\mathrm{T}}, \mu_{\mathrm{b}}^{\mathrm{C}}, \mu_{\mathrm{a}}^{\mathrm{C}}) = (0.7, 0.2, 0, 0)$
   im Szenario 2 untersucht.

2.3.2 Die untersuchten Tests

Es wurden drei statistische Tests verglichen: Der Pooling-Test, der Kombinationstest und ein modifizierter Kombinationstest mit Zusatzbedingung.

2.3.2.1 Der Pooling-Test

Der erste und einfachste Test besteht darin, die evtl. unterschiedlichen Verteilungen zu ignorieren und einen herkömmlichen Zweigruppenvergleich mit Hilfe des t-Tests durchzuführen. Diese Möglichkeit ist wahrscheinlich die in der Praxis am meisten verwendete und sie soll im Folgenden als *Pooling-Strategie* bezeichnet werden. Die Nullhypothese, die durch diesen Test geprüft wird, ist die anfangs genannte einseitige Fragestellung.

2.3.2.2 Der Kombinationstest

Der nächste Ansatz war der, dass das nach dem Amendment rekrutierte Patientenkollektiv eigenständig mit Hilfe des einseitigen t-Tests ausgewertet und aus den zwei Testentscheidungen für die einzelnen Phasen eine für das gesamte Kollektiv abgeleitet wird. Formal ergeben sich dadurch zwei Nullhypothesen

H_0^{b}: $\mu_{\mathrm{T}}^{\mathrm{b}} = \mu_{\mathrm{C}}^{\mathrm{b}}$ gegen $H_1^{\mathrm{b}} : \mu_{\mathrm{T}}^{\mathrm{b}} > \mu_{\mathrm{C}}^{\mathrm{b}}$ und
H_0^{a}: $\mu_{\mathrm{T}}^{\mathrm{a}} = \mu_{\mathrm{C}}^{\mathrm{a}}$ gegen $H_1^{\mathrm{a}} : \mu_{\mathrm{T}}^{\mathrm{a}} > \mu_{\mathrm{C}}^{\mathrm{a}}$
und eine globale Nullhypothese
H_0^{ab}: H_0^{b} *und* H_0^{a} gegen H_1^{ab}: H_1^{b} *oder* H_1^{a}

Eine sehr bekannte Möglichkeit, die globale Hypothese zu testen, bietet Fishers Kombinationstest (s. z.B. [61, 62]). Die Notation auch [61] verwendend, ist dieser Test so begründet: Liegen die p-Werte $p_1, \ldots, p_k$ von k unabhängigen Testentscheidungen vor, dann können diese durch Produktbildung kombiniert werden. Wie in [61] weiter ausgeführt wird, gibt es nun zwei entscheidende Fakten: Erstens ist der p-Wert selbst eine zufällige Größe und unter der Nullhypothese des angewendeten Tests uniform auf dem Intervall [0,1] verteilt. Zweitens gibt es einen Zusammenhang zwischen der uniformen Verteilung und der χ^2-Verteilung. Dieser lautet: Wenn eine Zufallsvariable U uniform verteilt ist, dann folgt der Term $-2 \log U$ einer χ_2^2-Verteilung. Weiterhin ist die Summe unabhängiger χ^2-verteilter Zufallsvariablen wieder χ^2-verteilt ist und zwar mit der Summe der Freiheitsgrade als neuem Freiheitsgrad. Daraus folgt, dass unter der Nullhypothese der Term

$$K := -2 \log(p_1 \cdot \ldots \cdot p_k) = -2 \log p_1 - \cdots - 2 \log p_k$$

unter H_0^{ab} eine χ_{2k}^2-Verteilung aufweist. Ein signifikanter Kombinationstest zum Niveau α ergibt sich also, wenn die Größe K das $1-\alpha$-Quantil der χ_{2k}^2-Verteilung überschreitet.

In der vorliegenden Untersuchung wurde die klinische Studie nur in zwei Teile zerlegt, daher lag immer der Kombinationstest für $k = 2$ vor. Die Vorgehensweise lässt sich leicht für mehr als ein Amendment (d.h. mehr als zwei Phasen) übertragen. Die verwendete Prozedur soll im Folgenden mit *Kombination* bezeichnet werden.

2.3.2.3 Der modifizierte Kombinationstest

Der letzte betrachtete Test war eine Modifikation des Kombinationstests. Da ein Kombinationstest signifikant werden kann, obwohl kein einziger der einzelnen p-Werte für sich betrachtet signifikant ist, wurde zusätzlich zu einem signifikanten Ergebnis im Kombinationstest auch Signifikanz mindestens eines Einzeltests (einer Phase) gefordert. Dieses Konzept kann dabei helfen, der Forderung von regulatorischer Seite nach einem Nachweis der Wirksamkeit in mindestens einer Phasen gerecht zu werden. Es entsteht dadurch jedoch eine Situation mit multiplen Tests, da jetzt nicht mehr nur die globale Hypothese H_0^{ab} des Kombinationstests, sondern auch die zwei Elementarhypothesen H_0^{b} und H_0^{a} getestet werden.

Unter solchen Umständen ist darauf zu achten, dass die *familienweise* Fehlerwahrscheinlichkeit erster Art nicht vergrößert wird. Bei der beschriebenen Vorgehensweise ist jedoch die Struktur eines Abschlusstests [63] erkennbar, bei dem folgende Vorgehensweise [64] angemessen ist: Es ist zunächst die Schnitthypothese H_0^{ab} auf dem vorgegebenen α-Niveau zu prüfen, anschließend werden, sofern die Schnitthypothese zum (lokalen) α-Niveau abgelehnt werden konnte, die Einzelhypothesen H_0^{b} und H_0^{a} ebenfalls zum (lokalen) α-Niveau getestet. Die Nullhypothese H_0^{a} kann zum Niveau α abgelehnt werden, wenn sich für den Test gegen H_0^{ab} als auch den gegen H_0^{a} ein p-Wert kleiner oder gleich α ergibt. Entsprechendes gilt für die Hypothese H_0^{b}.

3 Ergebnisse

3.1 Reale Studien

3.1.1 Beschreibung der analysierten Studien

Die für diese Arbeit erstellte Datenbank an Studien wurde aus einer internen Projektdatenbank und einer Literaturdatenbank gewonnen.

In der Datenbank waren insgesamt 534 Projekte verzeichnet, von denen 483 als Forschungsprojekte identifiziert werden konnten. Insgesamt konnten 204 von diesen wiederum als mögliche Kandidaten markiert wurden. Nur 40 davon erfüllten alle Kriterien, die benötigt wurden, um die Studie analysieren zu können. In Tabelle 3.1 sind die wichtigsten Merkmale zusammengefasst.

Mit 29 der 40 Studien machen die randomisierten, kontrollierten Studien den Hauptanteil aus. Es wurden neun patientenbasierte Beobachtungsstudien und zwei epidemiologische Studien aufgenommen.

Es gab eine hohe Anzahl an Interventionsstudien, die keiner Phase der Medikamentenentwicklung zugeordnet werden konnten. Dies sind Studien, die in der prä-GCP-Ära durchgeführt wurden oder Studien, deren Daten durch Internetrecherche gefunden wurden. Bei älteren Studien war anhand der Dokumente oft nicht mehr festzustellen, welcher Phase sie (wenn überhaupt einer) zugeordnet waren. Es wurde für die Auswertung nur dann eine Phase zugeordnet, wenn dies im Protokoll, dem Bericht oder einer Publikation zur Studie explizit erwähnt wurde. Bei den Internetdaten ergab teilweise selbst weitergehende Nachforschung (z.B. E-Mail-Anfragen) keine zusätzlichen Informationen zur Phase der Studie. Bei den zwei Studien mit unbekannter Randomisierung handelte es sich um diejenigen Internet-Daten, bei deren Beschreibung diese Information fehlte und auch eine Nachfrage kein Ergebnis erbrachte. Da diese beiden Studien aber Placebo als Kontrolltherapie verwendeten, war davon auszugehen, dass sie zumindest einfach blind waren.

Alle placebokontrollierten Studien waren verblindet (bis auf eine, deren Status unbekannt war), und alle aktiv kontrollierten waren unverblindet, so dass eine eindeutige Zuordnung des Effekts zu einem der beiden Kriterien nicht möglich war.

3.1.2 Rekrutierung

Da die Geschwindigkeit der Rekrutierung und ihre Gleichmäßigkeit als Einflussfaktoren für Varianzunterschiede im Verlauf der Studien betrachtet wurden, werden sie an dieser Stelle abweichend von der Reihenfolge im Methodenteil zuerst behandelt.

Merkmal der Studie	Ausprägung	Anzahl	Prozent
Art der Studie	IS*, RCT	29	72.5
	NIS**, pat.-basiert	9	22.5
	NIS**, bev.-basiert	2	5.0
ab hier nur Interventionsstudien			
Entwicklungsphase	unbekannt	11	37.9
	nicht zutreffend	9	31.0
	III	5	17.2
	IV	4	13.8
Chronizität	akut	3	10.3
	chronisch	25	86.2
	beides möglich	1	3.4
Anzahl Zentren	unizentrisch	9	31.0
	$2-10$ Zentren	12	41.4
	$11-$ Zentren	8	27.6
Randomisierung	unbekannt	2	6.9
	ja	27	93.1
Verblindung	unbekannt	1	3.4
	ja	21	72.4
	nein	7	24.1
Jahr der Studie	vor GCP	14	48.3
	seit GCP	15	51.7
Art der Kontrollen	Placebo	22	75.9
	aktiv	7	24.1
Amendment	unbekannt	2	6.9
	nein	21	72.4
	ja	6	20.7
Art der Intervention	Medikament	23	79.3
	sonst	6	20.7
Anzahl Arme	2	27	93.1
	3	2	6.9
Rekrutierungsdauer	bis 1 Jahr	9	31.0
	mehr als 1 Jahr	20	69.0

* IS: Interventionsstudie
** NIS: Nichtinterventionsstudie

Tabelle 3.1:
Beschreibung der Metadaten der untersuchten Studien

3.1.2.1 Rekrutierungsgeschwindigkeit

Die Rekrutierungsgeschwindigkeit über alle Studien hinweg war sehr variabel, s. Tabelle 3.2. Der Median der Geschwindigkeit entsprach einer Rate, bei der ca. ein Proband in drei Tagen rekrutiert wurde. Die langsamste Rekrutierung wurde mit durchschnittlich einem Patienten in 37 Tagen bei einer Nichtinterventionsstudie, die schnellste mit über 5 Patienten pro Tag bei einer Interventionsstudie bestimmt.

Bei der Betrachtung der Geschwindigkeit wurden untere und obere Quartile berechnet und dann der Anteil an Studien im jeweiligen Quartil bestimmt. Die Quartile bezogen sich einerseits auf die Gesamtmenge der Studien (für die Tabellenzeilen *Gesamt* und *Art der Studie*) oder auf die Subgruppe Nichtinterventionsstudien (für alle anderen Parameter). Bei gleichmäßiger Verteilung innerhalb eines Parameters sollte der Anteil immer bei 25% liegen, so wie es in der Zeile *Gesamt* der Fall ist.

Bei der Untersuchung der Eigenschaft **Art der Studie** fielen aus der Gruppe der Interventionsstudien nur 17% in das obere Quartil, während es bei den Nichtinterventionsstudien fast die Hälfte war. Die Interventionsstudien und die Nichtinterventionsstudien unterschieden sich bei der medianen Rekrutierungsgeschwindigkeit in dem Ausmaß (Anstieg um 30%), dass der Fisher-Pitman-Permutationstest einen p-Wert von 0.0949 aufwies. Die Nichtinterventionsstudien zeigen zwar eine größere mittlere und mediane Rekrutierungsgeschwindigkeit, die Variabilität in dieser Gruppe war jedoch auch höher. Die Unterschiede in der Geschwindigkeit können nicht durch eine größere Anzahl Zentren erklärt werden, denn das Hauptgewicht bei den Nichtinterventionsstudien lag mit 55% bei den unizentrischen Studien, während bei den Interventionsstudien diese Klasse nur 30% ausmachte. Ein möglicher Grund könnte sein, dass es zwar einen vergleichbaren Anteil an „kurz“ (max. 1 Jahr) rekrutierenden Studien gab (30 bzw. 36%), bei dem aber bei den Interventionsstudien naturgemäß weniger Patienten rekrutiert werden mussten. Im Mittel waren dies nämlich 145, während bei den kurz rekrutierenden Nichtinterventionsstudien durchschnittlich 452 Patienten rekrutiert wurden.

Bei der **Entwicklungsphase** ergab sich, dass von den fünf Phase-III-Studien nur eine in das obere Quartil bei der Geschwindigkeit fiel, dass aber dafür keine einzige im unteren Quartil zu finden war. Bei den Phase IV-Studien waren je 50% in den beiden Quartilen vertreten, allerdings waren dies auch nur 4 Studien und die Schätzer damit nicht sehr verlässlich. Phase-IV-Studien wiesen sowohl einen höheren Median (Faktor 3) als auch Mittelwert (Faktor 4) auf. Der t- und der FP-Test waren nicht signifikant ($p = 0.39, 0.26$). Dies kann auch an den starken Streuungsunterschieden in den beiden Gruppen liegen: Der Quartilsabstand bei den Phase-IV-Studien war 24fach höher und die Standardabweichung um den Faktor 6.

Bei der Eigenschaft der **Chronizität** der Indikation konnte ein Anstieg der medianen Geschwindigkeit von den akuten zu den akut/chronischen hin zu den chronischen Erkrankungen beobachtet werden. Betrachtete man nur die beiden extremen Gruppen, ergab sich ein Anstieg auf das Dreifache beim Median und auch beim Mittelwert. Sowohl der t-Test als auch der FP-Test blieben jedoch nicht signifikant ($p = 0.47, 0.46$). Die Anteile in den Quartilen waren angesichts der jeweiligen Fallzahlen nicht bedeutsam von 25% entfernt.

Der Verlauf über die Kategorien für die **Anzahl Zentren** war nicht eindeutig: Während die

unizentrischen Studien mit einer medianen Geschwindigkeit von 0.23 Patienten/Tag rekrutierten, war es bei den Studien mit zwei bis 10 Zentren nur noch 0.16 und bei solchen mit 11 oder mehr Zentren mit 0.54 wieder größer. Keiner der durchgeführten Tests auf Gruppenunterschiede wurde jedoch signifikant. Betrachtete man die Korrelation zwischen der Anzahl Zentren und der Rekrutierungsgeschwindigkeit, ergab sich für die Pearsonkorrelation ein Wert von 0.39 ($p = 0.0127$) und eine Spearmankorrelation von 0.42 ($p = 0.0075$). Dieser Widerspruch zur kategorisierten Variante des Vergleiches ist mit dem damit verbundenen Informations- und damit Powerverlust zu erklären. Es ist zu vermuten, dass die Rekrutierungsgeschwindigkeit mit der Anzahl der Zentren zunimmt, was auch unmittelbar einleuchtet. Weiterhin deuten die Anteile, die in das obere bzw. untere Quartil fallen in eine ähnliche Richtung: Während die unizentrischen verhältnismäßig oft bei den langsam rekrutierenden vertreten sind und die Studien mit 2-10 Zentren seltener bei den schnellen zu finden sind, gilt für die Studien mit 11 oder mehr Zentren, dass sie zur Hälfte bei den schnell rekrutierenden sind, aber gar nicht bei den langsamen einsortiert wurden.

Beim Kriterium **Randomisierung** fiel keine Studie in die Referenzgruppe, so dass sich hier nur der globale Trend wiederholt.

Für den Einflussfaktor **Verblindung** konnte eine um die Hälfte höhere mediane und nahezu doppelt so große mittlere Rekrutierungsgeschwindigkeit der verblindeten Studien beobachtet werden, die jedoch nicht signifikant höher war (t-, FP-Test: $p = 0.26, 0.45$). Die Verteilung auf die Gruppen der aktiv und placebokontrollierten Studien war nahezu identisch zu der Verteilung aufgrund der Art der Verblindung, so dass die gleichen Ergebnisse für die **Art der Kontrollen** gelten. Die Anteile in den Quartilen der schnellen und langsamen Studien entsprachen dem erwarteten Viertel.

Das **Jahr der Studie** –eingeteilt nach GCP-Ära– scheint höchstens eine schwache Rolle bei der Geschwindigkeit der Rekrutierung gespielt zu haben: die Anteile in den beiden äußeren Quartilen waren nah an 25%, es gab aber beim Übergang zu den jüngeren Studien eine leichte Verschiebung zu schneller Rekrutierung. Die medianen Geschwindigkeiten lagen nah beieinander und der FP-Test lieferte einen p-Wert von 0.51. Hob man die Kategorisierung der Jahre auf, wiesen sowohl die Pearson-Korrelation als auch die Spearman-Korrelation einen Wert von 0.31 auf, beide p-Werte waren ca. 0.1, so dass sich auch hier höchstens leichter Zusammenhang abzeichnet, bei dem ein späteres Jahr der Studie mit einer höheren Rekrutierungsgeschwindigkeit einher ging.

Es waren drei der sechs Studien mit **Amendment** unter den schnell rekrutierenden und keine bei den langsam rekrutierenden. Bei den Studien ohne Amendment waren dagegen nur 1/5 unter den schnellen und 1/3 bei den langsamen. Der Grund könnte eine intendierte Beschleunigung der Rekrutierung durch das Amendment sein. Die mediane Rekrutierungsgeschwindigkeit bei den Studien mit Amendment war dreifach erhöht, die mittlere war hingegen identisch. Trotz dieser Unterschiede wurde weder der t- noch der FP-Test signifikant (beide: $p = 0.99$).

Die **Art der Intervention** schien keine Rolle zu spielen, denn die Anteile in den Quartilen waren nahe bei 25% (man beachte aber die geringe Zahl der Studien, bei denen kein Medikament untersucht wurde) und die mediane Rekrutierungsgeschwindigkeit war nahezu

identisch, die mittlere um 50% erhöht.

Bei der **Anzahl der Behandlungsgruppen** fiel auf, dass die beiden dreiarmigen Studien zum Quartil der langsam rekrutierenden gehörten, eine Tatsache, die sich auch bei der medianen und mittleren Geschwindigkeit zeigte: der Wert belief sich auf 0.06 Patienten pro Tag und stellt somit das Minimum aller Rekrutierungsgeschwindigkeiten bei den Interventionsstudien dar. Allerdings war die Fallzahl in dieser Gruppe sehr klein. Bei den zweiarmigen Studien waren die Anteile in den Quartilen wie erwartet.

Bei der **Dauer der Studie** dagegen zeigten sich deutliche Unterschiede hinsichtlich der Rekrutierungsgeschwindigkeit. Die Studien mit Rekrutierungsdauer bis maximal ein Jahr waren mit einem Anteil von 67% im oberen Quartil der Rekrutierungsgeschwindigkeit vertreten, während es bei den länger rekrutierenden nur 5% waren, der FP-Test auf Gruppenunterschiede ergab einen p-Wert von 0.0067. Wählte man eine feinere Unterteilung der Rekrutierungsdauer, zeichnete sich folgender Trend ab: bei den Studien mit einer Dauer von maximal drei Monaten waren alle im oberen Quartil der Rekrutierungsgeschwindigkeit, bei restlichen bis zu 12 Monaten waren es noch 29%, bei den verbleibenden bis zu drei Jahren nur noch 14% und bei den anderen war es keine einzige mehr. Ein entsprechend entgegengesetzter Trend konnte für die Zugehörigkeit zum langsamen Quartil der Studien beobachtet werden. Der Pearson-Korrelationskoeffizient der Rekrutierungsdauer gegen die Rekrutierungsgeschwindigkeit war -0.50 mit einem p-Wert von 0.0059, der Spearman-Korrelationskoeffizient war -0.82 mit einem p-Wert von $5 \cdot 10^{-8}$ signifikant von Null verschieden. Auch dies spiegelt die Beobachtung wider: Je länger die Dauer der Studie war (genauer: ihre Rekrutierungsphase), desto langsamer war die Rekrutierung. Da die Fallzahl bei Studien oft im Voraus feststeht, muss der Zusammenhang eher so formuliert werden: Je langsamer die Rekrutierung, desto länger ist die Rekrutierungsphase.

3.1.2.2 Ungleichmäßige Rekrutierung

Bei der ungleichmäßigen Rekrutierung wurde der Anteil bereits rekrutierter Patienten zu einen bestimmten Zeitpunkt mit dem bis dahin verstrichenen Anteil Zeit an der Gesamtdauer der Studie verglichen. Wie schon in der Methodik dargelegt, bedeutet ein positiver Wert eine anfänglich höhere Rekrutierungsgeschwindigkeit als es bei gleichmäßiger Rekrutierung der Fall gewesen wäre. Eigenschaften der Verteilung von δ_{rek} in Abhängigkeit von den Metainformationen sind in Tabelle 3.3 abgedruckt.

Der Median der Größe δ_{rek} über alle Studien betrug 0.07, der Quartilsabstand war 0.19. Damit herrscht ein leichtes Übergewicht bei den anfänglich schnell rekrutierenden Studien vor. Der minimale Wert von δ_{rek} über alle Studien betrug -0.17 (bei einer Interventionsstudie) und der maximale 0.33 (bei einer patientenbasierten Nichtinterventionsstudie), was ebenfalls auf eine leichte Verschiebung in zu positiven Werten deutet. Der Mittelwert von δ_{rek} betrug wie der Median 0.07 und war signifikant von Null verschieden (p-Wert des Ein-Stichproben-t-Tests betrug 0.0007).

Bei der **Art der Studie** hatten unter den Interventionsstudien 10 der 29 ein negatives δ_{rek},

v_{rek} in Pat/Tag	N	Median ($Q3-Q1$)	Mittel ± SD	Min − Max	Anteil (%) $> Q3$	Anteil (%) $< Q1$
gesamt						
gesamt	40	0.33(0.99)	0.84 ± 1.25	$0.03 - 5.29$	25.0	25.0
Art der Studie						
IS	29	0.29(0.45)	0.64 ± 1.04	$0.06 - 5.29$	17.2	24.1
NIS	11	0.38(2.47)	1.38 ± 1.61	$0.03 - 4.94$	45.5	27.3
Entwicklungsphase						
unbekannt	11	0.35(0.48)	0.58 ± 0.75	$0.06 - 2.36$	18.2	36.4
nicht zutreffend	9	0.38(0.26)	0.42 ± 0.33	$0.07 - 1.16$	22.2	11.1
III	5	0.19(0.13)	0.37 ± 0.41	$0.12 - 1.09$	20.0	0.00
IV	4	0.61(3.12)	1.64 ± 2.48	$0.06 - 5.29$	50.0	50.0
Chronizität						
akut	3	0.12(0.42)	0.23 ± 0.23	$0.08 - 0.49$	0.00	33.3
beides möglich	1	0.19(0.00)	$0.19 \pm .$	$0.19 - 0.19$	0.00	0.00
chronisch	25	0.35(0.54)	0.71 ± 1.11	$0.06 - 5.29$	28.0	24.0
Anzahl Zentren						
unizentrisch	9	0.23(0.40)	0.47 ± 0.56	$0.06 - 1.65$	22.2	44.4
$2 - 10$ Zentren	12	0.16(0.35)	0.40 ± 0.64	$0.06 - 2.36$	8.33	25.0
$11-$ Zentren	8	0.54(0.79)	1.18 ± 1.70	$0.23 - 5.29$	50.0	0.00
Randomisierung						
unbekannt	2	0.37(0.04)	0.37 ± 0.03	$0.35 - 0.39$	0.00	0.00
ja	27	0.23(0.60)	0.66 ± 1.08	$0.06 - 5.29$	25.9	25.9
Verblindung						
unbekannt	1	0.41(0.00)	$0.41 \pm .$	$0.41 - 0.41$	0.00	0.00
ja	21	0.29(0.45)	0.74 ± 1.20	$0.06 - 5.29$	23.8	23.8
nein	7	0.19(0.63)	0.39 ± 0.41	$0.06 - 1.16$	28.6	28.6
Jahr der Studie						
prä GCP	14	0.23(0.45)	0.51 ± 0.67	$0.06 - 2.36$	14.3	28.6
GCP	15	0.35(0.96)	0.76 ± 1.31	$0.06 - 5.29$	33.3	20.0
Art der Kontrollen						
Placebo	22	0.32(0.45)	0.72 ± 1.17	$0.06 - 5.29$	22.7	22.7
aktiv	7	0.19(0.63)	0.39 ± 0.41	$0.06 - 1.16$	28.6	28.6
Amendment						
unbekannt	2	0.37(0.04)	0.37 ± 0.03	$0.35 - 0.39$	0.00	0.00
nein	21	0.23(0.45)	0.66 ± 1.20	$0.06 - 5.29$	19.1	33.3
ja	6	0.69(0.93)	0.66 ± 0.50	$0.15 - 1.16$	50.0	0.00
Art der Intervention						
Medikament	23	0.29(0.47)	0.69 ± 1.15	$0.06 - 5.29$	21.7	26.1
sonst	6	0.28(0.54)	0.44 ± 0.42	$0.07 - 1.16$	33.3	16.7
Anzahl Arme						
2	27	0.35(0.54)	0.68 ± 1.07	$0.06 - 5.29$	25.9	18.5
3	2	0.06(0.00)	0.06 ± 0.00	$0.06 - 0.07$	0.00	100
Rekrutierungsdauer						
max. 1 Jahr	9	1.12(1.07)	1.53 ± 1.56	$0.23 - 5.29$	66.7	0.00
mehr als 1 Jahr	20	0.16(0.29)	0.24 ± 0.19	$0.06 - 0.70$	5.00	35.0

Tabelle 3.2:
v_{rek}: Rekrutierungsgeschwindigkeit in Patienten / Probanden pro Tag, Q3-Q1: Interquartilsabstand. SD: Standardabweichung, Anteil (%) $< Q1$ bzw. $> Q3$: Anteile an langsam bzw. schnell rekrutierenden Studien, „.“ markiert Werte, die nicht bestimmt werden konnten

bei den patientenbasierten Nichtinterventionsstudien waren es nur zwei der neun untersuchten und bei den bevölkerungsbasierten Nichtinterventionsstudien war es eine der beiden. Bei beiden Studientypen war das mittlere und mediane δ_{rek} nahezu gleich, ebenso verteilten sich jeweils etwa 25% auf das untere und obere *globale* Quartil. Hinsichtlich der Ungleichmäßigkeit der Rekrutierung scheinen sich die Studientypen nicht zu unterscheiden, in beiden Fällen lag eine anfänglich leicht erhöhte Rekrutierungsgeschwindigkeit vor (p-Werte des t-Tests bzw. FP-Tests waren 0.58 und 0.54)

Die **Entwicklungsphase** schien einen schwachen Einfluss auf die Gleichmäßigkeit der Rekrutierung gehabt zu haben, denn die Phase III Studien wiesen ein medianes δ_{rek} von -0.02 und ein mittleres von 0 auf, was augenscheinlich eine gleichmäßige Rekrutierung nahe legt. Die Verteilung bei dieser Kategorie war allerdings bimodal und es fielen dabei insgesamt 60% der Werte von δ_{rek} in das obere und untere Quartil. Die Phase III Studien warteten damit mit dem größten Anteil an ungleichmäßig rekrutierenden Studien auf.

Die drei Studien zu akuter Indikation wiesen ein negatives mittleres und medianes δ_{rek} auf, während dies bei chronischer Indikation nicht so war, der FP-Test war aber nicht signifikant. Die Verteilung auf die Quartile war in beiden Gruppen der **Chronizität** nicht gleich: Während alle drei Studien mit akuter Indikation zu den 50% um den Median der gesamten Verteilung zählten, waren bei den chronischen Erkrankungen die Anteile in den äußeren Quartilen wie erwartet.

Bei der **Anzahl an Studienzentren** ergab sich, dass die unizentrischen Studien im Mittel und beim Median von δ_{rek} näher an Null (0.02) lagen, als die anderen, die einen Median von ca. 0.09 aufwiesen. Die Differenz war aber nicht groß genug, um signifikant zu sein (FP-Test lieferte p=0.31). Sowohl die Pearson- als auch die Spearmankorrelation waren nahe Null und negativ, so dass bei bei ansteigender Anzahl Zentren eher mit anfänglich langsamer Rekrutierung zu rechnen ist. Dieser Zusammenhang macht umgekehrt mehr Sinn: Wenn eine Studie anfänglich langsam rekrutierte, könnte mit der Eröffnung neuer Zentren reagiert worden sein. Die Studien mit zwei bis zehn Zentren fielen auch hier auf: sie hatten eine im Mittel stark ungleichmäßige und anfänglich schnelle Rekrutierung. Bei der Rekrutierungsgeschwindigkeit nahmen diese Studien ebenfalls eine Sonderstellung ein, denn dabei waren sie die langsamsten. Das führt zu dem Bild, dass diese Studien zwar anfänglich schnell, aber doch insgesamt die langsamsten waren. Entweder wurde die langsame Rekrutierung durch die Eröffnung neuer Zentren zwar beschleunigt, aber nicht so stark, dass die Geschwindigkeit der anderen Klassen erreicht wurde, oder eine anfänglich schnelle Rekrutierung wurde durch Schließen von Zentren verzögert. Eine grafische Darstellung widersprach der Annahme eines nichtlinearen Zusammenhanges.

Zur **Randomisierung** kann aufgrund der fehlenden Vergleichsgruppe wie zuvor keine Aussage getroffen werden.

Unverblindete Studien schienen eine größere Ungleichmäßigkeit aufzuweisen als verblindete, denn sie waren im Median 16% weiter als sie bei gleichmäßiger Rekrutierung hätten sein dürfen, während es bei den verblindeten nur 7% waren. Ebenso waren sie mit 57% überdurchschnittlich oft im oberen Quartil der beobachteten Werte. Trotzdem blieb sowohl der FP- als auch der t-Test nicht signifikant (p=0.59 bzw. 0.60). Für die **Art der Kontrollen**

galt dasselbe wie für die **Verblindung**, da hier nahezu dieselbe Einteilung der Studien vorlag.

Die Gruppen von Studien vor und nach der Einführung von GCP (**Jahr der Studie**) unterschieden sich kaum im Median oder dem Mittelwert von δ_{rek}, die Verteilung dieses Wertes war jedoch in beiden Gruppen unterschiedlich: während die Studien vor GCP eine unimodale Verteilung mit Hauptgewicht im positiven Bereich hatten, wiesen die Studien nach GCP eine bimodale Verteilung auf, bei der ein Gipfel im negativen und einer im positiven Bereich lag. Dies zeigte sich auch am Anteil von 1/3 im oberen wie im unteren Quartil bei den post-GCP-Studien. Die prä-GCP-Studien hatten in beiden Quartilen auch den gleichen Anteil, er betrug aber nur jeweils 14.3%. Eine Korrelationsanalyse der Variablen zum Jahr der Studie gegen δ_{rek} lieferte 0.14 (Pearson) bzw. 0.19 (Spearman), beide Werte waren nicht signifikant von Null verschieden.

Der Median der Größe δ_{rek} war für die Studien mit **Amendment** größer als der für die Studien ohne Amendment. Das bedeutet, dass die Studien mit Amendment vor dem Zeitpunkt des Amendments schneller rekrutierten als es bei gleichmäßiger Rekrutierung der Fall gewesen wäre, denn für Studien mit Amendment wurde die Studie am Zeitpunkt des Amendments in zwei Hälften geteilt. Die Tests auf Unterschiede zwischen den beiden Gruppen, bei denen der Median der Studien mit Amendment um sieben Prozentpunkte über dem derjenigen ohne Amendment lag, wiesen keine Signifikanz auf (FP-Test: $p = 0.78$, t-Test: $p = 0.79$). Bei den Studien ohne Amendment lag eine eingipflige Verteilung mit leichtem Übergewicht bei den positiven Werten vor, während bei den Studien mit Amendment der Bereich um die Null schwach besetzt war und die extremen Werte (negativ wie positiv) vorherrschten. Das lässt auf eine Steuerungswirkung in beiden Richtungen schließen: In einigen Studien wurde die Rekrutierung gebremst und in anderen wurde sie beschleunigt. Beispielsweise wurde bei einer Studie mit anfänglich schwacher Rekrutierung das Höchstalter angehoben und zusätzliche Komorbiditäten erlaubt. Diese Vermutung wird auch dadurch gestützt, dass in den beiden äußeren Quartilen der Gesamtverteilung von δ_{rek} je 1/3 der Studien mit Amendment liegt. Zu erwarten war hier jeweils ein Anteil von 1/4. Leicht kleinere Anteile der Studien ohne Amendment liegen dagegen in den äußeren Quartilen, wobei diese mit 23.8% und 19.1% nur knapp unter den erwarteten 25% liegen.

Die 23 Medikamentenstudien folgten bei δ_{rek} dem globalen Trend. Die verbleibenden sechs Studien mit anderer Behandlung rekrutierten anfänglich im Median schneller, im Mittel waren die beiden Gruppen gleich schnell. Jedoch waren die Hälfte der Nichtmedikamentenstudien im Quartil der anfänglich schnellen Studien und 1/3 im Quartil der anfänglich langsamen. Beide Vergleiche auf Gruppenunterschiede bei der **Art der Intervention** lieferten jedoch p-Werte nahe Eins.

Die Größe δ_{rek} verteilte sich auch sehr gleichmäßig über das Kriterium der **Anzahl an Behandlungsarmen** der Studie. Es gab jedoch nur zwei dreiarmige Studien, von denen die eine anfänglich schnell und die andere langsam rekrutierte.

Bei der **Rekrutierungsdauer** befanden sich hinsichtlich δ_{rek} gleich viele Studien im oberen wie im unteren Quartil, wenn man innerhalb der kurz und lang rekrutierenden Studien verglich. Bei den länger laufenden Studien war dieser Anteil jedoch mit 30% höher als bei den kurz laufenden Studien, bei denen sich nur 10% im ersten bzw. dritten Quartil befanden. Die

Mittelwerte waren nahezu gleich, nur die kürzer rekrutierenden Studien wiesen einen doppelt so großen Median auf. Tests auf Gruppenunterschiede blieben aber nicht signifikant, ebenso waren die Korrelationskoeffizienten mit 0.04 (Pearson) und 0.03 (Spearman) nicht signifikant von Null verschieden. Sie deuten allenfalls einen sehr schwachen Zusammenhang an, bei dem die Rekrutierung bei länger laufenden Studien eher zur Ungleichmäßigkeit zu tendieren scheint als es bei kurzen Studien der Fall ist. Eine anfänglich schleppende Rekrutierung ist dabei genauso möglich ist wie eine anfänglich schnelle.

3.1.3 Varianzunterschiede

3.1.3.1 Deskription der Varianzunterschiede

In diesem Abschnitt sollen die Unterschiede in der Variabilität beschrieben werden. Dabei wird die hierarchische Struktur der Daten zunächst außer Acht gelassen. Insgesamt wurden 182 Variablen aus den 40 untersuchten Studien bewertet. Wichtige beschreibende Statistiken können in Tabelle 3.4 abgelesen werden.

Aus der Tabelle erkennt man, dass die große Masse der Beobachtungen sich um den Wert Eins gruppiert, d.h. dass im Mittel und auch im Median die Variabilität in der ersten und zweiten Phase identisch war. Es gab aber auch Variablen, deren Standardabweichung sich mehr als vervierfacht hatte (sowohl bei einer Nichtinterventionsstudie als auch bei einer Interventionsstudie), sowie solche, bei denen es einen Rückgang dieses Streumaßes auf 7%, d.h. ca. um den Faktor 14, gegeben hat (bei einer Interventionsstudie). 90% aller Werte bewegten sich zwischen den Grenzen 0.64 und 1.69, was einer Verminderung um 1/3, bzw. Steigerung um 2/3 der Standardabweichung entspricht. Der positive Wert der Schiefe weist auf eine Verteilung hin, die links steil ist und nach rechts ausläuft, während die relativ große Kurtosis von 28.8 (eine Normalverteilung hat den Wert 3) auf eine sehr spitze Verteilung mit mehr Masse in den Randbereichen hindeutet, als es bei einer Normalverteilung zu erwarten wäre. Tatsächlich ist das Verhältnis der Standardabweichung die Wurzel aus dem Verhältnis der Varianzen der beiden Phasen und weist somit eine Verteilung auf, die einer (gemischten und möglicherweise nichtzentralen) F-Verteilung ähnlich sieht.

In Abbildung 3.1 sind Boxplots für alle Studien gezeichnet. Um mögliche Einflussfaktoren zu finden, wurden diese Boxplots gemäß Gruppenzugehörigkeit sortiert. Die ausgewählte Abbildung z.B. zeigt eine Gruppierung nach der Art der Studie.

Es ist bei den Medianen ein leichtes Übergewicht bei den Studien mit einem Rückgang der Variabilität in der zweiten Phase zu erkennen. Bei Einteilung in die Kategorien der verschiedenen Metadaten zeigten sich diese Tendenzen:

Bei den Nichtinterventionsstudien (**Art der Studie**) war das Gewicht noch weiter in Richtung der Studien verschoben, die in der zweiten Phase weniger Variabilität aufwiesen. Die Streubreite, die anhand der Länge der Box aus dem Boxplot abgelesen wurde, erschien bei diesen Studien generell kleiner zu sein. Bei den Interventionsstudien waren die Mediane des Verhältnisses der Standardabweichungen bei 12 der 29 Studien größer als Eins, was bei diesen Studien auf eine Zunahme der Variabilität hindeutet. Außerdem waren die Mediane

δ_{rek}	N	Median ($Q3 - Q1$)	Mittel ± SD	Min − Max	Anteil (%) >Q3	Anteil (%) <Q1
gesamt						
gesamt	40	0.07(0.19)	0.07 ± 0.12	$-0.17 - 0.33$	25.00	25.00
Art der Studie						
IS	29	0.07(0.18)	0.06 ± 0.12	$-0.17 - 0.28$	24.14	24.14
NIS	11	0.09(0.23)	0.09 ± 0.13	$-0.10 - 0.33$	27.27	27.27
Entwicklungsphase						
unbekannt	11	0.07(0.16)	0.07 ± 0.12	$-0.10 - 0.28$	18.18	18.18
nicht zutreffend	9	0.10(0.18)	0.08 ± 0.10	$-0.06 - 0.20$	33.33	22.22
III	5	−0.02(0.22)	0.00 ± 0.14	$-0.17 - 0.17$	20.00	40.00
IV	4	0.09(0.21)	0.09 ± 0.13	$-0.07 - 0.23$	25.00	25.00
Chronizität						
akut	3	−0.02(0.04)	-0.01 ± 0.02	$-0.02 - 0.02$	0.00	0.00
beides möglich	1	−0.17(0.00)	$-0.17 \pm .$	$-0.17 - -0.17$	0.00	100.00
chronisch	25	0.07(0.20)	0.08 ± 0.11	$-0.10 - 0.28$	28.00	24.00
Anzahl Zentren						
unizentrisch	9	0.02(0.11)	0.03 ± 0.09	$-0.09 - 0.16$	11.11	33.33
2 − 10 Zentren	12	0.10(0.24)	0.09 ± 0.14	$-0.17 - 0.28$	41.67	16.67
11− Zentren	8	0.09(0.17)	0.06 ± 0.11	$-0.10 - 0.20$	12.50	25.00
Randomisierung						
unbekannt	2	−0.01(0.17)	-0.01 ± 0.12	$-0.10 - 0.07$	0.00	50.00
ja	27	0.07(0.20)	0.07 ± 0.12	$-0.17 - 0.28$	25.93	22.22
Verblindung						
unbekannt	1	−0.04(0.00)	$-0.04 \pm .$	$-0.04 - -0.04$	0.00	100.00
ja	21	0.07(0.15)	0.06 ± 0.11	$-0.10 - 0.28$	14.29	19.05
nein	7	0.16(0.26)	0.09 ± 0.15	$-0.17 - 0.23$	57.14	28.57
Jahr der Studie						
prä GCP	14	0.07(0.15)	0.07 ± 0.10	$-0.09 - 0.28$	14.29	14.29
GCP	15	0.07(0.23)	0.05 ± 0.13	$-0.17 - 0.23$	33.33	33.33
Art der Kontrollen						
Placebo	22	0.07(0.16)	0.05 ± 0.11	$-0.10 - 0.28$	13.64	22.73
aktiv	7	0.16(0.26)	0.09 ± 0.15	$-0.17 - 0.23$	57.14	28.57
Amendment						
unbekannt	2	−0.01(0.17)	-0.01 ± 0.12	$-0.10 - 0.07$	0.00	50.00
nein	21	0.07(0.17)	0.07 ± 0.11	$-0.09 - 0.28$	23.81	19.05
ja	6	0.14(0.26)	0.06 ± 0.15	$-0.17 - 0.17$	33.33	33.33
Art der Intervention						
Medikament	23	0.07(0.18)	0.06 ± 0.11	$-0.10 - 0.28$	17.39	21.74
sonst	6	0.11(0.25)	0.06 ± 0.15	$-0.17 - 0.20$	50.00	33.33
Anzahl Arme						
2	27	0.07(0.18)	0.06 ± 0.11	$-0.17 - 0.28$	22.22	22.22
3	2	0.08(0.29)	0.08 ± 0.21	$-0.06 - 0.23$	50.00	50.00
Rekrutierungsdauer						
max. 1 Jahr	9	0.12(0.12)	0.08 ± 0.09	$-0.10 - 0.16$	11.11	11.11
mehr als 1 Jahr	20	0.06(0.23)	0.06 ± 0.13	$-0.17 - 0.28$	30.00	30.00

Tabelle 3.3:
δ_{rek}, SD: Standardabweichung, Q3-Q1: Interquartilsabstand. Anteile: Proportion im ersten oder vierten Quartil

Statistik:	Wert	Statistik:	Wert
Mittelwert:	1.05	Standardabweichung:	0.48
Median:	0.97	Interquartilsabstand:	0.22
unteres Quartil:	0.88	oberes Quartil:	1.11
5%-Perzentil:	0.64	95%-Perzentil:	1.69
Minimum:	0.07	Maximum:	4.78
Schiefe:	4.36	Kurtosis:	28.8

Tabelle 3.4:
Deskription der Verhältnisse der Standardabweichungen

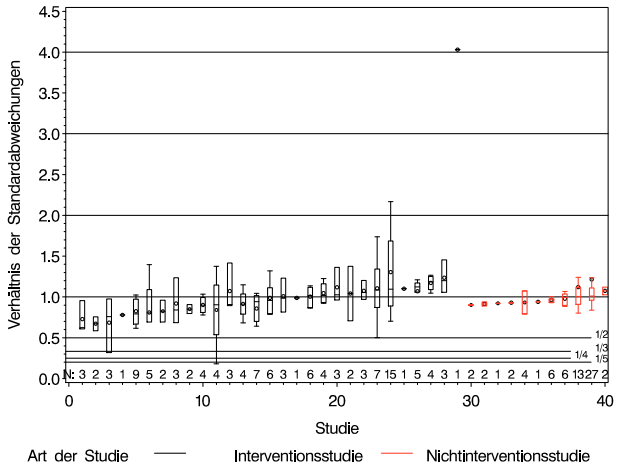

Abbildung 3.1:
Verteilung der Verhältnisse der Standardabweichungen. Die Reihe an Zahlen über der horizontalen Achse gibt an, wieviele Variablen bei der jeweiligen Studie berücksichtigt worden sind. Der Kreis kennzeichnet den Mittelwert, die waagerechte Linie im Boxplot den Median der Verhältnisse der Standardabweichungen. Die linke ansteigende Gruppe von Boxplots in schwarz zeigt die Interventionsstudien, die rechte in rot die Nichtinterventionsstudien. Die Studien wurden in beiden Gruppen jeweils aufsteigend nach dem Median sortiert.

der Verhältnisse bei den Nichtinterventionsstudien näher um die Eins gruppiert als es bei den Interventionsstudien der Fall war. Von den fünf kleinsten Verhältnissen gehörte nur eine Variable zu einer Nichtinterventionsstudie während es bei den fünf größten Werten drei waren.

Während bei den **Phase**-III-Studien leicht mehr Studien (drei der fünf) in der zweiten Phase eine größere Variabilität aufwiesen, war es bei den Phase-IV-Studien nur eine der vier. Die Studien mit unbekannter Phase verhielten sich passend zum globalen Bild.

Unter den Studien zu **chronischer Indikation** waren elf der 25 in der zweiten Phase mit höherer Variabilität. Von den drei Studien zu einer akuten Indikation war der Median von nur einer größer als Eins. Zusätzlich waren die Streubreiten bei den akuten Indikationen geringer.

Bei der **Anzahl Zentren** war ein Anstieg der Studien mit Median größer als Eins zu beobachten, wenn die Anzahl der Zentren stieg. Die Streubreiten erschienen für die multizentrischen Studien mit weniger als 11 Zentren am größten, diese Gruppe ist auch schon bei der Rekrutierung aufgefallen.

Die Einteilung nach **Verblindung** (und damit **Art der Kontrollen**) zeigte keine nennenswerten Unterschiede zwischen den Gruppen außer, dass bei den verblindeten / placebokontrollierten Studien eine leicht höhere Streubreite zu beobachten war.

Das **Jahr der Studie** schien ebenfalls keinen Einfluss auf den Anteil von Studien mit Median größer Eins zu haben. Es war allerdings zu beobachten, dass die Verteilungen der Verhältnisse bei den Studien vor Einführung von GCP eher einzelne große Werte aufwiesen, während es bei den anderen Studien vereinzelte kleine Werte gab.

Unter den sechs Studien mit **Amendment** war bei dreien der Median des Verhältnisses größer als Eins. Die Variabilität hat also bei der Hälfte dieser Studien in der zweiten Phase zugenommen. Die Streubreiten in den beiden Gruppen waren vergleichbar groß.

War die neue **Therapie** ein Medikament, war der Anteil an Medianen größer Eins mit dem globalen Anteil vergleichbar, während es bei den sonstigen Interventionen etwas weniger war (zwei der sechs untersuchten Studien). Die Streubreiten waren vergleichbar groß.

Die Unterschiede in den Streubreiten lassen sich allerdings auch teilweise durch eine unterschiedlich große Anzahl bewerteter Variablen pro Studie erklären, so dass nicht zwingend die geringere Variabilität der wahren Werte der Grund sein muss.

3.1.3.2 Vergleich mit der Mischverteilung

Abbildung 3.2 zeigt die Verteilung der Varianzquotienten in Form eines Histogrammes, bei dem zusätzlich ein Kernschätzer für die Dichte der Varianzquotienten und die theoretische Dichte eingefügt sind. Die theoretische Gesamtverteilung wurde als Mischung der einzelnen theoretischen F-Verteilungen unter der Hypothese *keines* Unterschieds zwischen den Phasen bestimmt.

Die theoretische Verteilung zeichnet sich gegenüber der tatsächlich beobachteten durch eine größere Kurtosis aus, die empirische Verteilung hat mehr Gewicht in den Rändern als die theoretische. Die theoretische Verteilung hat aufgrund ihrer Konstruktion nahe der Eins ihr Maximum, dies trifft auch in etwa auf die empirische zu. Das deutet darauf hin, dass in den meisten Fällen kein großer Unterschied in den Varianzen in den beiden Phasen vorlag. Die

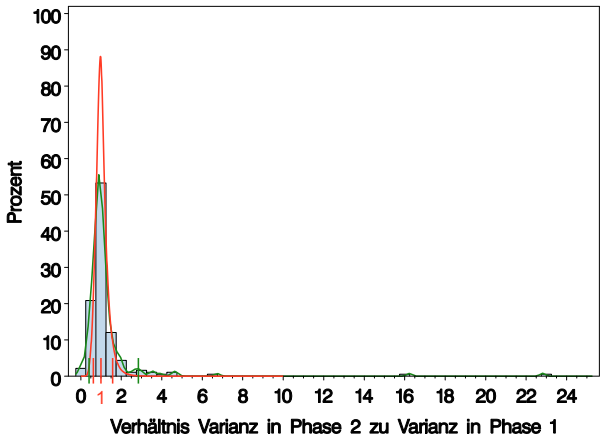

Abbildung 3.2:
Histogramm der Varianzverhältnisse (blaue Balken). Die rote glatte Kurve zeigt die theoretische Mischverteilung der Verhältnisse. Die grüne zweite Kurve entstand durch Kern-Schätzung der Dichte und spiegelt die realen Daten wieder. Die kurzen senkrechten äußeren Linien in grün markieren links das 5%- und rechts das 95%-Perzentil der vorliegenden Verteilung, während die jeweils zur Eins in der Mitte folgenden roten Linien die entsprechenden Perzentile der theoretischen Verteilung markieren.

etwas breitere Verteilung kommt auch in den Perzentilen zum Ausdruck: das 5%-Perzentil der theoretischen Verteilung wurde als 0.62 bestimmt, während das der realen Daten sich durch Quadrieren des Wertes aus Tabelle 3.4 als 0.41 errechnet. Das theoretische 95%-Perzentil ist 1.57 und das der realen Daten beträgt 2.85. Somit ist das 5%-Perzentil ca. um den Faktor 2/3 kleiner und das 95%-Perzentil knapp um den Faktor zwei größer.

Der Schluss aus diesen Überlegungen ist, dass sich im Allgemeinen die Varianzen nicht stark geändert haben, wobei es mehr Ausnahmen gibt, als mit der theoretische Verteilung vereinbar sind. Diese Ausnahmen sind sowohl bei besonders kleinen Werten als auch bei besonders großen Werten zu finden.

Der Referenzwert des modifizierten Kolmogorov-Smirnov-Tests auf Unterschied von Verteilungsfunktionen, der für einen signifikanten Test mit Niveau von 1% überschritten werden muss, lautet 1.628 [54]. Mit dem vorgefundenen Wert von 1.856 lehnt Kolmogorov-Smirnov-Test die Hypothese der Gleichheit der beobachteten und der theoretischen Verteilung unter der Annahme konstanter Varianz über die Phasen hinweg ab. Er liefert damit ein weiteres Indiz dafür, dass sich es mehr Varianzverhältnisse mit extremen Werten gibt, als es durch den Zufall zu erklären ist.

Dass die realen Varianzquotienten nicht zu der Hypothese gleicher Varianzen über die Phasen passen, zeigt auch der P-P-Plot in Abbildung 3.3, denn im unteren wie im oberen Bereich liegt mehr Gewicht als der vorgegebenen theoretischen Verteilung.

3.1.3.3 Untersuchung mit hierarchischem Modell

In Tabelle 3.5 sind die Ergebnisse von Anpassungen hierarchischer Modelle aufgelistet. Die Varianzkomponente $\hat{\tau}_{00}$ für den zufälligen Faktor *Studie* betrug für die Gesamtheit aller 40

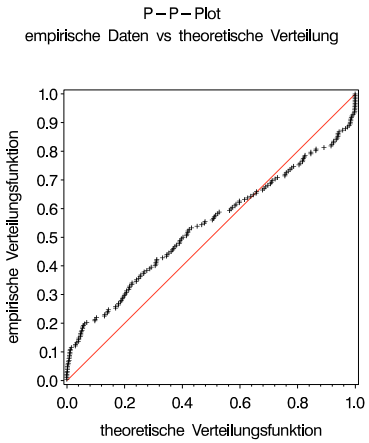

Abbildung 3.3:
P-P-Plot der empirischen gegen die theoretischen Werte für die Verhältnisse der Varianzen der beiden Phasen. Die rote Linie markiert die Winkelhalbierende, auf der die Punkte liegen sollten, wenn die realen Werte der theoretischen Verteilung folgen würden.

Studien nur 0.008633 und war damit nicht signifikant von Null verschieden ($p = 0.2054$), im Gegensatz zur Reststreuung $\hat{\sigma}^2 = 0.2176$ ($p < 0.0001$). Dies deutet darauf hin, dass die Gesamtstreuung der Daten sich *nicht* gut durch die Clusterung in Studien erklären lässt, solange Interventions- und Nichtinterventionsstudien *gemeinsam* betrachtet wurden. Die sich dadurch ergebende *Intraklassen-Korrelation* $\hat{\rho}$ von 0.033 besagt, dass nur ca. 3% der Gesamtstreuung durch die hierarchische Struktur der Studien erklärt werden kann. Der Wert des Gütekriteriums zur Modellanpassung (-2log L) ist vergleichsweise hoch (244.1), was für eine schlechtere Modellanpassung als bei den anderen Modellen spricht. Der Gesamtmittelwert in diesem Modell wurde auf 1.0346 geschätzt, was darauf hindeutet, dass *im Mittel* unter Berücksichtigung der Clusterstruktur nahezu keine Varianzveränderung zu verzeichnen ist. Wurde in dieses Modell die Art der Studie als unabhängige Variable eingeschlossen, wurde dadurch die Modellanpassung nur minimal besser (-2log $L = 243.8$) und die Intraklassen-Korrelation wurde sogar noch geringer (0.02).

Ein anderes Bild ergab sich, wenn die Betrachtungen auf die Interventionsstudien eingeschränkt wurde. Es ergab sich dann $\hat{\tau}_{00} = 0.168$ ($p = 0.0163 \approx 0.02$) bei einer Reststreuung $\hat{\sigma}^2 = 0.10$ ($p < 0.001$) und somit $\hat{\rho} = 0.62$, was auf einen sehr viel höheren Anteil erklärter Streuung und eine höhere Intraklassen-Korrelation bedeutet, wenn die Clusterung innerhalb dieser Subgruppe von Studien betrachtet wurde. Das Gütekriterium $-2 \log L$ zur Modellanpassung wurde deutlich besser (121.0). Wurden nun jedoch zusätzlich die Metadaten in das Modell eingeschlossen, zeigte sich kaum noch eine Veränderung gegenüber dem (unkonditionalen) Modell ohne Metadaten. Die Modellanpassung wurde durch die Hinzunahme auch nicht merklich verbessert, was sich auch in der Nichtsignifikanz der Modellparameter *aller* erhobenen Metadaten wiederholte, welche als fixe Effekte spezifiziert wurden. Die einzigen Metavariablen, die zu einer *Verkleinerung* der Varianzquotienten führten, waren die *Art der Studie* (um 0.05), die *Rekrutierungsdauer* (um 0.07) und *Chronizität* (um 0.17). Die

Ausprägungen *randomisiert*, *verblindet*, *(post) GCP*, *placebokontrolliert*, *Amendment*, *medikamentöse Intervention*, *Phase III-Studie* und *2-armige Studie* führten zu kleineren oder größeren *Zunahmen* der Varianzquotienten. Das Metadatum *Entwicklungsphase* trat durch einen Wert von $\hat{\rho} = 0.90$ heraus, was aber der kleinen Zahl an Studien (5 und 4) zuzuschreiben ist.

Die einzigen Variablen, die verhältnismäßig kleine p-Werte für die Modellparameter aufwiesen, waren *Amendment* und *Rekrutierungsungleichheit*.

Beschreibung	$\hat{\tau}_{00}$	$\hat{\tau}^{\delta}_{00}$	$\hat{\sigma}^2$	p_{F}	p_{τ}	β	$-2\log L$	$\hat{\rho}$
Alle Studien	0.007	.	0.22	.	0.22	.	244.1	0.03
Art der Studie (IS vs NIS)	0.005	.	0.22	0.52	0.29	-0.05	243.8	0.02
Alle Interventionsstudien	0.168	.	0.10	.	0.02	.	121.0	0.62
Rekrutierungsdauer (≤ 1 vs > 1a)	0.169	-0.001	0.10	0.69	0.02	-0.07	120.8	0.62
Chronizität (akut vs chron.)	0.173	-0.028	0.10	0.57	0.02	-0.17	117.4	0.63
Anzahl Zentren (multi- vs unizentr.)	0.156	0.074	0.10	0.26	0.02	0.21	119.6	0.60
Randomisierung (ja vs unbek.)	0.168	0.001	0.10	0.84	0.02	0.07	121.0	0.62
Verblindung (ja vs nein)	0.158	0.062	0.11	0.53	0.03	0.13	119.2	0.58
Jahr der Studie (GCP vs prä GCP)	0.169	-0.002	0.10	0.56	0.01	0.10	120.7	0.62
Art der Kontrollen (Plac. vs akt.)	0.166	0.016	0.10	0.56	0.02	0.12	120.7	0.61
Amendment (ja vs nein)	0.171	-0.013	0.11	0.13	0.01	0.35	115.2	0.62
Art d. Intervention (Med. vs sonst)	0.165	0.019	0.10	0.54	0.02	0.13	120.6	0.61
Entwicklungsphase (III vs IV)	0.771	-3.576	0.09	0.35	0.03	0.60	37.4	0.90
Anzahl Arme (2 vs 3)	0.169	-0.001	0.10	0.94	0.02	0.03	121.0	0.62
Rekrutierungsgeschwindigkeit	0.165	0.018	0.10	0.62	0.02	0.06	120.8	0.61
Rekrutierungsungleichheit	0.137	0.184	0.11	0.09	0.03	0.26	117.7	0.56

Tabelle 3.5:
Ergebnisse der hierarchischen Modellierung mit gemischten Modellen. $\hat{\tau}_{00}$ ist die Varianzkomponente, die die Variabilität zwischen den Studien quantifiziert, $\hat{\tau}^{\delta}_{00}$ ist die relative Veränderung zum Modell ohne fixe Effekte in Zeile 3, $\hat{\sigma}^2$ ist die Reststreuung, p_{F} gehört zum Test des Regressionsparameters β des fixen Effekts gegen der Wert 0. An β kann auch die Richtung der Veränderung durch Übergang von der letztgenannten zur erstgenannten Ausprägung bei den Metadaten abgelesen werden. p_{τ} gehört zum Test der Varianzkomponente τ_{00} gegen den Wert 0, $-2\log L$ ist das Modellanpassungskriterium „-2 Log-Likelihood“ (ein kleinerer Wert bedeutet bessere Anpassung). ρ beziffert die Intraklassen-Korrelation der Varianzkoeffizienten und spiegelt somit ihre Clusterung innerhalb der Studien wider.

3.2 Verblindete Fallzahladaption

In diesem Abschnitt werden die Ergebnisse der Simulationsstudien zur verblindeten Fallzahladaption zuerst für die normalverteilten und dann für die binären Endpunkte gezeigt.

Es wird bei der Untersuchung auf die Fehlerraten 1. Art eingegangen, sowie die erreichten Fallzahlen und deren Variabilität. Bei den erreichten Fallzahlen wurde die Lage untersucht, aber auch, wie weit die Fallzahlen von den Werten einer konventionellen Planung mit der tatsächlichen Varianz entfernt lagen. Die Variabilität der Fallzahlen wurde in Form des Variationskoeffizienten dargestellt, welcher die Standardabweichung auf den jeweiligen Mittelwert bezieht. Damit wird der Tatsache Rechnung getragen, dass mit steigenden Fallzahlen

auch mit einer steigenden Standardabweichung zu rechnen ist. Die Untersuchung der Variabilität hat den Hintergrund, dass neben der zu erwartenden Anzahl Patienten in einer adaptiv geplanten Studie auch die Präzision dieser Vorhersage von Interesse ist.

3.2.1 Stetige Endpunkte

Die Darstellung der Ergebnisse in diesem Abschnitt wird so gegliedert, dass von verschiedenen Varianz(fehl)annahmen ausgegangen und für jede (Fehl-)Annahme die Konsequenzen für die verschiedenen Studiengrößen beleuchtet werden.

3.2.1.1 Korrekte Varianz

Der optimale Fall ist der einer richtig spezifizierten Varianz. In Tabelle 2.2 auf Seite 25 entspricht dies den Szenarien 1 bis 4. Abbildung 3.4 zeigt die empirischen α-Levels aller acht Prozeduren für die geplanten Fallzahlen. Alle zugehörigen Konfidenzintervalle überdeckten den angestrebten Wert 0.05, es lässt sich jedoch aus der Graphik eine schwache Trennung in zwei Gruppen erkennen. Die Gruppe, die zu den unrestringierten Prozeduren gehörte, hatte Kurven, die tendenziell über denen der anderen lagen, insbesondere bei den mittleren Studiengrößen. Die unrestringierten Prozeduren konnten größere Abweichungen vom Idealwert 5% haben (maximal absolut 0.11%) als die restringierten (maximal 0.06%). Die höchste Fehlerrate 1. Art über alle Fallzahlen und Prozeduren hinweg war 5.11%(uoscc bei geplantem $N_{\text{plan}} = 200$), die niedrigste war 4.96% (rkfcc bei $N_{\text{plan}} = 400$). Bei der Fallzahl $N_{\text{plan}} = 200$ erzeugten alle Prozeduren mehr als 5% an Fehlern 1. Art, während bei $N_{\text{plan}} = 400$ alle unterhalb dieser Schranke lagen. Mit empirischen α-Fehlern von 5.038%, 5.106% und 5.060% für die drei kleineren Studiengrößen war die Basisprozedur uosnc immer unter denen mit den größten Werten. Beim größten Studientyp dagegen war der Fehler 4.940% und damit der zweitkleinste.

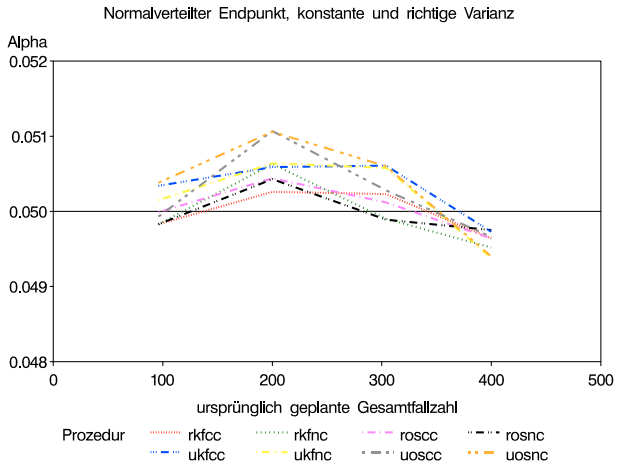

Abbildung 3.4:
Anteil an empirischen Fehlern 1. Art bei den acht Adaptionsprozeduren im Falle normalverteilter Endpunkte mit konstanter Varianz über den Studienverlauf, die bei der Planung richtig eingeschätzt worden ist. Die horizontale Referenzlinie markiert die angestrebten 5%.

Die erreichten mittleren Fallzahlen (siehe 3.5) aller Prozeduren verhielten sich linear zu den ursprünglich geplanten und unterschieden sich über alle Studiengrößen hinweg um maximal 21.2 Patienten bzw. maximal 15.6% bezogen auf die korrekte Fallzahl, wobei die unrestringierten immer niedrigere Fallzahlen erbrachten. Die Differenz innerhalb eines u/r-Paares (d.h. mit gleicher Eigenschaft für Control-Chart und KF-Korrektur) war mindestens 6.5 Patienten ($N_{\text{plan}} = 96$), höchstens 18.1 ($N_{\text{plan}} = 400$). Prozentual waren die restringierten Prozeduren um mindestens 2.0% ($N_{\text{plan}} = 400$) höher gelegen (bezogen auf die unrestringierten Prozeduren) und höchstens 14.1% ($N_{\text{plan}} = 96$) (siehe auch Tabelle 3.8). Kaum Unterschiede hingegen gab es bei den Paaren bezüglich Control-Chart (-0.73 bis +1.13%). Die KF-Korrektur brachte dagegen eine bis zu 8.2% kleinere Fallzahl als ihre Nichtverwendung, in absoluten Zahlen konnten es maximal 7.7 Patienten weniger sein.

Die finalen mittleren Fallzahlen der Prozeduren uosnc und uoscc wichen in ihren Extremen am wenigsten von der tatsächlich benötigten Fallzahl ab (s. Tabelle 3.9, S. 56), der *Betrag* der höchsten Abweichung war 1.38%. Am weitesten wichen die Prozeduren ukfcc und ukfnc nach unten ab (bis zu -9.44%), während rosnc und roscc bis zu +6.17% zu viel Patienten rekrutieren konnten.

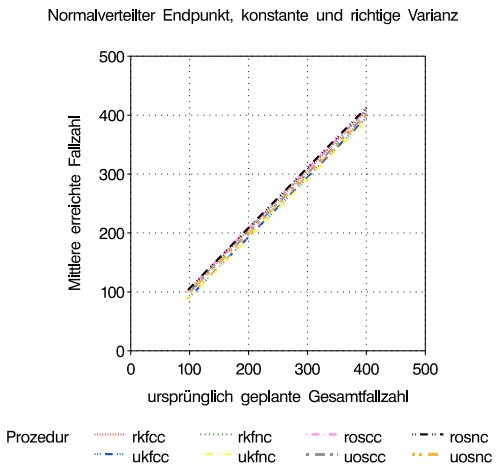

Abbildung 3.5:
Ursprüngliche aus fixer Fallzahlplanung errechnete und mittlere tatsächlich erreichte Fallzahlen im Falle normalverteilter Endpunkte mit konstanter Varianz, die bei der Planung richtig eingeschätzt worden ist.

Die Variationskoeffizienten der erreichten Fallzahl zum Szenario dieses Abschnittes sind in Abbildung 3.6 dargestellt. Der Variationskoeffizient fiel, je größer die Fallzahl der Studie war, die Standardabweichung (ohne Abbildung) selbst hatte einen steigenden Trend. Die unrestringierten Prozeduren bildeten mit Variationskoeffizienten von 8.2 bis 20.8% eine Gruppe, die immer oberhalb der restringierten Prozeduren lag, deren Werte zwischen 3.9 und 10.4% blieben. Unter den unrestringierten Prozeduren wies die Basisprozedur uosnc immer die kleinsten Werte für den Variationskoeffizienten auf. *Innerhalb* der zwei Gruppen beim Variationskoeffizienten, die durch das Merkmal Restriktion entstehen, war eine Umgruppierung bei den Prozeduren über die Studiengrößen zu beobachten: Während bei kleinen Studiengrößen Paare aufgrund der gleichen Eigenschaft hinsichtlich des Merkmals kf/os gebildet

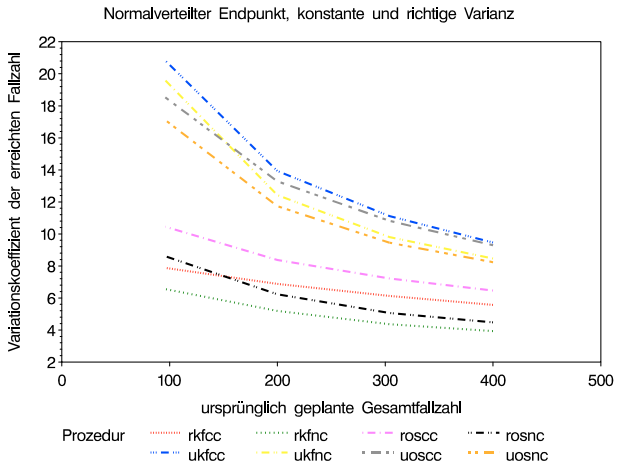

Abbildung 3.6:
Variationskoeffizienten in % für die Adaptionsprozeduren bei normalverteilten Endpunkten mit konstanter Varianz, die bei der Planung richtig eingeschätzt worden ist.

werden, ist bei den größeren Studien eine Paarbildung aufgrund der gleichen Ausprägung des cc/nc-Merkmals zu beobachten.

3.2.1.2 Überschätzte Varianz

Der erste untersuchte Fall einer fehlspezifizierten Varianz ist der einer in Wahrheit niedrigeren Varianz, wobei für die Simulation eine um die Hälfte kleinere Varianz von 0.5 generiert wurde. Der Faktor für die Standardabweichung war dementsprechend $\sqrt{1/2} \approx 0.71$ (Szenarien 5 bis 8 in Tabelle 2.2, Seite 25).

Den Fehler 1. Art für die verschiedenen Prozeduren und Fallzahlen für dieses Szenario kann man in Abbildung 3.7 sehen. Die Basisprozedur uosnc lieferte empirische Fehler 1. Art von 5.105%, 4.852%, 5.015% und 5.029% mit aufsteigender Studiengröße, wobei der zweite zur initialen Größe von 200 Patienten gehörende Wert sich signifikant von 5% unterschied. Insgesamt waren fünf Fälle zu beobachten, in denen das Konfidenzintervall um den empirischen Fehler 1. Art nicht mehr die angestrebten 5% einschloss. Diese Fälle betrafen nur unrestringierten Prozeduren: Bei der kleinsten Studiengröße waren der empirische Fehler 1. Art bei den Prozeduren ukfnc und uoscc zu groß, während beim nächst größeren Studientyp die Prozeduren ukfcc, ukfnc und uosnc einen zu kleinen Anteil aufwiesen. Die Prozedur ukfnc tauchte damit sogar in beiden Gruppen auf.

Die unrestringierten Prozeduren wiesen bei den kleineren Fallzahlen größere Abstände vom Idealwert 5% auf als die restringierten Prozeduren (Maxima der Abstände von 5% zwischen den Gruppen waren 1.5-fach bis zu 74fach höher), bei den letzteren mit 400 Patienten allerdings war der Abstand geringer (70%). Weiterhin ist bemerkenswert, dass alle restringierten Prozeduren identische empirische Fehler 1. Art für jede der untersuchten Fallzahlen lieferten.

In der Graphik mit den erreichten tatsächlichen Fallzahlen (Abbildung 3.8) ist eine ähnliche Beobachtung zu machen. Es ist deutlich zu erkennen, dass die restringierten Prozeduren genau auf den ursprünglich geplanten Fallzahlen verharrten (die mittleren Fallzahlen waren

identisch zu den geplanten), während die unrestringierten erwartungsgemäß niedrigere Fallzahlen lieferten. Diese waren im Mittel ca. halb so groß wie die der restringierten Prozeduren, was mittlere Differenzen von 47.6 bis 207.6 Patienten bedeutete.

Neben der erwarteten Deckungsgleichheit der Fallzahlkurven bei den restringierten Prozeduren ist es noch auffällig, dass auch die der unrestringierten nahe beieinander lagen. Unter ihnen waren die der kf-Prozeduren unterhalb der os-Prozeduren und zwar in allen Fällen um ca. 7.5 Patienten, was einen Prozentanteil von 3.7% bei den größten Studien bzw. 15.5% bei den kleinster Studien bedeutete.

Control-Charts bewirkten eine größere mittlere Fallzahl, als wenn sie nicht verwendet wurden, die Erhöhung war in Mittel aber moderat mit maximal 1.9 Patienten bzw. 1.36% (Tabelle 3.8, S. 55).

Die tatsächlich benötigte Fallzahl war in diesem Szenario halb so groß wie in der ursprünglichen Planung, daher war die Abweichung der restringierten Prozeduren vom Idealwert im Mittel +100%, was je nach Studiengröße 48 bis 200 Patienten bedeutete (Tabelle 3.9, S. 56). Die unrestringierten Prozeduren teilten sich in zwei Gruppen: die kf-Varianten wichen um bis zu 15.5% (7 - 8 Patienten) nach unten von der tatsächlich benötigten Fallzahl ab. Die os-Varianten zeigten im Mittel kleinere Abweichungen, wobei die mittleren Fallzahlen von uoscc im Mittel leicht größer (max. 1.06%), die von uosnc dagegen im Mittel leicht kleiner (maximal 0.17% zu tief) als der tatsächlich benötigte Umfang war. Für beide waren die absoluten Abweichungen aber mit maximal 1.7 Patienten gering.

Bei der Variabilität der erreichten Fallzahlen (Abbildung 3.9) wiederholte sich das zuvor gezeichnete Bild: Die Standardabweichung und folglich auch der Variationskoeffizient der Fallzahlen aller restringierten Prozeduren war gleich Null und dies über alle Fallzahlen hinweg. Bei den unrestringierten Prozeduren war der Verlauf des Variationskoeffizienten dem im vorigen Kapitel mit der konstanten und korrekt spezifizierten Varianz sehr ähnlich und sogar die Größenordnung stimmte überein (vorher: startend bei 17-21%, endend bei 8-9%, jetzt startend bei 22-27%, endend bei 11-14%). Wie schon im Abschnitt über die korrekt spezifizierte Varianz bildeten bei den kleinen Fallzahlen die kf-Prozeduren ein Paar mit höherer Variabilität, während bei den größeren Fallzahlen die Gruppe aus den cc-Prozeduren den größeren Variationskoeffizienten aufwies. Die Standardprozedur uosnc hatte mit den Werten 22.0, 12.6, 15.5 und 11.0% (sortiert nach aufsteigender initialer Studiengröße) den kleinsten Varianzkoeffizienten unter den unrestringierten Prozeduren.

3.2.1.3 Unterschätzte Varianz

Beim Szenario einer Varianz, die bei der Planung zu niedrig eingeschätzt wurde, war der spezifizierte Wert 1, während in Wahrheit eine Varianz von 2 vorlag. Dies entsprach einer Vergrößerung der Standardabweichung um $\sqrt{2} \approx 1.41$ (Szenarien 9-12 in Tabelle 2.2, Seite 25).

Keine der Prozeduren verletzte das Niveau und alle blieben abgesehen vom größten Studientyp antikonservativ (Abbildung 3.10). Es bildeten sich Paare mit nahezu identischem α-Fehler (maximale Differenz: 0.002%-Punkte), welche sich immer aus einer restringierten und der da-

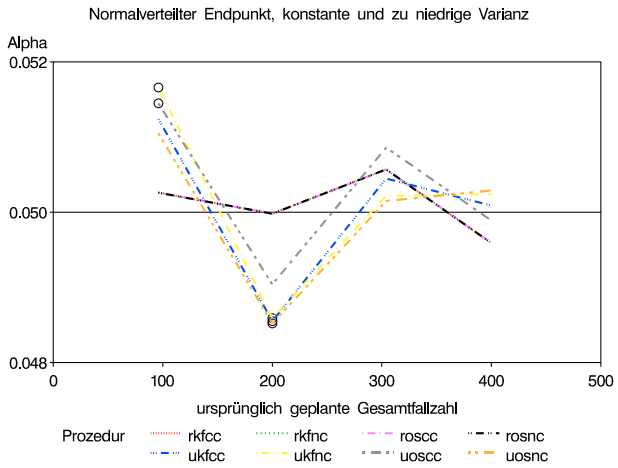

Abbildung 3.7:
Empirischer Fehler 1. Art für eine Varianz, die bei der Planung zu hoch angesetzt wurde. Die Kreise markieren diejenigen Punkte, in denen das zugehörige Konfidenzintervall den Idealwert 5% nicht eingeschlossen hat. Diese Punkte betrafen die Prozeduren ukfnc und uoscc bei der kleinsten Studiengröße und die Prozeduren ukfcc, ukfnc und uosnc bei der nächst größeren Studiengröße.

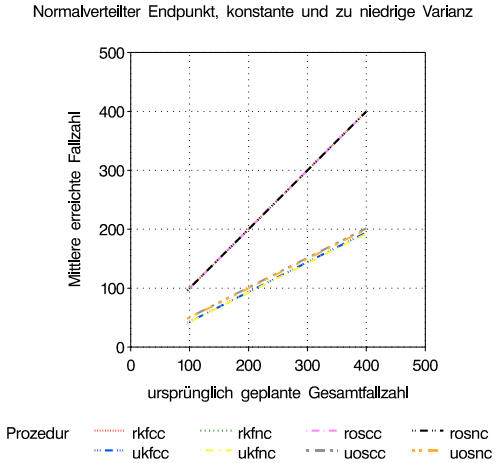

Abbildung 3.8:
Ursprüngliche aus fixer Fallzahlplanung errechnete und mittlere tatsächlich erreichte Fallzahlen im Falle normalverteilter Endpunkte mit einer Varianz, die bei der Planung doppelt so hoch eingeschätzt worden ist. Es ist deutlich zu erkennen, dass die unrestringierten Prozeduren die Fallzahl auf die Hälfte korrigiert haben.

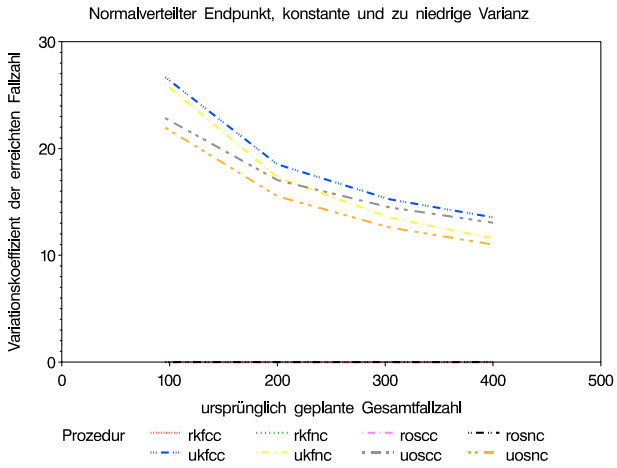

Abbildung 3.9:
Variationskoeffizienten der erreichten Fallzahlen für das Szenario mit überschätzter Varianz. Die restringierten Prozeduren weisen keinerlei Variabilität auf, die der unrestringierten fällt mit steigender Fallzahl. Die Basisprozedur uosnc hatte unter den unrestringierten Prozeduren immer den kleinsten CV.

zu passenden unrestringierten Prozedur zusammensetzten (d.h. mit gleichen Ausprägung der Merkmale KF-Korrektur und Control-Chart). Auch insgesamt ergaben sich ähnliche Fehlerraten, sie bewegten sich über alle Fallzahlen hinweg zwischen 4.946 und 5.093%. Die vier Prozeduren mit Control-Charts bewegten sich näher an den idealen 5% als die andere Gruppe. Die Basisprozedur uosnc hatte empirische Fehler 1. Art von 5.034, 5.036, 5.066 und 4.946% (sortiert nach steigender Studiengröße). Keine der Eigenschaften führte zu konsistent kleineren Abweichungen vom angestrebten Fehler 1. Art von 5%. Die Fallzahlen wurden durch

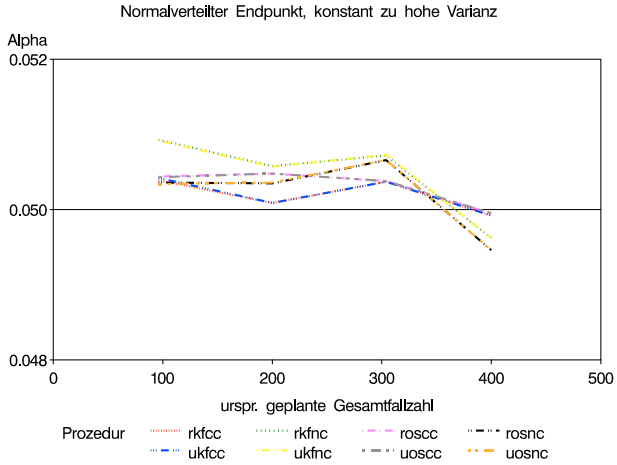

Abbildung 3.10:
Anteil an empirischen Fehlern 1. Art bei den acht Adaptionsprozeduren im Falle normalverteilter Endpunkte mit unterschätzter Varianz. Die horizontale Referenzlinie markiert die angestrebten 5%. Keine Prozedur lieferte einen Fehler, der sich signifikant von 5% unterschied. Es ist eine Gruppenbildung von restringierten und unrestringierten Varianten mit gleicher Eigenschaft für KF-Korrektur und Control-Chart zu erkennen.

alle Prozeduren auf ca. das Doppelte der ursprünglich geplanten Fallzahl erhöht. Der maxi-

male Unterschied der Fallzahlen bewegte sich für die verschiedenen Studiengrößen zwischen 9 ($N_{\text{plan}} = 96$) und 13 ($N_{\text{plan}} = 400$) Patienten. Die restringierten Prozeduren wurden im Gegensatz zum Szenario mit der real zu kleinen Varianz nun nicht mehr vom Adaptieren abgehalten und konnten sich nun genauso wie die unrestringierten Prozeduren verhalten. Der Unterschied zwischen einer restringierten Prozedur und ihrem unrestringierten Gegenstück betrug im Mittel maximal 0.07 Patienten. Außerdem waren die Zahlen der kf-Prozeduren immer kleiner als die der zugehörigen os-Prozeduren, der Unterschied innerhalb der kf/os-Paare war bei allen Studiengrößen ca. 7 Patienten, was einer prozentualen Verringerung um 1% bis 4% entsprach. Außerdem lieferten die cc-Prozeduren im Mittel leicht größere Fallzahlen als ihre nc-Pendants (Tabelle 3.8, S. 55). Dieser Effekt wurde größer mit steigender Studiengröße, beginnend bei 1.4 Patienten und endend bei 6.0, relativ gesehen bewegten sich die cc-Prozeduren zwischen 0.7% und 0.8% über den nc-Prozeduren. Da die mittleren Fallzahlen der restringierten und unrestringierten Prozeduren jeweils nahezu gleich waren, ergaben sich Unterschiede hauptsächlich durch die anderen Eigenschaften: KF-Korrektur verursachte eine systematische Unterschätzung der Fallzahl. Die kfnc-Verfahren wichen dabei am stärksten nach unten ab (bis zu 5.4%, bzw. 10 Patienten), bei den kfcc-Verfahren war die Abweichung kleiner (4.7%, 9 Patienten). Die os-Verfahren hatten noch einmal kleinere Abweichungen von der benötigten Fallzahl. Während die nc-Varianten die Fallzahl ebenso systematisch unterschätzten (ca. 1.6%, 3 Patienten), war bei den oscc-Verfahren sowohl Unter- als auch Überschätzung möglich (-0.8% bis +0.4%, bzw. -1.5 bis 3.2 Patienten, s. Tabelle 3.9, S. 56).

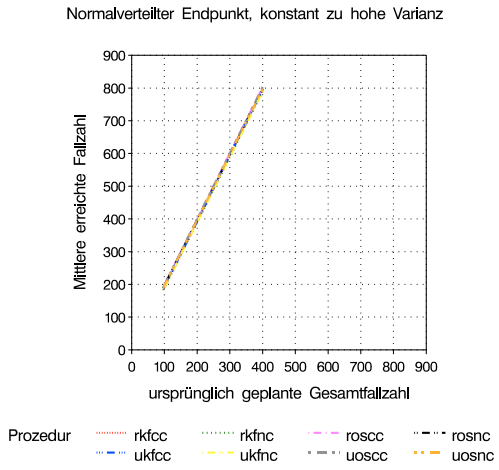

Abbildung 3.11:
Ursprüngliche aus fixer Fallzahlplanung errechnete und mittlere tatsächlich erreichte Fallzahlen im Falle normalverteilter Endpunkte mit einer initial zu niedrig eingeschätzten Varianz. Alle Prozeduren korrigierten die Fallzahl auf ca. das Doppelte, was einer fixen Planung mit der richtigen Varianz entsprach. Die Fallzahlen der einzelnen unterschieden sich um maximal 13 Patienten.

Die Variabilität der erreichten Fallzahlen war keiner aus den anderen Szenarien ähnlich (s. Abbildung 3.12), außer, dass der Variationskoeffizient mit steigender Studiengröße kleiner wurde. Anders als bei den anderen Varianzspezifikationen entstanden zwei Gruppen aufgrund des Merkmals Control-Chart, wobei die Prozeduren mit Control-Chart durchgehend die größere Variabilität aufwiesen: bei der kleinsten Studiengröße war dies 14.4 - 15.2 versus

12.4 - 13.3, bei der größten 9.49 - 9.54 versus 5.97 - 6.05. Innerhalb dieser Gruppen bildeten die Prozeduren mit KF-Korrektur bei der kleinsten Studiengröße eine Gruppe mit leicht größeren Variabilitäten (eine kf-Prozedur unterschied sich um bis zu 0.8 von der zugehörigen os-Prozedur), diese sehr kleine Differenz war aber schon ab dem nächst größeren Studientyp vernachlässigbar. Die kleinste Variabilität hatte durchgehend die Prozedur rosnc, immer gefolgt durch uosnc, wobei diese beiden Prozeduren fast gleiche Werte aufwiesen.

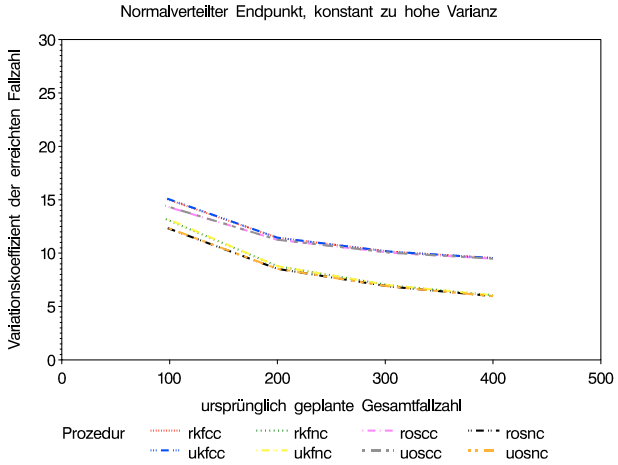

Abbildung 3.12:
Variationskoeffizienten in % für die Adaptionsprozeduren bei normalverteilten Endpunkten mit konstanter Varianz, die bei der Planung zu niedrig eingeschätzt worden ist. Es tritt anders als bei den anderen Varianzverläufen eine Gruppierung nach dem Kriterium *Control-Chart* auf.

3.2.1.4 Zusammenfassende Tabellen

Tabelle 3.6 zeigt die minimalen und maximalen Anteile an Fehlern 1. Art für die Basisprozedur uosnc und der anderen Prozeduren (zusammengefasst) in den drei Varianzszenarien.

Prozedur	Wahre Varianz	Empirischer Fehler 1. Art in %
Basisprozedur	wie erwartet	4.940 - 5.106
(uosnc)	größer als erwartet	4.946 - 5.066
	kleiner als erwartet	4.852 - 5.105
modifiziert	wie erwartet	4.938 - 5.107
	größer als erwartet	4.946 - 5.093
	kleiner als erwartet	4.855 - 5.166

Tabelle 3.6:
Minimaler und maximaler mittlerer Anteil an Fehlern 1. Art für die Basisprozedur uosnc und ihre sieben Modifikationen in den drei verschiedenen Szenarien der Varianz(fehl)spezifikation.

In Tabelle 3.7 werden alle Fälle zusammengefasst, in denen der Fehler 1. Art signifikant von den angestrebten 5% abwich.

Anhand der Tabelle 3.8 werden Fallzahlen für alle drei Varianzspezifikationen dargestellt, dabei werden Paare untersucht, die aus dem Vergleich der Auswirkung des Vorhandenseins

Prozedur	Initiale Fallzahl	Wahre Varianz...	Empirischer Fehler 1. Art in %
uosnc	200	kleiner als erwartet	4.852
ukfcc	200	kleiner als erwartet	4.855
ukfnc	200	kleiner als erwartet	4.859
uoscc	96	kleiner als erwartet	5.145
ukfnc	96	kleiner als erwartet	5.166

Tabelle 3.7:
Fälle mit empirischem α-Fehler, der signifikant von 5% unterschiedlich ist. In fünf Fällen war der empirische Fehler 1. Art signifikant von 5% verschieden. Dies war nur in dem Szenario mit überschätzter Varianz der Fall und betraf nur unrestringierte Prozeduren (u.a. auch die Basisprozedur uosnc) bei den kleineren zwei Studiengrößen. Es gab Abweichungen sowohl über als auch unter die angestrebten 5%.

eines Merkmals mit der des Nichtvorhandenseins entstehen. So werden beim Vergleich der Wirkung der KF-Korrektur immer eine kf-Prozedur mit einer os-Prozedur verglichen, wobei die anderen zwei Merkmale cc und re gleich sein müssen. Dann wurden die extremen absoluten und relativen Differenzen bestimmt, wobei die Prozedur *ohne* das betrachtete Merkmal den Referenzwert lieferte. So ergibt sich z.B. der Wert 2.00% für Restriktion bei konstanter und korrekter Varianz als das Minimum der Werte $\frac{\hat{N}_{\mathrm{re}}-\hat{N}_{\mathrm{nr}}}{\hat{N}_{\mathrm{nr}}}$. In diesem Szenario wurden die vier Paare uosnc/rosnc, uoscc/roscc, ukfnc/rkfnc und ukfcc/rkfcc über die vier Studiengrößen verglichen.

Varianz	Merkmal	Min. Diff. in %	Max. Diff. in %	Min. Diff.	Max. Diff.
wie erwartet	Restriktion	2.00	14.05	6.5	18.1
	Control-Chart	-0.73	1.13	-3.0	4.3
	KF-Korrektur	-8.18	-0.64	-7.7	-2.6
niedriger als erwartet	Restriktion	97.91	136.7	47.6	208
	Control-Chart	0.00	1.36	0.0	1.9
	KF-Korrektur	-15.5	0.00	-7.5	0.0
höher als erwartet	Restriktion	-0.00	0.04	-0.0	0.1
	Control-Chart	0.76	0.81	1.4	6.1
	KF-Korrektur	-3.97	-0.90	-7.6	-7.2

Tabelle 3.8:
Maximale und minimale Differenzen der mittleren finalen Fallzahlen bei verschiedenen Varianzverläufen und den drei Merkmalen. Die Differenzen wurden aus der mittleren finalen Fallzahl der Prozedur mit dem Merkmal aus Spalte 2 und der mittlerem Fallzahl der Prozedur ohne dieses Merkmal bestimmt. Die Extremwerte wurden über die verschiedenen Studiengrößen ermittelt. Sowohl relative wie absolute Differenzen wurden berechnet, wobei diese *nicht* aus dem gleichen Szenario zu stammen brauchen.

In der Tabelle 3.9 werden die Extremwerte der Abweichungen der finalen Fallzahlen der einzelnen Prozeduren vom tatsächlich benötigten Umfang dargestellt. Es werden wieder prozentuale und absolute Abweichungen gezeigt. Die Bestimmung der Extremwerte allein über die Ausprägungen eines Merkmals wie in Tabelle 3.8 schien nicht sinnvoll, da die Ausprägungen der jeweils anderen Merkmale unberücksichtigt blieben. Als Referenzwert für diese Tabelle

wurde die tatsächlich benötigte Fallzahl verwendet und es ergab sich beispielsweise der Wert 1.72% als Minimum der Werte $\frac{\hat{N}_{\text{rkfcc}} - N_{\text{tats}}}{N_{\text{tats}}}$ über alle Studiengrößen im Szenario mit korrekter Varianz.

Varianz...	Prozedur	Min. Diff. in %	Max. Diff. in %	Min. Diff.	Max. Diff.
wie erwartet	rkfcc	1.72	3.19	3.1	6.9
	rkfnc	2.34	3.29	3.2	9.4
	roscc	2.37	6.11	5.9	9.5
	rosnc	3.13	6.17	5.9	12.5
	ukfcc	-8.72	-1.09	-8.4	-4.4
	ukfnc	-9.44	-2.18	-9.1	-8.7
	uoscc	-0.65	0.36	-0.6	1.4
	uosnc	-1.38	-0.35	-1.5	-1.3
niedriger als erwartet	rkfcc	100.0	100.0	48.0	200
	rkfnc	100.0	100.0	48.0	200
	roscc	100.0	100.0	48.0	200
	rosnc	100.0	100.0	48.0	200
	ukfcc	-14.6	-2.82	-7.0	-5.6
	ukfnc	-15.5	-3.80	-7.7	-7.4
	uoscc	0.83	1.06	0.4	1.7
	uosnc	-0.17	-0.06	-0.3	-0.0
höher als erwartet	rkfcc	-4.68	-0.51	-9.0	-4.1
	rkfnc	-5.42	-1.27	-10	-10
	roscc	-0.76	0.39	-1.5	3.2
	rosnc	-1.55	-0.36	-3.0	-2.9
	ukfcc	-4.71	-0.51	-9.0	-4.1
	ukfnc	-5.46	-1.27	-10	-10
	uoscc	-0.77	0.39	-1.5	3.2
	uosnc	-1.57	-0.36	-3.0	-2.9

Tabelle 3.9:
Maximale und minimale Differenzen der mittleren finalen Fallzahlen bei verschiedenen Varianzspezifikationen und den acht Prozeduren. Die Differenzen wurden aus der mittleren finalen Fallzahl der Prozedur und der tatsächlich benötigten Fallzahl bestimmt. Die Extremwerte wurden dann über die verschiedenen Studiengrößen ermittelt. Sowohl relative wie absolute Differenzen wurden berechnet, wobei diese *nicht* aus dem gleichen Szenario zu stammen brauchen.

3.2.2 Binäre Endpunkte

Bei den Simulationen zu binären Zufallsgrößen konnte die Varianz einen von drei verschiedenen Werten annehmen, von denen der kleinste mit einer PER von 0.7 (Varianz=0.21), die mittlere mit einer PER von 0.6 (Varianz=0.24) und die größte mit einer PER von 0.5 (Varianz=0.25) einher ging. Dabei steht PER für die erwartete Anzahl an „Erfolgen“ in der Gesamtstichprobe (*Pooled Event Rate*).

Unter den drei Möglichkeiten wurde das Verhalten der Adaptionsprozeduren bei korrekter vorheriger Spezifizierung dieser Streumaße untersucht, aber auch wenn die mittlere Streuung angenommen worden ist, aber in Wirklichkeit eine der beiden anderen vorlag.

3.2.2.1 Korrekte, niedrige Varianz

Der erste untersuchte Fall einer Studie mit binärem Endpunkt ist der einer konstant niedrigen Varianz (PER=0.7), deren Wert bei der Planung auch so vorhergesehen wurde.

Es gab nur eine einzige Situation (Abbildung 3.13), in der der empirische α-Fehler unter 5% lag, nämlich die der Prozedur rnc bei $N_{\text{plan}} = 200$ mit 4.993%. Die gleiche Prozedur erreichte bei $N_{\text{plan}} = 304$ einen Anteil von 5.000%. Ansonsten waren alle anderen Fehlerraten darüber. Schon bei diesem Szenario gab es Prozeduren, die signifikant zu oft die gültige Nullhypothese abgelehnt haben. Die betroffenen Fälle waren bei der niedrigsten Fallzahl und bei den Prozeduren mit Control-Charts zu beobachten (5.226%, 5.196%). Bei kleineren Studien wiesen die cc-Prozeduren einen höheren Fehleranteil auf, bei den größeren war dieser Unterschied geringer. Die Standardprozedur lieferte Fehlerraten 1. Art zwischen 5.04 und 5.08%.

Die vier Prozeduren führten zu mittleren Fallzahlen (Abbildung 3.14), die sich um weniger als 12 Patienten unterschieden. Die unrestringierten Prozeduren konnten potentiell kleinere Zahlen liefern und lagen daher im Mittel unter den restringierten. Die Control-Chart-Prozeduren kamen unter den nc-Prozeduren zu liegen, beides bestätigt sich auch in Tabelle 3.11 auf S. 63: Die re-Prozeduren lagen im Mittel um bis zu 4.8% bzw. 8.2 Patienten über den nr-Prozeduren, während die cc-Variante mit bis zu -1.9% bzw. -7.7 Patienten unterhalb der nc-Variante liegen konnte, aber nicht darüber.

Beim Vergleich mit der tatsächlich benötigten Fallzahl (Tabelle 3.12, Seite 64) zeigte rnc mit mindestens +2% bzw. +4.1 Patienten und höchstens +4.3% bzw. 4.3 Patienten die größte mittlere Abweichung vom Idealwert, rcc überschätzte diesen ebenfalls, jedoch nicht so stark. Die Prozedur ucc blieb systematisch unterhalb der richtigen Fallzahl mit bis zu -1.8% bzw. -3.2 Patienten. Die Basisprozedur unc dagegen wich sowohl nach oben als auch nach unten mit einem maximalen Betrag von 0.7% bzw. 2.2 Patienten ab.

Die Variationskoeffizienten der Fallzahlen teilten die Prozeduren in zwei Gruppen: Die der unrestringierten Prozeduren bewegten sich zwischen 12.4 und 13.1% bei der kleinsten Fallzahl, während sich die restringierten Prozeduren dort zwischen 5.1 und 6.3% aufhielten. Beide Gruppen folgten einem allgemeinen Abwärtstrend mit steigender geplanter Fallzahl und endeten bei 4.4-5.2% bzw. 2.3-2.8%. Innerhalb der Gruppen auf Basis der Eigenschaft *Restriktion* hatten bei den kleineren Fallzahlen jeweils die cc-Prozeduren eine größere relative Streuung, während bei den größeren Fallzahlen dies bei den nc-Prozeduren (und damit auch für die Basisprozedur) der Fall war. Die Prozedur unc wies Werte zwischen 5.2 und 12.4% auf.

3.2.2.2 Korrekte, mittlere Varianz

Bei mittelgroßer Streuung (PER=0.6) bildeten sich beim Fehler 1. Art wie im Fall der kleinen Varianz (PER=0.7) zwei Gruppen an Prozeduren heraus: Einerseits die cc-Prozeduren, andererseits die nc-Prozeduren. In beiden gab es Verletzungen des Niveaus, wobei bei der kleinsten Fallzahl nur die Prozedur rnc antikonservativ war (5.138%) und bei den höheren Fallzahlen die zwei cc-Prozeduren von 5% signifikant unterschiedliche Fehleranteile aufwiesen:

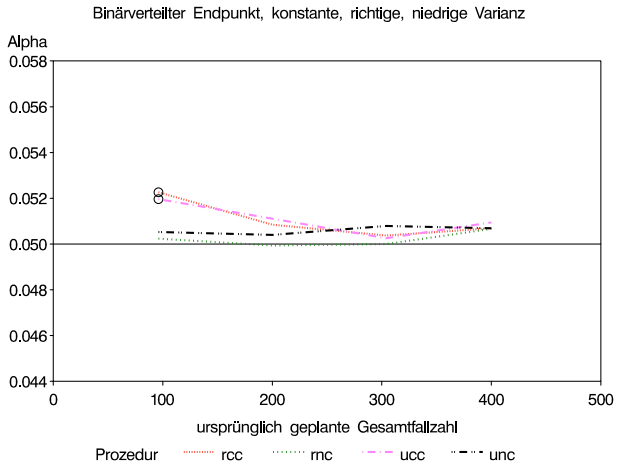

Abbildung 3.13:
Empirischer Fehler 1. Art der vier Prozeduren bei den vier Studiengrößen. Die Kreise markieren Stellen, an denen das Konfidenzintervall die gewünschten 5% nicht überdeckt.

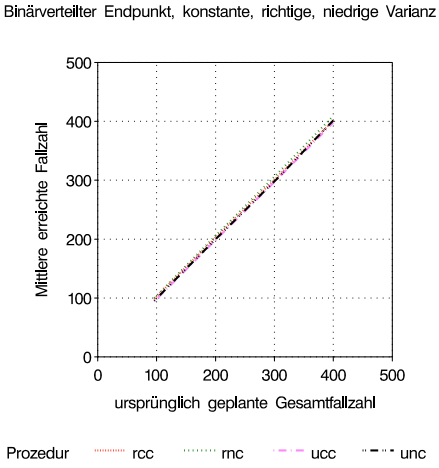

Abbildung 3.14:
Ursprünglich geplante und mittlere finale Fallzahlen der vier Prozeduren bei den vier Studiengrößen.

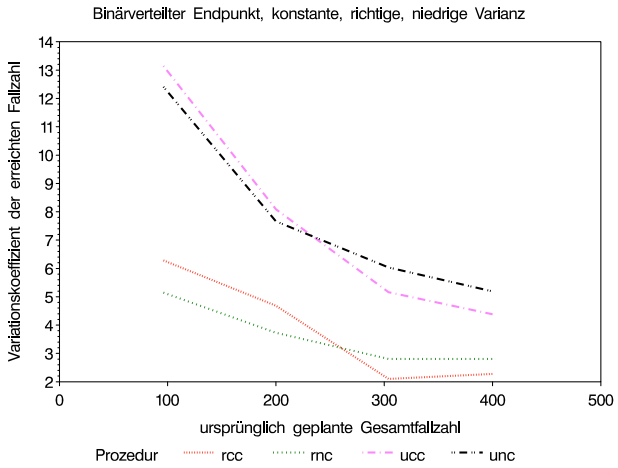

Abbildung 3.15:
Variationskoeffizient (in %) der finalen Fallzahlen der vier Prozeduren bei den vier Studiengrößen.

bei geplanten 304 Patienten waren sie mit mindestens 5.293% antikonservativ, bei geplanten 400 Patienten mit höchstens 4.824% zu konservativ. Die Basisprozedur unc verursachte Fehler 1. Art zwischen 4.96 und 5.1%, die alle nicht von 5% signifikant verschieden waren.

Die erreichten Fallzahlen unterschieden sich über die Studiengrößen hinweg nur um 3-5 Patienten, wobei sich re/nr-Paare um 0.1% bis 2.1% unterschieden (re-Prozeduren endeten wieder höher), was in absoluten Zahlen zwischen 0.4 und 4.2 Patienten bedeutete (Tabelle 3.11, Seite 63). Die Verwendung von Control-Charts bedeutete keine systematische Abweichung nach unten gegenüber der nc-Variante wie im Falle der kleinen Varianz, dennoch war die maximale mittlere Überschreitung mit nur 0.1%, bzw. 0.3 Patienten nicht so stark wie die maximal mögliche Unterschreitung, welche -1.9% bzw. -4.5 Patienten betrug.

Als einzige Prozedur endete rcc mit mittleren Fallzahlen, die numerisch genau mit den tatsächlich benötigten übereinstimmten (Tabelle 3.12, Seite 64). Im Gegensatz dazu wich rnc systematisch nach oben (0.9-2.0%) und ucc systematisch nach unten (-1.5% bis -0.1%) ab. Bei der Basisprozedur unc waren Abweichungen in beide Richtungen möglich (max. 0.6%).

Dieses Szenario wartete mit einer geringeren Variabilität (CV) der finalen Fallzahlen vor als beim vorhergegangenen aus dem Abschnitt mit der kleinen Varianz. Beim letzten Szenario waren die höchsten Variationskoeffizienten durch ca. 14% nach oben beschränkt, in diesem Szenario durch 7%. Weiterhin war die Variabilität bei den restringierten Prozeduren wieder niedriger als bei den unrestringierten und die der Prozedur rcc war sogar Null. Die zwei restringierten Prozeduren hatten wie vorher einen fallenden Variationskoeffizienten mit steigender Studiengröße. Dieser fiel von 5.2-5.7% bei der kleinsten Studiengröße auf 1.2-2.5% bei der größten. Die unrestringierten Prozeduren hingegen verliefen nahezu gleichbleibend, aber auf einem niedrigeren Niveau: Der Variationskoeffizient der Prozedur rcc verlief konstant bei Null und rnc blieb leicht fallend im Bereich zwischen 2.0 und 1.1%. Ein Wechsel der Plätze der cc/nc-Paare innerhalb einer re/nr-Gruppe fand im Gegensatz zum ersten Szenario nicht statt, sondern die nc-Prozeduren blieben immer über den cc-Prozeduren. Die Basisprozedur unc hatte mit 2.5-5.8% durchgehend den größten Variationskoeffizienten für die erreichte

Fallzahl.

3.2.2.3 Korrekte, hohe Varianz

In diesem Szenario wurde die höchste Variabilität der Daten erzeugt (PER=0.5) und eine Planung unterstellt, die diesen Wert antizipierte. Die beobachteten Fehler 1. Art waren in 12 von 16 Fällen so weit von 5% entfernt, dass die Konfidenzintervalle diesen Wert nicht mehr einschlossen. Alle Prozeduren unterschieden sich bei den kleineren drei Studiengrößen signifikant von 5%, wobei sie bei den Fallzahlen 96 und 200 zu antikonservativ (Anteil Fehler 1. Art war mindestens 5.197%) und bei 304 zu konservativ waren (Anteil Fehler 1. Art war höchstens 4.704%). Bei der geplanten Fallzahl von 400 dagegen waren die Fehlerraten aller Prozeduren nicht signifikant von 5% verschieden. Die in diesem Szenario beobachteten Abstände der empirischen α-Fehler waren mit minimal 4.641% bis maximal 5.629% die höchsten unter denen mit korrekt spezifizierter Varianz. Die Basisprozedur bewegte sich zwischen den Anteilen 4.7 und 5.6%.

Die Entwicklung der Fallzahlen glich den schon beschriebenen Fällen mit kleineren Varianzen: Die ursprünglich geplanten Fallzahlen wurden im Mittel erreicht. Bemerkenswert ist, dass sich die maximalen und minimalen Werte der verschiedenen Prozeduren um nur noch 0.4 bis 0.6 Patienten unterschieden. Damit erzeugt das Szenario die kleinsten Unterschiede zwischen den Fallzahlen aller Prozeduren. Es setzt sich also der Trend fort, denn die maximalen Unterschiede zwischen den Prozeduren waren bei mittlerer Varianz ebenfalls kleiner als bei der kleinsten.

Die restringierten Prozeduren unterschieden sich von ihren unrestringierten Gegenstücken um maximal +0.5%, die cc-Prozeduren von ihren nc-Gegenstücken um +0.2%. In beiden Fällen bedeutete das Vorhandensein des jeweiligen Merkmals eine Erhöhung der finalen mittleren Fallzahl (Tabelle 3.11, Seite 63).

Die beiden restringierten Prozeduren endeten im Mittel mit genau der tatsächlich benötigten Fallzahl, bei den unrestringierten Prozeduren war hingegen eine systematische Abweichung nach unten zu erkennen, wobei diese mit höchstens -0.5% sehr gering war (Tabelle 3.12, Seite 64).

Die zwei restringierten Prozeduren wiesen außerdem keinerlei Variation in den erreichten Fallzahlen auf und die beiden unrestringierten erzeugten Werte für den Variationskoeffizienten zwischen 1.8 und 2.4% bei $N_{\text{plan}} = 96$ und 0.1 bis 0.5% bei $N_{\text{plan}} = 400$. Wie im Szenario mit kleiner und korrekter Varianz tauschten die cc-Prozeduren die Plätze, denn bei kleinen Studien hatte die Prozedur ucc den größeren Variationskoeffizienten, bei den größeren Studien war es dann unc. Die letztgenannte Basisprozedur hatte Werte zwischen 0.5 und 1.8% und war damit ab der geplanten Studiengröße von 200 Patienten die Prozedur mit dem größten Variationskoeffizienten.

3.2.2.4 Überschätzte Varianz

Dies ist der erste der beiden Fälle, in denen die Variabilität bei der Planung fehlspezifiziert wurde. Dennoch ist er hinsichtlich der Trennschärfe der weniger kritische Fall, denn die auf-

zudeckende Differenz zwischen den Proportionen der beiden Gruppen wurde so gewählt, dass sie für eine höhere Varianz (PER=0.6) hätte mit einer Power von 85% nachgewiesen werden können.

Das Szenario zeichnet sich durch die geringe Abweichung des Anteils an Fehlern 1. Art von den erwünschten 5% aus. Es gibt hier nur zwei Fälle, in denen eine Prozedur signifikant antikonservativ war. Dies war die Prozedur ucc bei den Fallzahlen 200 (5.14%) und 304 (5.168%). Bei allen anderen Prozeduren lagen zwar meist leicht antikonservative Ergebnisse vor, diese waren jedoch nie signifikant von 5% verschieden. Die restringierten Prozeduren hatten nahezu identische Anteile an Fehlern 1. Art, die Basisprozedur unc solche zwischen 4.96 und 5.12%.

In diesem Szenario zeigte sich auch erstmals, dass die Prozeduren unterschiedlich reagiert haben: Bei einer Varianz, die kleiner war als erwartet, sollte die Fallzahl entsprechend gesenkt werden. Wie erwartet konnten die restringierten Prozeduren die Fallzahl jedoch nicht herabsetzen und blieben mit sehr kleiner Spannbreite (maximale Abweichung nach oben war 0.1 Patienten) auf den zuvor spezifizierten Werten, während die unrestringierten in allen Fällen kleinere mittlere Fallzahlen lieferten.

Eine restringierte Prozedur unterschied sich von ihrem unrestringierten Gegenstück um mindestens +11.1 und höchstens um +15.6%, in absoluten Fallzahlen: +12.1 bis +51.4 Patienten (Tabelle 3.11, Seite 63).

Während die unrestringierten cc-Varianten bei kleineren Studiengrößen nahezu identische Fallzahlen erbrachten (cc: 84.0, nc: 83.3 bei $N_{\text{plan}} = 96$), gab es bei den größeren sichtbare Unterschiede (cc: 359.5, nc: 348.6 bei $N_{\text{plan}} = 400$). Die Prozedur ohne Control-Chart benötigte hier eine etwas geringere Fallzahl. So reduzierte der Einsatz eines Control-Charts höchstens nur sehr leicht die Fallzahlen (-0.1%), während Vergrößerungen um +3.1% gegenüber der passenden nc-Variante möglich waren.

Verglichen mit einer fixen Fallzahlplanung (Tabelle 3.12, Seite 64) auf Basis der korrekten Varianz wurden durch die restringierte Prozeduren zwischen 13.7% und 14.8% erhöhte Stichprobenumfänge bestimmt, da eine Verkleinerung nicht möglich war. Bei der Prozedur ucc wurden leicht zu große Umfänge berechnet, denn sie lag im Mittel mindestens 0.2% bis höchstens 3.0% über der richtigen Fallzahl. In absoluten Zahlen waren mittlere Überschätzungen von 10.5 Patienten möglich. Die Basisprozedur unc dagegen endete mit systematisch zu kleinen Fallzahlen, aber das Ausmaß der Unterschätzung mit höchstens -0.7%, bzw. -0.6 Patienten fiel eher gering aus.

Bei der relativen Streuung (CV) der Fallzahlen hatten die restringierten Prozeduren Werte nahe Null, über alle Fallzahlen hinweg betrug der maximale Variationskoeffizient 0.7%. Die unrestringierten Prozeduren dagegen wiesen Werte von 13.6 und 15.8% bei der kleinsten geplanten Fallzahl und abfallend bei der höchsten Fallzahl Werte von 5.5 bis 11.0% auf, wobei die cc-Variante immer die mit dem höheren der zwei Werte war, sodass die Basisprozedur unc unter den unrestringierten durchgehend den kleinsten Variationskoeffizienten besaß. Bei den restringierten Prozeduren wirkte dieser Faktor genau umgekehrt, da die cc-Varianten mit einem über die Studiengrößen konstanten Variationskoeffizienten von Null und die nc-Varianten mit Werten zwischen Null und einem Prozent beobachtet wurden.

3.2.2.5 Unterschätzte Varianz

Der zweite Fall einer zur Planungszeit fehlspezifizierten Variabilität ist der einer real zu hohen Streuung der untersuchten Variablen. Für die Simulation wurde eine minimal nachzuweisende Differenz der Proportionen verwendet, die zu einer geringeren Variabilität gehörte. Während die Differenz bei einer PER von 0.6 noch mit 85% Power nachweisbar gewesen wäre, ist sie bei einer PER von 0.5 mit einer kleineren Power verbunden. Die Prozeduren mussten also die Fallzahl entsprechend nach oben korrigieren.

Das Bild der Fehler 1. Art war wie beim Szenario mit hoher und korrekt spezifizierter Varianz sehr uneinheitlich mit starken Schwankungen über und unter die gewünschten 5%, in Extremfällen sind 4.484 bis 5.721% aufgetreten. Alle Prozeduren verletzten signifikant das Niveau bei den kleineren drei Studiengrößen und nur die beiden nc-Prozeduren bei der Studiengröße 304 und die cc-Prozeduren bei der Studiengröße 400 hielten das Niveau ein. Aufgrund des Kriteriums des Control-Charts fand eine Teilung beim Fehler 1. Art statt, die Basisprozedur unc sowie die Prozedur rnc hatten meist die gleichen Anteile an Fehlern 1. Art, sie bewegten sich zwischen 4.5 und 5.7%.

Die nc-Prozeduren konnten wie erwartet im Mittel stärker auf die zu hohe Variabilität reagieren. Bei der kleinsten geplanten Fallzahl unterschieden sich die Prozeduren um maximal 3.7 Patienten und bei der höchsten Fallzahl von 400 unterschied sich die höchste mittlere Fallzahl von der niedrigsten um 15.5 Patienten.

Beim Vergleich der re/nr-Paare (Tabelle 3.11, Seite 63) war kaum ein Unterschied zu beobachten, denn der minimale Unterschied war 0, der maximale 0.13%, bzw. 0.1 Patienten. Die cc/nc-Paare dagegen zeigten größere Differenzen, wobei die Verwendung des Control-Charts immer zu kleineren Fallzahlen führte als die Nichtverwendung. Die Abweichungen betrugen zwischen -3.6 und -4.7%, in absoluten Zahlen -3.6 bis -15 Patienten.

Die tatsächlich benötigte Fallzahl wurde durch die cc-Prozeduren systematisch unterschätzt, es ergaben sich hier Unterschiede von bis zu -4.4% bzw. -15.5 Patienten. Die nc-Prozeduren konnten sowohl nach oben wie nach unten abweichen, aber der Betrag war nur maximal 0.27% / 0.9 Patienten (Tabelle 3.12, Seite 64).

Eine Gruppenbildung aufgrund der cc-Eigenschaft zeigte sich ebenfalls in der Variabilität der Fallzahlen, welche bei allen Prozeduren sehr gering ausfiel (der CV betrug maximal 1.7%). Die rcc-Prozedur wies bei keiner Studiengröße einen Variationskoeffizienten ungleich Null auf, während die ucc-Prozedur von 1.2 auf 0.01% fiel. Die anderen zwei Prozeduren unc und rnc fielen ebenfalls von 1.7 bzw. 1.2 auf den gemeinsamen Wert von 0.4%.

3.2.2.6 Zusammenfassende Darstellungen

Tabelle 3.10 zeigt die minimalen und maximalen Anteile an Fehlern 1. Art für die Basisprozedur unc und der anderen Prozeduren (zusammengefasst) in den fünf Varianzszenarien.

Insgesamt gab es 80 Kombinationen aus Prozedur (4), Studiengröße (4) und Varianzszenario (5). In 33 Fällen überdeckte das 95%-Konfidenzintervall nicht die erwünschten 5%.

Prozedur	Wahre Varianz...	Empirischer Fehler 1. Art in %
Basisprozedur	wie erwartet, niedrig	5.040 - 5.079
(unc)	wie erwartet, mittel	4.961 - 5.099
	wie erwartet, hoch	4.704 - 5.615
	kleiner als erwartet	4.962 - 5.121
	größer als erwartet	4.486 - 5.720
modifiziert	wie erwartet, niedrig	4.993 - 5.226
	wie erwartet, mittel	4.824 - 5.294
	wie erwartet, hoch	4.641 - 5.629
	kleiner als erwartet	4.921 - 5.168
	größer als erwartet	4.484 - 5.721

Tabelle 3.10:
Minimaler und maximaler mittlerer Anteil an Fehlern 1. Art für die Basisprozedur unc und ihre drei Modifikationen in den fünf verschiedenen Szenarien der Varianz(fehl)spezifikation.

Varianzverlauf	Merkmal	Min. Diff. in %	Max. Diff. in %	Min. Diff.	Max. Diff.
richtige, niedrige Varianz	Restriktion	0.89	4.82	3.5	8.2
	Control-Chart	-1.89	-0.34	-7.7	-1.0
richtige, mittlere Varianz	Restriktion	0.09	2.06	0.4	4.2
	Control-Chart	-1.92	0.09	-4.5	0.3
richtige, hohe Varianz	Restriktion	0.00	0.48	0.0	0.6
	Control-Chart	0.00	0.19	0.0	0.6
zu niedrige Varianz	Restriktion	11.12	15.56	12.1	51.4
	Control-Chart	-0.12	3.13	-0.1	10.9
zu hohe Varianz	Restriktion	0.00	0.13	0.0	0.1
	Control-Chart	-4.69	-3.58	-15	-3.6

Tabelle 3.11:
Maximale und minimale Differenzen der mittleren finalen Fallzahlen bei verschiedenen Varianz(fehl)annahmen und den zwei Merkmalen. Die Differenzen wurden aus der mittleren finalen Fallzahl der Prozedur mit dem Merkmal aus Spalte 2 und der mittlerem Fallzahl der Prozedur ohne dieses Merkmal bestimmt. Die Extremwerte wurden über die verschiedenen Studiengrößen ermittelt. Sowohl relative wie absolute Differenzen wurden berechnet, wobei diese *nicht* zur gleichen Studiengröße gehören müssen.

Varianzverlauf	Prozedur	Min. Diff. in %	Max. Diff. in %	Min. Diff.	Max. Diff.
richtige, niedrige Varianz	rcc	0.27	2.63	0.8	2.8
	rnc	1.99	4.28	4.1	9.0
	ucc	-1.76	-0.56	-3.2	-1.6
	unc	-0.72	0.35	-2.2	1.4
richtige, mittlere Varianz	rcc	0.00	0.00	0.0	0.0
	rnc	0.86	1.96	1.9	4.5
	ucc	-1.45	-0.09	-1.4	-0.4
	unc	-0.19	0.60	-0.7	1.8
richtige, hohe Varianz	rcc	0.00	0.00	0.0	0.0
	rnc	0.00	0.00	0.0	0.0
	ucc	-0.40	-0.00	-0.4	-0.0
	unc	-0.48	-0.15	-0.6	-0.4
zu niedrige Varianz	rcc	13.77	14.73	12.2	51.0
	rnc	13.77	14.74	12.3	51.0
	ucc	0.18	3.02	0.1	10.5
	unc	-0.74	-0.11	-0.6	-0.3
zu hohe Varianz	rcc	-4.44	-3.63	-15	-3.8
	rnc	-0.15	0.27	-0.2	0.9
	ucc	-4.44	-3.63	-15	-3.8
	unc	-0.28	0.27	-0.3	0.8

Tabelle 3.12:
Maximale und minimale Differenzen der mittleren finalen Fallzahlen bei verschiedenen Varianz(fehl)annahmen und den vier Prozeduren. Die Differenzen wurden aus der mittleren finalen Fallzahl der Prozedur und der tatsächlich benötigten Fallzahl bestimmt. Die Extremwerte wurden dann über die verschiedenen Studiengrößen ermittelt. Sowohl relative wie absolute Differenzen wurden berechnet, wobei diese *nicht* zur gleichen Studiengröße gehören müssen.

3.3 Ein Kombinationstest bei geänderter Variabilität

3.3.1 Fehler 1. Art

In beiden Szenarien haben die Pooling- und die Kombinationsstrategie das Niveau eingehalten (Abbildung 3.16). Die Testprozedur Komb&1, bei der noch zusätzlich Signifikanz in einer der beiden Phasen verlangt wurde, damit sie insgesamt signifikant wurde, erzeugte durchgehend eine kleinere Fehlerrate als 5% und blieb immer bei ca. 4.5%.

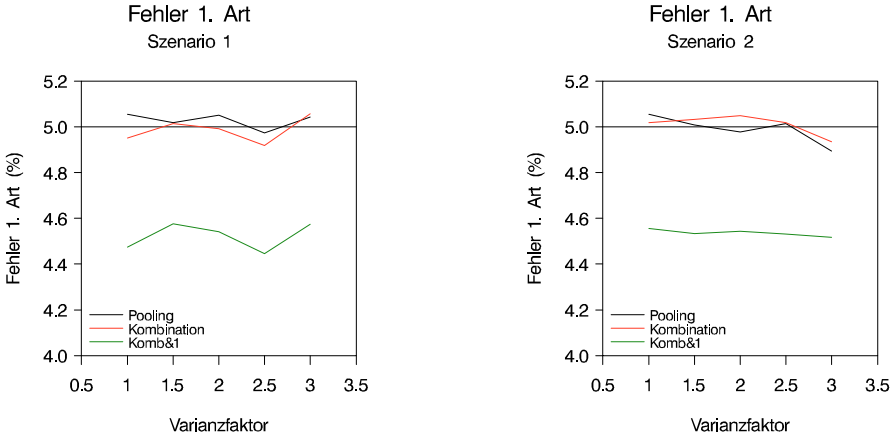

Abbildung 3.16:
Empirische Fehler 1. Art für die drei Prozeduren Pooling, Kombination, Komb&1.
Die horizontale Linie markiert das angestrebte 5% Niveau.

Weder das Szenario für die Rekrutierung noch die Stärke der Inflation scheinen in den Verläufen über die Varianzinflation einen größeren Einfluss zu haben. Die 95%-Konfidenzintervalle deckten bei der Pooling- und reinen Kombinationsstrategie die 5% in allen Fällen ab, während sie beim Abschlusstest durchgehend komplett darunter lagen.

3.3.2 Trennschärfe

3.3.2.1 Konstante Mittelwerte

Abbildung 3.17 zeigt die Trennschärfe des jeweiligen Tests bei *kleiner* Mittelwertdifferenz. Alle Tests haben bei diesem kleinen Effekt eine sehr geringe Trennschärfe. Für beide Szenarien zusammen betrachtet, ist das Maximum bei 16.82%, das Minimum bei 9.83%. In beiden Szenarien sind die relativen Lagen der Trennschärfe ähnlich. Findet keine Varianzinflation statt, dann ist die Reihenfolge bezüglich der Trennschärfe mit der höchsten beginnend wie folgt: Pooling-Test, Kombinationstest, Abschlusstest. Wird die Varianz erhöht, ändert sich die Reihenfolge dahingehend, dass der Pooling-Test an die zweite Stelle fällt. Die Abschlusstestprozedur bleibt die mit der geringsten Power, welche parallel zu dem des Kombinationstests verläuft.

Die Simulationsergebnisse deuten an, dass der Punkt, an dem der Kombinationstest den Poolingtest überholt, beim Szenario 2 früher als beim Szenario 1 ist. Denn im ersteren findet der Übergang zwischen den Werten 1.5 und 2 für den Varianzfaktor statt, im letzteren ist er zwischen den Werten 2 und 2.5.

Beim Vergleich der Pooling- und der Kombinationsstrategie zeigt sich in Szenario 1, dass Pooling im besten Fall um maximal 1.3 Prozentpunkte höher lag und im schlechtesten Fall um 0.66 Prozentpunkte niedriger. Für das Szenario 2 waren diese zwei Differenzen 0.9 Prozentpunkte bzw. 0.7 Prozentpunkte.

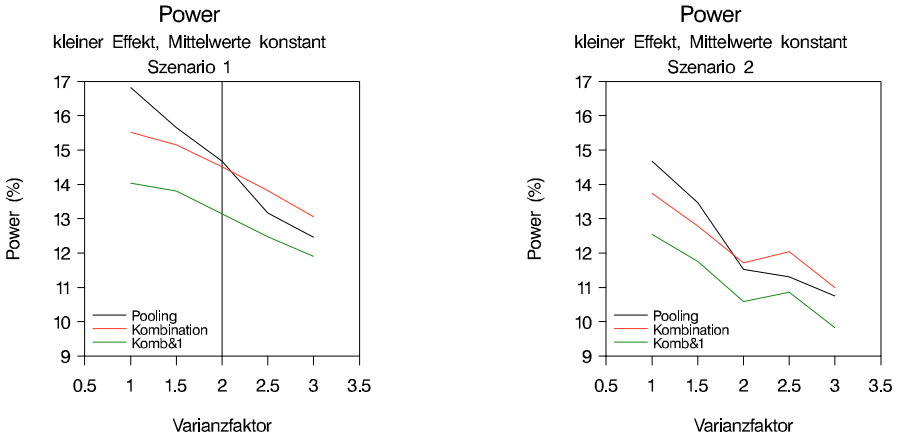

Abbildung 3.17:
Trennschärfe der drei untersuchten Prozeduren unter den beiden Rekrutierungsszenarien bei mittelgroßem Mittelwertunterschied. Zur besseren Abschätzung der Stelle, an der die Kombinationsprozedur die Pooling-Prozedur übertrifft, wurde beim Varianzfaktor 2 eine Hilfslinie gezogen.

Abbildung 3.18 entspricht der gleichen Situation für eine *mittelgroße* Mittelwertdifferenz, wobei zwei Eigenschaften der Verläufe sofort auffallen: Erstens ist der Verlust an Power im zweiten Szenario generell stärker ausgeprägt als im ersten, zweitens fällt die Power des Poolingtests bei ansteigendem Varianzfaktor unter die des Abschlusstests (dies war bei kleiner Mittelwertdifferenz nicht so).

Zur ersten Beobachtung ist festzustellen, dass die Poolingprozedur mit der größten Power startet und mit der kleinsten endet, so dass die über die Varianzinflationsfaktoren den größten Verlust aufweist. Dieser beträgt im ersten Szenario 16.8 Prozentpunkte im zweiten 28.5. Beide Kombinationsprozeduren verlieren nur 9.1 bzw. 19.5 Prozentpunkte.

Die Poolingprozedur startete im Falle keiner Varianzinflation (Faktor 1) oberhalb der beiden anderen Prozeduren, sie lag 0.7 Prozentpunkte über der Kombinationsprozedur und 1.9 über der Abschlusstestprozedur. Bei der extremsten Varianzinflation (Faktor 3) fällt der Poolingtest jedoch 7 Prozentpunkte unter den Kombinationstest und 4.9 unter den Abschlusstest. Die Stelle, ab der die beiden Kombinationsprozeduren die Poolingprozedur an Power übertreffen war in beiden Szenarien zwischen 1.5 und 2 beim Kombinationstest bzw. 2 und 2.5 beim Abschlusstest.

Wie in der vorigen Situation verlaufen die Powerkurven des beiden Kombinationsprozeduren weitgehend parallel, wobei die Abschlusstestprozedur eine kleinere Power hatte als die Kombinationsprozedur. Ihre Trennschärfen unterschieden sich um maximal 2.2 bzw. 3.7 Prozentpunkte in den beiden Szenarien.

Bei den Powerverläufen zur *größten* betrachteten Mittelwertdifferenz (Abbildung 3.19) ist zu erkennen, dass alle Prozeduren eine sehr hohe Power aufweisen und die Unterschiede sehr

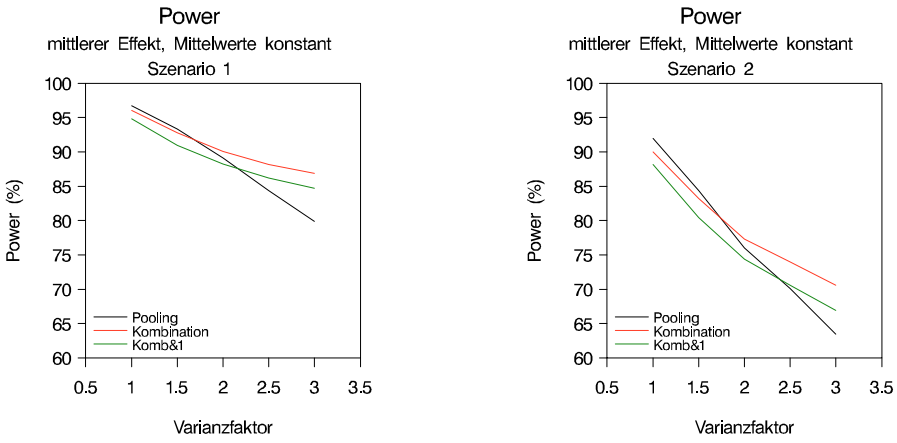

Abbildung 3.18:
Trennschärfe der drei untersuchten Prozeduren unter den beiden Rekrutierungsszenarien bei mittelgroßem Mittelwertunterschied.

gering sind. Im Szenario 1 bewegen sich alle beobachteten Werte zwischen 99.94 und 100%, im Szenario 2 zwischen 98.84 und 100%. Obwohl die Unterschiede extrem klein waren, zeigte sich, dass die Pooling-Prozedur hier nie besser (höchstens gleichwertig) war als die beiden anderen Testverfahren.

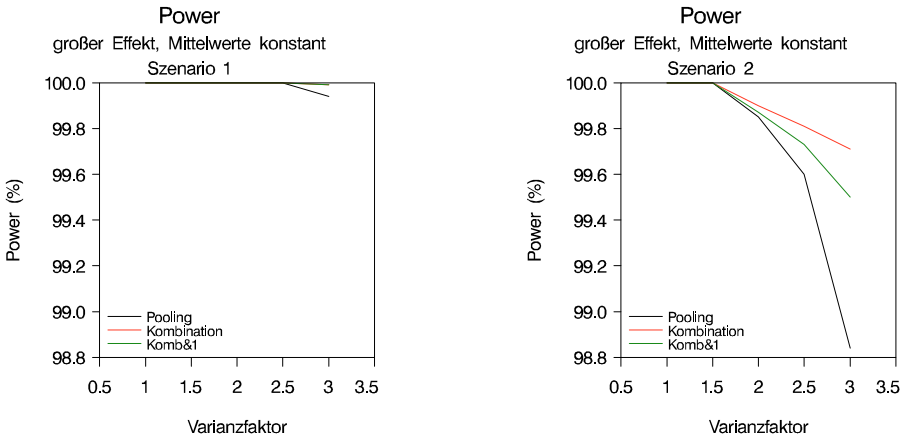

Abbildung 3.19:
Trennschärfe der drei untersuchten Prozeduren unter den beiden Rekrutierungsszenarien beim größten Mittelwertunterschied.

3.3.2.2 Variable Mittelwerte bei konstanter Differenz

Eine potentiell andere Situation lag vor, wenn sich die Mittelwerte über die Zeit verändern konnten. Zuerst wurde die einfache Möglichkeit einer gleichbleibenden Mittelwertdifferenz untersucht. Da sich im letzten Kapitel gezeigt hat, dass sich relevante Unterschiede bei mittelgroßem Effekt ergaben, wurde dieser hierfür verwendet. Die Mittelwerte der Behandlungsgruppen wurden über die Phasen hinweg (A) leicht um 0.1, (B) mittel um 0.5 oder (C) stark um 1.0 verschoben (Abbildung 3.20).

Dabei ist zu beachten, dass die beiden Testverfahren mit Verwendung der Kombination

von diesen Verschiebungen theoretisch nicht tangiert werden durften, da für diese die beiden Phasen jeweils getrennt ausgewertet wurden und das Testergebnis nur von der Differenz der Mittelwerte abhing, welche sich in dieser Situation nicht änderte. In der Tat sieht man kaum Unterschiede bei den Unterszenarien (A), (B) und (C) was die Power der Kombinations- und Abschlusstestprozedur angeht. Außerdem zeigt sich das gleiche Bild wie beim Szenario mit konstanten Mittelwerten: in Szenario 2 ist der Verlust an Power generell höher. Die Begründung aus dem vorigen Szenario kann auch hier angewendet werden.

Eine weitere Beobachtung bezüglich der Pooling-Strategie scheint erwähnenswert: Sie verhält sich nicht gleich unter den verschiedenen Verschiebungen, denn mit größerer Verschiebung wird ihre Trennschärfe geringer. Startet sie bei einer schwachen Verschiebung (A) und konstanter Varianz noch oberhalb der beiden anderen Testprozeduren, ist ihre Trennschärfe bei starker Verschiebung (C) im Szenario 1 auf dem Niveau der Abschlusstestprozedur und im Szenario 2 zwischen den beiden anderen Prozeduren. Der Grund für dieses Bild liegt in der Mischverteilung, bei der durch diesen Shift ein konstanter Gruppenunterschied „verwischt“ wird.

Zusätzlich fällt die Trennschärfe der Pooling-Prozedur auch stärker mit steigendem Shift. Dieser Effekt ist jedoch sehr schwach ausgeprägt und könnte auch ein durch die Simulation begründeter Zufall sein: Im Szenario 1 ist bei einem Shift von 0.1 der Unterschied der Power von Null-Inflation zu dreifacher Varianzinflation 16.11 Prozentpunkte, bei einem Shift von 0.5 ist diese Differenz 16.38 Prozentpunkte, bei einem Shift von 1.0 ist sie 17.93 Prozentpunkte. Im Szenario 2 sind die Werte für den Verlust an Trennschärfe bei der Poolingprozedur zwar insgesamt höher, es lässt sich aber nicht der gleiche Trend wie in Szenario 1 erkennen, hier sind die Unterschiede 28.03, 28.31 und 27.88 Prozentpunkte.

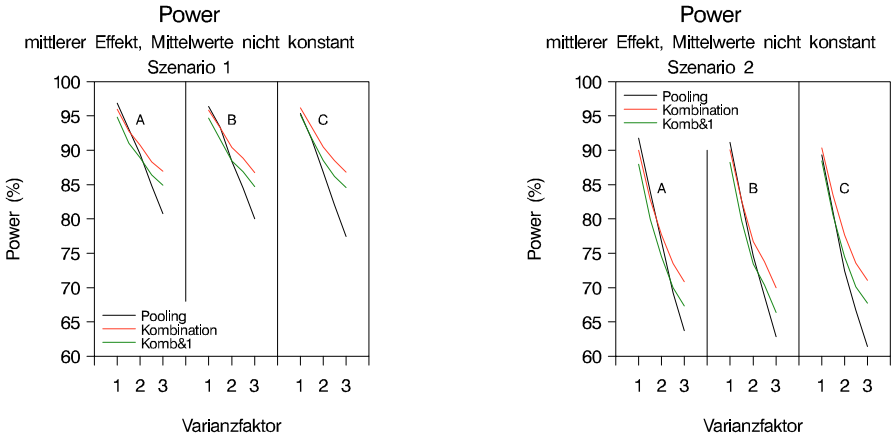

Abbildung 3.20:
Power der drei untersuchten Prozeduren bei konstantem mittelgroßem Effekt von 0.5 und drei Verschiebungen der Mittelwerte von Verum- und Kontrollgruppe. (A) steht für eine kleine Verschiebung von 0.1, (B) für eine mittelgroße von 0.5 und (C) für eine große um 1.0.

3.3.2.3 Nichtkonstante Differenz

Zuletzt durfte auch die Mittelwertdifferenz zwischen den Phasen variieren. Es wurde für die beiden Szenarien von einer Differenz ausgegangen, unter der bei Null-Inflation eine Power von ca. 80% vorlag, wie sie bei klinischen Studien mindestens angestrebt wird. Es wurde in diesem Fall eine Verringerung der Mittelwertdifferenz simuliert, denn bei klinischen Studien, bei denen ein Amendment die Rekrutierungsgeschwindigkeit erhöhen soll, ist durch das heterogenere Patientenkollektiv eher mit einem kleineren Behandlungseffekt in der zweiten Phase zu rechnen. Darüber hinaus stellt dieses Szenario den ungünstigeren Verlauf hinsichtlich der Trennschärfe dar und es ist zu klären, wie stark die statistischen Prozeduren davon beeinträchtigt werden.

Abbildung 3.21 zeigt den Effekt der Varianzinflation, wenn die Kontrolltherapie konstant einen Wert von Null liefert, während Verum unter Szenario 1 im Mittel in der ersten Phase einen Effekt von 0.7 aufweist und dann in der zweiten Phase auf 0.2 abfällt. In Szenario 2 wurde von einem anfänglichen Verumeffekt von 0.5 ausgegangen, welcher dann ebenfalls auf 0.2 sank.

Die Pooling-Prozedur war viel stärker durch die Varianzinflation betroffen als die beiden anderen Tests, welche wie in den Kapiteln 3.3.2.1 und 3.3.2.2 nahezu parallele Verläufe zeigen. Die Pooling-Prozedur macht mit steigender Varianz in der zweiten Phase einen Verlust an Power von 26.4 Prozentpunkten in Szenario 1 durch, in Szenario 2 einen Verlust von 28.7 Prozentpunkten. Der Kombinationstest und der Abschlusstest werden weniger stark beeinflusst, denn ihre Power fällt in beiden Szenarien um nur ca. 6 Prozentpunkte. Die beiden Testverfahren unterscheiden sich um einen Betrag von 1.5 bis 2 Prozentpunkten und ab dem Varianzfaktor 1.5 kann bei ihnen eine höhere Power (7.1 bzw. 5.3 Prozentpunkte) als bei der Pooling-Prozedur verzeichnet werden, der Unterschied wird dann immer größer und ist schließlich bei einem Varianzfaktor 3 bei beiden Prozeduren über 20 Prozentpunkte groß.

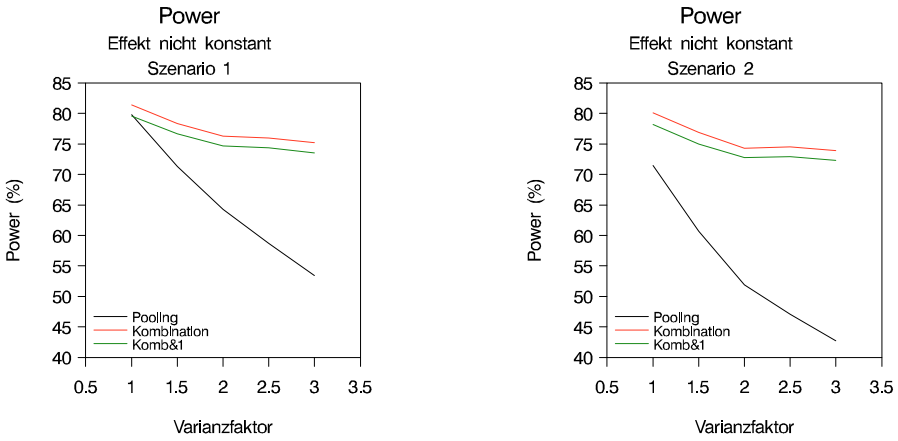

Abbildung 3.21:
Power der drei Testverfahren unter der Bedingung sich über die Phasen ändernder Mittelwertdifferenzen und sich ändernder Varianzen. Im Szenario 1 (2) war der Zeitpunkt der Varianzinflation bei der Hälfte (einem Drittel) der rekrutierten Patienten, die Mittelwertdifferenz fiel von 0.7 (0.5) auf 0.2.

4 Diskussion

4.1 Reale Studien

4.1.1 Rekrutierungsgeschwindigkeit und -ungleichheit

Rekrutierungsgeschwindigkeit Die mittlere Rekrutierungsgeschwindigkeit der einzelnen Studien erwies sich als sehr variabel, sie schwankte bei den Interventionsstudien zwischen einem Patienten pro Monat bis zu knapp fünf Patienten pro Tag. In [4] wurden Rekrutierungsraten für 10 multizentrische Studien berichtet, die beim Cooperative Studies Program Coordinating Center (CSPCC) am VA Medical Center, Perry Point, Maryland, zu verschiedenen Indikationen (z.B. Behandlung von Alkoholismus, Diabetes oder Epilepsie) durchgeführt wurden. Die Rekrutierungsraten betrugen für diese Studien zwischen durchschnittlich 0.0013 und 0.88 Patienten pro Tag. Obwohl diese Studien damit nicht so schnell rekrutierten wie die hier untersuchten, liegt dort ebenfalls eine große Spannbreite vor.

Die Untersuchung von Zusammenhängen von Rekrutierungsgeschwindigkeit und den erhobenen Metadaten ergab signifikante Einflüsse durch die Art Studie (Nichtinterventionsstudien waren zu einem großen Anteil unter den schnell rekrutierenden) und die Anzahl der Zentren (je mehr Zentren desto höher die Geschwindigkeit). Die Dauer der Studie zeigte ebenfalls einen signifikanten Einfluss, welcher angesichts der Tatsache, dass die Fallzahl meist vorgeplant ist, offensichtlich erscheint.

Variablen, die aufgrund der Datenlage kaum belastbare Informationen erbringen konnten, waren die Chronizität der Erkrankung (nur 3 Studien zu chronischen Erkrankungen), Randomisierung (nahezu alle Interventionsstudien waren randomisiert), die Anzahl der Behandlungsarme (fast alle Interventionsstudien hatten zwei Behandlungsgruppen). Außerdem konnte der Einfluss der Verblindung nicht von dem der Verwendung von Placebokontrollen getrennt werden, da bei den untersuchten Studien die beiden Eigenschaften nur gemeinsam auftraten. Folgende schwache Trends konnten beobachtet werden: Phase-III-Studien waren eher um den Median versammelt, Phase IV eher in den Extremen der Rekrutierungsgeschwindigkeit. In der Reihenfolge *akute Indikation*, *beides möglich*, *chronische Indikation* gab es einen Anstieg der mittleren Geschwindigkeit, verblindete bzw. placebokontrollierte Studien schienen im Mittel und im Median schneller als die anderen zu sein, ebenso wie post-GCP-Studien waren Studien mit Amendment etwas schneller, genauso wie zweiarmige im Vergleich zu dreiarmigen Studien. Bei der Art der Intervention schienen die Nichtmedikamentenstudien schneller zu sein.

Die Frage, ob eine chirurgische Intervention eine Rolle bei der Rekrutierung spielt, wird in der Literatur widersprüchlich beantwortet [48, 49]. Da unter den untersuchten Studien zu

wenige mit chirurgischer Intervention vertreten sind, kann keine der beiden Beobachtungen bestätigt werden. Dagegen konnten die Rekrutierungsraten bei medikamentöser Intervention in [48] mit den vorliegenden verglichen werden. Während dort die mediane Dauer zwischen zwei Patientenaufnahmen *pro Zentrum* 34.8 Tage betrug (Q1:11.7, Q3:98.0), war diese bei den hier untersuchten Studien kürzer: 13.2 Tage (Q1:3.7, Q3: 32.7), was darin begründet sein könnte, dass die Studien in [48] sämtlich in Großbritannien durchgeführt wurden und dort aufgrund des Gesundheitssystems andere Voraussetzungen für die Rekrutierung herrschen als in Deutschland.

Bei den Betrachtungen der vorliegenden Studien ist zusätzlich mit einem (Selektions-)Bias zu rechnen, denn Studien, die wegen zu langsamer Rekrutierung abgebrochen werden mussten, konnten nicht mehr in die Analyse gelangen, wodurch die Stärke des Einflusses der Metavariablen auf die Rekrutierungsgeschwindigkeit unterschätzt werden könnte. Dass der Anteil wegen schleppender Rekrutierung abgebrochener Studien beachtlich sein kann, sieht man an der Aufstellung in [4], in der von elf Studien drei wegen nicht mehr erfüllbaren Rekrutierungszielen abgebrochen wurden. An anderer Stelle wird ein Anteil von 25 von 195 Studien (13%) berichtet [45].

Ungleichmäßige Rekrutierung Ungleichmäßige Rekrutierung wurde in dieser Arbeit mit Hilfe der Größe δ_{rek} quantifiziert. Sowohl der Median als auch der von Null signifikant unterschiedliche Mittelwert zeigten an, dass die Studien eher zu einer schnellen anfänglichen Rekrutierung neigen. Dies kann jedoch ebenso durch einen Selektionseffekt zustande gekommen sein: Studien, die anfangs eher langsam oder gar schleppend rekrutierten liefen eher Gefahr, deswegen abgebrochen zu werden und in dieser Aufstellung zu fehlen – bei der Recherche nach Studien tauchte der Abbruch wegen schleppender Rekrutierung des Öfteren als Grund für das Nichtvorhandensein von Daten auf.

Bei keiner der Metaeigenschaften konnte ein signifikanter Zusammenhang zur ungleichmäßigerer Rekrutierung gefunden werden. Es galten weiterhin die gleichen Einschränkungen für die möglichen Schlüsse, die auch schon für die Rekrutierungsgeschwindigkeit genannt wurden. Folgende Trends zeichneten sich schwach ab: drei der fünf Phase-III-Studien fanden sich in den äußeren zwei Quartilen der Gesamtverteilung der Interventionsstudien. Die Studien mit chronischer Indikation hatten die erwartete Verteilung, von den akuten dagegen befand sich keine in den äußeren Quartilen. Unizentrische Studien schienen gleichmäßiger zu rekrutieren. Bei den aktiv kontrollierten und unverblindeten Studien (es waren dieselben Studien) fand sich ein höherer Anteil unter den anfangs schnellen. Ein uneinheitlicheres Bild boten die Studien mit den Eigenschaften *Studienstart nach GCP-Einführung*, *Amendment*, *Nichtmedikamentenstudie* und *dreiarmigen Studien*, denn sie hatten größere Anteile sowohl bei den anfänglich langsamen als auch bei den anfänglich schnellen Studien. Keine Unterschiede dagegen konnten aufgrund der Art der Studie festgestellt werden.

Bei Reviews [45] wurde der Anteil an Studien mit zu langsamer Rekrutierung auf die Hälfte geschätzt. Allerdings könnte hier ein Publication-Bias hin zur Überrepräsentierung dieser Problemberichte vorliegen, denn in Veröffentlichungen zu Studien wurde öfter erwähnt, dass es Rekrutierungsprobleme gab, als dass die Rekrutierung plangemäß voran schritt [2]. Um doch

noch die nötige Anzahl an Patienten einzuschließen, musste dann die Rekrutierungsdauer verlängert werden oder die Bemühungen auf anderem Wege gesteigert werden. Das konnte die Eröffnung neuer Zentren oder auch die Verwendung anderer Kanäle zur Kontaktierung der potentiellen Studienteilnehmer sein. Außerdem wiesen die Autoren auch auf Fälle hin, bei denen ein überproportional großer Anteil der Studienteilnehmer gegen Ende der Rekrutierungsphase in die Studie eingeschlossen wurden.

4.1.2 Varianzunterschiede

Für die Untersuchung von Varianzunterschieden wurden die Rohdaten von 40 real durchgeführten Studien untersucht. Darunter waren klassische randomisierte kontrollierte klinische Studien, Anwendungsbeobachtungen und epidemiologische Studien. Von den meisten Studien gab es mehrere analysierte Variablen. In einer ersten Näherung wurden alle 182 Variablen aus diesen Studien als unabhängig betrachtet und einfache deskriptive Statistiken ihrer Varianzverhältnisse berechnet. Es ergab sich eine Verteilung, die rechtsschief (linkssteil) war, d.h. sie war unsymmetrisch und nach rechts hin flach auslaufend. Da es sich bei den Daten um Verhältnisse aus Varianzen / Standardabweichungen handelte, war eine Verteilungsform wie die der F-Verteilung auch zu erwarten. Der Gipfel des Histogramms war um den Wert Eins zu finden, was eine Häufung des Falles keiner Varianzänderung über die Zeit hinweg bedeutet. Auch die zentralen Lagemaße Median und Mittelwert waren nahe Eins. Der positive Wert der Schiefe von 4.36 bestätigte den visuellen Eindruck der Rechtsschiefe der Verteilung, und ihre Steilgipfligkeit, die bereits im Histogramm sehr deutlich wird, wurde durch die deutlich positive Kurtosis von fast 30 untermauert. Dieser Wert -um 3 vermindert- vergleicht jedoch nur das Gewicht der Verteilung mit der Normalverteilung und daher kann er nicht zur Beantwortung der Frage verwendet werden, ob nun mehr oder weniger extreme Varianzverhältnisse als zufällig vorliegen.

Zur Beantwortung dieser Frage musste erst die theoretische Verteilung der Varianzverhältnisse hergeleitet werden, wenn man die Gültigkeit der Nullhypothese keines Varianzunterschiedes über die Zeit bei den Studien annimmt. Diese theoretische Verteilung musste eine Mischung aus mehreren $F_{m,n}$-Verteilungen sein. Um die tatsächliche Verteilung mit der theoretischen zu vergleichen, wurden mehrere Wege beschritten:

Zuerst wurden die beiden Verteilungen graphisch dargestellt. Dabei ergab sich, dass bei der theoretischen Verteilung eine noch viel höhere Kurtosis hätte vorliegen müssen und dass es scheinbar mehr Varianzquotienten als unter der Nullhypothese erwartet in den Ausläufern (d.h. den extremen Bereichen) der vorgefundenen Verteilung gab. Diese Beobachtung passte auch zum Vergleich der Perzentile der theoretischen und empirischen Verteilung: Die der empirischen Verteilung lagen deutlich weiter vom „Normwert“ Eins entfernt als die theoretischen.

Auch die zweite Möglichkeit des graphischen Vergleichens, der P-P-Plot, wies auf eine Unterschiedlichkeit der Verteilungen hin, da der Plot deutlich von der Winkelhalbierenden abwich. Der P-P-Plot ist dem Q-Q-Plot verwandt, leidet aber unter dem Defizit, auch sensitiv für lineare Transformationen bei Verteilungen zu sein. Der Vorteil des P-P-Plots in der gege-

benen Situation war, dass die Wahrscheinlichkeiten der Quantile der empirischen Verteilung durch gewichtete Summation der Verteilungsfunktionen der zu mischenden F-Verteilungen zu bestimmen war. Das hatte zur Folge, dass sowohl die Quantile als auch die Wahrscheinlichkeiten mit maximal möglicher numerischer Präzision bestimmt waren. Für den Q-Q-Plot wäre die Bestimmung der theoretischen Quantile nur über Interpolation bzw. Iteration möglich gewesen, was eine zusätzliche Quelle für numerische Fehler dargestellt hätte.

Drittens wurde mit Hilfe des Kolmogorov-Smirnov-Tests geprüft, ob der optisch festgestellte Unterschied auch statistisch signifikant war. Dabei wurden die real beobachteten Varianzquotienten mit der theoretischen Mischverteilung unter der Nullhypothese verglichen. Mit einem p-Wert von unter 0.01 lieferte auch dieser Test einen Hinweis, dass die Verteilung nicht der theoretischen entsprach.

Aus der deskriptiven Statistik der Varianzverhältnisse kann man entnehmen, welche Werte man in Studien allgemein zu erwarten hat: Das untere Quartil fiel auf eine 12-prozentige Abnahme der Standardabweichung, während das obere auf eine 11-prozentige Zunahme deutete. Es fielen 90% der Werte in den Bereich zwischen einer Abnahme um 35%- und einer Zunahme um 69%. In Extremfällen verringerte sich die Standardabweichung um 93% bzw. vergrößerte sich auf das fast 5-fache. Diese Beobachtungen bildeten eine Richtschnur für die späteren Kapitel, in denen Varianzanstiege gefunden werden, Gegenmaßnahmen ergriffen und schließlich ein alternativer Test verwendet werden sollte.

Die Unterschiede zur erwarteten F-Mischverteilung können auch zu einem gewissen Maß durch die Nichterfüllung idealisierender Annahmen zustande gekommen sein: (1) beruht die Herleitung auf auf der Normalverteilung der Rohdaten und (2) wurde die Trennung in die Phasen 1 und 2 nur aufgrund der zeitlichen Reihenfolge der Rekrutierung vorgenommen, d.h. dass die Behandlungsgruppen nicht zwingenderweise gleich groß waren.

Zuletzt wurden die Varianzverhältnisse mit Hilfe eines gemischten Modells untersucht. Dieses gemischte Modell verwendete die Variable *Studie* als zufälligen Effekt und berücksichtigte damit die hierarchische Struktur der Daten. Der Anteil an erklärter Streuung aufgrund der Clusterung der Daten belief sich auf ca. 60%. Die Hinzunahme der Meta-Eigenschaften der Studien brachte nahezu keine Änderung dieses Wertes mit sich, welcher dann als partieller (d.h. um den Einfluss der Metavariablen bereinigter) Intraklassen - Korrelationskoeffizient zu lesen ist. Keine der zu den o.g. Eigenschaften korrespondierende Variable war im jeweiligen Modell signifikant von Null verschieden, so dass sich dadurch auch kein Einfluss ablesen ließ. Die kleinsten p-Werte betrugen 0.09 (*Rekrutierungsungleichheit*) und 0.13 (*Amendment*). Die Regressionsparameter dieser beiden Einflussfaktoren betrugen 0.26 und 0.35, woraus insbesondere geschlossen werden kann, dass größere Rekrutierungsungleichheit und Amendments den Varianzquotienten steigen lassen, was plausibel erscheint. Generell ließen alle Metavariablen außer der Rekrutierungsdauer und der Übergang von chronischer zu akuter Indikation den Wert für den Varianzquotienten und damit die Varianz in Phase 2 gegenüber Phase 1 steigen.

Bei den beiden Variablen *Rekrutierungsungleichheit* und *Amendments* wurde auch inhaltlich bereits eine Verbindung zur Varianzdynamik vermutet. Möglich ist auch, dass sie auf demselben kausalen Pfad liegen: Amendments bewirken eine Rekrutierungsungleichheit, was

wiederum zu heterogeneren Patientengruppen und damit zu ungleicher Varianz in den Phasen führen kann.

Jedoch bleibt festzuhalten, dass sich an dem durch die Clusterung erklärten Anteil von ca. 60% durch die Hinzunahme weiterer möglicher Prädiktoren fast nichts geändert hat und dass deshalb noch 40% unerklärter Varianz verbleiben.

4.1.3 Einschränkungen

Bei der Suche nach Rohdaten zu real durchgeführten Studien erwiesen sich die Quellen außer der Projektdatenbank aus verschiedenen Gründen als problematisch. In der Publikationsliste war die Zuordnung zum konkreten Projekt oft schwer möglich und zusätzlich konnte aus dem Vorhandensein einer Publikation unter Institutsmitarbeit nicht darauf geschlossen werden, dass auch Daten dazu vorlagen, außerdem konnten mehrere der insgesamt 1337 verzeichneten Publikationen auf demselben Projekt basieren und dies möglicherweise auch zu verschiedenen Entwicklungsphasen des Projektes (zu erkennen an z.B. variierenden Fallzahlen).

Im Internet fanden sich einige Studiendaten, wobei diese oft nicht den Anforderungen entsprachen, um in die hier angestellten Analysen einbezogen zu werden. Ein Fall der des Öfteren eintrat war das Fehlen einer Datumsvariablen. *Wenn* eine solche Variable vorzufinden war, dann aus dem Grund, dass diese in eine Überlebenszeitauswertung einbezogen werden musste – dann wiederum war oft keine alternative Zielgröße in der Datenbank enthalten. Bei Rohdaten aus der Literatur fehlten aus Platzökonomie oft die Datumsvariable und zusätzlich konnte es vorkommen, dass die Datenbanken um Patienten gekürzt worden waren. Wenn Letzteres nicht der Fall war, war die Fallzahl meist zu gering. Diese Studiendaten stammten fast ausschließlich aus Quellen, bei denen sie als Beispieldaten zu statistischen Verfahren dienten. Es wurde auch schon durch andere Autoren beobachtet [65], dass es sehr selten ist, öffentlich zugängliche Daten aus Zulassungsstudien vorzufinden.

Des Weiteren ist bei der Recherche nach Daten das Problem des *Medienbruches* zu erwähnen. In diese Untersuchung wurden nur elektronisch verfügbare Daten einbezogen. Es ist aber davon auszugehen, dass zu älteren Studien sehr viel weniger Daten auf elektronischen Trägern zur Verfügung standen. Dies könnte zu einem Selektionsbias bei den älteren Studien geführt haben, wobei offen ist, in welche Richtung dieser gewirkt haben könnte.

Die 40 bzw. 29 Studien, deren Daten geeignet waren, reichten bei vielen Fragestellungen und Auffälligkeiten nicht für statistische Signifikanzen, der Teil über die Analyse der real durchgeführten Studien ist jedoch als explorativ zu betrachten. Dieses Problem wird weiter verschärft durch die Tatsache, dass die ermittelten Eigenschaften der Studien keinesfalls unabhängig sind. Das Paar aus Verblindung und der Art der Kontrollen z.B. kann nicht vollkommen unkorreliert sein, denn Studien mit Placebokontrolle sind nur verblindet sinnvoll. Ein weiteres Paar betrifft das Jahr der Studie und die Phase der Entwicklung. Die Einteilung in die Phasen I-IV wurde zwar schon in den 1970er Jahren verwendet, aber in den vorliegenden Daten sind Informationen dazu erst nach dem Verbindlichwerden der GCP-Richtlinien verfügbar gewesen, so stammten alle neun Studien, die sicher in die Phasen III oder IV eingeordnet werden konnten, aus diesem Zeitraum.

Zur Rekrutierungsungleichheit lässt sich einschränkend sagen, dass zwar ein numerischer Wert für die Ungleichheit hergeleitet werden konnte, dass aber damit noch nicht klar ist, welche Grenzen für eine *relevant* ungleichmäßige Rekrutierung gesetzt werden müssen. Die Verwendung von Quartilen an dieser Stelle erreicht nur eine Einteilung *innerhalb* der untersuchten Menge von Studien und lässt eine Übertragung auf andere Studien prinzipiell nicht zu.

4.2 Verblindete Fallzahladaption

Im Abschnitt über die verblindete Fallzahladaption wurden Simulationsstudien unter der Nullhypothese keines Unterschieds durchgeführt, um Eigenschaften der mehrfachen verblindeten Fallzahladaption zu ermitteln. Diese Eigenschaften betrafen den Fehler 1. Art des statistischen Tests am Ende der Studie, die tatsächlich erreichte Fallzahl und ihre Variabilität. Untersucht wurden sowohl normalverteilte als auch binäre Zielgrößen unter verschiedenen Bedingungen für die Studiengröße und die Varianzannahme bei der Planung.

Die vorgestellten Prozeduren basieren im Falle einer stetigen Zielgröße hauptsächlich auf der Arbeit von Kieser und Friede [19] und im Falle einer dichotomen Zielgröße hauptsächlich auf der von Friede und Kieser [22]. Die untersuchten Methoden erweitern letztere in dem Sinne, als dass die Auswirkungen mehrmaliger Adaption und im stetigen Falle der Verwendung des zusätzlichen Kriteriums *Control-Chart* untersucht wurden.

4.2.1 Stetige Endpunkte

4.2.1.1 Fehler 1. Art

Bei der Fallzahlrekalkulation ist besonders auf den Fehler 1. Art des Tests zu achten. Steins Prozedur [8] hält das Niveau sogar exakt ein [21], was allerdings nicht nur zwischenzeitliche Entblindung erfordert, sondern auch nur gilt, wenn über alle möglichen Werte der Interimschätzung der Varianz gemittelt wird [66]. Das bedeutet, dass der Fehler 1. Art inflationiert werden kann, wenn sich die Varianz im zweiten Teil der Studie stark erhöht. Somit kontrolliert Steins Prozedur nur den *unbedingten* Fehler 1. Art.

Die ebenfalls entblindende und auf dem Verfahren von Stein basierende Prozedur mit interner Pilotstudie von Wittes und Brittain [9] kann ebenfalls das Niveau verletzen, wobei dieser Effekt für praxisnahe Situationen nicht sehr stark ist. Es wurden bei den untersuchten Verfahren, die die Eigenschaft *Restriktion* aufweisen, Fehlerraten von 5.1% und auch 5.2% gefunden, wenn die Varianz 1.5-mal bzw. doppelt so groß wie vorher angenommen. Die Autorinnen empfehlen daher, die Größe der Inflation des α-Fehlers bei einer Rekalkulation abzuschätzen.

Birkett und Day [57] suchen nach einer Regel für die Größe der Interimstichprobe und berichten dabei auch Ergebnisse zum Fehler 1. Art der ursprünglichen wie auch einer unrestringierten Variante des Verfahrens nach Wittes und Brittain [9]. Die unveränderte Prozedur erzeugte bei überspezifizierter Varianz (reale niedriger als angenommen) α-Fehler zwischen 4.7 und 5.0%, bei korrekt spezifizierter Varianz solche zwischen 4.9 und 5.1% und bei unterspe-

zifizierter Varianz Fehleranteile zwischen 5.1 und 5.3%. Im für die Rekalkulationsprozeduren zentralen motivierenden Fall der Unterspezifizierung konnte demnach die Antikonservativität bestätigt werden. In der unrestringierten Variante war bei allen Spezifikationen der Varianz eine deutliche Antikonservativität zu verzeichnen, wenn die Pilotstichprobe sehr klein war. So wurde bei Unterspezifizierung der Varianz und einer Pilotstichprobe nach $n = 5$ Patienten pro Gruppe ein α-Fehler von 5.3% gefunden. Dieser wurde größer, wenn die Stichprobe weiter verkleinert wurde.

Dass die Methode der internen Pilotstudie von Wittes und Brittain das Niveau verletzen muss, wurde auch durch numerische Integration gezeigt [67]. Es wurden α-Fehler zwischen 4.92 und 5.33% für die unrestringierte Prozedur und zwischen 4.85 und 7.9% für die restringierte Prozedur berechnet. Bei der restringierten Prozedur fiel die Inflation des α-Fehlers mit steigendem Verhältnis aus vermuteter Varianz und Behandlungseffekt, außerdem schienen die Fälle mit unterspezifizierter Varianz problematisch. Für die unrestringierte Prozedur waren die Einflüsse der Designparamter auf das Niveau komplizierter, z.B. konnte eine kleine Größe der Pilotstichprobe zu einem übermäßig großen α-Fehler führen. Es wurde empfohlen, in Fällen mit großer zu erwartender Inflation des Fehlers 1. Art die kritischen Grenzen des t-Tests entsprechend (z.B. 1.99 statt 1.96) anzupassen.

Im Artikel wurden mehrere Kombinationen aus dem Verhältnis des Quadrats der vermuteten Varianz zur Mittelwertdifferenz, des Anteils, den die Interimstichprobe an der ursprünglich geplanten Gesamtfallzahl ausmachte und dem Verhältnis der vermuteten zur realen Varianz untersucht. Die entsprechenden Parameter der vorliegenden Untersuchung können nicht direkt auf die aus der genannten Literatur abgebildet werden. Wurden die α-Fehler *ähnlicher* Konfigurationen (keine Control-Charts oder KF-Korrektur, ähnliche Studiengröße und Varianzfehlspezifikation) verglichen, war zu erkennen, dass die hier untersuchten Verfahren zu kleineren Fehlern tendierten, wobei dieser Unterschied bei den kleinen Fallzahlen und unrestringierten Prozeduren auffälliger war.

Zucker et al. [20] wiederholten die Untersuchungen aus [67] mit einer leicht modifizierten Prozedur nach Wittes und Brittain. Es wurde ebenfalls numerisch integriert, jedoch Steins Rekalkulationsformel und ein einseitiger Test mit einem Signifikanzniveau von 2.5% verwendet. In verschiedenen Situationen wurden α-Fehler zwischen 2.50 und 3.95% für die unrestringierten und Fehler zwischen 2.50 und 2.66% für die restringierten Prozeduren ermittelt.

In [68] wurde ein stark erhöhter Anteil an Fehlern 1. Art durch die Methode von Wittes und Brittain verursacht, wenn sie unrestringiert verwendet wurde. Die Erhöhung des Fehlers trat zusammen mit einer niedrigeren Trennschärfe insbesondere dann auf, wenn die Fallzahl pro Gruppe in der Pilotstudie *sehr* klein war (≤ 5). Bei der durch Denne und Jennison [10] optimierten Variante der Prozedur lagen die empirischen Fehlerraten 1. Art näher an den nominellen bzw. wiesen eine kleinere Antikonservativität auf. Der α-Fehler betrug beispielsweise für die größte Interimstichprobe bei zweifacher Unterschätzung der tatsächlichen Varianz 4.9%, bei korrekter Varianz 5.0% und bei zweifacher Überschätzung 5.1%, während beim Verfahren nach Wittes durchgehend 5.1% auftraten. Kieser und Friede [18] leiteten schließlich für die Prozedur eine Methode zur Adjustierung des α-Levels her, indem sie die maximale Inflation des Fehlers 1. Art abschätzten.

Gould und Shih stellen in [12] ein Verfahren vor, das ohne zwischenzeitliche Entblindung der Daten auskommt. Es werden zwei Möglichkeiten angegeben, um die Varianz verblindet zu schätzen: Einerseits eine approximative Variante der Adjustierung, wie sie auch von Zucker et al. [20] und Kieser und Friede [19] angegeben wurde. Andererseits wurde der EM-Algorithmus verwendet. Beide Methoden führten zu Rekalkulationsprozeduren, die den α-Fehler kaum veränderten, wie aus einer expliziten Darstellung des Bias abgeleitet werden konnte. In einer Simulationsstudie [13] ermittelte Gould für sein Verfahren den empirischen α-Fehler von 5.04% (95%-KI: 4.74% - 5.34%, Min: 3.9%, Max: 5.9%), wenn die Rekalkulation nach 1/4 der initialen Fallzahl stattfand. Wenn nach der Hälfte rekalkuliert wurde betrug der Fehler 5.11% (95%-KI: 4.73% - 5.50%, Min: 3.6%, Max: 6.3%). Zucker et al. [20] kategorisieren die Methode von Gould und Shih als eine, die im Gegensatz zur Prozedur von Stein und Wittes / Brittain auch das auf die finale Fallzahl bedingte Niveau einhält, was über ein Permutationsargument begründet wird.

In [1] werden ebenfalls Ergebnisse zu α-Fehlern des Verfahrens nach Gould & Shih dargestellt. Diese bewegten sich zwischen 4.8 (95%-KI: 4.22-5.43%, da auf 5000 Replikationen basierend) und 5.3% (95%-KI: 4.70-5.96%). Es lässt sich dabei weder ein Trend über die Größe der Pilotstichprobe noch über die Größe der Fehlspezifikation der Varianz erkennen.

Die Ergebnisse zur Prozedur nach Gould und Shih sind nur eingeschränkt mit denen aus dieser Arbeit vergleichbar, da es eine Reihe von wichtigen Unterschieden bei den Verfahren gab, so z.B. die Beschränkung der Gesamtfallzahl durch das Doppelte der initialen Fallzahl, sowie die Bedingung, dass eine Rekalkulation nur dann zu einer neuen Fallzahl führte, wenn die beobachtete Varianz mindestens 5% (oder 30%) größer als die für die initiale Planung verwendete war.

Die **für diese Arbeit** untersuchten Verfahren leiten sich primär aus den Betrachtungen von Kieser und Friede [19], die durch numerische Integration gezeigt haben, dass die einfache Rekalkulation das Niveau exakt einhält. Daher wurde vermutet, dass dies auch bei der mehrfachen Rekalkulation der Fall war und für die meisten Szenarien wurde dies auch bestätigt. Im Szenario mit korrekt spezifizierter Varianz gab es keinen konsistenten Trend über die Studiengrößen hinweg und die Eigenschaften der Restriktion, des Control-Charts und der Adjustierung der Varianz schienen sich nicht auf den Fehler 1. Art auszuwirken. In der Situation mit fehlspezifizierter Varianz waren sowohl Verletzungen des vorgegebenen Niveaus von 5% als auch Gruppierungen von Prozeduren zu erkennen, die sich abhängig von der Situation nach unterschiedlichen Kriterien bildeten.

So gab es eine ausgeprägte Gruppenbildung bei zu niedriger und zu hoher Varianz, wobei es bei den Szenarien mit zu kleiner angenommener Varianz eine Einteilung in restringierte und unrestringierte Verfahren zu beobachten war. Beim Szenario mit zu hoher Varianz dagegen verhielten sich Prozeduren mit und ohne Restriktion nahezu gleich und bildeten Paare. Die Tatsache, dass in den Gruppen teilweise *identische* Fehler 1. Art beobachtet wurden, kann damit erklärt werden, dass aus einer gleichen Fallzahl bei zwei verschiedenen Prozeduren folgte, dass sie auch auf der *identischen* Realisation an Zufallszahlen operierten. Aufgrund dessen ist leicht einzusehen, dass in Fällen, in denen wegen zu kleiner realer Varianz die Fallzahl hätte herab gesetzt werden sollen, die restringierten Prozeduren alle auf ihren zu großen Fallzahlen

fixiert wurden und somit die gleichen Daten im t-Test verwendet wurden, was zum gleichen p-Wert und dann zum gleichen Anteil an signifikanten Replikationen führte. Die Paarbildung bei der zu großen realen Varianz hat einen ähnlichen Grund: jede unrestringierte Prozedur und ihr restringierter Gegenpart lieferten die gleiche Fallzahl, da die Restriktion in diesem Falle keine Rolle spielte.

Signifikante Abweichungen vom 5%-Fehlerniveau wurden nur bei unrestringierten Verfahren (einmal auch bei der Basisprozedur uosnc) beobachtet und zwar bei zu kleiner realer Varianz und kleineren Studientypen, wobei es Abweichungen nach oben als auch nach unten gab. Diese Beobachtungen decken sich zwar nicht mit denen Erwartungen, andererseits ist bei einer Familie von 96 Konfidenzintervallen (8 Prozeduren, 3 Szenarien, 4 Studiengrößen, zur Vereinfachung Unabhängigkeit voraussetzend) mit einer jeweiligen Überdeckungswahrscheinlichkeit von 95% auch mit ca. $96 \cdot 5/100 = 4.8$ Fällen zu rechnen, in denen zwar der theoretische Anteil vorliegt, aber das Intervall diesen nicht überdeckt. Damit entspricht die beobachtete Anzahl von fünf Fällen in etwa der Erwartung. Nach Bonferroni-Holm-Adjustierung [64] der 95%-Konfidenzintervalle um die empirischen α-Fehler gab es keinen einzigen signifikanten Fall mehr. Die Tatsache, dass die Abweichungen keinem festen Trend folgten, spricht ebenso für einen Zufallsfund.

4.2.1.2 Erreichte Fallzahl

Für die nachfolgenden Betrachtungen der in der Literatur berichteten Ergebnisse war meist die Alternativhypothese gültig – außer in den Fällen, in denen explizit auf das Gegenteil hingewiesen wird.

In Steins Artikel [8] wird eine Abschätzung für die Fallzahl nach der Rekalkulation hergeleitet. Stein kommt zum Schluss, dass die finale Fallzahl nur leicht über der einer fixen Planung mit der wahren Varianz ist, solange die Pilotstichprobe mindestens 30 Beobachtungen umfasst und eine "moderate" Varianz vorliegt.

Die Prozedur von Wittes und Brittain [9] hat die Eigenschaft der Restriktion wie sie auch in der vorliegenden Arbeit zum Einsatz kam. Die berichteten finalen mittleren Fallzahlen lassen Ähnlichkeiten zu den hier untersuchten restringierten Prozeduren zu erkennen: Verharren auf der initialen Fallzahl bei Überspezifizierung der Varianz, leichte Erhöhung (93.2 vs 86 Patienten, d.h. +8.4%) gegenüber fixer Planung bei korrekter Spezifizierung und schließlich proportionales Ansteigen mit der tatsächlichen Varianz bei Unterspezifizierung. In [57] wurde das Verfahren von Wittes und Brittain auch hinsichtlich der erwarteten finalen Fallzahl unter der Alternativhypothese mit im Wesentlichen denselben Ergebnissen untersucht. Wurde die initiale Fallzahlplanung weggelassen und bei festen Fallzahlen eine Interimstichprobe untersucht, wurden je nach Größe dieser Stichprobe Überschätzungen von 1.7 bis 66.7% beobachtet. Das Verfahren wurde unter dem Namen *pooled variance* auch in [69] untersucht. In einem Beispiel wurden bei korrekt spezifizierter Varianz und verschiedenen Größen der Pilotstichprobe finale Fallzahlen beobachtet, die maximal 3.0% über der tatsächlich benötigten lagen.

Für das Verfahren nach Denne und Jennison [10] wurde das Verhältnis aus erwarteter und

tatsächlich benötigter Fallzahl numerisch berechnet. Die finalen Fallzahlen sind demnach am nähesten bei den tatsächlichen, wenn die Fallzahl initial leicht unterschätzt wurde. Bei stark unter- oder überschätzter Fallzahl waren dagegen die finalen Fallzahlen deutlich erhöht (s. Figure 1). In praxisnahen Situationen wurden optimale Fallzahlen ermittelt, die 4%, 9% oder 2% über der tatsächlich benötigten Fallzahl lagen, wenn die Varianz korrekt, doppelt bzw. halb so groß wie die tatsächliche angenommen wurden.

Gould und Shih stellen fest, dass die erwartete Fallzahl nach Rekalkulation mit ihrer Prozedur eine rechtsschiefe Verteilung aufweist [12]. Zucker et al. [20] berechneten für das Verfahren von Gould und Shih mittlere Fallzahlen, die sowohl im unrestringierten wie im restringierten Fall durchgehend über denen einer konventionellen Fallzahlplanung mit der korrekten Varianz lagen. Die Gruppe mit dem kleinsten angenommenen Effekt (ca. 0.4 Standardabweichungen) wies dabei mit +1.2 bis +200.5% die kleinsten Fallzahlinflationen auf.

In [1] berichtet Gould für eine gültige Nullhypothese indirekt das Verhältnis von tatsächlich benötigter und mittlerer finaler Fallzahl seiner auf dem EM-Algorithmus basierenden Prozedur. Aus den Angaben zum Verhältnis $f_{\text{var}} := \text{CV}^2_{\text{true}}/\text{CV}^2_{\text{Des}}$ des angenommenen und des tatsächlichen Variationskoeffizienten und dem Verhältnis $f_{\text{ss}} := n_{\text{final}}/n_{\text{init}}$ der erreichten zur initialen Gesamtfallzahl können die Verhältnisse $n_{\text{final}}/n_{\text{tats}}$ als $f_{\text{ss}}/f_{\text{var}}$ bestimmt werden. Für eine unterspezifizierte Varianz wurde aufgrund der Restriktion dieser Prozedur beim Doppelten der initial bestimmten Fallzahl gestoppt, bei korrekter Spezifizierung ergab sich eine Überschätzung um 1%, bei 1.5- bzw. 2-facher höherer tatsächlicher Varianz blieb die Prozedur 3.3% bzw. 5% unterhalb der Fallzahl einer fixen Planung.

Friede und Kieser [69] untersuchten die auf dem EM-Algorithmus von Gould basierende Prozedur, wobei in Abhängigkeit von der Größe der Pilotstichprobe die tatsächlich benötigte Fallzahl um 1.25% unter- und bis zu 4.69% überschätzt wurde.

Im Falle der korrekten Spezifizierung der Varianz endeten alle **hier untersuchten Verfahren** mit Fallzahlen in der Größenordnung, die sich auch bei fixer Fallzahlplanung ergeben hätten. Die Verwendung der mehrmaligen Rekalkulation war retrospektiv in diesem Fall nicht nötig. Sie verursachte mittlere Abweichungen von der idealen Fallzahl um maximal fast 10% nach unten bzw. maximal 6% nach oben. In absoluten Fallzahlen war im Mittel bis zu 9 Patienten zu wenig bzw. 13 Patienten zu viel rekrutiert worden.

Die Basisprozedur uosnc gehörte dabei zu den Verfahren mit den kleineren mittleren Abweichungen von der tatsächlich benötigten Fallzahl im Mittel etwas zu kleine Fallzahlen, die Abweichung war jedoch mit maximal -1.5 Patienten (-1.4%) gering.

In den Fällen mit fehlspezifizierter Varianz konnte die Abweichung bei allen Rekalkulationsverfahren größer sein. Dies war insbesondere bei restringierten Prozeduren und überschätzter Varianz der Fall, wobei im erzeugten Fall einer real halb so großen Varianz die Fallzahl entsprechend doppelt so groß war wie eigentlich benötigt, was in Einklang mit der zentralen Fallzahlformel steht. Ansonsten waren mittlere Abweichungen um -15.5% und -10 Patienten nach unten und solche um +1% und 3.2 Patienten nach oben möglich. In diesen Fällen bewegten sich die finalen Fallzahlen damit in Größenordnungen, die sich auch unter Kenntnis der wahren Varianz mit fixer Fallzahlplanung ergeben hätten.

Die Prozedur uosnc wies im Fall einer real niedrigeren Varianz – verglichen mit den anderen

Prozeduren – die kleinsten Abweichungen zur fixen Fallzahlplanung mit der korrekten Varianz auf. Im Fall einer real zu großen Varianz wurden nur bei den beiden oscc-Verfahren kleinere Abweichungen nach unten beobachtet.

Restriktion führte, wie schon ausgeführt, aufgrund der Bedingung an die Fallzahl immer zu größeren mittleren finalen Studienumfängen als es bei nicht restringierten Verfahren der Fall war. Naturgemäß war diese Differenz sehr groß, wenn die Fallzahl aufgrund einer kleinen realen Varianz hätte heruntergesetzt werden müssen. In den anderen Fällen war der maximale Abstand zwischen unrestringierten und restringierten Prozeduren kleiner. Er fiel von 100% auf 14.5% bei korrekter Varianzannahme und auf 0.04% bei unterschätzter Varianz. In absoluten Fallzahlen waren dies 200, 18 und 0 Patienten.

Die *KF-Korrektur* besteht in einer Verkleinerung des Varianzschätzers aus den os-Prozeduren, so dass folglich mit einer kleineren Fallzahl als bei den zweitgenannten Verfahren zu rechnen war. Die Korrektur führt jedoch nur dann zu einem biasfreien Varianzschätzer, wenn der Gruppenunterschied korrekt spezifiziert wurde [19]. Die untersuchten Szenarien wurden jedoch alle unter der Nullhypothese *keines* Gruppenunterschieds betrachtet, so dass eine systematisch negativ verzerrte Schätzung des benötigten Stichprobenumfanges resultieren musste. Nach [69] beträgt die erwartete Fallzahl pro Gruppe

$$E(\hat{N}_{\mathrm{ADJ}}) = 2 \cdot \frac{(z_\alpha + z_\beta)^2 \sigma^2}{\Delta^{*2}} \left[1 + \frac{n_1}{2(2n_1 - 1)} \left(\left(\frac{\Delta}{\sigma} \right)^2 - \left(\frac{\Delta^*}{\sigma} \right)^2 \right) \right] = N I_\sigma,$$

d.h. die erwartete Fallzahl pro Gruppe ist das I_σ-fache der Fallzahl einer fixen Planung. Im Spezialfall einer gültigen Nullhypothese vereinfacht sich dies zu

$$E_0(\hat{N}_{\mathrm{ADJ}}) \approx N - \frac{1}{2}(z_\alpha + z_\beta)^2.$$

Es war demnach, unabhängig von der tatsächlichen Fallzahl, mit insgesamt ca. $2 \cdot 1/2 \cdot (z_\alpha + z_\beta)^2 \approx 7.2$ Patienten zu wenig zu rechnen. Dies entspricht in etwa der beobachteten Unterschätzung bei der Prozedur ukfnc bei der Annahme der korrekten Varianz (-8.7 bis -9.1 Patienten). Der Faktor I_σ verursacht im Falle einer korrekt spezifizierten Varianz für ukfnc eine Unterschätzung um -1.8 bis -7.6%, beobachtet wurden -2.2 bis -9.4%. In beiden Fällen war die Unterschätzung durch die mehrfache Rekalkulation etwas größer als bei der einfachen Rekalkulation.

Die obige Abschätzung für $E_0(\hat{N}_{\mathrm{ADJ}})$ ist unabhängig von der Studiengröße und der Art der tatsächlichen Varianz, weswegen eine Unterschätzung gleichen Umfangs bei der einfachen Rekalkulation zu vermuten war. Dagegen wurde bei der mehrfachen Rekalkulation durch ukfnc statt der erwarteten 7.2 bei real zu kleiner Varianz 7.4 bis 7.7 Patienten zu wenig rekrutiert. Bei real zu großer Varianz waren es durchgehend 10 Patienten zu wenig.

Die relativen Abweichungen bei einfacher Rekalkulation betrugen -2.8 bis -15.5% und entsprachen damit in etwa den theoretischen von -3.6 bis -15.2% bei real zu niedriger Varianz. War die Varianz höher als erwartet, wich ukfnc im Mittel zwischen -1.3 und -5.5% von der fixen Fallzahl ab, bei einfacher Rekalkulation waren -0.9 bis -3.8% zu erwarten.

Es ist also bei der dreimaligen Rekalkulation unter der Nullhypothese und bei korrekter

Annahme über die Varianz zu beobachten, dass sowohl beim adjustierten wie beim kruden Varianzschätzer eine leicht größere Zahl an Patienten zu wenig rekrutiert werden als tatsächlich benötigt werden. Dies kann numerische Gründe aber auch strukturelle wie die dreimalige Anwendung der Adaption haben. Ein zufälliger Fehler aufgrund der Simulation ist ebenso möglich.

Die mittlere finale Fallzahl war, wie zu erwarten, bei korrigierter Varianzschätzung immer kleiner als bei unkorrigierter Schätzung und es waren maximale Abweichungen von bis zu -15.5% zu verzeichnen. Der Unterschied war im Mittel nicht größer als ca. -7.7 Patienten. Weiterhin war damit zu rechnen, dass dieses Verhalten aus zwei Gründen bei kleinen Studien eine größere Rolle spielte: Erstens ist bei kleineren Studien der Gruppenunterschied für die Planung größer gewesen, was zu einem größeren Korrekturterm führen musste (siehe Korrekturformel in Abschnitt 2.2.1.2, Seite 21). Zweitens war die Größenordnung des Unterschieds immer gleich und daher der Anteil bei kleinen Studien entsprechend größer.

Die Verwendung von *Control-Charts* hatte den Sinn, eine zu häufige Rekalkulation zu verhindern. Abweichungen von der Fallzahl einer fixen Planung mit der wahren Varianz sind durch Control-Charts geringfügig verkleinert worden. Im Falle einer korrekt spezifizierten Varianz wurde die Überschätzung durch die restringierten Prozeduren und die Unterschätzung durch die unrestringierten kleiner, wenn Control-Charts verwendet wurden. Bei Fehlspezifikation der Varianz war durchgehend eine Erhöhung der Minima und Maxima der mittleren Fallzahlen zu beobachten, die einzige Ausnahme bildeten die Fälle der restringierten Prozeduren bei überschätzter Varianz, da die Restriktion stärker wirkte.

4.2.1.3 Variabilität der Fallzahl

Aufgrund der Abhängigkeit der Rekalkulation von den Daten ist die finale Fallzahl eine zufällige Größe geworden, deren Variabilität mit Hilfe des Variationskoeffizienten untersucht worden ist. In der Literatur sind entsprechende Untersuchungen selten. Friede und Kieser [69] berichten für die Prozedur nach Wittes und Brittain (*pooled variance*) einen Variationskoeffizienten von 9.9 - 32.7%, der mit zunehmender Größe der Pilotstichprobe kleiner wird. Ein gleicher Trend war für die EM-Prozedur von Gould zu beobachten, der Variationskoeffizient blieb für diese Prozedur zwischen 10.4 - 31.8%. Für das Verfahren mit KF-Korrektur wurden Werte zwischen 10.3 und 33.9% berichtet.

Der Variationskoeffizient für **die hier untersuchten Prozeduren** fiel bei allen angenommenen Varianzen mit steigender Studiengröße. Dabei war der Wert des Variationskoeffizienten nie größer als 26.7%. In Szenarien mit zu kleiner realer Varianz war die Variabilität der Fallzahlen bei den restringierten Prozeduren praktisch immer Null, was sich sofort aus der Restriktion begründen lässt. Des Weiteren bildeten die restringierten Prozeduren in allen Szenarien außer dem mit zu hoher Varianz eine Gruppe mit kleinerer Fallzahlvariabilität. Dies ist eine Folge davon, dass die finale Fallzahl dieser Prozeduren prinzipiell nur über der initialen liegen konnte. Beim genannten Sonderfall fand eine Teilung in zwei Gruppen aufgrund des Merkmals *Control-Chart* statt.

Bei den zuerst genannten Szenarien gab es innerhalb der Gruppierung nach dem Merk-

mal der Restriktion noch wechselnde Untergruppierungen, die mehr oder weniger stark ausgeprägt waren: bei kleineren Studiengrößen folgte die Untergruppierung dem Merkmal der KF-Korrektur, wobei die Prozeduren *mit* diesem Merkmal eine höhere Fallzahlvariabilität aufwiesen. Bei größeren Studientypen trat eine Untergruppierung nach dem Merkmal des Control-Charts an ihre Stelle. Eine Erklärung hierfür könnte sein, dass durch die mehrfache Rekalkulation im Gegensatz zur Vorlage die Größe der Interimstichprobe selbst eine Zufallszahl geworden ist. In [19] war die Größe der Interimstichprobe durch die initiale Schätzung und einen festgelegten Anteil determiniert. Dies trifft auch bei der mehrfachen Rekalkulation für die *erste* Interimstichprobe zu, die zweite und dritte sind dagegen Zufallszahlen.

Mit steigender Studiengröße unterschied sich die Fallzahlvariabilität zwischen den kf- und os-Prozeduren nur noch wenig, denn der Term $\frac{n}{2(2n-1)}$ nimmt für kleine Werte von n größere Werte als bei großen n an, so dass bei kleinen Interimstichproben und KF-Korrektur mit einer größeren Variabilität des Varianzschätzers aufgrund der Variabilität in der Größe der Interimstichprobe zu rechnen ist. Hinzu kommt, dass bei großen Studien eine vergleichsweise kleine Mittelwertdifferenz zugrunde gelegt werden konnte, was den Einfluss der Variabilität aufgrund der Größe der Interimstichprobe weiter verkleinert hat.

Im Sonderfall mit konstant zu großer Varianz ist zu erkennen, dass die Restriktion wie erwartet keine Rolle spielt und auch zwischen kf- und os-Prozeduren kaum ein Unterschied in der Variabilität der Fallzahl besteht. Als Grund für Letzteres wird die im ersten Schritt auf der Basis einer Interimstichprobe mit deterministischem Umfang bestimmte verdoppelte Gesamtstichprobengröße vermutet, die dann zu vergleichsweise großen zweiten und dritten Interimstichproben führte, welche zu einer vergleichbaren Variabilität der finalen Fallzahl bei den kf- und os-Prozeduren führte. Die Unterschiedlichkeit bei der Streuung der erreichten Fallzahl, die aufgrund der Control-Charts entstand, ist aber unabhängig von den Studiengrößen und blieb somit als einzig sichtbarer Faktor übrig, bei dem die höhere Variabilität mit der Verwendung des Control-Charts assoziiert war.

Bei der Anwendung der Idee des Control-Charts musste eine zusätzliche Bedingung erfüllt werden, so dass es insgesamt weniger oft als bei den anderen Prozeduren zu einer Rekalkulation kam. Ob die Bedingung erfüllt war, wurde durch den aktuellen Wert der Power bestimmt und ob dieser die vorgegebenen Grenzen verlassen hatte. Es kam also ein zusätzliches zufälliges Element in die Entscheidung, ob überhaupt korrigiert wurde, was zu einer größeren Variabilität der erreichten Fallzahlen führte. Durch die zu hohe reale Varianz war eine Korrektur der Fallzahl im ersten Schritt *immer* notwendig. Außerdem war die zu erreichende Power bei den cc-Prozeduren nicht nur durch einen einzelnen Wert (85%), sondern durch einen Bereich (80%-90%) gegeben.

4.2.1.4 Sonstige Bemerkungen

Stein machte keine Vorschläge zum Zeitpunkt der Interimanalyse zu seiner Prozedur [8], Birkett und Day [57] untersuchten diese Frage und kamen zum Schluss, dass nicht der Anteil, sondern die absolute Fallzahl bei der Interimstichprobe den entscheidenden Einfluss auf die Fehler 1. und 2. Art sowie die erwartete Gesamtfallzahl habe. Dabei wurde $n_1 = 10$ Patienten

pro Gruppe als minimale Fallzahl für die Pilotstichprobe ermittelt. Die für diese Arbeit untersuchten Verfahren gehen einen anderen Weg mit einer relativ kleinen ersten (1/4 der initialen Fallzahl) und einer relativ großen (3/4 der tatsächlich benötigten Fallzahl) letzten Interimstichprobe. Für die tatsächlich durchgeführten Simulationen war die minimale Größe einer Interimstichprobe (96/8=)12 Patienten pro Gruppe, welche somit auch die Empfehlung für das verwandte Verfahren erfüllte.

Das Verfahren von Gould und Shih entblindet die Daten nicht, sondern versucht die Modellparameter der Mischverteilung mit dem EM-Algorithmus zu bestimmen und zu vergleichen. Die Frage, ob dadurch nicht *doch* eine implizite Entblindung stattfindet, wurde in [13] behandelt. Dort wurde geschlossen, dass dies sehr unwahrscheinlich ist, da der Schätzer der Intragruppen-Varianz dafür viel zu ungenau sei. Jedoch wird in [70] darauf hingewiesen, dass auch ohne die o.g. Verfahren in Studien mit Blockrandomisierung konsistente Schätzer für Effektgröße und auch Intragruppen-Varianz existieren, wobei es keine Rolle spielt, ob die Blöcke feste Länge haben oder, wie in [15] empfohlen, variable Länge aufweisen.

4.2.2 Binäre Endpunkte

4.2.2.1 Fehler 1. Art

Die für diese Arbeit verwendete Rekalkulationsprozedur für binäre Endpunkte geht auf die Arbeit von Gould und Shih [11] zurück. Es werden dort Varianten vorgestellt, die entweder auf der Differenz, dem relativen Risiko oder dem Quotenverhältnis basieren. Für den Fehler 1. Art wurden die drei zugehörigen empirischen Verteilungsfunktionen gezeichnet und insgesamt Fehler zwischen 4.8 und 5.4% berichtet. Während die auf Quotienten beruhenden Varianten nahe bei der theoretischen Verteilung blieben, wich die Prozedur mit Verwendung der Differenz weiter ab. Die Kurve blieb aber nah an derjenigen, die zur Situation ohne Rekalkulation gehörte. Gould und Shih schlossen aus einer logistischen Regression, dass die PER der wichtigste Einflussfaktor für den Fehler 1. Art ist. Der Ansatz wurde auch in [1] untersucht, wobei kein Zusammenhang zwischen dem Fehler 1. Art und den Designparametern bzw. der Proportion in der Referenzgruppe festgestellt wurde. Die Fehler blieben zwischen 4.4 und 6.1%. Friede und Kieser [22] leiteten für die auf Differenzen beruhende Prozedur den tatsächlichen Fehler 1. Art her (s. S. 84).

Herson und Wittes [33] beschreiben eine Prozedur, die das Verfahren von Wittes und Brittain auf binäre Endpunkte erweitert. Dabei wird die Proportion in der Kontrollgruppe und daraufhin eine neue Fallzahl für den z-Test ermittelt. Die empirischen α-Fehler bewegten sich für alle Zeitpunkte und die zwei Regeln zur Fallzahlbeschränkung zwischen 4.34 und 5.28%. Ein Vergleich mit dem Fall keiner Pilotstichprobe führte zum Schluss, dass weniger die Rekalkulation, als vielmehr die Diskretheit der Verteilung zu den Abweichungen von nominellen 5% Niveau führten. In [1] wurde das Verfahren ebenfalls in einer Simulationsstudie verwendet und Fehleranteile 1. Art zwischen 4.4 und 5.6% gefunden. Auch für diese Prozedur konnte kein Zusammenhang zwischen Designparametern und dem Fehler ausgemacht werden.

Shih und Zhao [14] verwendeten eine Dummy-Stratifizierung, um die Verblindung in der Pilotstichprobe aufrecht zu erhalten. Die Prozedur lieferte Anteile an Fehlern 1. Art, die

konsistent über 5% lagen, die maximale Abweichung konnte dabei +1.42% betragen. Diese α-Inflation wird begründet der "teilweisen" Entblindung, die im Zuge der Prozedur stattfindet.

In allen Szenarien **bei den hier untersuchten Verfahren** gab es mindestens eine Prozedur, die das α-Level verletzte. Jede der vier Prozeduren wurde unter 20 Szenarien (5 Varianzen $\times$ 4 Studiengrößen) durchgeführt, wobei die Basisprozedur unc mit sechs die kleinste Anzahl und ucc mit elf die größte Anzahl an Niveauverletzungen zeigte. Die Verfahren rcc und rnc lagen mit neun und sieben Verletzungen dazwischen. In der Mehrzahl der Szenarien hatten die beiden cc-Prozeduren sehr ähnliche empirische Fehler 1. Art, die aber nicht systematisch ober- oder unterhalb der nc-Prozeduren lagen, aber tendenziell öfter das Niveau verletzten. Bei insgesamt 80 Datenpunken aufgrund der Simulationen war (zur Vereinfachung Unabhängigkeit voraussetzend) mit ca. $80 \cdot 0.05 = 4$ Verletzungen des Niveaus durch Zufall zu rechnen, was durch die insgesamt 33 signifikant von 5% unterschiedlichen Fehleranteile weit überschritten wird. Selbst Bonferroni-Holm-Adjustierung für multiples Testen ließ 30 signifikante Anteile übrig.

Wie in [22] ist auch hier zu vermuten, dass die Liberalität des Pearson-χ^2-Tests mehr Einfluss auf die Niveaueinhaltung gehabt hat, als die Rekalkulation und ihre cc- bzw. re-Varianten. Das tatsächliche Niveau des Pearson χ^2-Tests wurde in [21] angegeben als

$$\alpha_{\text{fix}}^{\text{act}} = \sum_{x_1=0}^{n_1} \sum_{x_2=0}^{n_2} \binom{n_1}{x_1} \binom{n_2}{x_2} \pi^{x_1+x_2} (1-\pi)^{n-x_1-x_2} \mathbb{1}_{\chi^2 \geq \chi^2_{1-\alpha}},$$

wobei x_i die Anzahl der *Erfolge* in der Gruppe i ist $(i = 1, 2)$, n_i die zugehörigen Gruppengrößen mit $n = n_1 + n_2$, $\pi = \pi_1 = \pi_2$ die PER, χ^2 die gewöhnliche Pearsonsche Teststatistik, χ^2_γ das γ-Quantil der χ^2-Verteilung mit einem Freiheitsgrad und $\mathbb{1}_{(.)}$ die Indikatorfunktion ist. Die wirklichen Fehlerraten 1. Art des Tests wurden in Abbildung 4.1 für die hier verwendeten Werte der PER und für Gesamtfallzahlen zwischen 40 und 500 Patienten berechnet.

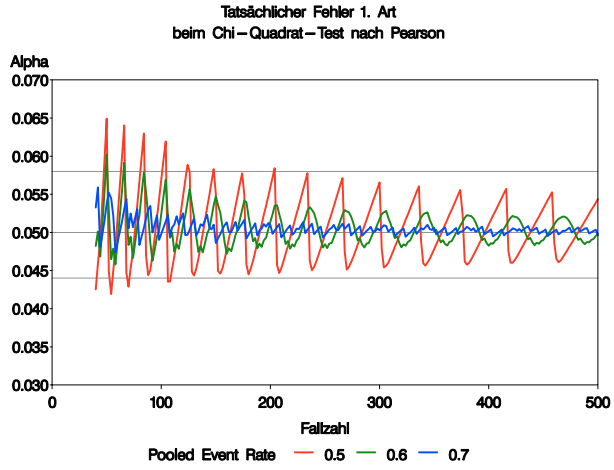

Abbildung 4.1:
Tatsächlicher Fehler 1. Art beim χ^2-Test nach Pearson. Die horizontalen Linien markieren den Ausschnitt der α-Achse, der bei den Simulationsergebnissen verwendet wurde, die mittlere den Wert 5%.

Der exakte α-Fehler des χ^2-Tests schwankt extrem, wobei auf sehr kleine Werte sehr große

folgen können, wenn die Fallzahl nur wenig erhöht wird. Die Amplitude der Schwankung nimmt mit der Fallzahl, aber noch viel stärker mit der Entfernung der PER vom Wert 0.5 ab. Dieses Bild hat sich auch bei den Simulationsergebnissen gezeigt: Die größten Schwankungen lagen bei der PER von 0.5 vor, die kleinsten bei der von 0.7.

Die exakten α-Fehler des χ^2-Tests wurden mit denen verglichen, die sich bei den mittleren finalen Fallzahlen der einzelnen Verfahren ergeben haben. Es blieben bei diesem Vergleich nur noch 13 signifikante Abweichungen übrig, wobei ein Teil dadurch entstanden sein kann, dass bei den finalen Fallzahlen wegen der simulierten 1:1-Randomisierung auf *geradzahlige* Gesamtfallzahlen gerundet werden musste, aber schon kleine Änderungen in der Fallzahl zu großen Änderungen des tatsächlichen Fehlers 1. Art führen konnten. Damit bleibt immer noch eine kleine Anzahl an signifikanten Abweichungen vom theoretischen α-Level, die potentiell der Rekalkulation bzw. ihren Varianten zugeschrieben werden könnte.

4.2.2.2 Erreichte Fallzahl

Gould [11] berichtet mediane finale Fallzahlen unter der Nullhypothese für auf Quotienten basierende Rekalkulationen. Wurde die Fallzahl beschränkt, lagen die finalen Fallzahlen um 32 bis 60% (keine Fallzahlbeschränkung) bzw. um 30 bis 50% (neue Fallzahl maximal doppelt so groß) über denen, die mit Kenntnis des wahren Quotenverhältnisses berechnet worden wären. Die Überschätzung wurde größer mit steigendem Quotenverhältnis. In [1] wurde das Verfahren für die Differenzen unter der Nullhypothese in einer Simulationsstudie untersucht. Bei korrekter Spezifizierung wichen die mittleren finalen Fallzahlen um 0 bis +7.5% von den Fallzahlen einer konventionellen Planung mit der tatsächlichen PER ab. Bei Überspezifizierung der PER konnten Abweichungen von nur 0 bis 0.2%, bei Unterspezifizierung solche von -16.1 bis +7.5% beobachtet werden. Wie bei den stetigen Endpunkten wurde auch bei den binären ein zusätzlicher Design-Parameter eingeführt, aufgrund dessen entschieden wurde, ob die rekalkulierte Fallzahl wirklich die initiale ersetzen sollte.

Herson und Wittes [33] berichten ebenfalls erreichte mittlere Fallzahlen unter der Nullhypothese. Für das Verfahren, das dem unrestringierten ähnlich ist, konnten die finalen Fallzahlen sowohl über- wie auch unterhalb denen aus einer fixen Planung mit den wahren Parametern liegen (+57%, -14%). Beim Verfahren mit Restriktion konnte die Fallzahl noch weiter über- aber genauso stark unterschätzt werden (+92%, -14%). Das Verfahren unterscheidet sich in mehreren Punkten stark von dem hier untersuchten, so dass Vergleiche nur eingeschränkt sinnvoll sind. So wies z.B. die restringierte Prozedur auch eine *obere* Schranke für die Fallzahl auf und es wurde der z-Test verwendet. Außerdem fand die Rekalkulation auf Basis des relativen Risikos statt, was deutliche Unterschiede in den rekalkulierten Fallzahlen verursachen kann. Gould [1] fand für die Prozedur mittlere finale Fallzahlen, die bei korrekter Spezifizierung der PER unter der Nullhypothese nah bei den tatsächlich benötigten Fallzahlen lagen. Die Abweichung bewegte sich zwischen 0 und +8.1% (verschiedene Designs). Bei Unterspezifizierung der PER wichen die Fallzahlen zwischen -8.7 und 0% ab, bei Überspezifizierung gab es keine Abweichung. Die Prozedur wurde von Gould an die von ihm vorgeschlagene angeglichen, so dass andere Grenzen für die rekalkulierte Fallzahl und auch ein anderer Test

als bei Herson und Wittes [33] zur Anwendung kamen.

Für die Prozedur mit Dummy-Stratifizierung nach Shih und Zhao [14] wurden finale mediane Fallzahlen auch unter der Nullhypothese tabelliert. Es wurden bei einer Restriktion der finalen Fallzahl auf den Bereich zwischen 60 und 410% (Table I) der initialen Fallzahl mediane finale Fallzahlen erzeugt, die abhängig von der PER durchgehend über dreimal höher lagen als die aus der *initialen* Planung, sie konnten sogar über viermal höher liegen als die *tatsächlich* benötigte Fallzahl. Bei einer kleineren Obergrenze (Table II) wurden entsprechend kleinere Fallzahlen gefunden, die immer noch über doppelt so groß wie die *initialen* und auch die *tatsächlich* benötigten Fallzahlen waren.

Für **die hier untersuchten Methoden** konnten bei korrekter Spezifikation der Variabilität Abweichungen der mittleren finalen Fallzahlen vom Ergebnis einer fixen Fallzahlplanung um bis zu 4.3% bzw. bis zu 9 Patienten nach oben und um bis zu 1.8% bzw. 3.2 Patienten nach unten beobachtet werden. Diese beiden Extremwerte stammten aus dem Szenario mit der kleinsten Varianz, was auf einen allgemeineren Trend hinweist: Bei der nächst größeren Variabilität unterschieden sich die mittleren Fallzahlen um +2.0% / +4.5 Patienten bis -1.5% / -1.4 Patienten, bei der größten Varianz gab es keine Abweichungen nach oben und solche um maximal 0.5% / 0.6 Patienten nach unten.

Ebenso waren Unterschiede der finalen Fallzahlen über alle Prozeduren größer in Szenarien mit niedriger Varianz als mit hoher (weiterhin vorausgesetzt, dass die PER richtig spezifiziert war). Während der maximale Unterschied nach Studiengrößen getrennt im Szenario mit PER=0.7 noch 11.3 Patienten betrug, war er bei einer PER=0.6 nur noch maximal 5.1 und bei PER=0.5 schließlich 0.6.

Die Basisprozedur unc endete bei korrekt niedriger und mittlerer Spezifikation der Varianz mit Fallzahlen, die sowohl über als auch unter der korrekten liegen konnten. Bei der höchsten Varianz unterschätzte sie diese systematisch um im Mittel mindestens 0.15% bzw. 0.4 Patienten.

Restriktion führte bei korrekter Varianzannahme wie erwartet zu einer Überschätzung der Fallzahl, wobei ebenfalls der Grad der Überschätzung mit steigender Variabilität kleiner wurde und bei der höchsten Varianz sogar auf Null sank. Unterschätzungen gab es im Mittel für diese Prozeduren nicht.

Die Varianten mit Control-Chart vergrößerten die Unterschätzung und verkleinerten die Überschätzung in den meisten Fällen mit korrekter Varianzannahme - bis auf einen Fall mit der höchsten Varianz. Dies spiegelt sich auch beim Vergleich der cc/nc-Paare wider: Mit ansteigender korrekt spezifizierter Varianz, lagen die Fallzahlen der cc-Prozedur unter (PER=0.7), waren vergleichbar groß (PER=0.6) und schließlich über (PER=0.5) denen der des nc-Gegenparts.

In Szenarien mit fehlspezifizierter PER konnten größere Unterschiede auftreten, da z.B. die restringierten Prozeduren keine kleineren Fallzahlen als die initialen ergeben durften. Die Basisprozedur unc lieferte bei überschätzter Varianz leicht zu kleine Fallzahlen (maximal 0.7% / 0.6 Patienten), bei unterschätzter Varianz waren im Mittel Unterschätzungen wie Überschätzungen möglich, deren Betrag mit weniger als einem Patienten bzw. 1% aber ebenso klein waren.

Die Restriktion führte bei überschätzter Varianz zu einer entsprechend zu großen Fallzahl (fast +15%, bzw. +51 Patienten). Bei unterschätzter Varianz spielte dagegen die Restriktion keine Rolle mehr, es traten dann eher Unterschiede durch die Verwendung von Control-Charts auf.

Die Verwendung von Control-Charts dämpfte die Anpassung der Fallzahl bei falscher Annahme der PER ab. War die Varianz zu niedrig und die initiale Fallzahl damit zu groß, sorgten Control-Charts dafür, dass die finale oberhalb der tatsächlich benötigten Fallzahl lag (ucc lag zwischen +0.2 bis +3.0% oberhalb, bzw. 0.1 bis 10.5 Patienten). Bei zu großer Varianz endeten die cc-Prozeduren dagegen zu niedrig (rcc und ucc beide -4.4 bis -3.6%, bzw. -15 bis -3.8 Patienten). Damit wich die Wirkung von Control-Charts bei den binären Endpunkten von der bei stetigen Endpunkte ab, weil bei stetigen ihre Verwendung fast durchgehend zu einer Erhöhung der Fallzahl führte. Das Verhalten bei den binären Endpunkten entsprach damit eher der Erwartung, nach der während der Rekalkulation eine Power von 85% angestrebt wurde, bei den cc-Prozeduren aber eine Power von 80-90% akzeptabel war. Dadurch sollte eine Verringerung der Fallzahl enden, sobald 90% von oben her erreicht würde, und eine Steigerung, wenn 80% von unten her erreicht würde.

Die beiden Fälle mit Fehlspezifikation der PER unterschieden sich im Ausmaß der Falschannahme. Bei der Unterschätzung der Varianz wurde eine PER von 0.6 angenommen, obwohl tatsächlich der Wert 0.5 richtig gewesen wäre. Dies führt (s. Tabelle 2.4) zu einer angenommenen Varianz von 0.24 gegenüber einer wirklichen von 0.25. Bei der Überschätzung der Varianz wurde dagegen eine PER von 0.6 angenommen während 0.7 richtig gewesen wäre. Die tatsächliche Varianz ist in diesem Fall 0.21 und damit weiter vom vermuteten Wert entfernt als im anderen Fall.

Die für die binären Endpunkte verwendete Fallzahlformel basiert auf einer Vereinfachung einer Annäherung mit Hilfe der Normalverteilung [22], aufgrund welcher damit zu rechnen war, dass die damit bestimmten Fallzahlen etwas zu groß waren, wenn sie mit einer exakten Fallzahlplanung verglichen würden. Das Ausmaß der Überschätzung ist dabei höchstens $z_{1-\alpha}^2 + 2z_{1-\alpha/2}z_{1-\beta}$, welches sich auf 6.77 Patienten insgesamt beläuft, wenn man einen zweiseitigen Test mit $\alpha = 0.05$ und $\beta = 0.15$ zugrunde legt (z_γ ist hierbei wieder das γ-Quantil der Normalverteilung). Allerdings gibt es noch eine Verzerrung in die Gegenrichtung, bei der die erwartete Fallzahl nach der Rekalkulation *um* bis zu $N_{\text{tats}}/N_{\text{pilot}}$ unterhalb der tatsächlich benötigten Fallzahl liegen kann. Da im vorliegenden Fall vereinfachend angenommen werden kann, dass die Größe der letzten Interimstichprobe ungefähr 3/4 der tatsächlich benötigten Fallzahl beträgt, ist also der Umfang der Verzerrung nach unten kleiner als Eins, so dass die Verzerrung nach oben durch die Annäherung stärker ins Gewicht fallen wird. Dieser Zusammenhang kann aber nicht für den Vergleich zwischen fixer und wiederholter Fallzahlplanung herangezogen werden, da für beide die gleiche Formel verwendet wurde.

4.2.2.3 Variabilität der Fallzahl

Die Variabilität der erreichten Fallzahl von Rekalkulationsprozeduren für binäre Endpunkten wird in der Literatur selten numerisch berichtet.

Für die Prozedur Gould und Shih [11] werden durch Gould [1] mittlere erreichte Fallzahlen nebst Standardabweichungen berichtet. Aus diesen können Variationskoeffizienten zwischen 0 und 20.0% berechnet werden. Auffällig waren die Situationen, in denen eine Referenzproportion von 0.1 angenommen wurde, aber in Wahrheit eine Proportion von 0.3 vorlag. In diesen Fällen war der Variationskoeffizient mindestens 9.6%.

Für die Prozedur nach Herson und Wittes [33] wurden in [1] ebenfalls mittlere Fallzahlen und zugehörige Standardabweichungen aufgelistet. Dabei konnten Variationskoeffizienten zwischen 0 und 17.3% abgeleitet werden, wobei nur für die Szenarien mit vermuteter Referenzproportion von 0.1 Variationskoeffizienten ungleich Null vertreten waren. Diese wiederum waren immer mindestens 6.8%.

Shih und Zhao [14] geben die Variabilität der Fallzahl nicht numerisch an, sagen jedoch aus, dass die Variabilität der rekalkulierten Fallzahl hoch sei und empfehlen die Verwendung der Prozedur nicht für kleinere Studientypen wie sie in Phase I oder II üblich sind.

Die Variabilität (in Form des Variationskoeffizienten) der finalen Fallzahl der **hier untersuchten Methoden** wurde wie bei den stetigen Endpunkten kleiner mit steigender Studiengröße und wurde ebenfalls kleiner mit steigender Varianz des Endpunkts. Letztere Beobachtung galt auch unabhängig davon, ob die PER bei der Planung richtig vorhergesehen wurde oder nicht, denn der maximale Wert bei der kleinen Varianz betrug 13.1% (bei *zu kleiner* Varianz 15.8%) und fiel dann mit steigender Variabilität auf 5.7% bzw. auf 2.4% (bei *zu großer* Varianz 1.8%).

Wie erwartet zeigten die unrestringierten Prozeduren eine Variabilität, die mindestens so groß war wie die der restringierten. Dies galt bis auf den Fall mit zu hoher realer Varianz, bei dem die cc-Prozeduren mit steigender Studiengröße auf nahezu Null fielen und die nc-Varianten bei 0.4% endeten.

Die Basisprozedur unc wies nur im Falle einer überschätzten Variabilität durchgehend die kleinsten Werte für den CV unter den unrestringierten Prozeduren auf. Andernfalls wurden ihre Werte größer als die der anderen, wenn Studien mit größeren initialen Fallzahlen betrachtet wurden.

Die cc-Prozeduren spielten bei den binären Endpunkten eine nicht so einheitliche Rolle wie bei den stetigen Endpunkten hinsichtlich der Variabilität der finalen Fallzahl. Während die Verwendung von Control-Charts bei den stetigen meist mit einer erhöhten Variabilität der finalen Fallzahl einher ging, konnte die der cc-Prozeduren bei binären Endpunkten innerhalb des gleichen Szenarios bei kleinen Studien oberhalb der der nc-Prozeduren liegen, und bei größeren darunter fallen. Es war aber auch eine unterschiedliche Lage über die Szenarios hinweg zu beobachten.

4.2.3 Ausblick

In den Simulationsstudien wurden Eigenschaften der mehrfachen verblindeten Fallzahlrekalkulation und einiger ihrer Variationen untersucht. Dies geschah immer unter der Annahme, dass die Nullhypothese gültig war, so dass nur Ergebnisse zur Niveaueinhaltung und der Fallzahl möglich waren. Bei der Fallzahl war die Frage, ob und wenn ja, wieviele Patienten

zusätzlich oder weniger rekrutiert werden mussten, obwohl kein Unterschied zwischen den Gruppen vorlag.

Diese Betrachtungen können in mehrere Richtungen fortgesetzt werden: Ein nahe liegender erster Schritt wäre, die Verfahren unter der Alternativhypothese zu analysieren, um Erkenntnisse hinsichtlich der Trennschärfe bzw. Fallzahl zu erhalten. Außerdem könnten die analytischen Betrachtungen aus den beiden Artikeln von Kieser und Friede [19, 22] auf die mehrfache Rekalkulation ausgeweitet werden, wobei ein wichtiger Unterschied wäre, dass sich die Größen der Interimstichproben nun zufällig ergeben. Weiterhin könnte im binären Fall der Einsatz von exakten Tests untersucht werden, so dass die durch den χ^2-Test bedingte Liberalität beseitigt werden kann. Zu diesem Zweck könnten auch die Verwendung von Abschätzungen wie in [18] zur mehrfachen Fallzahlkorrektur erwogen werden.

Als weitere Ansatzpunkte für zukünftige Forschung könnte auch die Frage nach der Anzahl und den Zeitpunkten der Interimstichproben betrachtet werden. Hier wurden exemplarisch drei sich dynamisch ergebende Zeitpunkte gewählt und es müsste untersucht werden, wann und ob weniger oder mehr sinnvoll wären. Denkbar wäre auch auch eine Erweiterung, bei der die Anzahl der Monitoringzeitpunkte nicht zu Anfang festgelegt werden muss.

Außerdem könnten andere Randomisierungsverhältnisse betrachtet werden, da in den untersuchten Fällen die beiden Behandlungsgruppen immer gleich groß gewählt worden sind. Es könnten auch alternative Teststatistiken (nichtparametrische Tests, Überlebenszeitanalysen), sowie Äquivalenztests und Tests auf Nichtunterlegenheit verwendet werden.

4.3 Ein Kombinationstest bei geänderter Variabilität

Im Abschnitt über den Kombinationstest wurde ein einfach anzuwendendes Verfahren untersucht, das erlaubt, einen statistischen Test auszuführen, selbst wenn bekannt ist, dass die Variabilität im Laufe der Studie nicht konstant war, weil sie zu einem Zeitpunkt durch ein Amendment geändert wurde.

Im Artikel von Chow und Shao [38] wird eine Möglichkeit der Analyse in solch einer Situation aufgezeigt. Sie basiert auf der Anwendung von gewichteter linearer Regression. Dazu wird der primäre Endpunkt pro Gruppe und Phase bestimmt. Als *Phase* wird der Zeitraum bezeichnet, in dem ein bestimmtes Amendment gültig war. Im Beispiel [38] gab es zwei Amendments, so dass drei Phasen entstanden. Allgemein erhält man bei K Amendments auf diese Weise $K+1$ Phasen (im Beispiel gilt $K=2$), und bei einer Studie mit zwei Behandlungsgruppen somit $2 \cdot (K+1)$ Mittelwerte respektive Streuungen. Das genannte Verfahren beruht darauf, dass man die Mittelwerte der Gruppen und Phasen mit Hilfe der im Amendment veränderten Variablen modelliert, d.h. pro Phase und Gruppe sind auch die wichtigen Kovariablen im Amendment zusammenzufassen. Im Beispiel aus dem Artikel könnte man die Baselinewerte der FEV_1-Variablen als Kovariable verwenden. Bezeichnet

- Δ_0^{T} die Differenz der FEV1-Werte (Hauptzielvariable) der Behandlungsgruppe vor allen Amendments

- Δ_0^{C} die entsprechenden Werte der Kontrollgruppe
- Δ_1^{T} bzw. Δ_1^{C} die Werte nach dem ersten Amendment und so fort
- x_0^{T} die Werte der Kovariablen der Behandlungsgruppe vor allen Amendments
- x_0^{C} die Werte der Kovariablen der Kontrollgruppe vor allen Amendments usw.,

dann kann aus den mittleren Differenzen $\bar{\Delta}_i^{\mathrm{Grp}}$ und mittleren Kovariablen $\bar{x}_i^{\mathrm{Grp}}$ Punktepaare für zwei Regressionsgeraden gebildet werden.

Die Gerade für die Interventionsgruppe basiert auf den Punkten $(\bar{x}_0^{\mathrm{T}}, \bar{\Delta}_0^{\mathrm{T}})$, $(\bar{x}_1^{\mathrm{T}}, \bar{\Delta}_1^{\mathrm{T}})$, …, $(\bar{x}_K^{\mathrm{T}}, \bar{\Delta}_K^{\mathrm{T}})$, die der Referenzgruppe auf den Punkten $(\bar{x}_0^{\mathrm{C}}, \bar{\Delta}_0^{\mathrm{C}})$, $(\bar{x}_1^{\mathrm{C}}, \bar{\Delta}_1^{\mathrm{C}})$, …, $(\bar{x}_K^{\mathrm{C}}, \bar{\Delta}_K^{\mathrm{C}})$. Werden nun diese Mittelwerte mit der Anzahl der zugrunde liegenden Beobachtungen (d.h. Patienten) gewichtet, kann aus der Lage der beiden Regressionsgeraden abgeleitet werden, ob sich die beiden Gruppen hinsichtlich der Hauptzielvariable bei bestimmten Werten der Kovariablen signifikant unterscheiden.

Die Autoren geben die Möglichkeit an, aus diesen Geraden statistische Schlüsse z.B. über den Behandlungseffekt in der Phase 0 abzuleiten. Wendet man das beschriebene Verfahren in einer Situation mit nur einem Amendment an, resultiert dies in zwei Regressionsgeraden, die nur durch jeweils zwei Punkte verlaufen. Der Test, der aus dem Vergleich der Regressionsgeraden zum Mittelwert der Kovariablen in einer der Phasen resultiert, reduziert sich allerdings zum normalen t-Test, bei dem nur die Daten aus der jeweiligen Phase verwendet werden. In aller Regel ist eine Inferenz für alle Patienten auf der Basis von Patienten einer Phase nicht erwünscht und es ist möglich, eine beliebige Funktion der Mittelwerte zu schätzen und darauf einen Test der globalen Schnitthypothese aufzubauen. Die Wahl der Funktion und die der Kovariablen in den Regressionen bleibt dabei dem Anwender überlassen.

Die in dieser Arbeit vorgestellte Methode unterscheidet sich vom oben beschriebenen Ansatz in mehreren Punkten: Erstens *müssen* keine Kovariablen zur Berechnung der primären Variablen herangezogen werden und es muss keine Modellannahme über die Abhängigkeit der Variablen getroffen werden. Es ist auch kein einzelner Punkt im (u.U. mehrdimensionalen) Raum der Kovariablen zu spezifizieren, an dem der Wirksamkeitsvergleich durchgeführt werden soll. Zur Thematik der Kovariablen ist noch zu bemerken, dass sie im Verfahren von Chow und Shao nötig sind, aber ihre Anzahl durch die Anzahl der Amendments begrenzt ist – eine Beschränkung, die aus der Algebra des Verfahrens stammt und die durch Verwendung von verallgemeinerten Matrixinversen aufgehoben werden kann. Im hier beschriebenen Kombinationsverfahren kann ganz auf Kovariablen verzichtet werden, durch Anwendung kovarianzanalytischer Methoden kann jedoch eine beliebige Anzahl berücksichtigt werden.

Beiden Verfahren ist gemeinsam, dass die Anzahl der Amendments beliebig ist und dass sie trotz der Verwendung von Teilmengen der gesamten Studiendaten *keine* Interimanalysen darstellen, denn die Tests werden nicht während des Studienverlaufes durchgeführt, sondern erst, wenn die Daten komplett erhoben sind. Es ist allerdings möglich und sinnvoll, schon im Studienprotokoll oder im statistischen Analyseplan die Möglichkeit bzw. Notwendigkeit einer Auswertung auf die beschriebene Art festzulegen. Dies kann kann bei der Einführung des Amendments geschehen, denn bei diesen sollten deren statistischen Konsequenzen dargelegt werden [35].

In den ersten zwei Kapiteln wurden entsprechende Indikatoren und Methoden zum Identifizieren einer solchen Situation beschrieben. Wenn das Amendment jedoch auf entblindeten Daten basiert, sind die p-Werte der einzelnen Phasen nicht mehr unabhängig voneinander und das Verfahren somit nicht mehr anwendbar (vgl. [71]).

Bei den durchgeführten Simulationsstudien ergab sich, dass alle drei Testverfahren aus Kapitel 2.3 das Niveau bei verschiedenen Faktoren und Rekrutierungsannahmen einhielten. Dies war auch so zu erwarten, da die Gruppengrößen für den Vergleich zwischen Verum und Kontrolle identisch waren, die Verteilungen immer Symmetrie und gleiche Streuung aufwiesen. Unter diesen Bedingungen führt die Abweichung von der Normalverteilung nur zu sehr kleinen Fehlern, wenn sie trotzdem bei der t-Statistik in den Gruppen unterstellt wird [72].

Beim Verfahren Komb&1, bei dem zusätzlich zur Signifikanz des Kombinationstests noch die des Tests in mindestens einer der Phasen verlangt wurde, lag der mittlere Fehler 1. Art deutlich unterhalb der angesetzten 5%. Dies ist nicht überraschend, weil durch die Zusatzbedingung dieser zweiten Signifikanz die Nullhypothese weniger oft abgelehnt werden kann als beim Kombinationstest, man denke z.B. an die Situation, in der die p-Werte der beiden Phasen 0.08 bzw. 0.09 betragen, womit der keiner der beiden t-Tests in den einzelnen Phasen signifikant würde, der Kombinationstest jedoch eine Teststatistik von $-2\log(0.08{\cdot}0.09) = 9.87$ aufweisen würde, welche mit einer Wahrscheinlichkeit von 0.043 von einer χ^2_4-Verteilung überschritten wird, was Signifikanz auf dem 5%-Niveau für den Kombinationstest bedeutet. Formaler kann man dies auch so einsehen: Die Zusatzbedingung der Ablehnung einer der Hypothesen H_0^{a} oder H_0^{b} über die Ablehnung von H_0^{ab} hinaus, welche auch als Abschlusstest [63] bezeichnet werden kann, muss hier zu einer konservativeren Testprozedur führen, da die Ablehnung von H_0^{ab} allein schon eine solche von 5% herbeiführt. Bei der Diskussion der adaptiven Designs durch Bauer und Köhne [41] wurde ebenfalls auf [63] hingewiesen und es wurde in ganz ähnlicher Weise argumentiert. Die Beobachtung der größeren Konservativität spiegelt sich auch in der Power des Tests wider, denn sie war immer kleiner als die des Kombinationstests, da die Zusatzbedingung auch unter der Alternative greift.

Die Trennschärfe beim zweiten Szenario war generell kleiner als die beim ersten, was einerseits mit der kleineren Studiengröße in diesem Szenario und andererseits mit dem größeren Anteil an Beobachtungen mit erhöhter Varianz (nämlich 2/3 gegenüber 1/2 im Szenario 1) erklärt werden kann. Es ist daher u.U. mit einem doppelten Effekt zu rechnen, wenn die Einschlusskriterien aufgeweitet und/oder die Ausschlusskriterien gelockert werden, um die Rekrutierung zu beschleunigen: Das spätere Patientenkollektiv wird nicht nur eine größere Varianz aufweisen, sondern evtl. auch die größere Gruppe darstellen.

Eine weitere Beobachtung aus den Powerverläufen deckt sich mit bereits bekannten Einsichten: Wenn der Varianzfaktor Eins war, unterschieden sich die Kollektive vor und nach dem Amendment hinsichtlich der Varianz nicht. In diesen Fällen war so gut wie immer der Pooling-Test besser als die beiden auf dem Fisher-Test basierenden. Dies ist auch schon bei der Betrachtung von adaptiven Designs [41] festgestellt worden. Dort wurde in Abschnitt 3.1 quantifiziert, mit wieviel Verlust an Power durch sie zu rechnen ist. Der Betrag, um den sich die Power unterscheidet, wurde in Abhängigkeit vom Verhältnis der beiden Stichprobengrößen und der Power des zugehörigen gepoolten u.m.p.-Test (uniformly most powerful

test) tabelliert. In dieser Arbeit waren die relativen Stichprobengrößen der ersten Phase 1/2 und 1/3. Für ähnliche Werte ergaben sich dort ein Verlust an Trennschärfe für den Fisher-Test von 1.2%-Punkten im Falle einer 50:50-Rekrutierung und einem Verlust von 1.0% im Falle einer 30:70-Rekrutierung (also nicht 1:2 wie in der vorliegenden Arbeit), wenn man eine Power von 95% ansetzte. Die in dieser Arbeit vorgefundenen Differenzen für einen *ungefähr* gleichen Wert für die Power des Pooling-Tests waren: 96.71%-96.05%=0.66%-Punkte und 91.93%-90.02%=1.91%. Man sieht, dass die Verluste an Power in etwa vergleichbar sind. In [41] wurden die Differenzen mit steigender Power immer kleiner, so dass auch hier leicht kleinere zu erwarten waren, da die Power des Pooling-Tests oberhalb des tabellierten Wertes von 95% lag.

Weiterhin ist bei Bauer und Köhne zu erkennen, dass bei gleicher Trennschärfe der Verlust mit steigender Ungleichheit der Gruppengrößen ebenfalls ansteigt. Auch dies kann in der durchgeführten Simulationsstudie bestätigt werden, obwohl es Fälle mit konstanten Mittelwerten gab, in denen die Pooling-Prozedur eine kleinere Power zeigte: Der maximale Unterschied in diesen Fällen betrug 0.34%-Punkte.

Wurde in den für diese Arbeit durchgeführten Simulationen zugelassen, dass sich die Mittelwerte bei gleichem Effekt über die Phasen änderten, konnte schon bei einem Varianzfaktor von Eins die Poolingprozedur eine *niedrigere* Power als die Kombinationsprozedur ergeben, so hatte z.B. bei einem kleinen Effekt (0.1) und starker Verschiebung (1.0) die Kombinationsprozedur bereits eine um 3.46%-Punkte höhere Power. Wenn sich zusätzlich die Mittelwertdifferenz in den Gruppen änderte, konnten bis zu 8.62%-Punkte höhere Trennschärfen bei der Kombinationsprozedur beobachtet werden (Übergang vom Effekt 0.7 zum Effekt 0.2).

Bei kleiner Varianzinflation ist der Pooling-Test besser als die Kombinationstests. Dies ist auch zu erwarten gewesen, denn die Kollektive vor und nach der Varianzänderung unterscheiden sich kaum und der Verlust an Power bei der Kombinationsprozedur, wie in [41] beschrieben, kommt deutlicher zum Tragen. In Situationen, in denen eine kleine Mittelwertdifferenz vorlag, waren zwar sichtbare Unterschiede in der Trennschärfe zu beobachten, aber die absoluten Werte waren für alle beobachteten Prozeduren und Szenarien so gering, dass Vorteile nur bedingt relevant erscheinen. Bei sehr großen Mittelwertdifferenzen dagegen war die Power aller Prozeduren sehr groß und die Vorteile hatten nur Dimensionen von Bruchteilen von Prozenten für alle beobachteten Varianzinflationen. Die Unterschiede mit größerer Relevanz ergaben sich bei mittelgroßen Mittelwertunterschieden. Während hier die Pooling-Prozedur im Falle konstanter Varianzen noch einen kleinen Vorteil hatte, war dieser bei einer Varianzvergrößerung auf das 1.5- bis 2fache egalisiert und bei Werten darüber war die Kombinationsprozedur oft besser. Der Poolingtest wurde bei noch größeren Varianzunterschieden sogar vom Abschlusstest überholt, dessen Power wie schon ausgeführt immer unterhalb der reinen Kombinationsprozedur lag.

Die Erklärung für die beobachteten Vorteile der Kombinationsprozeduren liegt meist in der Mischverteilung, die entsteht, wenn Zufallszahlen mit verschiedenen Streuungen und/oder Mittelwerten zu einer Stichprobe vereinigt werden (siehe Abbildung 4.2). Da die Kombinationsprozeduren die Phasen getrennt betrachten, sind sie durch die Mischung der Verteilungen höchstens durch die kleineren Fallzahlen beeinflusst, während die Pooling-Prozeduren mit

einer Abweichung von ihren Voraussetzungen konfrontiert werden. Der t-Test setzt Normalverteilungen in den Untersuchungsgruppen voraus, während dort jedoch Mischverteilungen vorliegen.

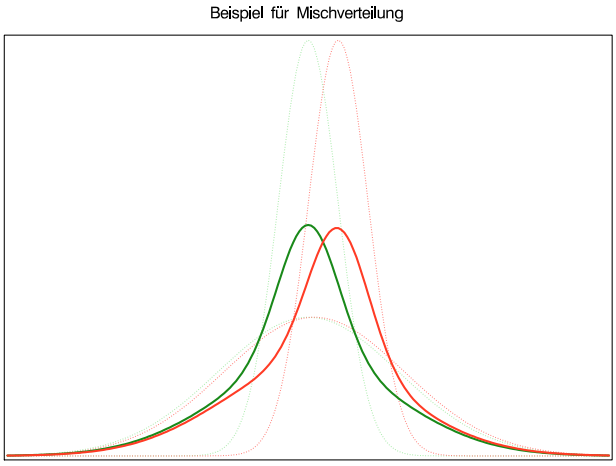

Abbildung 4.2:
Dichten für das Beispiel einer Situation, in der durch Varianzanstieg Mischverteilungen entstanden sind. Die Farben deuten die verschiedenen Behandlungsgruppen an. Die feineren Linien stehen für die Verteilungen in den einzelnen Phasen. Die dickeren Linien bilden die Dichte der Mischverteilung ab.

Die beschriebene Kombinationsprozedur lässt auch auf beliebige andere Tests und Skalenniveaus übertragen, so z.B. auf nicht-normalverteilte Zufallsgrößen (Mann-Whitney-U-Test), binäre Endpunkte (χ^2-Test) oder Überlebenszeiten (Log-Rank-Test). Bei binären Endpunkten könnten mit ihrer Hilfe Phänomene wie das Simpson-Paradoxon [73], das beim Poolen entstehen kann, umgangen werden.

Ein Problem, welches bei vielen Ansätzen mit Kombinationstests auftritt, ist das der Parameter- bzw. Effektschätzung. In klinischen Studien ist nicht nur von Interesse, *ob* eine Behandlung besser ist als eine andere, sondern auch, um *wieviel* sie besser ist. In [74] wird dargelegt, dass die Schätzung des Effekts dann problematisch sein kann, wenn durch eine geplante oder ungeplante Zwischenanalyse der resultierende Umfang der Gesamtstichprobe zu einer Zufallsvariablen wird. Dann kann der Schätzer für den Effekt erheblich verzerrt oder seine Variabilität sehr groß sein. Brannath et al. [74] erwähnen andererseits auch, dass die üblichen Schätzer und Konfidenzintervalle für das fixe Design unverzerrt bzw. valide bleiben, wenn die Fallzahl der Studie unveränderbar ist. Im vorliegenden Fall wird keine Fallzahladaption vorgenommen, so dass die üblichen Effektschätzer und Konfidenzintervalle anwendbar bleiben sollten. Dennoch könnten die Methoden aus [74] angewendet werden, um zu sinnvollen Effektschätzern zu gelangen, wenn über die Phasen verschiedene Effekte beobachtet wurden.

In den Simulationen wurden nur Varianz*vergrößerungen* erzeugt. Dies stellt keine starke Einschränkung dar, denn aufgrund der Tatsache, dass sowohl die Pooling- als auch die beiden Kombinationsprozeduren invariant gegenüber der zeitlichen Abfolge der Phasen sind, können die Erkenntnisse sofort auf Verkleinerungen der Varianz übertragen werden. Eine un-

eingeschränkte Übertragbarkeit auf Varianzverkleinerungen gilt allerdings nur im Falle eines Amendments zur Hälfte der rekrutierten Patienten, da hier die Gewichte der Mischverteilung gleich groß sind.

Folgende Einschränkungen für die Verwendung der Kombinationsprozeduren sind zu nennen: Erstens wurden in dieser Simulationsstudie nur Szenarien untersucht, in denen sich die Varianz beider Gruppen im Laufe der Zeit gleichartig veränderte, d.h., wenn z.B. von einer Varianzinflation auf das Doppelte in der zweiten Phase ausgegangen wurde, dann bedeutete dies immer, dass sich sowohl die Verum- als auch die Kontrollgruppe von einem gemeinsamen Wert in der ersten Phase (Eins) auf den *gleichen* Wert (hier Zwei) vergrößert hatte. Damit lag immer Homoskedastizität vor. Der Fall von Heteroskedastizität wurde hier nicht untersucht und könnte somit Gegenstand weiterführender Untersuchungen sein. Durch Anwendung entsprechender Testverfahren (z.B. Satterthwaite-Variante des t-Tests), sind unter ähnlichen Bedingungen wie den hier untersuchten höhere Trennschärfen des Kombinationstests zu erwarten.

Weiterhin darf keine Entblindung der Daten im Verlauf der Studie stattfinden, da ansonsten die beobachteten p-Werte nicht mehr unabhängig sein könnten, was aber eine Voraussetzung für die Anwendbarkeit ihrer Kombination nach Fisher ist. Ist also eine weiterreichende Flexibilität erforderlich oder gewünscht, sollte z.B. ein adaptives Verfahren wie es von Bauer und Köhne [41] beschrieben wurde oder noch flexiblere Methoden auf der Basis der *conditional error rate*, wie sie von Müller und Schäfer [71] erarbeitet wurden, zum Einsatz gebracht werden.

Verallgemeinerungen des Testverfahrens Komb&1 auf eine beliebige Anzahl von Amendments sind prinzipiell möglich, Abschlusstestprinzip muss aber weiterhin eingehalten werden, d.h. zu jeder signifikanten Elementarhypothese müssen auch alle zugehörigen Schnitthypothesen abgelehnt werden. So entstehen z.B. bei zwei Amendments drei Elementarhypothesen $H_0^{\mathrm{a}}, H_0^{\mathrm{b}}, H_0^{\mathrm{c}}$. Würde jetzt durch den Kombinationstest die globale Schnitthypothese $H_0^{\mathrm{abc}} = H_0^{\mathrm{a}} \wedge H_0^{\mathrm{b}} \wedge H_0^{\mathrm{c}}$ auf dem 5%-Niveau abgelehnt, muss in Analogie für die Signifikanz einer der Hypothesen, sagen wir H_0^{a}, auf dem 5%-Niveau gefordert werden, dass auch die Hypothesen $H_0^{\mathrm{ab}} = H_0^{\mathrm{a}} \wedge H_0^{\mathrm{b}}$ und $H_0^{\mathrm{ac}} = H_0^{\mathrm{a}} \wedge H_0^{\mathrm{c}}$ zusätzlich zur Hypothese H_0^{a} selbst auf dem 5%-Niveau abgelehnt werden. Somit ist die Anwendung dieser Testvariante zwar nicht unmöglich, aber sie wird mit zunehmender Anzahl an Amendments komplexer.

5 Zusammenfassung

Die hier vorgelegte Arbeit befasst sich im Wesentlichen mit Varianzunterschieden und Varianzfehlspezifikationen von Zielgrößen klinischer Studien.

Im Mittel und im Median waren nur sehr geringe Abweichungen der Varianz in der zweiten Hälfte realer Studien zu erkennen. Die vorgefundene Verteilung der Varianzquotienten unterschied sich jedoch signifikant von der unter der Annahme *keines* Unterschieds hergeleiteten theoretischen Mischverteilung. Grafische Methoden deuteten auf eine größere Anzahl an extremen Verhältnissen hin, was durch das 5%-Perzentil (Faktor 2/3) und das 95%-Perzentil (Faktor 1.8) bestätigt wurde. Die Untersuchung mit einem hierarchischen Modell korrigierte den Gesamtmittelwert der Verhältnisse der Standardabweichungen auf 1.03 und zeigte, dass Variablen innerhalb einer Studie zu gleichartigem Verhalten tendierten. Bei den erhobenen Eigenschaften könnte nur durch *ungleichmäßige Rekrutierung* ($p = 0.09$) und *Amendments* ($p = 0.13$) ein Einfluss auf die Varianzungleichheit vermutet werden.

Anschließend wurden statistische Eigenschaften von neuen Fallzahladaptionsprozeduren für normalverteilte (t-Test) und binäre (χ^2-Test) Endpunkte unter der Nullhypothese keines Gruppenunterschieds bestimmt. Die Prozeduren sollten im Verlauf der Studie *mehrfach* und verblindet die Varianzannahme überprüfen und ggf. die Fallzahl korrigieren. Im Falle stetiger Endpunkte wurde das Niveau eingehalten, bei binären ist zu vermuten, dass die Liberalität des χ^2-Tests für die Niveauverletzungen verantwortlich ist. Korrektur für Verblinden führte erwartungsgemäß zu einer leicht unterschätzten Fallzahl, Kontrollgrenzen für die Power verhinderten eine zu häufige Rekalkulation, bewirkten aber u.U. eine verzerrte Fallzahlschätzung. Abhängig vom Vorhandensein anderer Merkmale konnte die Variabilität der Fallzahl durch sie erhöht aber auch verringert werden. Mindestfallzahlen führten teilweise zu starker Überschätzung der Fallzahl, wirkten sich aber ansonsten nicht nachteilig aus. Die Auswirkungen für binäre Variablen konnten sich von denen für die stetigen unterscheiden.

Im dritten Teil wurde eine Möglichkeit der alternativen Auswertung von Studien vorgestellt, in denen eine Veränderung der Varianz der Zielgröße angenommen wurde. Anhand der Ergebnisse aus dem ersten Teil konnten verschiedene Szenarien Varianzinflation und/oder Mittelwertverschiebung über den Zeitverlauf einer Studie simuliert und der Fehler 1. und 2. Art der gewöhnlichen Pooling-Prozedur, des Kombinationstests von Fisher und einer Abschlusstestprozedur ermittelt werden. Keines der Verfahren verhielt sich antikonservativ. Die Prozeduren auf Basis des Kombinationstests wiesen verglichen mit der Poolingprozedur bei mittelgroßer Effektstärke eine um bis zu 7%-Punkte größere Power auf, bei veränderten Mittelwerten lagen sie bis zu 20%-Punkten höher. Bei großen und kleinen Effektstärken war der Vorteil dagegen gering, oder der Poolingtest war geringfügig besser.

Literaturverzeichnis

[1] Gould, A.L.: Sample size re-estimation: recent developments and practical considerations. In: *Stat Med* 20 (2001), Nr. 17-18, S. 2625–2643

[2] Lovato, L.C. ; Hill, K. ; Hertert, S. ; Hunninghake, D.B. ; Probstfield, J.L.: Recruitment for controlled clinical trials: Literature summary and annotated bibliography. In: *Control Clin Trials* 18 (1997), Nr. 4, S. 328–352

[3] Carew, B.D. ; Ahn, S.A. ; Boichot, H.D. ; Dierenfeldt, B.J. ; Dolan, N.A. ; Edens, T.R. ; Weiner, D.H. ; Probstfield, J.L.: Recruitment strategies in the studies of left ventricular dysfunction (SOLVD): strategies for screening and enrollment in two concurrent but separate trials. The SOLVD Investigators. In: *Control Clin Trials* 13 (1992), Nr. 5, S. 325–338

[4] Collins, J.F. ; Williford, W.O. ; Weiss, D.G. ; Bingham, S.F. ; Klett, C.J.: Planning patient recruitment: fantasy and reality. In: *Stat Med* 3 (1984), Nr. 4, S. 435–443

[5] Deyo, R.A. ; Bass, J.E. ; Walsh, N.E. ; Schoenfeld, L.S. ; Ramamurthy, S.: Prognostic variability among chronic pain patients: implications for study design, interpretation, and reporting. In: *Arch Phys Med Rehabil* 69 (1988), Nr. 3, S. 174–178

[6] Hunninghake, D.B. ; Knoke, J. ; LaDouceur, M. ; Peterson, F.: Population characteristics according to recruitment source. In: *Circulation* 66 (1982), Nr. 6 Pt 2, S. IV46–IV48

[7] Lee, Y.J.: Interim recruitment goals in clinical trials. In: *J Chronic Dis* 36 (1983), Nr. 5, S. 379–389

[8] Stein, C.: A Two-Sample Test for A Linear Hypothesis Whose Power Is Independent of the Variance. In: *Ann Math Stat* 16 (1945), Nr. 3, S. 243–258

[9] Wittes, J. ; Brittain, E.: The role of internal pilot studies in increasing the efficiency of clinical trials. In: *Stat Med* 9 (1990), Nr. 1, S. 65–71

[10] Denne, J.S. ; Jennison, C.: Estimating the sample size for a t-test using an internal pilot. In: *Stat Med* 18 (1999), Nr. 13, S. 1575–1585

[11] Gould, A.L.: Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. In: *Stat Med* 11 (1992), Nr. 1, S. 55–66

[12] Gould, A.L. ; Shih, W.J.: Sample-Size Reestimation Without Unblinding for Normally Distributed Outcomes with Unknown-Variance. In: *Commun Stat A-Theor* 21 (1992), Nr. 10, S. 2833–2853

[13] Gould, A.L.: Issues in blinded sample size re-estimation. In: *Commun Stat B-Simul* 26 (1997), Nr. 3, S. 1229–1239

[14] Shih, W.J. ; Zhao, P.L.: Design for sample size re-estimation with interim data for double-blind clinical trials with binary outcomes. In: *Stat Med* 16 (1997), Nr. 17, S. 1913–1923

[15] International Conference on Harmonisation E9 Expert Working Group: ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials. In: *Stat Med* 18 (1999), Nr. 15, S. 1905–1942

[16] Friede, T. ; Kieser, M.: On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation. In: *Stat Med* 21 (2002), Nr. 2, S. 165–176

[17] Waksman, J.A.: Assessment of the Gould-Shih procedure for sample size re-estimation. In: *J Biopharm Stat* 6 (2007), Nr. 1, S. 53–65

[18] Kieser, M. ; Friede, T.: Re-calculating the sample size in internal pilot study designs with control of the type I error rate. In: *Stat Med* 19 (2000), Nr. 7, S. 901–911

[19] Kieser, M. ; Friede, T.: Simple procedures for blinded sample size adjustment that do not affect the type I error rate. In: *Stat Med* 22 (2003), Nr. 23, S. 3571–3581

[20] Zucker, D.M. ; Wittes, J.T. ; Schabenberger, O. ; Brittain, E.: Internal pilot studies II: Comparison of various procedures. In: *Stat Med* 18 (1999), Nr. 24, S. 3493–3509

[21] Friede, T. ; Kieser, M.: Sample size recalculation in internal pilot study designs: a review. In: *Biom J* 48 (2006), Nr. 4, S. 537–555

[22] Friede, T. ; Kieser, M.: Sample size recalculation for binary data in internal pilot study designs. In: *Pharm Stat* 3 (2004), Nr. 4, S. 269–279

[23] Friede, T. ; Mitchell, C. ; Muller-Velten, G.: Blinded sample size reestimation in non-inferiority trials with binary Endpoints. In: *Biom J* 49 (2007), Nr. 6, S. 903–916

[24] Coffey, C.S. ; Muller, K.E.: Exact test size and power of a Gaussian error linear model for an internal pilot study. In: *Stat Med* 18 (1999), Nr. 10, S. 1199–1214

[25] Kieser, M. ; Friede, T.: Blinded sample size reestimation in multiarmed clinical trials. In: *Drug Inf J* 34 (2000), Nr. 2, S. 455–460

[26] Bolland, K. ; Sooriyarachchi, M.R. ; Whitehead, J.: Sample size review in a head injury trial with ordered categorical responses. In: *Stat Med* 17 (1998), Nr. 24, S. 2835–2847

[27] Whitehead, J. ; Whitehead, A. ; Todd, S. ; Bolland, K. ; Sooriyarachchi, M.R.: Mid-trial design reviews for sequential clinical trials. In: *Stat Med* 20 (2001), Nr. 2, S. 165–176

[28] Lake, S. ; Kammann, E. ; Klar, N. ; Betensky, R.: Sample size re-estimation in cluster randomization trials. In: *Stat Med* 21 (2002), Nr. 10, S. 1337–1350

[29] Zucker, D.M. ; Denne, J.: Sample-size redetermination for repeated measures studies. In: *Biometrics* 58 (2002), Nr. 3, S. 548–559

[30] Coffey, C.S. ; Muller, K.E.: Properties of internal pilots with the univariate approach to repeated measures. In: *Stat Med* 22 (2003), Nr. 15, S. 2469–2485

[31] Wüst, K. ; Kieser, M.: Blinded sample size recalculation for normally distributed outcomes using long- and short-term data. In: *Biom J* 45 (2003), Nr. 8, S. 915–930

[32] Wüst, K. ; Kieser, M.: Including long- and short-term data in blinded sample size recalculation for binary endpoints. In: *Comput Stat Data An* 48 (2005), Nr. 4, S. 835–855

[33] Herson, J. ; Wittes, J.: The use of interim analysis for sample size adjustment. In: *Drug Inf J* 27 (1993), S. 753–760

[34] Lösch, C. ; Neuhäuser, M.: The statistical analysis of a clinical trial when a protocol amendment changed the inclusion criteria. In: *BMC Med Res Methodol* 8 (2008), Nr. 1

[35] Cleophas, T.J. ; Zwinderman, A.H. ; Cleophas, T.F.: *Statistics Applied to Clinical Trials*. 3. Springer, 2006

[36] Chow, S.C. ; Chang, M. ; Pong, A.: Statistical consideration of adaptive methods in clinical development. In: *J Biopharm Stat* 15 (2005), Nr. 4, S. 575–591

[37] US Department of Health and Human Services: *Guideline for Format and Content of the Clinical and Statistical Sections of New Drug Applications*. Rockville, MD : Food and Drug Administration, 1988

[38] Chow, S.C. ; Shao, J.: Inference for clinical trials with some protocol amendments. In: *J Biopharm Stat* 15 (2005), Nr. 4, S. 659–666

[39] Svolba, G. ; Bauer, P.: Statistical quality control in clinical trials. In: *Control Clin Trials* 20 (1999), Nr. 6, S. 519–530

[40] Dubertret, L. ; Sterry, W. ; Bos, J.D. ; Chimenti, S. ; Shumack, S. ; Larsen, C.G. ; Shear, N.H. ; Papp, K.A.: Clinical experience acquired with the efalizumab (Raptiva) (CLEAR) trial in patients with moderate-to-severe plaque psoriasis: results from a phase III international randomized, placebo-controlled trial. In: *Br J Dermatol* 155 (2006), Nr. 1, S. 170–181

[41] Bauer, P. ; Köhne, K.: Evaluation of experiments with adaptive interim analyses. In: *Biometrics* 50 (1994), Nr. 4, S. 1029–1041

[42] Vickers, A. ; Rees, R. ; Zollman, C. ; Smith, C. ; Ellis, N.: Acupuncture for migraine and headache in primary care: a protocol for a pragmatic, randomized trial. In: *Complement Ther Med* 7 (1999), Nr. 1, S. 3–18

[43] McCarney, R. ; Fisher, P. ; Haselen, R. van: Accruing large numbers of patients in primary care trials by retrospective recruitment methods. In: *Complement Ther Med* 10 (2002), Nr. 2, S. 63–68

[44] Neuhäuser, M. ; Senske, R.: The analysis of multicentre clinical trials when there is heterogeneity between centres. In: *J Stat Comput Sim* 79 (2009), Nr. 11, S. 1381–1387

[45] Prescott, R.J. ; Counsell, C.E. ; Gillespie, W.J. ; Grant, A.M. ; Russell, I.T. ; Kiauka, S. ; Colthart, I.R. ; Shepherd, S.M. ; Ross, S. ; Russell, D.: Factors that limit the quality, number and progress of randomised controlled trials. In: *Health Technol Assess* 3 (1999), Nr. 20, S. 1–143

[46] Sickmüller, B. (Hrsg.): *Klinische Arzneimittelprüfungen in der EU. Grundsätze für Standards der Guten Klinischen Praxis (GCP) bei der Durchführung von Studien mit Arzneimitteln am Menschen in der EU-CPMP/ICH-GCP Guideline*. 4. Aulendorf : Editio Cantor Verlag, 1998 (pharmind Serie dokumentation)

[47] Schwarz, J.A.: *Leitfaden Klinische Prüfungen von Arzneimitteln und Medizinprodukten. Good Clinical Practice, Planung, Organisation, Durchführung und Dokumentation*. 3. Aulendorf : Editio Cantor Verlag, 2005 (Der Pharmezeutische Betrieb)

[48] Cook, J.A. ; Ramsay, C.R. ; Norrie, J.: Recruitment to publicly funded trials - Are surgical trials really different? In: *Contemp Clin Trials* 29 (2008), Nr. 5, S. 631–634

[49] McCulloch, P. ; Taylor, I. ; Sasako, M. ; Lovett, B. ; Griffin, D.: Randomised trials in surgery: problems and possible solutions. In: *Br Med J* 324 (2002), Nr. 7351, S. 1448–1451

[50] Pitman, E.J.G.: Significance tests which may be applied to samples from any population. In: *J Roy Stat Soc Suppl* 4 (1937), Nr. 2, S. 119–130

[51] Mehta, C. ; Patel, N ; Corporation, CYTEL S. (Hrsg.): *StatXact Users Manual*. Cambridge, MA : CYTEL Software Corporation, 2007

[52] Pruscha, H.: *Statistisches Methodenbuch: Verfahren, Fallstudien, Programmcodes*. Springer, 2006

[53] Büning, H. ; Trenkler, G. ; Gruyter, Walter de (Hrsg.): *Nichtparametrische statistische Methoden*. 2. Berlin : Walter de Gruyter, 1994

[54] Stephens, M.A.: Edf Statistics for Goodness of Fit and Some Comparisons. In: *J Am Stat Assoc* 69 (1974), Nr. 347, S. 730–737

[55] Wilk, M.B. ; Gnanades, R.: Probability Plotting Methods for Analysis of Data. In: *Biometrika* 55 (1968), Nr. 1, S. 1–17

[56] Singer, J.D.: Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. In: *J Educ Behav Stat* 23 (1998), Nr. 4, S. 323–355

[57] Birkett, M.A. ; Day, S.J.: Internal pilot studies for estimating sample size. In: *Stat Med* 13 (1994), Nr. 23-24, S. 2455–2463

[58] Weihs, C. ; Jessenberger, J.: *Statistische Methoden zur Qualitätssicherung und -optimierung in der Industrie*. Weinheim, New York, Chichester, Brisbane, Singapore, Toronto : Wiley-VCH, 1999

[59] Hartung, J. ; Elpelt, B. ; Klösener, K.-H.: *Statistik. Lehr- und Handbuch der angewandten Statistik*. 6. München : R. Oldenbourg Verlag, 1987

[60] Julious, S.A.: Tutorial in biostatistics - Sample sizes for clinical trials with normal data. In: *Stat Med* 23 (2004), Nr. 12, S. 1921–1986

[61] Hedges, L.V. ; Olkin, I.: *Statistical methods for meta-analysis*. Boston : Academic Press,Inc., 1985

[62] Hartung, J. ; Knapp, G. ; Bimal, K.S.: *Statistical Meta-Analysis with Applications*. 1. Hoboken, New Jersey : J.Wiley & Sons, 2008 (Wiley Series in Probability and Statistics)

[63] Marcus, R. ; Peritz, E. ; Gabriel, K.R.: Closed Testing Procedures with Special Reference to Ordered Analysis of Variance. In: *Biometrika* 63 (1976), Nr. 3, S. 655–660

[64] Horn, M. ; Vollandt, R.: *Multiple Tests und Auswahlverfahren*. Stuttgart : Gustav Fischer Verlag, 1995 (biometrie)

[65] Vickers, A.J.: Whose data set is it anyway? Sharing raw data from randomized trials. In: *Trials* 7 (2006)

[66] Proschan, M.A.: Sample size re-estimation in clinical trials. In: *Biom J* 51 (2009), Nr. 2, S. 348–357

[67] Wittes, J. ; Schabenberger, O. ; Zucker, D. ; Brittain, E. ; Proschan, M.: Internal pilot studies I: type I error rate of the naive t-test. In: *Stat Med* 18 (1999), Nr. 24, S. 3481–3491

[68] Jennison, C ; Turnbull, B.W.: *Group sequential methods with applications to clinical trials*. Boca Raton : Chapman and Hall/CRC, 2000

[69] Friede, T. ; Kieser, M.: A comparison of methods for adaptive sample size adjustment. In: *Stat Med* 20 (2001), Nr. 24, S. 3861–3873

[70] van der Meulen, E.A.: Are we really that blind? In: *J Biopharm Stat* 15 (2005), Nr. 3, S. 479–489

[71] Müller, H.H. ; Schäfer, H.: A general statistical principle for changing a design any time during the course of a trial. In: *Stat Med* 23 (2004), Nr. 16, S. 2497–2508

[72] Gayen, A.K.: Significance of difference between the means of two non-normal samples. In: *Biometrika* 37 (1950), Nr. 3-4, S. 399–408

[73] Simpson, E.H.: The Interpretation of Interaction in Contingency Tables. In: *J Roy Stat Soc B Met* 13 (1951), Nr. 2, S. 238–241

[74] Brannath, W. ; König, F. ; Bauer, P.: Estimation in flexible two stage designs. In: *Stat Med* 25 (2006), Nr. 19, S. 3366–3381

Danksagung

Ich möchte mich bei allen bedanken, die mich beim Anfertigen dieser Arbeit unterstützt haben, besonders Prof. Dr. Markus Neuhäuser für die Hilfe bei der Themenfindung als auch bei der wissenschaftlichen Untersuchung dieser interessanten Fragestellung. Prof. Dr. Karl-Heinz Jöckel gab mir einerseits die Möglichkeit, neben der Tätigkeit am Institut für Medizinische Informatik, Biometrie und Epidemiologie diese eigenen Untersuchungen durchzuführen, während ich andererseits in ihm immer einen kompetenten Ansprechpartner fand. Mit den Kolleginnen und Kollegen aus den Arbeitsgruppen Biometrie und Epidemiologie konnte ich viele interessante und inspirierende Fachgespräche führen, wobei ich Dr. Nils Lehmann und Dr. André Scherag besonders erwähnen will.

Ich danke der Deutschen Forschungsgemeinschaft für ihre Unterstützung beim Projekt "Varianzunterschiede in klinischen Studien", in dem viele Ergebnisse der vorgelegten Arbeit gewonnen wurden und ich danke Anja Marr für die unschätzbare Hilfe in diesem Projekt.

Allen Mitarbeiterinnen und Mitarbeitern des Instituts, die mir bei der Beschaffung der entsprechenden Roh- und Metadaten der untersuchten Studien behilflich gewesen sind, sei herzlich gedankt - wertvolle Beiträge lieferten vor allem Claudia Ose und Hildegard Lax.

Meiner Freundin Claudia danke ich für ihr Verständnis und ihre Geduld sowie die wertvolle Hilfe beim Korrekturlesen.

Abschließend danke ich meinen Eltern, die mich immer unterstützt und die mir zusammen mit meinem Bruder Wolfgang stets den nötigen Rückhalt gegeben haben.

Lebenslauf

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten