Dissertation

Optimale nicht-gewichtete Zwischenauswertungen zur Fallzahlneuberechnung

zur Erlangung des Grades eines Doktors der Naturwissenschaften
der Fakultät Statistik der TU Dortmund

am
17. Juli 2017

von
Katja Brandau

aus Wuppertal vorgelegt

Gutachter:

Prof. Dr. Guido Knapp

Prof. Dr. Jörg Rahnenführer
Fakultät Statistik
TU Dortmund

Prof. Dr. Hans-Joachim Trampisch
Medizinische Fakultät
Ruhr-Universität Bochum

Inhaltsverzeichnis

1 Kurzfassung

Die ethische Forschung an Lebewesen verlangt eine angemessene Fallzahl. Die Unvollkommenheit des derzeit konventionellen Vorgehens wie auch anderer Vorschläge zur Bestimmung einer angemessenen Fallzahl erläutern Kapitel 2.1 beziehungsweise Kapitel 2.3. Das Ziel der Arbeit ist die Reduktion des Defizits durch die Einführung eines alternativen Verfahrens zur Fallzahlbestimmung.

Den Kern bildet die Adaption des finalen Stichprobenumfangs im Studienverlauf (vgl. Kapitel 2.2). Sie erfolgt durch die Iteration der verlaufsunabhängigen Quantifizierung der Unsicherheit bezüglich der Fallzahl anhand sämtlicher vorliegender Daten und der darauf abgestimmten Erhebung zusätzlicher Daten (vgl. Kapitel 3.1). Die ausreichende Gewissheit in Bezug auf eine (geplante) angemessene Fallzahl oder auf die Gültigkeit der Nullhypothese leitet das Iterationsende ein. Gegebenenfalls werden noch ausstehende Beobachtungen erhoben. Der abschließende statistische Test der kumulierten Daten wird durch einen universalen kritischen Bereich definiert. Der gesamte Prozess wird in der Planungsphase der Studie durch die Optimierung eines Bayes-Risikos vollständig festgelegt. Kapitel 3.3 enthält einen Leitfaden für die Umsetzung des Konzepts.

Die Optimalität bezieht sich auf eine zu definierende Verlustfunktion zur Beurteilung flexibler Studiendesigns (vgl. Kapitel 3.2). Hierfür wird in Kapitel 4.2 mangels geeigneter Kandidaten (vgl. Kapitel 4.1) ein Grundgerüst entwickelt und nach einer Begutachtung in Kapitel 4.3, unter anderem für gruppensequenzielle Verfahren, als geeignet befunden. Infolge dessen wird zudem gezeigt, dass das vorgeschlagene Vorgehen zur Fallzahlbestimmung selbst unter starken Einschränkungen ein Gewinn sein kann (vgl. Kapitel 4.4).

2 Hintergrund

2.1 Das Kernproblem der Fallzahlplanung

In der medizinischen Forschung kommen statistische Tests zur Absicherung gegen die Einführung von Behandlungen ohne Zusatznutzen zum Einsatz. Zu dem Zweck können statistische Tests nicht-vorteilhafte Therapien mit der gewünschten Sicherheit als nicht-vorteilhaft erkennen. Ein sinnvoller Einsatz setzt darüber hinaus voraus, dass ein statistischer Test auch eine nützlichere Behandlung als solche identifizieren kann. Der Test gelte dann als valide. Seine Validität hängt maßgeblich vom Stichprobenumfang ab. Je mehr Beobachtungen vorliegen, umso weniger fallen die zufälligen Schwankungen der einzelnen Beobachtungen ins Gewicht. Das erleichtert die Identifikation eines etwaigen Zusatznutzens (als Abgrenzung zu einem Zufallsprodukt).

Nun konkurrieren der Stichprobenumfang und damit die Validität auf ökonomischer Ebene mit der Verfügbarkeit von Ressourcen. Die begrenzte Menge verfügbarer Probanden könnte in einer Studie die maximal mögliche Aussagekraft oder aber in mehreren Studien eine akzeptable Validität hergeben. Die Ressourceneffizienz ist aber auch unabhängig von konkurrierenden Studien ein Thema. Eine größere Stichprobe erhöht direkte wie auch indirekte Kosten. Zu den direkten Kosten können Studienmedikamente gezählt werden. Indirekte Kosten kommen zum Beispiel durch einen verzögerten Markteintritt eines Medizinproduktes aufgrund einer verlängerten Studienlaufzeit zustande. Unter anderem erhöhen größere Fallzahlen die Studiendauer mindestens dann, wenn eine konstante oder abnehmende Rekrutierungsrate vorliegt. Eine frühestmögliche Verfügbarkeit der Studienergebnisse ist auch aus ethischer Sicht angezeigt. Schließlich könnte den Patienten eine bessere Behandlung vorenthalten

werden. Das trifft insbesondere auf die überzähligen Studienteilnehmer zu, die unnötig mit einer unterlegenen Therapie behandelt werden. Hinzu kommen stets die vermeidbaren Belastungen durch die Studienteilnahme. Davon sind alle überreichlichen Patienten betroffen. Folglich verbieten ethische Argumente, zu denen auch die Ressourceneffizienz zählt, Studien mit einer mehr als gerade noch akzeptablen Validität. Die gleichen Argumente sprechen ebenso gegen eine zu geringe Teilnehmerzahl, womit zweifelhafte Ergebnisse riskiert werden. Die Ressourcen wären nicht zielführend eingesetzt.

Unter einer akzeptablen Validität einer Studie zur Absicherung gegen die Einführung von Behandlungen ohne Zusatznutzen sei eine Power von gerade eben 90 % ⟨5b⟩ zu verstehen, wo die maximal zugestandene Wahrscheinlichkeit für die fehlerhafte Einführung, bezeichnet als Signifikanzniveau, fest ist. Als Power werde eine Wahrscheinlichkeit bezeichnet, mit der der statistische Test eine im speziellen Ausmaß bessere Behandlung korrekt als besser anzeigt. Je größer die Wirksamkeit ist, umso wahrscheinlicher gelingt die Entkräftung der Unwirksamkeit ⟨5⟩ ⟨5d⟩. Deswegen muss das spezielle Ausmaß für die Begriffsklärung der akzeptablen Validität angegeben werden.

In der Literatur wird das spezielle Ausmaß mit der minimal relevanten Verbesserung gleichgesetzt ⟨5a⟩ ⟨5b⟩. Jedoch müsste eine allgemeingültig gerechtfertigte minimale Relevanz in die Formulierung der Testhypothese eingehen. Die Erfüllung der minimalen Relevanz entscheidet schließlich über die Einführung der neuen Therapie und ebendiese Entscheidung soll der statistische Test absichern. Für den besagten Schluss, ob mindestens die minimal relevante Verbesserung gegeben ist, ist die Validität zu beurteilen. Dafür kann die minimale Relevanz nicht auch noch herangezogen werden. Ohnehin ist sie ein Werturteil über die Behandlung. Die Validität ist aber eine Bewertung des statistischen Tests.

Ein Werturteil über die Validität anhand der Wahl eines speziellen Ausmaßes ist im medizinischen Kontext nicht gefragt. Wenn die neue Therapie besser ist, muss die wahre Verbesserung die Validität definieren. Mit der darauf abgestimmten Fallzahl gelingt die Einführung von 90 % aller überlegenen Behandlungen. Eine abweichende und damit unangemessene Fallzahl infolge der Wahl eines anderen speziellen Ausmaßes ist wegen der obigen Gründe zu vermeiden. Das ist übrigens ein weiteres, schwerwiegendes Argument gegen die Verwendung des minimal relevanten Werts, der selten mit dem wahren Wert übereinstimmen dürfte. Wenn hingegen die neue Therapie in Wahrheit nicht besser ist, verbietet sich die Durchführung einer Studie. Die Beurteilung der Validität für eine Fallzahlplanung erübrigt sich.

Jedoch ist die Wahrheit, wohlgemerkt im Vorfeld der Studie, im Allgemeinen nicht ausreichend bekannt. Die theoriebasierte Annäherung an die Wahrheit leidet unter der Komplexität des menschlichen Organismus. Ein unvollständiges Verständnis lässt an deduktiven Schlüssen zweifeln. Eben auch deswegen werden Studien durchgeführt. Sie sollen oftmals erst die Wahrheit so weit in Erfahrung bringen wie bereits für ihre Planung hätte bekannt sein sollen. Das Problem kann nur bedingt durch das Einbeziehen vorbestehender Studien gelöst werden. Dieser Weg entfällt für Studien auf (empirisch) unerforschten Gebieten (Fall 1). Dabei dürften gerade diese Studien bevorzugt gefördert und damit durchgeführt werden, weil sie großen Wissenszuwachs versprechen. Das ergibt sich aus der geforderten Ressourceneffizienz. Aus dem gleichen Grund darf die Forschungsfrage einer neuen Studie nur entfernt der Fragestellungen vorliegender gut geplanter Studien gleichen. Folglich können bestehende Studienergebnisse, wenn es sie denn gibt, nicht ohne Weiteres für die Fallzahlplanung der neuen Studie übernommen werden (Fall 2). Für die Übertragung braucht es jedoch wieder theoretische Überlegungen. Sie beinhalten die Quantifizierung sämtlicher Verzerrungen mindestens der

vorliegenden Studienergebnisse. Der Gedankengang lässt gegenüber der Planbarkeit korrekter Fallzahlen skeptisch werden. Dabei wurde, nebenbei bemerkt, noch nicht einmal auf die Schwierigkeit eingegangen, die Vorinformationen in statistische Effektmaße umzusetzen. Davon sind besonders nicht-parametrische Tests und Methoden im Kontext mehrerer zusammenwirkender Einflussgrößen betroffen. Folglich besteht der Bedarf, alternative Vorgehen bei der Fallzahlbestimmung zu erwägen.

2.2 Ein Ansatz zur Problemlösung

Mein Vorschlag für ein neues Vorgehen bei der Fallzahlplanung setzt am noch unerwähnten dritten Fall an, der nicht auftreten soll. Wenn es eine Studie gibt, der es wegen verfehlter Planung an Aussagekraft mangelt, kann die Wiederholung der Studie mit angepasster Fallzahl in Betracht gezogen werden. Es handelt sich also um keine Replikation. Die neue Studie dürfte durch die zusätzliche Erkenntnis eine angemessenere Fallzahl haben.

Für die Beantwortung der Fragestellung sollte dann stets die gesamte vorliegende Information in Form der Beobachtungen der beiden homogenen Studien genutzt werden. Das ist das Leitmotiv sogenannter Meta-Analysen. Hiernach gehen in den statistischen Test der nächsten Studie ebenso die Daten der ersten Studie ein, an denen dort die Hypothese bereits (erfolglos) geprüft wurde (vgl. Kapitel 2.3). Den mit der kumulativen Testung, unabhängig von einer Fallzahlanpassung, verbundenen Gewinn veranschaulicht Abbildung 1. Dort sind insgesamt 170 Probanden pro Behandlung auf potenziell zwei aufeinanderfolgende, höchstens bezüglich der Fallzahl verschiedene Studien aufgeteilt. Das Scheitern der ersten Studie durch eine unzureichende Evidenz gegen die Unwirksamkeit der einzuführenden Behandlung veranlasse die zweite Studie. Abbildung 1 zeigt drei Spezialfälle für die Aufteilung der Studienteilnehmer

auf die beiden Studien. Die Fälle umfassen eine nahezu gleiche Aufteilung (84:86) und die Aufteilungen mit der größtmöglichen Imbalance, bei denen einmal die erste und das andere Mal die zweite Studie fast alle Probanden enthält. Die Power der separaten Testung, bei der der zweite Test im Gegensatz zur kumulativen Testung nicht auf den 170 Beobachtungen aus beiden Studien, sondern auf den 84, 168 oder 2 Beobachtungen der zweiten Studie basiert, ist für die beiden Extremfälle deckungsgleich, weshalb Abbildung 1 scheinbar nur einen Extremfall mit separater Testung zeigt. Hiervon unterscheidet sich die kumulative Testung bei einer Imbalance zugunsten der zweiten Studie (2:168) kaum. Nichtsdestotrotz erreicht sie in allen Fällen eine, durchaus auch wesentlich, höhere Power als eine separate Testung.

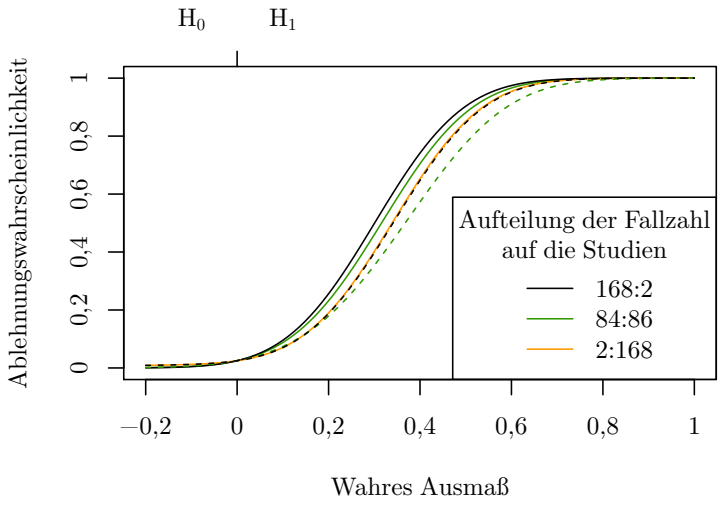

Abbildung 1: Wahrscheinlichkeit, die Nullhypothese H_0 im Setting des einseitigen Gauß-Tests (vgl. Seite 16) zum globalen Signifikanzniveau von 2,5 % mit insgesamt 170 Probanden je Behandlung abzulehnen, wenn zwei Studien lokal zu 1,25 % kumulativ (durchgezogene Linie) oder separat (gestrichelte Linie) getestet werden.

Im voranstehenden Beispiel wurden sowohl beim separaten als auch beim kumulativen Testen beide aufeinanderfolgende Tests lokal für ein niedrigeres Signifikanzniveau konstruiert. Das ist notwendig, damit das

globale Signifikanzniveau für das hinter den zwei Tests liegende multiple Testproblem eingehalten wird. Demnach müsste die erste Studie zu einer Fragestellung stets mit einem Test zu einem lokal reduzierten Signifikanzniveau geprüft werden, wenn eine ausreichende Fallzahl nicht garantiert werden kann. Andernfalls ist eine Wiederholung der Studie unter Einhaltung des globalen Signifikanzniveaus, außer in seltenen und daher vernachlässigbaren Sonderfällen, unmöglich. Nach der Überlegung in Kapitel 2.1 gibt es im Allgemeinen keine Garantie für eine angemessene Fallzahl. Allen betroffenen initialen Studien müsste der Power-Verlust eines niedrigeren Signifikanzniveaus angehängt werden. Im Grunde würde die Chance auf das Gelingen einer Studie verringert, weil sie gering sein könnte, um einem Fall Rechnung zu tragen, der nicht auftreten darf. Es scheint unrealistisch, dass es dafür einen Konsens gibt. Dabei hat jede sich dem widersetzende Studie mit unzureichender Fallzahl zur Folge, dass sich deren Fragestellung anschließend nicht mehr ohne Verletzung des Signifikanzniveaus beantworten lässt, sofern das initiale Scheitern die erneute Überprüfung veranlasst. Die Bedingung dürfte in der Regel gegeben sein. Das führt zu dem Schluss, dass nicht getestet werden darf, solange die Angemessenheit der Fallzahl zu ungewiss ist; auch nicht, wenn die Erhebung eines einzelnen Datensatzes für eine beabsichtigte Studie abgeschlossen ist. Weil solche Datensätze jeweils zweifelhaften Umfangs die Fragestellung nicht eigenständig, sondern in kumulierter Form gemeinsam beantworten sollen, werden sie mit dem Verständnis, dass eine Studie eine Fragestellung beantwortet, als Teile beziehungsweise Phasen einer einzelnen Studie aufgefasst. Die Erhebungen sollten wegen Synergieeffekten ohnehin von den gleichen Beteiligten möglichst nahtlos durchgeführt werden.

2.3 PUBLIZIERTE FLEXIBLE DESIGNS ZUR PROBLEMLÖSUNG

FALLZAHLMODIFIKATIONEN MITTELS CONDITIONAL REJECTION PROBABILITIES

Das aus Kapitel 2.2 resultierende Vorgehen gleicht dem adaptiver Verfahren. Bei adaptiven Designs wird eine Studie in mehrere Phasen unterteilt anstatt mehrere außer in der Fallzahl (und der Zeit) gleich geplanter Studien zu einer zu verbinden. Die Methodik adaptiver Verfahren dient der Einhaltung des Signifikanzniveaus trotz datenbasierter Studienverläufe wie bei Fallzahlanpassungen. Die Formulierung über Conditional Rejection Probabilities (CRPs) führten Müller und Schäfer 2004 ein ⟨12⟩. Darüber lassen sich die adaptiven Designs anderer Autoren darstellen, die über die Brownsche Bewegung oder p-Wert-Kombinationen argumentieren ⟨3a⟩.

Das CRP-Prinzip geht von einem statistischen Test $\varphi_K(\mathbf{X}_K) \in \{0,1\}$ zum Signifikanzniveau α_K mit den Daten $\mathbf{X}_K \in \mathbb{R}^{N_K}$ von N_K geplanten Studienteilnehmern aus. Im Vorfeld der ersten Datenerhebung zur Fragestellung sei $K = 1$ und α_1 gleiche dem Signifikanzniveau α. Anhand derselben Informationen $\mathcal{I}$, die für die voranstehende Planung verwendet werden, und gegebenenfalls, auch später eintreffenden, externen Informationen falle die Entscheidung für eine Zwischenauswertung mit $N_{K,1}$ Beobachtungen, $N_{K,1} \in (0, N_K]$. Durch die Aufteilung der Studie in zwei Phasen teilt sich der Zufallsvektor $\mathbf{X}_K$ in zwei Teilvektoren, $\mathbf{X}_K = (\mathbf{X}_{K,1}, \mathbf{X}_{K,2}) \in \mathbb{R}^{N_{K,1}+N_{K,2}}$.

Die Zwischenauswertung bezwecke infolge der Überlegungen vorangegangener Kapitel die Anpassung der Fallzahl N_K anhand der gegebenen Daten $\mathbf{X}_{K,1}$. Damit verfolge der weitere Studienverlauf die Erhebung des Zufallsvektors $\mathbf{X}_{K+1}(\mathbf{X}_{K,1}) = \mathbf{X}_{K+1} \in \mathbb{R}^{N_{K+1}(\mathbf{X}_{K,1})}$ statt des

Teilvektors $\mathbf{X}_{K,2} \in \mathbb{R}^{N_{K,2}}$. Hierdurch verändert sich auch der statistische Test am Studienende. Er hängt nun an einer weiteren Stelle von den Beobachtungen $\mathbf{X}_{K,1}$ ab: $\varphi_K\left((\mathbf{X}_{K,1}, \mathbf{X}_{K+1}(\mathbf{X}_{K,1}))\right)$. Es handelt sich nicht mehr um den ursprünglichen Test, der die Einhaltung des Signifikanzniveaus gewährleistet. Nach dem Beweis (2) von Müller und Schäfer kann der abschließende Test $\varphi_K\left((\mathbf{X}_{K,1}, \mathbf{X}_{K+1}(\mathbf{X}_{K,1}))\right)$ durch einen Test $\varphi_{K+1}\left((\mathbf{X}_{K,1}, \mathbf{X}_{K+1}\left(\mathbf{X}_{K,1}\right))\right)$ mit den zur Zwischenauswertung gleichen (oder kleineren) CRPs wie beim ursprünglich geplanten Test $\varphi_K\left(\mathbf{X}_K\right)$, und zwar

$$\alpha_{K+1}(\mathbf{X}_{K,1}) = \mathbb{E}_{H_0}\left(\varphi_K(\mathbf{X}_K)|\mathbf{X}_{K,1}\right), \tag{1}$$

ausgetauscht werden, um das Signifikanzniveau einzuhalten:

$$\begin{aligned} &\mathbb{E}_{H_0}\left(\varphi_{K+1}\left((\mathbf{X}_{K,1}, \mathbf{X}_{K+1}\left(\mathbf{X}_{K,1}\right))\right)\right) \\ &= \mathbb{E}_{H_0}\left(\mathbb{E}_{H_0}\left(\varphi_{K+1}\left((\mathbf{X}_{K,1}, \mathbf{X}_{K+1}\left(\mathbf{X}_{K,1}\right))\right)|\mathbf{X}_{K,1}\right)\right) \\ &\leq \mathbb{E}_{H_0}\left(\alpha_{K+1}\right) = \mathbb{E}_{H_0}\left(\mathbb{E}\left(\varphi_K(\mathbf{X}_K)|\mathbf{X}_{K,1}\right)\right) = \mathbb{E}_{H_0}\left(\varphi_K(\mathbf{X}_K)\right) \leq \alpha_K. \end{aligned} \tag{2}$$

Der Test $\varphi_{K+1}\left((\mathbf{X}_{K,1}, \mathbf{X}_{K+1}\left(\mathbf{X}_{K,1}\right))\right)$ zum Signifikanzniveau α_K werde als Test $\varphi_{K+1}\left(\mathbf{X}_{K+1}\right)$ zum Niveau α_{K+1} umformuliert. Hierfür lässt sich zum Zeitpunkt der Zwischenauswertung die Fallzahl planen. Es handelt sich um die noch ausstehende Teilnehmerzahl $N_{K+1}(\mathbf{X}_{K,1})$.

Mit $K = K+1$ kann das CRP-Prinzip wiederholt angewandt werden. Dafür stehen nun die Informationen $\mathcal{I} = \mathcal{I} \cup \{\mathbf{X}_{K-1,1}\}$ zur Verfügung. Vorschriften für die Bestimmung von $N_{K,1}$ und ein darüber impliziertes Iterationsende ($N_{K,1} = N_K$) schlug zum Beispiel Hartung vor ⟨8⟩.

Der Bewahrung der CRP (1) im Studienverlauf ist mit Skepsis zu begegnen. Die CRP drückt den Widerspruch der vorliegenden Beobachtungen $\mathbf{X}_{K,1}$ gegen die Nullhypothese vor dem Hintergrund der geplanten übrigen Daten $\mathbf{X}_{K,2}$ aus. Je stärker die gegebenen Beobachtungen der Nullhypothese widersprechen, umso größer sind die Wahrscheinlichkeiten, dass die Nullhypothese am Studienende abgelehnt wird. Sprechen

die gegebenen Beobachtungen gegen die gültige Nullhypothese, ist die CRP größer, wenn nur noch wenige Daten hinzukommen. Schließlich vermindern letztere den gemeinsamen Widerspruch unter der Gültigkeit der Nullhypothese höchstwahrscheinlich.

Signalisieren die vorliegenden Beobachtungen nun ein größeres Ausmaß als bei der Planung vermutet, werden daraufhin weniger Studienteilnehmer aufgenommen. Wäre die Teilnehmerzahl von Beginn an richtig geplant worden, müsste die CRP zur Zwischenauswertung höher sein als sie ist. Der Widerspruch der anfänglichen Beobachtungen wird mit der Bewahrung der CRP herabgespielt. Umgekehrt wird die Widerspruchskraft bei Beobachtungen überbewertet, wenn aufgrund des Signals eines unerwartet kleinen Ausmaßes mehr Probanden als ursprünglich geplant hinzukommen. Die zukünftigen Beobachtungen erhalten durch das CRP-Prinzip gegebenenfalls mehr beziehungsweise weniger Entscheidungsgewalt gegenüber den vorliegenden Daten als ihnen nach ihrem Informationsgehalt zustehen müsste.

Burman und Sonesson verdeutlichten an einem Zahlenbeispiel, dass die unverhältnismäßige Machtverteilung absurde Ergebnisse erzeugen kann ⟨2⟩. Darin wird eine Zwischenauswertung nach 100 von 1000 geplanten Beobachtungen durchgeführt. Die 100 Beobachtungen zeigen einen negativen Behandlungseffekt an. Daraufhin wird nur noch ein weiterer Studienteilnehmer beobachtet. Das CRP-Prinzip schreibt seinem Outcome die Entscheidungsgewalt von 900 Probanden zu. Deswegen genügt ein entsprechend geringer Einspruch einer einzelnen Beobachtung gegen einen nicht-positiven Behandlungseffekt, um ihn zu verwerfen. Dabei zeigt der Effektschätzer, der jede Beobachtung gleichwertig einbezieht, im Zahlenbeispiel einen negativen Effekt und somit das Gegenteil an. Die Erwägung eines nach unten verzerrten Schätzers (vgl. Punkt 7 auf Seite 34) ist kein Trost für das irritierende Studienergebnis. Die Situation könnte in

der Praxis auftreten, wenn die Vermutung einer unterlegenen neuen Behandlung die Aufnahme weiterer Probanden verbietet (vgl. Seite 6) und das Outcome (mindestens) eines bereits randomisierten Studienteilnehmers zum Zeitpunkt der Zwischenauswertung noch nicht vorlag. „In such a situation, it is accepted regulatory practice to base decision making on the final results of the trial (not the interim analysis)“ ⟨6⟩.

Wegen der unverhältnismäßigen Machtverteilung sollten jegliche Adaptionen mittels des CRP-Prinzips beziehungsweise mittels dessen Äquivalenten ⟨3a⟩ höchstens als Ausnahme in extremen Miseren zum Einsatz kommen. Die Etablierung des Prinzips zur Lösung des in Kapitel 2.1 beschriebenen Problems ist damit ausgeschlossen.

Gruppensequenzielle Tests

Auch gruppensequenzielle Tests sind keine einwandfreie Lösung im Umgang mit Unsicherheiten bei der Fallzahlplanung. Sie setzen zum Teil das in Kapitel 2.2 beanstandete Vorgehen um, Studien solange durchzuführen, bis eine kumulative Testung $\varphi_K\left(\{\mathbf{X}_\ell : 0 < \ell \leq K\}\right)$ zur Ablehnung der Nullhypothese führt ⟨18a⟩. Die Zahl der Tests ist aber, zumindest indirekt, begrenzt. Die Maximalzahl sei mit κ bezeichnet. Somit lässt sich das globale Signifikanzniveau α durch Tests mit niedrigeren, aber nicht unendlich kleinen lokalen Niveaus $0 < \alpha_1 \leq ... \leq \alpha_\kappa \leq \alpha$ ohne zusätzliche Modifikationen einhalten. Dazu gelte für den Hypothesentest φ_K der K-ten Auswertung, $K = 1, ..., \kappa$, unter Hinzunahme von $\alpha_0 = 0$

$$\mathbb{P}_{H_0}\Bigg(\left\{\varphi_K\left(\{\mathbf{X}_\ell : 0 < \ell \leq K\}\right) = 1\right\} \cap \bigcap_{L=1}^{K-1} \left\{\varphi_L\left(\{\mathbf{X}_\ell : 0 < \ell \leq L\}\right) \neq 1\right\}\Bigg) \leq \alpha_K - \alpha_{K-1}.$$

Für gleiche lokale Signifikanzniveaus und damit für ein gruppensequenzielles Design nach Pocock ⟨13⟩ spricht, dass andernfalls größere

Ausmaße benachteiligt würden. Das wäre befremdlich, wie die Übertragung auf Studien mit einmaliger Testung ohne Zwischenauswertungen verdeutlicht. Ein Studienplaner plane zwei Studien zu ein und derselben Erkrankung. Sie untersuchen auf die gleiche Weise jeweils eine andere Behandlung, die der Standardbehandlung gegenübergestellt wird. Folglich definiere der gleiche klinisch relevante Effekt die Hypothesen beider Studien. In beiden Fällen sei das wahre Ausmaß bekannt. Nun werde die wirksamere Behandlung zu einem Signifikanzniveau von einem statt 2,5 % Prozent getestet. Das werde aber nicht gemacht, weil die fälschliche Einführung einer unwirksamen Therapie bei dieser Erkrankung besonders schlimm sei. Dann müsste auch die andere Studie zu einem Signifikanzniveau von einem Prozent durchgeführt werden. Das Niveau der einen Studie werde stattdessen herabgesenkt, um der weniger wirksamen Therapie ein Signifikanzniveau von 4 % statt 2,5 % zuzugestehen. Die Hürden für eine Einführung der weniger wirksamen Behandlung wären damit niedriger. Das ist vergleichbar damit, die klinische Relevanz von der Wirksamkeit der Behandlung abhängig zu machen (und beide Prüfungen zum Niveau von 2,5 % durchzuführen). Der klinischen Relevanz würde ihre Bedeutung entzogen. Im gruppensequenziellen Design übersetzen sich demnach verschiedene lokale Signifikanzniveaus für die datengesteuert aufeinanderfolgenden Auswertungen in eine Abhängigkeit der klinischen Relevanz und damit der Hypothesen von der beobachteten Wirksamkeit der Behandlung.

Direkte datenbasierte Fallzahlanpassungen sind kein Bestandteil gruppensequenzieller Designs. Die kumulierten Fallzahlen der Auswertungen, $n_1, ..., n_\kappa$, müssen ohne die Nutzung studieneigener Daten festgelegt werden. Durch eine Ablehnung nach wenigen Auswertungen, $K < \kappa$, kann eine indirekte Fallzahlanpassung in Richtung kleinerer durchschnittlicher Teilnehmerzahlen stattfinden. Demnach müsste eine hohe Maximalfallzahl n_κ gewählt werden, wenn recht unterschiedliche Ausmaße wahr sein

könnten. Nur so ist auch bei einem kleineren wahren Ausmaß die Einhaltung des Signifikanzniveaus für die Beantwortung der Fragestellung gewährleistet (vgl. Kapitel 2.2). Dadurch werden aber für ein dann doch größeres wahres Ausmaß zunehmend mehr Probanden untersucht als nötig. Das Konzept gruppensequenzieller Tests ist nicht das Erreichen einer angemessenen Fallzahl, sondern die Ablehnung der Nullhypothese. In diesem Sinne streben gruppensequenzielle Verfahren eine Ablehnungswahrscheinlichkeit von 100 % an. Dafür werden alle Ressourcen verbraucht, die zur Verfügung stehen. Folglich erfüllen gruppensequenzielle Designs nicht die gewünschten Eigenschaften, um als Patentrezept bei mehr als nur geringfügigen Unsicherheiten bei der Fallzahlplanung gelten zu können.

3 Entwicklung des saturierten flexiblen Designs

3.1 Adaptierbare Elemente

Ausgehend vom Ansatz aus Kapitel 2.2 wird ein neues Vorgehen bei der Fallzahlplanung entwickelt. Er sieht die Hypothesenprüfung erst bei einer angemessenen Gesamtfallzahl vor. Dazu ist die Feststellung der Angemessenheit zu klären. Die Fallzahl unterliegt durch ihre, dem Ansatz getreue, empirische Bestimmung anhand der bislang erhobenen Beobachtungen zufälligen Schwankungen. Das ist unvermeidbar. Trotzdem sind zutreffende Fallzahlen möglich. Ob eine gegebene Schwankung zu bedeutend abweichenden Teilnehmerzahlen führen kann, hängt vom Zusammenspiel der Größe des unbekannten wahren Ausmaßes und weiterer Parameter der Verteilung der Daten ab, auf die das Hypothesenpaar nicht abzielt.

Als veranschaulichendes Beispiel diene der Gauß-Test zum Signifikanzniveau von $\alpha = 0{,}025$ für das einseitige Testproblem

$$H_0 : \mu_1 - \mu_2 \leq 0 \quad vs. \quad H_1 : \mu_1 - \mu_2 > 0$$

bezüglich der unbekannten Erwartungswerte von $2n$ stochastisch unabhängigen, normalverteilten Zufallsvariablen $Y_{ij} \sim \mathcal{N}(\mu_j, \sigma)$, $j \in \{1{,}2\}$, $i = 1, ..., n$, mit bekannter Standardabweichung $\sigma > 0$ ⟨5c⟩. Die Fallzahl pro Behandlungsgruppe j für eine Power von $[1 - \beta]$ für den wahren Behandlungseffekt $\theta = [\mu_1 - \mu_2]/\sigma$ berechnet sich durch

$$f(\theta; \alpha, \beta) = 2\left[\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right]^2/\theta^2, \qquad (3)$$
$$\theta > 0,\ 0 < \alpha < 1 - \beta < 1.$$

Dabei steht Φ^{-1} für die Quantilfunktion der Standardnormalverteilung. Nicht-ganzzahlige und somit nicht-realisierbare Fallzahlen sind auf die nächstgrößere ganze Zahl aufzurunden, um die geforderte Power zu gewährleisten.

Die Abhängigkeit der Fallzahl pro Gruppe vom wahren Behandlungseffekt θ zeigt Abbildung 2. Hervorgehoben sind die Wertepaare (0,275; 278) und (0,325; 199) sowie (0,475; 94) und (0,525; 77). Das Beispiel veranschaulicht, dass die gleiche Abweichung ($\pm 0{,}025$) bei einem kleineren Effekt ($\theta = 0{,}3$) zu unterschiedlicheren Fallzahlen (278 und 199 statt $f(0{,}3; 0{,}025; 0{,}1) = 234$ versus 94 und 77 statt $f(0{,}5; 0{,}025; 0{,}1) = 85$) führt als bei einem größeren Wert ($\theta = 0{,}5$). Das gilt nicht nur für den Gauß-Test ⟨5d⟩. Deshalb müssen kleinere Effekte präziser geschätzt sein, um eine vergleichbar verlässliche Fallzahlschätzung zu bieten.

Für die Fallzahlplanung sind nicht die Parameter der Verteilung der Daten an sich, sondern ihre Auswirkung auf die Teststatistik entscheidend. Das wird an nicht-parametrischen Verfahren besonders deutlich. Daher basiere die Beurteilung der Angemessenheit einer Fallzahl direkt auf einem Schätzer des unbekannten wahren Werts der Teststatistik. Die

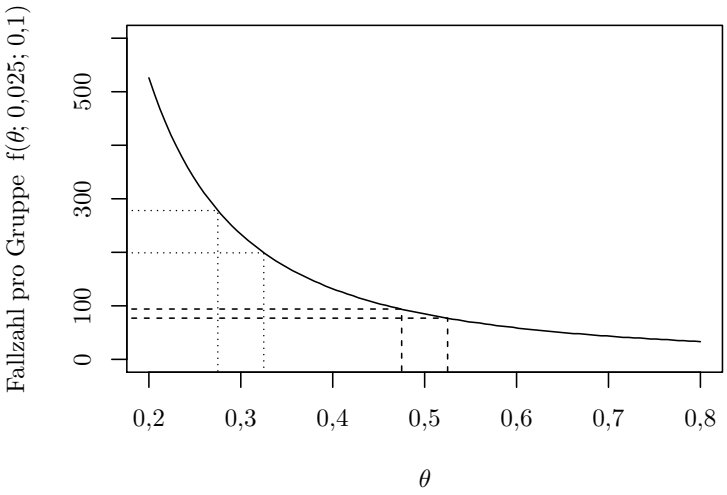

Abbildung 2: Auswirkung gleichgroßer Abweichungen (hier ±0,025) vom wahren Ausmaß θ auf die realisierbare Fallzahl pro Gruppe laut (3) für $\theta = 0{,}3$ und $\theta = 0{,}5$.

Sinnhaftigkeit zeigt sich darin, dass eine gegebene Fallzahl im konkreten Fall ausreicht, wenn die Nullhypothese aufgrund der Größe der beobachteten Teststatistik abgelehnt werden müsste. Die beobachtete Teststatistik ist ein Schätzer für den wahren Wert der Teststatistik gegeben der aktuellen Fallzahl. Es ist aber der Wert im Falle der angemessenen Fallzahl gefragt. Deshalb hänge die finale Teilnehmerzahl sowohl von der beobachteten Teststatistik als auch von der gegebenen Fallzahl ab. Die gegebene Fallzahl ist zudem für die Beurteilung der Unsicherheit als Folge der Schätzung heranzuziehen. Die Präzision einer sinnvollen Schätzung hängt im Rahmen der festen Eigenschaften des Schätzers und der gegebenen Verteilung der Daten von der Anzahl der zur Schätzung genutzten Beobachtungen ab. Wenn die gegebene Fallzahl zu klein ist, um sich der ausreichenden Präzision der Schätzung genügend sicher zu sein, empfiehlt es sich, weitere Beobachtungen für die Festlegung der finalen Fallzahl hinzuzunehmen. Dadurch werden weitere Studienphasen zur Überprüfung und Anpassung der finalen Fallzahl eingeführt. Die pauschale Durchführung einer festen Zahl von Überprüfungen wäre dagegen unattraktiv. Es

könnten gegebenenfalls Unsicherheiten oder die Mehrkosten unnötiger Zwischenauswertungen getragen werden müssen.

Unnütze Zwischenauswertungen drohen außerdem (auch bei ihrer adaptiven Einführung) dann, wenn die Probandenzahl für die nächste Studienphase nicht dem aktuellen Überprüfungsbedarf nachkommt. Das wird an einer feststehenden Abfolge von Fallzahlen deutlich, die der Studienplan für die aufeinanderfolgenden Phasen (noch unbekannter Zahl) vorschreibe. Unter Umständen ist die für die nächste Zwischenauswertung vorgesehene Fallzahl zu klein, als dass die Gewissheit bezüglich der Gesamtfallzahl relevant zunehmen könnte. Möglicherweise wäre eine direkte Endauswertung effizienter, obwohl Überprüfungsbedarf besteht. Das kann auch gelten, wenn die angemessene Gesamtfallzahl kleiner als die Anzahl der Probanden ist, die laut Plan für die nächste Zwischenauswertung erhoben werden müssten. Deshalb umfassen die Adaptionen im Studienverlauf die Entscheidung für eine weitere oder für die letzte Studienphase in Verbindung mit der dafür einzusetzenden Fallzahl. Die eigentliche Entscheidung falle für die Kombination der beiden Elemente. Sie basiere auf der Operationalisierung der Unsicherheit, bestehend aus der aktuell vorliegenden Teilnehmerzahl und der beobachteten Teststatistik.

Der zurückliegende Studienverlauf gehe nicht in die Beurteilung der Unsicherheit ein. Es fehlt ein inhaltlicher Grund für die Unterscheidung, ob die gegebene Sicherheit durch viele Probanden nach größerer Unsicherheit oder durch weniger viele Beobachtungen infolge geringerer Unsicherheit gewonnen wurde. Auch erübrigt sich die Berücksichtigung begrenzter verbliebener Entscheidungsmöglichkeiten, da prinzipiell keine Begrenzung vorgesehen ist.

Die Option, keine weiteren Studienteilnehmer mehr aufzunehmen, ist aus ethischer Sicht angezeigt, wenn mit ausreichender Gewissheit

die angemessene Fallzahl bereits vorliegt. Dazu müssen die Beobachtungen $\mathbf{X}_{N=n} \in \mathbb{R}^N$ einen für den gegebenen Stichprobenumfang n ausreichend großen Effekt anzeigen. Die Realisation der Statistik $T(\mathbf{X}_n)$ muss also eine Mindestgröße $c^*_{n,n}$ überschritten haben. Der Zustand bleibt für den Hypothesentest bestehen, weil keine weiteren Beobachtungen hinzukommen. Ein kritischer Wert c^*_n für die Statistik, der größer als die besagte Untergrenze ist, $c^*_n > c^*_{n,n}$, ist aus inhaltlichen Gründen abzulehnen. Er definierte eine Entscheidungsregel, bei der im Falle beinahe ausreichender Gewissheit (mindestens) ein weiterer Proband für den unmittelbar darauffolgenden Test aufgenommen wird,

$$T(\mathbf{X}_n) \leq c^*_{n,n} < c^*_n,$$

bei einem größeren Effekt die Nullhypothese sofort angenommen,

$$c^*_{n,n} < T(\mathbf{X}_n) \leq c^*_n,$$

und bei noch größerer Gewissheit die Nullhypothese abgelehnt würde,

$$c^*_{n,n} < c^*_n < T(\mathbf{X}_n).$$

Die Annahme der Nullhypothese bei sehr großen, aber nicht bei darunterliegenden Effekten ist widersinnig. Deswegen gelte die Mindestgröße für die Entscheidung, keine weiteren Studienteilnehmer mehr aufzunehmen, als Obergrenze für den kritischen Wert.

Bei einem höchstens genauso großen kritischen Wert wie der Mindestgröße ist die Entscheidung für ein Rekrutierungsende zwangsläufig mit der Ablehnung der Nullhypothese verbunden:

$$c^*_n \leq c^*_{n,n} \quad \wedge \quad T(\mathbf{X}_n) > c^*_{n,n} \quad \Rightarrow \quad T(\mathbf{X}_n) > c^*_n.$$

Ein kleinerer kritischer Wert bedeutet, dass ein und dieselben Beobachtungen mal als Widerspruch zur Nullhypothese und mal als dafür unzureichend gewertet werden. Das hängt von der Effektschätzung in der

vorangegangenen Studienphase ab. Im ersten Szenario erfolge nach einer Fallzahlüberprüfung mit n_1 Beobachtungen eine weitere Überprüfung nach insgesamt n_3, also $n_3 - n_1$ zusätzlichen Studienteilnehmern. Für eine Ablehnung der Nullhypothese muss im nächsten Studienteil $T(\mathbf{X}_{n_3}) > c^*_{n_3,n_3}\,(> c^*_{n_3})$ gelten. Hingegen genügt $T(\mathbf{X}_{n_3}) > c^*_{n_3}$ als Ablehnungskriterium, wenn die vorangegangene Studienphase mit n_2 Beobachtungen, $n_2 \neq n_1$, die endgültige Festlegung auf die Gesamtfallzahl n_3 ergab. Folglich hängt die Bewertung der Widerspruchskraft der gleichen Datenlage vom Studienverlauf ab. Dafür gibt es jedoch weiterhin inhaltlich keinen Grund. Es verletzte sogar das Suffizienzprinzip ⟨3b⟩. Demzufolge muss ein kritischer Wert der Untergrenze für die Entscheidung, keine weiteren Studienteilnehmer mehr aufzunehmen, gleichen.

Zudem müssen die kritischen Werte derart in Beziehung zueinanderstehen, dass keine Effekte benachteiligt werden (vgl. Kapitel 2.3). Die kritischen Werte seien deshalb gleich, $c = c^*_n\ \forall n$.

Der letzte Punkt behandelt die frühzeitige Annahme der Nullhypothese. Sie knüpft an den Gedanken aus Kapitel 2.1 an, dass sich die Durchführung einer Studie bei bekannter gültiger Nullhypothese verbietet. Der Verzicht auf die Durchführung ist ein Studienabbruch in der Planungsphase und kein vorzeitiger Studienabschluss. Es gibt mangels Beobachtungen keine statistisch gesicherte Entscheidung bezüglich der Nullhypothese. Das ist bei einem vorzeitigen Studienende mit Annahme der Nullhypothese als Teil eines datengesteuerten Verfahrens anders. Die Entscheidung wird an einem definierten Kriterium bezüglich der Beobachtungen festgemacht. Hiermit sollten die Studienpfade bekannt sein, die zwar im Verlauf das Kriterium für die vorzeitige Annahme der Nullhypothese erfüllen, aber am Ende doch zu einer Ablehnung geführt hätten. Hierdurch kann die gesamte Entscheidungsregel darauf abgestimmt werden, geringere Ablehnungswahrscheinlichkeiten zu kompensieren. Damit

sind etwaige Defizite bei der Ausschöpfung des Signifikanzniveaus und gegebenenfalls bei der Power minimal. Folglich ist eine vorzeitige Annahme der Nullhypothese ein im Gesamtkonzept statistisch gesicherter Schluss. Die vorzeitige Annahmeoption wird deshalb prinzipiell in die Entscheidungsregel aufgenommen.

Aus den gesamten obigen Begründungen ergibt sich eine Entscheidungsregel, bei der die Entscheidungen anhand der Kombination von drei Kriterien getroffen werden. Es handelt sich um die Anzahl der bislang erhobenen Beobachtungen, ob sie in einer Fallzahlüberprüfung oder einem abschließenden statistischen Test zum Einsatz kommen und in welchen Wertebereich die beobachtete Teststatistik fällt. Die Entscheidungsmöglichkeiten umfassen die Fallzahl für die nächste Studienphase und ob es sich dabei um die endgültige Gesamtfallzahl handelt oder nicht. Im Falle der Endgültigkeit kommen die Annahme und die Ablehnung der Nullhypothese als Wahlpflichtoptionen hinzu.

Für die Formalisierung der Entscheidungsregel sei $(n_i, i \in \mathbb{N})$ die (unendliche) Folge der Stichprobenumfänge, auf denen die Auswertungen basieren können. Im überschaubaren Beispiel in Tabelle 1 sind $n_1 = 20$, $n_2 = 40$, $n_3 = 60$ und $n_4 = 80$. Für jeden Stichprobenumfang n_i werde der Wertebereich der Teststatistik für die Entscheidungsfindung partitioniert. Dabei werde differenziert, ob es sich bei der aktuellen Auswertung um eine Fallzahlüberprüfung oder um die finale Auswertung handelt. Bei der finalen Auswertung besteht die Partition exakt aus zwei Bereichen. Eine beobachtete Teststatistik, die größer als der universale kritische Wert c ist, führe zu einer Ablehnung der Nullhypothese. Sonst werde die Nullhypothese angenommen. Das beinhalten im Beispiel die Zeilen 13, 14, und 18 bis 21 der Tabelle 1.

Bei einer Fallzahlüberprüfung werde der Bereich $(-\infty, c]$ weiter unterteilt. Es bleibe bei der Ablehnung der Nullhypothese im Falle einer

Zeile	Entscheidungskriterien			Entscheidung		
	Aktuelle Fallzahl	als finale Fallzahl	Untere Intervallgrenze	Zukünftige Fallzahl	als finale Fallzahl	Testentscheidung
1	20		$-\infty$	20	$*$	H_0
2	20		$c_{20,60}$	60		
3	20		$c_{20,40}$	40		
4	20		$c^*_{20,80}$	80	$*$	
5	20		$c^*_{20,60}$	60	$*$	
6	20		$c^*_{20,40}$	40	$*$	
7	20		c	20	$*$	H_1
8	40		$-\infty$	40	$*$	H_0
9	40		$c_{40,60}$	60		
10	40		$c^*_{40,80}$	80	$*$	
11	40		$c^*_{40,60}$	60	$*$	
12	40		c	40	$*$	H_1
13	40	$*$	$-\infty$	40	$*$	H_0
14	40	$*$	c	40	$*$	H_1
15	60		$-\infty$	60	$*$	H_0
16	60		$c^*_{60,80}$	80	$*$	
17	60		c	60	$*$	H_1
18	60	$*$	$-\infty$	60	$*$	H_0
19	60	$*$	c	60	$*$	H_1
20	80	$*$	$-\infty$	80	$*$	H_0
21	80	$*$	c	80	$*$	H_1

Tabelle 1: Entscheidungsprozedur mit eingeschränkter Rekrutierung, die mit einer Fallzahlüberprüfung (kein Stern [$*$] in Spalte 3) mit 20 Beobachtungen pro Behandlung (Spalte 2) beginnt und je nach Intervallzugehörigkeit der beobachteten Teststatistik (Spalte 4) eine Entscheidung (rechte Tabellenhälfte) anzeigt. Eine weitere Auswertung findet statt, wenn die zukünftige die aktuelle Fallzahl übersteigt. Dann schreiben Spalte 5 und 6 die Eigenschaften der nächsten Auswertung und somit die sodann vorliegenden Entscheidungskriterien in Spalte 2 und 3 vor. Die Entscheidungsprozedur endet mit einer Testentscheidung.

beobachteten Teststatistik oberhalb des Wertes c. Wenn die Teststatistik im Bereich $(c^*_{n_i,n_j}, c^*_{n_i,n_{j-1}}]$ liegt, $n_i < n_j$, falle die endgültige Entscheidung für insgesamt n_j Studienteilnehmer. Dass die damit verbundenen Beobachtungen ausschließlich in den finalen Test eingehen, zeige der Stern $(*)$ an. Schlüssig wird die Notation durch $c^*_{n_i,n_i} = c$. Im Beispiel in Tabelle 1 fällt die endgültige Entscheidung für 60 weitere Probanden, wenn die mit 20 Beobachtungen ermittelte Teststatistik im Bereich $(c^*_{20,80}, c^*_{20,60}]$ liegt (vgl. Zeile 4 der Tabelle 1). Nimmt sie dagegen einen Wert des Intervalls $(c^*_{20,40}, c]$ an, soll die Studie unumstößlich insgesamt 40 Studienteilnehmer enthalten (vgl. Zeile 6 der Tabelle 1). Die nächste und letzte Entscheidung fällt daraufhin zwangsläufig in den Zeilen 13 und 14 der Tabelle 1.

Eine weitere Fallzahlüberprüfung mit insgesamt n_j Beobachtungen folge auf eine beobachtete Teststatistik im Intervall $(c_{n_i,n_j}, c_{n_i,n_{j-1}}]$, $n_i < n_j < \max(n_\ell, \ell \in \mathbb{N})$. Dabei gelte $c_{n_i,n_i} = c^*_{n_i,\max(n_\ell, \ell\in\mathbb{N})}$. Damit findet im Beispiel nach einer Fallzahlüberprüfung mit 40 Beobachtungen eine Studienphase mit insgesamt 60 Beobachtungen zur Überprüfung statt, wenn die Teststatistik einen Wert zwischen $c_{40,60}$ und einschließlich $c^*_{40,80}$ annimmt (vgl. Zeile 9 der Tabelle 1). Eine Überprüfung mit den Beobachtungen von insgesamt 80 Studienteilnehmern kann nicht stattfinden, weil keine höheren Fallzahlen zur Wahl stehen. Die Notation berücksichtigt das durch $n_j < \max(n_\ell, \ell \in \mathbb{N})$.

Der Bereich $(-\infty, c_{n_i,\max\{n_j:\,(n_\ell,\,\ell\in\mathbb{N})\setminus\max(n_\ell,\,\ell\in\mathbb{N})\}}]$ komplettiert die Partition des Wertebereichs der Teststatistik für eine Überprüfung mit n_i Beobachtungen. Er ist mit dem Studienende mit einer vorzeitigen Annahme der Nullhypothese verbunden (vgl. Zeile 1, 8 und 15 der Tabelle 1).

Tabelle 1 lässt erkennen, dass das eingeführte Design ein gruppensequenzielles Design (vgl. Kapitel 2.3) als Spezialfall beinhaltet. Eine Studie ohne Zwischenauswertung kann wiederum als Spezialfall eines gruppensequenziellen Tests betrachtet werden. Tabelle 1 beschreibt einen

gruppensequenziellen Test mit vier Auswertungen jeweils nach 20 zusätzlichen Beobachtungen, wenn $c_{20,60} = c_{20,40} = -\infty$ und $c^*_{20,80} = c^*_{20,60} = c^*_{20,40} = c$, $c_{40,60} = -\infty$ und $c^*_{40,80} = c^*_{40,60} = c$ sowie $c^*_{60,80} = c$ gilt. Ansonsten ermöglichen die Zeilen 2, 3 und 9 der Tabelle 1 das Überspringen von Auswertungen mit vermutlich geringem oder inversem Zusatznutzen. Den Zeilen 4 bis 6, 10 und 11 sowie 16 kommt die Funktion zu, dem sonst gegebenen Streben nach einer Power von 100 % entgegenzuwirken (vgl. Kapitel 2.3).

Mit der konkreten Wahl der prinzipiell unendlich vielen Entscheidungsgrenzen befasst sich das nachfolgende Kapitel 3.2. Die Beschränkung auf eine endliche Anzahl von Auswertungen bis hin zu einer endlichen Teilnehmerzahl ist nicht damit zu vereinbaren, erst bei ausreichend gewisser finaler Gesamtfallzahl statistisch zu testen (vgl. Kapitel 2.2). Ausmaße an der Grenze zur Nullhypothese verlangen, auch bei einem Studiendesign ohne Zwischenauswertungen, unendlich viele Studienteilnehmer. Um sich der unendlichen Gesamtfallzahl sicher zu sein, braucht es selbst unendlich viele Probanden und Überprüfungen. Aufgrund der Fülle der prinzipiellen Optionen kann das eingeführte Design als saturiertes flexibles Design (SaFD) bezeichnet werden.

3.2 Automatisierung anhand einer Verlustfunktion

Eine flexible Entscheidungsregel zur Fallzahlanpassung ist vorab präzise zu definieren, sodass der Studienverlauf automatisch vonstattengehen kann. Die Forderung wird für adaptive Studiendesigns in der Literatur teils unter der Bezeichnung des Self-Designs umgesetzt ⟨8⟩ ⟨15⟩. Sie ist die Voraussetzung für die Beurteilung der Auswirkung der getroffenen Entscheidungen. Die Interpretierbarkeit der Studienergebnisse ist für die Akzeptanz des vorgeschlagenen Vorgehens unerlässlich. Zudem haben nur

so Vergleiche mit anderen Entscheidungsregeln Bestand. Nur die beste Entscheidungsregel darf in den Studienplan aufgenommen werden.

Die perfekte Entscheidungsregel liefert laut Kapitel 2.1 mit dem kleinstmöglichen Ressourcenaufwand einen gerade noch validen Schluss. Der Validitätsbegriff bezieht sich auf das wahre Ausmaß. Die unzureichende Kenntnis des wahren Ausmaßes behindert daher die Beurteilung der Perfektion einer Entscheidungsregel und damit die Studienplanung. Ein ausreichendes Wissen liege vor, wenn sich die Unkenntnis auf Ausmaße mit einem gleich hohen realen Ressourcenbedarf für einen gerade eben validen Schluss beschränkt. Dann ist für die Studienplanung irrelevant, welches Ausmaß das wahre ist. Genauso wäre bei größerer Unwissenheit eine Entscheidungsregel wünschenswert, die für alle potenziell wahren Ausmaße perfekt ist. Dem steht die Gegenläufigkeit des Strebens nach möglichst vielen (zu verschiedenen) Ausmaßen mit validen Schlüssen und dem dafür benötigten Ressourceneinsatz entgegen. Der Einsatz umfasst sowohl die Gesamtfallzahl als auch die Anzahl der Auswertungen. Zum Beispiel sei die korrekte Teilnehmerzahl für das kleinstmögliche Ausmaß geplant. Damit sind alle potenziell wahren Ausmaße mit validen Schlüssen verbunden. Dafür weicht der Ressourcenverbrauch für größere Ausmaße vom Ideal ab, indem die Fallzahl zu hoch ist und/oder durch Zwischenauswertungen angepasst wird. Im gegenteiligen Extremfall ist bloß die Testentscheidung für das größtmögliche Ausmaß valide. Das Risiko vergeblich aufgewandter Ressourcen ist entsprechend hoch, wenn nicht anderweitig Kosten durch Fallzahlanpassungen entstehen. Eine Entfernung von der perfekten Entscheidungsregel ist also unvermeidlich, sobald für mindestens ein mögliches Ausmaß eine andere Fallzahl als für andere potenziell wahren Ausmaße angemessen ist.

Welche imperfekte Entscheidungsregel in den Studienplan aufgenommen wird, bestimmt weiterhin das Gebot der Ressourceneffizienz. Dar-

unter kann aber nicht mehr nur der unter der Nebenbedingung einer Power von 90 % minimierte Ressourcenaufwand verstanden werden. Die Gewährleistung der Validität für alle möglichen Ausmaße erfordert in letzter Konsequenz einen beliebig hohen zusätzlichen Ressourcenaufwand (für das eine wahre Ausmaß). Auch sonst könnte ein unwesentlich höheres Risiko für vergebliche Aufwendungen sicheren Mehrkosten vorzuziehen sein. Demzufolge sind invalide Schlüsse mit einer effizienten Entscheidungsregel prinzipiell kompatibel. Die optimale Entscheidungsregel führt demnach den besten Kompromiss aus den Aufwendungen unter zwei verschiedenen Umständen herbei. Zum einen handelt sich um den Ressourceneinsatz, dessen vollständiger Verlust bei der Bewahrheitung der Ausmaße mit Ablehnungswahrscheinlichkeiten unter 90 % droht. Eine niedrigere Wahrscheinlichkeit bedeutet ein höheres Risiko. Auf der anderen Seite stehen die Mehrkosten für eine ausreichende Power den Kosten der perfekten Entscheidungsregel gegenüber.

Valide Schlüsse werden nach wie vor untereinander allein anhand der für sie benötigten Ressourcen verglichen. Das bedeutet, dass die Höhe der erreichten Power ab 90 % irrelevant ist. Andernfalls müsste von der bestehende Definition der Validität abgerückt werden (vgl. Kapitel 2.1). Die Definition der besten Entscheidungsregel zeigt also die perfekte Entscheidungsregel im Falle der ausreichenden Bekanntheit des wahren Ausmaßes an.

Die Findung der besten Entscheidungsregel setzt die Formulierung des genannten Kompromisses als Verlustfunktion mit eindimensionalem Wertebereich voraus. Eine zweikriterielle Optimierung konfrontierte den Studienplaner mit unvergleichlichen Entscheidungsregeln. Eine Verbesserung in einem Aspekt ist ausschließlich durch eine Verschlechterung des anderen Aspekts möglich (vgl. Seite 25). Die Optimierung der Verlustfunktion hat unter der Nebenbedingung des eingehaltenen Signifikanzni-

veaus zu erfolgen. Existieren mehrere global optimale Entscheidungsregeln, gebe die größere Menge potenzieller Effekte mit einer Power von mindestens 90 % den Ausschlag.

Die Verlustfunktion muss auf die speziellen Rahmenbedingungen einer Studie abgestimmt sein. Die monetären Kosten für eine Fallzahlprüfung variieren für jede Studie ebenso wie die ethische Dringlichkeit, Probanden oder Zeit einzusparen. Das bedeutet jedoch nicht zwangsläufig ein weniger idealistisches Vorgehen. Kosten für Überprüfungen, die kaum mehr Gewissheit bringen, sollten anderen Forschungsprojekten zukommen. Ebenso folgt die schnellere Beantwortung der Fragestellung aus ethischen Gründen Idealen. Zum Beispiel verlangt eine neuartige Epidemie eine zeitnahe Einführung von Therapien. Zugunsten einer kürzeren Studiendauer können Nachteile hinnehmbar sein, die in anderen Situationen weniger Akzeptanz finden.

Nichtsdestotrotz darf der Kompromiss nicht zu stark ausfallen. Ein unbrauchbares Ergebnis wird nicht wertvoller, indem es schnell ermittelt wurde. Es muss akzeptiert werden, dass manche Studien viel Zeit für verlässliche Ergebnisse benötigen. Der Erkenntnisgewinn hat Priorität. Das gilt unabhängig von der Dauer bis zur Erhebung des Endpunktes. Falls also einschließlich der Berücksichtigung der rein ethischen Anforderungen (an die Studiendauer) die beste Entscheidungsregel einen datengesteuerten Studienverlauf beinhaltet, ist diese Entscheidungsregel zu verwenden. Soll zum Beispiel die 10-Jahres-Mortalität untersucht werden, dauert es eben mindestens 10 Jahre bis zum Beginn der nächsten Studienphase. Das kann besser als eine zehnjährige Studie sein, die aber unbrauchbare Ergebnisse liefert.

Falls die Durchführung einer Studie mit einem datengesteuerten Verlauf nicht machbar ist, kann die Studie nicht durchgeführt werden. Das ist vergleichbar mit der Situation, in der die angemessene Fallzahl bekannt

ist, sie sich aber mangels Sponsoren nicht realisieren lässt. Die Durchführung mit geringeren Teilnehmerzahlen ist wegen des Risikos für einen vergeblichen Ressourceneinsatz mitsamt der zu erwartenden Verletzung des Signifikanzniveaus bei einer Folgestudie keine verantwortliche Option (vgl. Kapitel 2). Die mangelnde Machbarkeit ist also kein Argument gegen einen datengesteuerten Studienverlauf, wie manch eine Publikation denken lässt ⟨4⟩, sondern gegen die generelle Durchführung der Studie. Über die Datensteuerung entscheiden die ethischen Anforderungen. Die ethischen Aspekte kommen in der Verlustfunktion zusammen.

Die Abwägung der Ideale, unter anderem über die Sanktionierung der Auswertungszahl, sollte einer Ethikkommission obliegen. Präzedenzfälle wären wünschenswert. In jedem Fall begrenzte die Zuständigkeit die Zweckentfremdung aus rein wirtschaftlichen Interessen. Die Möglichkeit der Zweckentfremdung ist ein Argument für eine einheitliche Verlustfunktion, die aber eben den unterschiedlichen ethischen Ansprüchen verschiedener Studien nicht gerecht würde. Nichtsdestotrotz könnte es ein gemeinsames Grundgerüst geben, das die geforderte Individualität zulässt. Hiermit befasst sich Kapitel 4.

3.3 Leitfaden zur Umsetzung

Der folgende Leitfaden zur Umsetzung des entwickelten Studiendesigns dient sowohl der Zusammenfassung als auch der Ergänzung um praktische Aspekte und Anregungen für zukünftige Forschungstätigkeiten (◀). Von der Wiedergabe allgemeiner Handlungsempfehlungen, wie von der EMEA ⟨6⟩ und von der FDA ⟨7⟩, wird dabei abgesehen.

1. Den Ausgangspunkt für eine Studie mit einem SaFD bildet standardmäßig die Formulierung des Testproblems und die Wahl eines passenden statistischen Tests unter Berücksichtigung der Eigenschaften der zur Beantwortung der Fragestellung geeigneten Beobachtungen.

Wenn die multivariate Verteilung der damit verbundenen kumulativen Statistiken der jeweiligen Auswertungen nicht bestimmt werden kann, kann erwogen werden, die Studienplanung für den Gauß-Test durchzuführen (vgl. Seite 16). Die Ergebnisse gelten approximativ. Die Eignung des Behelfs, insbesondere vor dem Hintergrund von ◀ Überprüfungen nach wenigen zusätzlichen Studienteilnehmern, und alternativen Vorgehen bieten sich als Gegenstand zukünftiger Untersuchungen an.

2. An die gewohnte Initialisierung schließt sich die gewissenhafte Festlegung einer a-priori-Dichte π für das wahre Ausmaß θ der Alternativhypothese anhand des bestmöglichen Vorwissens an. Auch das SaFD verlangt hierbei die größte Sorgfalt. Es nimmt keine Arbeit ab. Die gegenteilige Auffassung läuft dem ethischen Gebot und dem eigenen Interesse eines geringen Ressourcenaufwands zuwider.

Es gibt eine Ausnahme, die nachträgliche Änderungen der a-priori-Dichte erlaubt. Sie betrifft das Vorliegen neuer Informationen, die nicht den gewonnenen Beobachtungen der eigenen Studie entstammen. Eine Korrektur der a-priori-Dichte kann höchstens den zukünftigen Studienverlauf beeinflussen. Selbst hierfür lassen sich unter Umständen nicht alle gebotenen Veränderungen effektiv umsetzen. Das verdeutlicht die Notwendigkeit, die a-priori-Dichte zu Beginn der Planungsphase so wahrheitsgetreu wie möglich festzulegen. Stellt sich ein Studiendesign als ungeeignet heraus, droht der Studienabbruch.

3. Der nächste Schritt der Planungsphase umfasst die Definition einer Verlustfunktion (vgl. Kapitel 3.2). Sie bewertet einen Entscheidungsprozess. Folglich ist die Optimierung mehrerer Verlustfunktionen mit anschließender Auswahl des “besten“ Ergebnisses sinnentleert. Solch ein Vorgehen zeugt von mangelndem Verständnis für die Bewertungen, die die Verlustfunktionen ausdrücken.

4. Die Verlustfunktion dient der Findung des besten Kompromisses. Manche Komponenten eines Studienplans sind aber unumstößlich. Hieraus ergeben sich Nebenbedingungen für die Optimierung der Entscheidungsregel. Zum Beispiel schränkt eine technisch notwendige Block-Randomisierung die möglichen Studienverläufe ein. Studienverläufe sind ausgeschlossen, wenn die dafür erforderlichen Entscheidungen durch leere Entscheidungsbereiche nicht getroffen werden können (vgl. Kapitel 3.1). Die Festlegung leerer Entscheidungsbereiche geht mit der Gleichsetzung der zugehörigen Entscheidungsgrenzen einher. Damit entfallen Variablen für die optimale Gestaltung einer Entscheidungsregel.

Allgegenwärtige Nebenbedingungen sind die Einhaltung des Signifikanzniveaus und die aufsteigende Ordnung der Entscheidungsgrenzen.

5. An diesem Punkt sollte das Optimierungsproblem, auch durch die Eigenschaften der Daten, vollständig beschrieben sein. Gesucht ist die Kombination aus denjenigen Entscheidungsgrenzen (vgl. Kapitel 3.1) und derjenigen Fallzahl der ersten Auswertung, die unter allen Kombinationen mit erfüllten Nebenbedingungen aus Punkt 4 das kleinste Bayes-Risiko aufweisen, das die a-priori-Dichte aus Punkt 2 und die Verlustfunktion aus Punkt 3 definieren. Wenn ihre Existenz sichergestellt ist, kann mit der Auffindung begonnen werden.

Aufgrund der hohen Komplexität des Optimierungsproblems liegt eine numerische Optimierung nahe. Doch auch ihrer Anwendung steht eine hohe (Rechen-)Komplexität im Wege. Deshalb wäre zu überlegen, ◀ inwieweit die möglichen Fallzahlen und die Anzahl der Auswertungen reduziert werden können. Die Umsetzung eines SaFDs ist ebenso wenig wie andere Verfahren davon ausgenommen, dass Sponsoren eine Finanzierung einer Studie mit immens vielen Teilnehmern nicht unter-

stützen (können). Daraus ergeben sich Einschränkungen. Sie dürfen nicht in die Definition der Verlustfunktion eingehen, weil die ethische Begründung dafür fehlt. Wenn die finanzierbaren Optionen nicht auszureichen scheinen (vgl. Kapitel 4.4), darf die Studie nicht realisiert werden. Trotz der Einschränkungen kann das Optimierungsproblem zu viele Variablen beinhalten. Möglicherweise genügt die Auswahl einiger Variablen, mit denen die anderen durch Interpolation beschrieben werden.

Für die numerische Optimierung muss ein Algorithmus ausgewählt werden. Zukünftige Forschungsarbeiten könnten sich damit befassen, welcher Algorithmus sich für welche, maßgeblich durch Punkt 1 definierten, Situationen anbietet und welche spezifischen Probleme auftreten. Kommt im Rahmen der Umsetzung ein Optimierungsalgorithmus in Frage, der unter Umständen mit Lösungen arbeitet, die die Nebenbedingungen verletzen, muss die Definition der Entscheidungsregel (vgl. Kapitel 3.1) erweitert werden. Sie muss dann auch für Entscheidungsgrenzen ohne aufsteigender Ordnung definiert sein. Ansonsten kann die numerische Optimierung anhand des ausgewählten Algorithmus durchgeführt und mit Punkt 6 auf Seite 34 fortgefahren werden. ◀

In der Formulierung der Entscheidungsregel aus Kapitel 3.1 definieren nicht-aufsteigend geordnete Entscheidungsgrenzen überschneidende Entscheidungsbereiche. Fällt der Wert der Statistik in solch eine Schnittmenge, gibt es keine eindeutige Entscheidung. Zur Veranschaulichung werde die Entscheidungsregel aus Tabelle 1 auf Seite 22 für eine Fallzahlüberprüfung mit 40 Beobachtungen mit konkreten Grenzen versehen (vgl. Tabelle 2). Liegt die beobachtete Teststatistik im Bereich (1,0; 1,3], konkurrieren demnach zwei Optionen. Bei Werten bis zu 1,7 muss die Nullhypothese vorzeitig angenommen werden. Gleichzeitig ist die Testung mit unumstößlich 80 Beobachtungen bei

Schätzungen zwischen 1,0 und einschließlich 2,0 vorgeschrieben. Die bisherigen Kriterien bilden im Falle nicht-disjunkter Entscheidungsbereiche keine vollständige Entscheidungsregel.

	Entscheidungskriterien			Entscheidung		
e	Aktuelle Fallzahl	als finale Fallzahl	Untere Intervallgrenze	Zukünftige Fallzahl	als finale Fallzahl	Testentscheidung
1	40		$-\infty$	40	$*$	H_0
2	40		1,7	60		
3	40		1,0	80	$*$	
4	40		2,0	60	$*$	
5	40		1,3	40	$*$	H_1

Tabelle 2: Zahlenbeispiel für den Ausschnitt der in Tabelle 1 auf Seite 22 dargestellten Entscheidungsregel ($n_j \in \{20, 40, 60, 80\}$) mit nichtdisjunkten Entscheidungsbereichen. Der Index $e \in \mathcal{E}_{40} = \{1, ..., 2 \cdot \#\{j : n_j > 40\} + 1\} = \{1, ..., 2 \cdot \#\{60, 80\} + 1\} = \{1, ..., 5\}$ bezieht sich auf die Handlungsoptionen bei der Fallzahlüberprüfung mit $n_i = 40$ Beobachtungen pro Behandlung.

Aus den Überlegungen in Kapitel 3.1 resultierten aber keine anderen als die gegebenen Entscheidungskriterien. Deshalb entscheide im Zweifelsfall der Zufall. Demnach hängt die Entscheidung über das Studienende mit der Annahme der Nullhypothese oder die Testung mit 80 Beobachtungen im Beispiel in Tabelle 2 von einem Münzwurf ab, wenn die beobachtete Teststatistik im Bereich (1,0; 1,3] liegt. Solch eine randomisierte Entscheidungsregel ist aus inhaltlicher Sicht abzulehnen. Sie diene nur als technischer Behelf für die Findung einer Entscheidungsregel mit aufsteigend geordneten Entscheidungsgrenzen mittels Optimierung unter der Nebenbedingung der Ordnung.

Die Einbindung des Zufalls erfolge über eine Zufallsvariable für jede Fallzahlüberprüfung, S_{n_i}. Sie ersetze die Teststatistik als direk-

tes Entscheidungskriterium. Die Teststatistik T_{n_i} bestimme aber die diskrete Verteilung der Zufallsvariablen S_{n_i}. Im Beispiel führt die Realisierung der Teststatistik T_{40} im Bereich (1,0; 1,3] dazu, dass die vorzeitige Annahme der Nullhypothese oder die Testung nach 80 Beobachtungen jeweils mit 50 %-iger und alle anderen Optionen mit nullprozentiger Wahrscheinlichkeit gewählt werden. Gilt stattdessen $T_{40} \in (-\infty; 1{,}0]$, fällt die Entscheidung eindeutig für die vorzeitige Annahme der Nullhypothese. Die Verteilung von S_{40} ist dann nicht wie zuvor durch den Wahrscheinlichkeitsvektor $(0{,}5; 0; 0{,}5; 0; 0)$, sondern durch $(1, 0, 0, 0, 0)$ gegeben.

Die Wahrscheinlichkeit für die Entscheidung mit dem Index $e \in \mathcal{E}_{n_i} = \{1, ..., 2 \cdot \#\{j : n_j > n_i\} + 1\}$ bei der überprüfenden Auswertung mit n_i Beobachtungen hänge also zum einen davon ab, ob die Teststatistik T_{n_i} in den zur Entscheidung gehörenden Entscheidungsbereich $\mathcal{C}_{e|n_i}$ fällt. Nur dann sei die Wahrscheinlichkeit größer als Null. Ihre Höhe hänge wiederum davon ab, in wie vielen Bereichen der Wert der Teststatistik liegt. Die konkurrierenden Entscheidungen haben alle die gleiche Wahrscheinlichkeit gewählt zu werden. Demnach gibt

$$p_{e,n_i}(T_{n_i}) = \frac{\mathbb{I}_{\mathcal{C}_{e|n_i}}(T_{n_i})}{\sum_{\epsilon \in \mathcal{E}_{n_i}} \mathbb{I}_{C_{\epsilon|n_i}}(T_{n_i})}$$

die Auswahlwahrscheinlichkeit der Entscheidung e gegeben der Teststatistik T_{n_i} an. Dabei dient die Indikatorfunktion

$$\mathbb{I}_{\mathcal{Q}}(q) = \begin{cases} 1, & \text{falls } q \in \mathcal{Q} \\ 0, & \text{falls } q \notin \mathcal{Q} \end{cases}$$

einerseits der Fallunterscheidung und andererseits dem Zählen der konkurrierenden Entscheidungen. Im Beispiel in Tabelle 2 liegt die Auswahlwahrscheinlichkeit für das sofortige Studienende mit Ablehnung der Nullhypothese nach 40 Beobachtungen, $e = 5$, im Falle von

$T_{40} = 2$ bei

$$p_{5,40}(2) = \frac{\mathbb{I}_{(1,3;\infty)}(2)}{\mathbb{I}_{(-\infty;1,7]}(2)+\mathbb{I}_{(1,7;1,0]}(2)+\mathbb{I}_{(1,0;2,0]}(2)+\mathbb{I}_{(2,0;1,3]}(2)+\mathbb{I}_{(1,3;\infty)}(2)} = \frac{1}{1+0+1+0+1} = \frac{1}{3}.$$

Die Erweiterung der Definition einer Entscheidungsregel für Entscheidungsgrenzen beliebiger Ordnung durch die Nutzung anderer Zufallsvariablen als Entscheidungskriterium hat ein abgewandeltes Optimierungsproblem zur Folge. Die Wahrscheinlichkeiten der Entscheidungspfade folgen der Verteilung $\mathbb{P}^{(S_{n_i}, i\in\mathbb{N})}$. Sie ergibt sich durch

$$\mathbb{P}^{(S_{n_i}, i\in\mathbb{N})} = \mathbb{P}^{(T_{n_i}, i\in\mathbb{N})} \prod_{j\in\mathbb{N}} \mathbb{P}^{S_{n_j}|(T_{n_i}, i\in\mathbb{N})}.$$

Das Gesetz der totalen Wahrscheinlichkeit ermöglicht die Nutzung der bedingten stochastischen Unabhängigkeit der Zufallsvariablen S_{n_i} gegeben der verteilungsrelevanten Teststatistiken T_{n_i}.

6. Nachdem die Studienplanung abgeschlossen ist, wird die Studie streng gemäß der geplanten Entscheidungsregel durchgeführt. Hierfür können in der Planungsphase Strategien erdacht werden, indem die Konsequenzen von Abweichungen von der optimalen Entscheidungsregel begutachtet werden (vgl. Seite 50).

7. Zum Studienende wird das wahre Ausmaß geschätzt. Die üblichen Schätzer sind für Stichproben mit festem Umfang konstruiert. Das vorgeschlagene Konzept nutzt dagegen dieselben Informationen zufälliger Beobachtungen bezüglich des wahren Ausmaßes zur Bestimmung der Fallzahl wie für die Schätzung. Daher gewährleistet ein erwartungstreuer Schätzer für datenunabhängige Fallzahlen keine erwartungstreuen Schätzungen im Fall von Fallzahlanpassungen ⟨16⟩. In der Literatur wird vorgeschlagen, trotzdem an den Schätzern festzuhalten und dann Korrekturmaßnahmen durchzuführen ⟨19⟩, mit

hohem numerischen Aufwand zu Rao-Blackwellisieren ⟨10⟩ oder auf mediantreue Schätzer auszuweichen ⟨1⟩. Mein Vorschlag ist dagegen allgemein und leicht umsetzbar. Der Studienplan schreibt immer eine Mindestfallzahl mindestens in Höhe der Fallzahl für die erste Studienphase vor. Sie hängt von keinen Beobachtungen ab. Am Studienende werde nun eine Vielzahl entsprechend großer Stichproben aus den vorliegenden Daten gezogen. Aus jeder Stichprobe ergibt sich mit einem erwartungstreuen Schätzer eine erwartungstreue Schätzung. Das arithmetische Mittel dieser erwartungstreuen Schätzer ist wegen der Linearität des Erwartungswerts selbst ein erwartungstreuer Schätzer, wenn die Anzahl der Stichproben datenunabhängig ist.

Die nähere Auseinandersetzung mit Schätzern samt ihren zugehörigen Konfidenzintervalle wird fortführenden Forschungen angetragen. Von besonderem Interesse ist die Gegenüberstellung des vorgeschlagenen Schätzers mit dem Rao-Blackwellisierten Schätzer hinsichtlich der Varianzen. Es ist zu klären, ob die vereinfachte Schätzung die damit verbundene Erhöhung der Varianz gegenüber der kleinsten Varianz unter allen sogenannten truncation-adaptable unbiased estimators (für den einzigen Parameter einer Verteilung aus der Exponentialfamilie), der des Rao-Blackwellisierten Schätzers ⟨10⟩, rechtfertigt. Davon abgesehen ließe sich zudem untersuchen, ob der Einsatz des vorgeschlagenen Schätzers die Leistung der adaptiven Designs nach dem CRP-Prinzip (vgl. Kapitel 2.3), wie zum Beispiel die von Hartung genannten Designs ⟨8⟩, verbessert. ◀

4 Verlustfunktionen zur Beurteilung flexibler Designs

4.1 Publizierte Verlustfunktionen

Kapitel 3.2 regte ein Grundgerüst für Verlustfunktionen zur Optimierung flexibler Designs an, auf dem die an selber Stelle geforderten Individualisierungen aufbauen könnten. Im Jahr 2016 listeten Zhang, Cui und Yang publizierte Funktionen zur Bewertung flexibler Designs auf ⟨21⟩. Keine davon berücksichtigt die benötigte Anzahl der Auswertungen.

Es ist kein Ersatz, wie in ⟨11⟩ die Funktionswerte von Designs mit der gleichen zugelassenen maximalen Auswertungszahl zu vergleichen. Zum Beispiel benötigt ein gruppensequenzieller Test nach Pocock (vgl. Kapitel 2.3) mit maximal 20 Studienphasen mit je zwei Probanden pro Behandlungsgruppe bei einem wahren Ausmaß von $\theta = 1$ durchschnittlich 7,64 Auswertungen. Jedes dritte Studienende ereignet sich nach mehr als neun Studienphasen. Ein adaptives Design nach dem CRP-Prinzip (vgl. Kapitel 2.3) wie auf Seite 43 nimmt dagegen von 100 maximal möglichen Auswertungen durchschnittlich 3,72 an, wenn mit einer Fallzahl von zwei Probanden pro Gruppe mit einem Gewicht von $\epsilon_1 = \frac{1}{11}$ begonnen wird und höchstens 5000 Probanden aufgenommen werden dürfen. Mehr als neun Studienphasen sind unwahrscheinlicher als 0,1 %. Die Argumentation mit der maximal möglichen Auswertungszahl von 100 ist hier realitätsfern. Die Maximalzahl gibt, wie gezeigt, ohnehin nur eingeschränkt die Befähigung für die breite Anwendung wieder, die für eine Etablierung eines Verfahrens relevant ist. Aber auch selbst wenn gleiche zugelassene maximale Auswertungszahlen einen Vergleich ansonsten verschiedener Designs zuließen, bliebe bei solch einem internen Vergleich

offen, welches der demnach besten Designs jeweils verschiedener maximaler Auswertungszahlen vorzuziehen wäre.

Hinzu kommen weitere Aspekte publizierter Bewertungskriterien, die den Optimierungsprozess fehlleiten. Zum Beispiel bezieht sich der sogenannte Average Performance Score

$$\mathrm{APS}_{\mathrm{ZCY}} = \int \pi(\theta) \left| \frac{\mathbb{E}_\theta(N)}{f(\theta;\alpha,\beta)} - 1 \right| d\theta \tag{4}$$

von Zhang, Cui und Yang, ein Bayes-Risiko mit der a-priori-Dichte π für das wahre Ausmaß θ, ausschließlich auf die durchschnittlich eingesetzte Gesamtfallzahl $\mathbb{E}_\theta(N)$ ⟨21⟩. Dem liegt folgende Unterstellung zu Grunde: „A design with an average sample size curve matching [the] gold standard will likely produce the targeted power with a sound average sample size [...]“. Mit dem Goldstandard ist die (nicht-zwingend realisierbare) Fallzahl $f(\theta;\alpha,\beta)$ für eine gerade eben valide einmalige Testung ohne Zwischenauswertungen zum Signifikanzniveau α gemeint. Nach diesem Score (4) gelten Entscheidungsregeln mit der gleichen erwarteten Fallzahl wie bei der Referenz mit $[1-\beta] = 90$ %-iger Power als optimal, ohne dass die Power wenigstens einmal 90 % erreichen muss. Darüber hinaus wird eine Entscheidungsregel negativ bewertet, die gegenüber der Referenz Probanden ohne Einbußen bei der Power spart.

Die Kritikpunkte treffen ebenfalls die Sample Size Discrepancy

$$\mathrm{SSD} = \int \pi(\theta) \left| \mathbb{E}_\theta(N) - f(\theta;\alpha,\beta) \right| d\theta$$

von Wu und Cui ⟨20⟩. Der einzige Unterschied zum Average Perfomance Score (4) besteht in der missachteten Abhängigkeit der Bedeutung einer Fallzahlabweichung vom wahren Ausmaß (vgl. Kapitel 3.1). Dadurch bestimmen die kleineren potenziell möglichen Ausmaße den Verlust stärker als von der a-priori-Dichte vorgesehen. Das Defizit findet sich auch beispielsweise bei der Average Sample Size

$$\mathrm{ASS} = \int \pi(\theta) \mathbb{E}_\theta(N)\, d\theta$$

von Jennison und Turnbull ⟨9⟩. Hiermit werden immerhin kleine Fallzahlen als Vorteil gewertet. Beiden Arbeiten kann im Gegensatz zu der von Zhang, Cui und Yang zugesprochen werden, dass die Power als Nebenbedingung wenigstens nicht gänzlich ignoriert wird. Wu und Cui genügt, dass die Ablehnungswahrscheinlichkeiten eines bestimmten Anteils der potenziell wahren Ausmaße oberhalb eines (um 10 Prozentpunkte) kleineren als den für die Validität geforderten Werts liegen. Das erzwingt keinen einzigen validen Schluss. Dagegen verlangen Jennison und Turnbull zwar Validität, aber nur für das wahre Ausmaß. Wäre es bekannt, interessierte die Bewertung anderer Ausmaße nicht. Aber auch die Forderung eines gewissen Prozentsatzes valider Schlüsse wäre nicht zufriedenstellend. Darin liegt der Widerspruch, die gewählten Ausmaße gemäß ihrer Möglichkeit wahr zu sein, einbeziehen zu wollen beziehungsweise zu müssen, aber dann doch einen Teil selektiv bezüglich der Validität zu ignorieren. Außerdem läuft die Forderung eines zunehmend großen Prozentsatzes von potenziell wahren Ausmaßen mit einer Power oberhalb einer festgelegten Grenze der gebotenen Ressourceneffizienz zuwider. Das brachte in Kapitel 3.2 den Kompromiss des überreichlichen Ressourcenverbrauchs aus zwei verschiedenen Quellen zur Sprache.

Am ehesten scheint der Average Performance Score (APS)

$$\mathrm{APS} = 100 \int \pi(\theta) \left[\frac{\mathbb{I}_{(f(\theta;\alpha,\beta),\infty)}\big(\mathbb{E}_\theta(N)\big)}{o-1} \left[\frac{\mathbb{E}_\theta(N)}{f(\theta;\alpha,\beta)} - 1 \right] + \frac{\mathbb{I}_{(-\infty,f(\theta;\alpha,\beta))}\big(f(\theta;\alpha,\beta_\theta)\big)\big[f(\theta;\alpha,\beta) - f(\theta;\alpha,\beta_\theta)\big]}{f(\theta;\alpha,\beta) - f(\theta;\alpha,1-[1-u]\,[1-\beta])} \right] d\theta, \tag{5}$$

von Liu, Zhu und Cui einen Kompromiss abzubilden ⟨11⟩. Der erste Summand bewertet eine über die Fallzahl der Referenz hinausgehende benötigte durchschnittliche Teilnehmerzahl als Verlust, $\mathbb{E}_\theta(N) > f(\theta;\alpha,\beta)$. Der zweite Summand bestraft zu niedrige erreichte Ablehnungswahrscheinlichkeiten, $[1-\beta_\theta] < [1-\beta]$ ($\Leftrightarrow f(\theta;\alpha,\beta_\theta) < f(\theta;\alpha,\beta)$). Die jeweils anders gerichtete Abweichung neutralisiert die auf Seite 33 definierte Indi-

katorfunktion $\mathbb{I}$. Damit endet das Streben nach geringen Teilnehmerzahlen bei der Fallzahl der Referenz. Die Menge der Entscheidungsregeln, die mit (unterschiedlich) niedrigeren Fallzahlen für eine ausreichende Power als gleichwertig deklariert werden, hängt von der Wahl der Referenz ab.

Die Übersetzung der erreichten Power $[1-\beta_\theta]$ in eine Fallzahl $f(\theta;\alpha,\beta_\theta)$ zur Bewertung ihrer Verhältnismäßigkeit bezüglich der tatsächlich dafür benötigten Fallzahl schlugen Jennison und Turnbull bereits zwei Jahre zuvor vor ⟨9⟩. Ihr Efficiency Index $\frac{f(\theta;\alpha,\beta_\theta)}{\mathbb{E}_\theta(N)}$ bewertet aber eben nur die Verhältnismäßigkeit. Die Wahl der als Referenz fungierenden Entscheidungsregel mit einer $[1-\beta_\theta] = 20$ %-igen Power wird mit einem Efficiency Index von Eins als ebenso gut befunden wie mit einer $[1-\beta_\theta] = 90$ %-igen Power ($f(\theta;\alpha,\beta_\theta) = \mathbb{E}_\theta(N)$).

Der APS (5) scheint bislang die einzige Funktion zu sein, die mit der Übersetzung der erreichten Power in eine Fallzahl zur Optimierung flexibler Designs an den Efficiency Index anknüpft. Angesichts der in der Bedeutung übereinstimmenden Bewertungen einer zu hohen Fallzahl und einer zu niedrigen Power stellt sich aber die Frage, warum der APS die Gewichtung der beiden Aspekte auf verschiedene Weise vornimmt. Die Bezugsgröße für die im Mittel zusätzlich benötigte Fallzahl gegenüber der Referenz ist

$$o \cdot f(\theta;\alpha,\beta) - f(\theta;\alpha,\beta), \quad o > 1.$$

Dagegen wird die Fallzahldifferenz, die den Power-Verlust beschreibt, durch

$$f(\theta;\alpha,\beta) - f(\theta;\alpha,1-[1-u]\,[1-\beta]), \quad u \in (0,1),$$

relativiert. Einmal erfolgt die Gewichtung also mit einem subjektiv festgelegten Faktor o außerhalb, das andere Mal mit einem anderen beliebigen Faktor u innerhalb der Funktion f, die die Fallzahl der Referenz beschreibt. Das steht der Anwendung der APS aber nicht in dem Maße

wie die Vernachlässigungen der Auswertungszahl sowie etwaiger Fallzahlersparnisse entgegen.

4.2 EINFÜHRUNG DER RELATIVE ADDITIONAL COSTS FOR VALIDITY

Die im vorangegangenen Kapitel 4.1 aufgezeigten Schwachstellen der recherchierten Verlustfunktionen stehen der Eignung der Funktionen zur Optimierung flexibler Designs entgegen. Insbesondere deshalb wird hier eine neue Verlustfunktion

$$v_\theta = \frac{w_\theta \cdot g\left(\mathbb{E}_\theta(N), \mathbb{E}_\theta(K)\right) - g\left(f\left(\theta;\, \min\left\{1-(1-\alpha)^{w_\theta}, 1-\beta\right\}, \beta\right),\, 1\right)}{f\left(\theta;\, \min\left\{1-(1-\alpha)^{w_\theta}, 1-\beta\right\}, \beta\right)}, \tag{6}$$

$$w_\theta = \max\left\{1, \frac{\log(\beta)}{\log(\beta_\theta)}\right\}$$

eingeführt. Sie erfüllt die Forderung aus Kapitel 3.2, eine Entscheidungsregel mit einem validen Schluss für das (potenziell) wahre Ausmaß θ anhand des Ressourcenverbrauchs g, nicht aber anhand der Höhe der ausreichenden Power zu bewerten. Gleichgültig der Höhe, mit der die erreichte Power $[1-\beta_\theta]$ die angestrebte Power $[1-\beta]$ einhält $(1-\beta_\theta \geq 1-\beta)$, gilt $w_\theta = 1$. Das ist die einzige Stelle, an der die Power in die Verlustfunktion (6) eingeht. Mit dem Wert von Eins ist w_θ in dem Sinne ein neutrales Element, als dass allein die nachfolgend erläuterte Bewertung des Ressourcenverbrauchs den Verlust bestimmt.

Sämtliche Aufwendungen für die Studie mit der zu begutachtenden Entscheidungsregel seien durch die Funktion g mit positivem Wertebereich beschrieben. Sie hänge im Interesse der Etablierung der Entscheidungsregel von der durchschnittliche Fallzahl $\mathbb{E}_\theta(N)$ und von der durchschnittlichen Zahl der Auswertungen $\mathbb{E}_\theta(K)$ ab. Die Zusammenführung der Kosten bildet die individuelle Komponente der Verlustfunktion (vgl. Kapitel 3.2) und ist deshalb bewusst nicht näher spezifiziert. Sie kann

unabhängig vom (potenziell) wahren Ausmaß θ definiert werden und so die faktischen Gesamtkosten ausdrücken.

Die Berücksichtigung des wahren Ausmaßes bei der Bewertung der Kosten obliegt der Verlustfunktion. Die Ablehnung der Nullhypothese benötigt im Allgemeinen weniger Ressourcen, je größer der wahre Behandlungseffekt ist (vgl. Kapitel 3.1). Die Bedeutung eines Verlusts muss aber für alle Ausmaße vergleichbar sein, damit die jeweiligen Verluste gemeinsam ein taugliches Bayes-Risiko ergeben. Hierfür werde die (nichtzwingend realisierbare) Fallzahl für einen gerade eben validen Schluss mit der ressourcenschonendsten einmaligen Testung ohne Fallzahlüberprüfungen zur Normierung herangezogen. Das Signifikanzniveau α entspreche dem der zu bewertenden Entscheidungsregel. Der Ressourcenaufwand umfasse exakt eine Auswertung auf der Basis von $f(\theta; \alpha, \beta)$ Studienteilnehmern. Die Notation beinhaltet der Allgemeinheit wegen nur die grundlegenden Parameter für Fallzahlbestimmungen. Andere müssen bei Bedarf hinzukommen. Aus der Division der Kosten der zu bewertenden Entscheidungsregel mit denen der Referenzregel resultieren Verluste, deren Bedeutung für alle Ausmaße gleich ist. Die relativen Kosten können jedoch nun nicht mehr gegeneinander aufgerechnet werden, wie es für ein zielführendes Bayes-Risiko nötig wäre. Deshalb definiere die relative Kostendifferenz in (6) den Verlust. Er werde mit RACV (Relative Additional Costs for Validity) abgekürzt. Das Bayes-Risiko lässt sich als gewichtetes Mittel der referenzbezogenen relativen Differenzen der Kosten für valide Schlüsse interpretieren.

Die konzipierte Bewertung bezieht sich bislang auf Entscheidungsregeln mit validen Schlüssen. Dem Zugang für eine zu bewertende Entscheidungsregel ohne validen Schluss für das Ausmaß θ ($w_\theta > 1$) liege die Vorstellung zu Grunde, dass die Entscheidungsprozedur in w_θ separaten Studien zum Einsatz kommen muss, damit die Wahrscheinlichkeit

für mindestens eine Studie mit abgelehnter Nullhypothese 90 % beträgt ($1 - \beta_\theta^{w_\theta} = 1 - \beta$). Damit steigen die Kosten um den Faktor w_θ. Der mehrfache Einsatz hat eine Verletzung des globalen Signifikanzniveaus zur Folge (vgl. Kapitel 2.2). Dem wird Rechnung getragen, indem der als Referenz herangezogenen Entscheidungsregel eine garantiert ebenso starke Verletzung zugestanden wird. Dadurch genügt der Referenzregel eine niedrigere Fallzahl. Auf diese Weise wird ein fairer Vergleich hergestellt. Die Vorstellung unabhängiger Wiederholungen der Studie mit der betreffenden Entscheidungsregel folgt im Gegensatz zu einer kumulativen Replikation dem Gedanken, dass es nur einen einzelnen Versuch für eine Studie gibt (vgl. Kapitel 2). Die kumulative Replikation bewertet einen Verbund anstelle der eigentlichen Entscheidungsregel als Individuum.

4.3 Begutachtung der RACV

Zur Diskussion der in Kapitel 4.2 definierten Verlustfunktion (6) komme sie zur Optimierung publizierter Entscheidungsregeln für die auf Seite 16 beschriebene Situation für den Gauß-Tests zum Einsatz. Als a-priori-Verteilung des wahren Ausmaßes θ werde die stetige Gleichverteilung auf dem Intervall [0,3; 0,7] gewählt. Das Bayes-Risiko werde anhand der gemittelten Verluste der 41 Stützstellen für θ im Abstand von 0,01 approximiert. Die Gesamtkosten ergeben sich der Einfachheit halber durch

$$g = \mathbb{E}_\theta(N) + \nu\,\mathbb{E}_\theta(K).$$

Demnach beansprucht jede Auswertung Fixkosten im Wert von ν Probanden. In der Realität dürfte die erste Auswertung mehr als die anderen kosten.

Es werden ausschließlich Entscheidungsregeln mit maximal $\kappa = 10$ Auswertungen und einem Signifikanzniveau von 2,5 % betrachtet. Sie basieren beim gruppensequenziellen Test nach Pocock (vgl. Ka-

pitel 2.3) auf der gleichen Anzahl von Beobachtungen. Zur Wahl stehen maximale Gesamtfallzahlen pro Behandlungsgruppe von 10 bis 550 in Schritten der Größe 10. Von der frühzeitigen Annahme der Nullhypothese wird hier, wie im Originalartikel ⟨13⟩, abgesehen. Dafür wären zusätzliche Spezifizierungen nötig. Daher wird die Studie fortgesetzt, bis eine kumulative Teststatistik den zugehörigen kritischen Wert überschreitet oder die maximale Anzahl an Auswertungen κ erreicht ist.

Als adaptive Entscheidungsregel werde Hartungs Self-Design von 2006 herangezogen ⟨8⟩, das hier anhand des CRP-Prinzips nach Müller und Schäfer formuliert wird (vgl. Kapitel 2.3). Den Kern der Vorschrift für die Wahl der Fallzahl für die nächste Auswertung zur Fallzahlanpassung bildet bei Hartung die Formel

$$\epsilon_{K+1} = 0{,}5 + 0{,}5 \cdot \begin{cases} \frac{\alpha_{K-1}-\alpha_K}{\alpha_{K-1}}, & \text{falls } \alpha_{K-1} \geq \alpha_K \\ \frac{\alpha_{K-1}-\alpha_K}{1-\alpha_{K-1}}, & \text{falls } \alpha_{K-1} < \alpha_K \end{cases}, \qquad \epsilon_\kappa = 1.$$

Dabei ist α_K die CRP der aktuellen Studienphase $K \in \{1, ..., \kappa\}$, $\alpha_1 = \alpha = 0{,}025$. Die Fallzahl der nächsten Phase ergibt sich durch $m_{K+1,1} = \left\lceil \epsilon_{K+1} \cdot f\left(\hat{\theta}_K;\, \alpha_K, \beta\right) \right\rceil$. Durch $m_{K+1,1} > 0$ basiert die nächste Auswertung, die es nach einer Fallzahlüberprüfung mittels CRP-Prinzip im Gegensatz zu gruppensequenziellen Verfahren geben muss (vgl. Kapitel 2.3), auf mindestens einer Beobachtung. Der nächste ist gleichzeitig der letzte Studienabschnitt, falls

- die vorletzte Studienphase erreicht ist, $K = \kappa - 1$,
- die Beobachtungen ausreichend stark gegen die Nullhypothese sprechen, $m_{K+1,1} = m_{K+1} = \left\lceil f\left(\hat{\theta}_K;\, \alpha_K, \beta\right) \right\rceil$, oder
- mit der ausstehenden Fallzahl keine weitere Studienphase ohne Überschreitung der maximalen Gesamtfallzahl $\bar{m}$ möglich ist, $m_{K+1,1} \geq \bar{m} - \sum_{k=1}^{K} m_{k,1}$.

Bei der finalen Auswertung werden die letztmalig hinzugekommenen m_{K+1} Beobachtungen zum Signifikanzniveau von α_{K+1} getestet.

Die Optimierung bezieht sich auf Maximalfallzahlen $\bar{m}$ zwischen 50 und 550 in 10-er Schritten. Zudem stehen die Fallzahlen $m_{1,1} \in \{10, 15, ..., 100\}$ für die erste Fallzahlanpassung mit, in Analogie zum gruppensequenziellen Design, $\epsilon_1 = \frac{1}{\kappa}$ zur Wahl. Für jede der genannten Einstellungen, bestehend aus der maximalen Auswertungsanzahl κ, der Maximalfallzahl $\bar{m}$ und der Startfallzahl $m_{1,1}$, werden die drei für den Verlust relevanten Eigenschaften, die erreichte Power, die durchschnittliche Fallzahl und die durchschnittliche Anzahl der Auswertungen, bezüglich jedes der 41 Ausmaße θ in 100.000 Simulationsdurchläufen ermittelt.

Die Abbildungen 3 bis 6 stellen die genannten Designtypen mit ihren jeweils besten Einstellungen für Fixkosten von $\nu = 25$ Einheiten pro Auswertung gegenüber. Die Referenz, der perfekte Gauß-Test mit der stets angemessenen Fallzahl, ist mit grauen Punkten eingezeichnet. Für die a-priori-Verteilung $\mathcal{R}[0{,}3;\ 0{,}7]$ sind für den Gauß-Test 81 Probanden pro Behandlungsgruppe RACV-optimal. Die Existenz einer optimalen Fallzahl zeigt, dass mit den RACV mehr als nur die Verhältnismäßigkeit bezüglich der Power, die der Gauß-Test als Referenz zwangsläufig erfüllt, bewertet wird. Mit 81 Studienteilnehmern wird beinahe für das mittlere der gleichermaßen potenziell wahren Ausmaße, und zwar für rund $\theta = 0{,}5094$ gepowert. Für solche Situationen müssen sich die Kostenfunktion und die a-priori-Dichte in spezieller Weise fügen. In Abbildung 3 ist die beabsichtigte asymmetrische Bewertung der Verluste invalider und valider Schlüsse trotz der annähernden Balance zu sehen. Bei höheren Fixkosten von $\nu = 75$ würden die invaliden Schlüsse stärker bestraft. Die optimale Menge der potenziellen Ausmaße mit zu kleiner Power reduzierte sich von $[0{,}3000; 0{,}5094)$ auf $[0{,}3000; 0{,}4703)$. Wenn eine Studie abseits der fallzahlbezogenen Kosten teuer ist, ist ein Scheitern vor dem Hintergrund konkurrierender Studien mit diesbezüglich niedrigeren Ansprüchen umso verlustreicher.

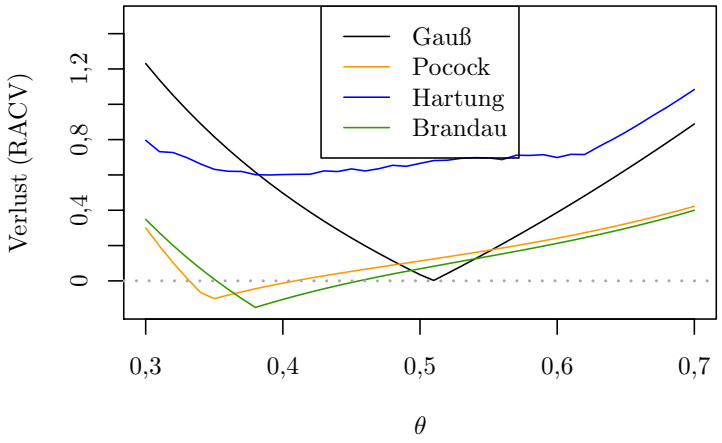

Abbildung 3: Einzelverluste von RACV-optimierten Designtypen für die a-priori-Verteilung $\mathcal{R}[0{,}3; 0{,}7]$ gegeben der Gesamtkostenfunktion $g = \mathbb{E}_\theta(N) + 25\,\mathbb{E}_\theta(K)$.

Der optimale Gauß-Test hat ein höheres Bayes-Risiko als das optimale gruppensequenzielle Design nach Pocock mit maximal viermal ($210/4 \approx$) 52 Probanden pro Behandlung ($0{,}48 > 0{,}15$). Er ist nur im Bereich von $(0{,}4847; 0{,}5524)$ besser, in dem die Power absolut um weniger als rund vier Prozentpunkten von der geforderten Power von 90 % abweicht (vgl. Abbildungen 3 und 4).

Die Verlustfunktion spiegelt die Eigenschaften gruppensequenzieller Tests gut wider. Durch die Möglichkeiten für eine frühzeitige Ablehnung der Nullhypothese können im Durchschnitt Probanden gespart werden (vgl. Abbildung 5). Zufällig größere Abweichungen von einem wahren Wert werden als entsprechend großer Effekt gedeutet. Die Ersparnis bedeutet eine vorteilhafte Abweichung von der Verhältnismäßigkeit. Der Vorteil des nicht-selektiven Mechanismus äußert sich in einer geringeren Steigung der Einzelverluste in Abbildung 3 im Vergleich zum Gauß-Test. Die Ersparnis ist im Bereich $(0{,}3300; 0{,}4081)$ mit dem bestmöglich gepowerten Ausmaß $\theta = 0{,}3441$ (vgl. Abbildung 4) sogar sichtlich so groß,

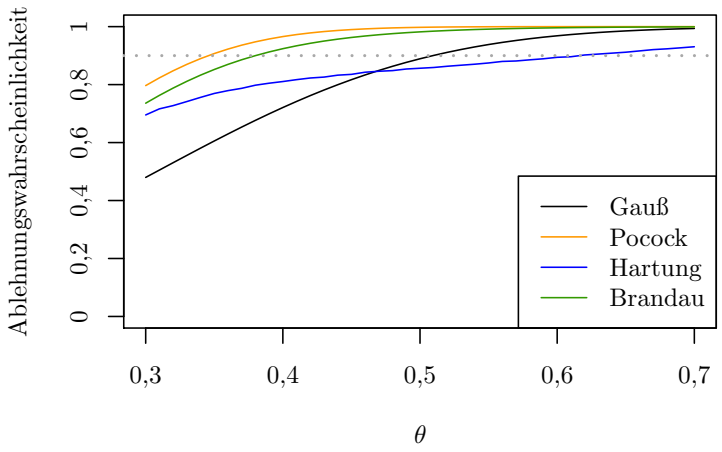

Abbildung 4: Power von RACV-optimierten Designtypen für die a-priori-Verteilung $\mathcal{R}[0{,}3; 0{,}7]$ gegeben der Gesamtkostenfunktion $g = \mathbb{E}_\theta(N) + 25\,\mathbb{E}_\theta(K)$.

dass trotz der Kosten der zusätzlichen Zwischenauswertungen ein Gewinn gegenüber dem perfekten Gauß-Test vorliegt.

Je kleiner das wahre Ausmaß hingegen ist, umso seltener kommt es zu frühzeitigen Ablehnungen (vgl. Abbildung 6). Mit zunehmender Nutzlosigkeit der Zwischenauswertungen wird deutlich, dass sie zulasten der Power gehen. Der Power-Verlust kommt zu den Fixkosten für die vergeblichen Zwischenauswertungen hinzu. Es würde besser direkt die gesamte Fallzahl in einem Gauß-Test eingesetzt. Das Kosten-Nutzen-Verhältnis verschlechtert sich durch die Doppelbelastung zusehends (vgl. Abbildung 3). Der optimale gruppensequenzielle Test gibt dem folglich nur wenig Raum und nutzt umso mehr seine Stärke bei großen Ausmaßen. Dadurch sind 89 % der potenziell wahren Ausmaße mit validen Schlüssen verbunden (vgl. Abbildung 4).

Bei adaptiven Designs nach dem CRP-Prinzip sind zwei Auswertungen unvermeidlich. Das ist vor allem ein Nachteil, wenn die Startfallzahl für das wahre Ausmaß ausreicht oder zu groß ist. Es muss ein Mini-

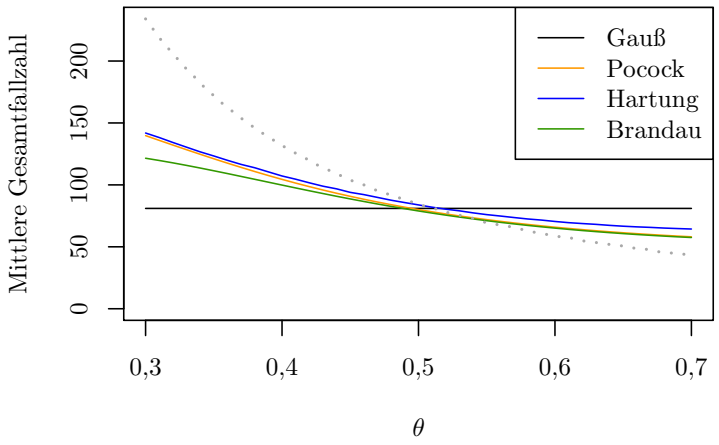

Abbildung 5: Durchschnittliche Gesamtfallzahl von RACV-optimierten Designtypen für die a-priori-Verteilung $\mathcal{R}[0{,}3; 0{,}7]$ gegeben der Gesamtkostenfunktion $g = \mathbb{E}_\theta(N) + 25\,\mathbb{E}_\theta(K)$.

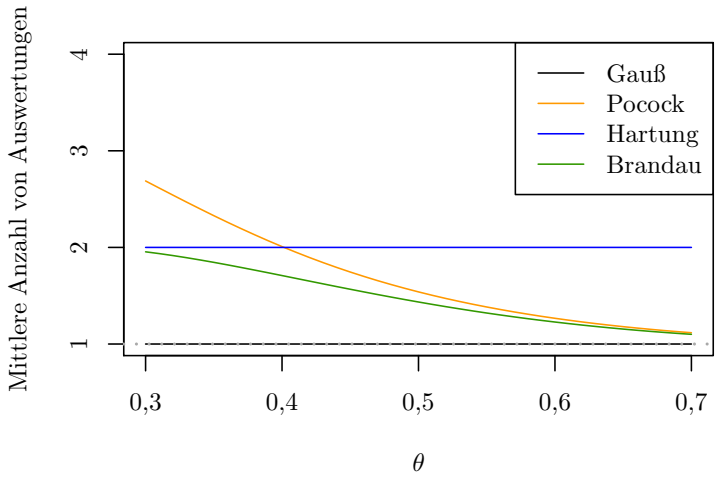

Abbildung 6: Durchschnittliche Anzahl von Auswertungen von RACV-optimierten Designtypen für die a-priori-Verteilung $\mathcal{R}[0{,}3; 0{,}7]$ gegeben der Gesamtkostenfunktion $g = \mathbb{E}_\theta(N) + 25\,\mathbb{E}_\theta(K)$.

mum an weiteren Studienteilnehmern beobachtet werden. Schwerer dürften aber die Kosten der generellen zusätzlichen Auswertung wiegen. Das zeigt sich in den Optimierungsergebnissen. Die durchschnittlichen Fallzahlen des optimalen adaptiven Designs nach Hartung ähneln denen des optimalen gruppensequenziellen Designs nach Pocock (vgl. Abbildung 5). Mit ihnen sind niedrige Verluste möglich. Die Verluste sind aber über den gesamten Bereich der möglichen Ausmaße erhöht (vgl. Abbildung 3) und ergeben ein Bayes-Risiko von 0,72. Die langsam zunehmende Power ist zwar erkennbar, setzt sich aber kaum durch (vgl. Abbildung 4). Es dominieren die Kosten der durchschnittlichen Anzahl der Auswertungen. Im optimalen Design müssen und können nur zwei Auswertungen durchgeführt werden (vgl. Abbildung 6). Damit sollen noch höhere Auswertungszahlen verhindert werden. Die Entscheidungsregel von Hartung scheint sogar für das größtmögliche Ausmaß $\theta = 0{,}7$ auf etwas über vier Auswertungen im Mittel hinauszulaufen. In Abbildung 7 ist eine dahingehende Stabilisierung bei zunehmenden maximalen Auswertungszahlen zu erkennen.

Folglich läuft das adaptive Design nach Hartung dem unflexiblen Gauß-Test höchstens bei niedrigen Auswertungskosten den Rang ab. Mit Fixkosten von $\nu = 5$ beträgt das Bayes-Risiko des optimalen adaptiven Designs mit einer Startfallzahl von 20 Probanden und maximal $\bar{m}$ =170 Probanden pro Behandlungsgruppe in höchstens $\kappa = 5$ Auswertungen 0,29. Das Bayes-Risiko des optimalen Gauß-Tests mit 72 Studienteilnehmern pro Behandlungsgruppe beläuft sich dagegen auf 0,40. Ohne den dominanten Kosten für die Auswertungen zeigt auch der kleinste Verlust des adaptiven Designs (-0,23) circa bei $\theta = 0{,}35$ den Übergang der Power zu den geforderten 90 % an. Das Verhalten ließ sich bereits bei den nicht-adaptiven Designs in Abbildung 3 und in Abbildung 4 beobachten. Damit legt die Begutachtung der RACV insgesamt deren Eignung nahe.

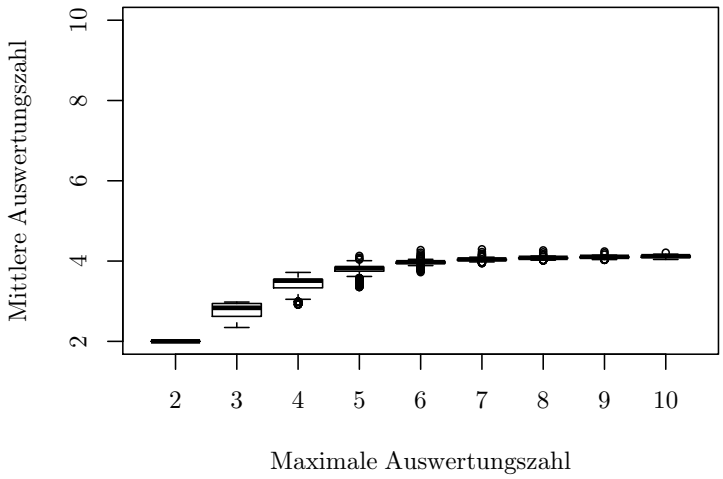

Abbildung 7: Boxplots der durchschnittlichen Anzahl von Auswertungen im Falle von $\theta = 0{,}7$ von allen für die Optimierung evaluierten adaptiven Designs nach Hartung in Abhängigkeit von der maximalen Auswertungszahl.

Wegen der Eignung rechtfertigt ein niedriges Bayes-Risiko die Nutzung eines gruppensequenziellen Designs, weil sich die Nachteile des Designs auf die RACV auswirken (vgl. Kapitel 2.3). Der Drang nach einer Power von 100 % führt zu größeren als den jeweils angemessenen Fallzahlen (vgl. Kapitel 5). Hingegen beseitigt ein geringes Bayes-Risiko nicht den Einwand bezüglich des CRP-Prinzips (vgl. Kapitel 2.3). Bei der Abwägung der Nutzung flexibler Designs ist zu beachten, dass das Bayes-Risiko die exakte Umsetzung wiedergibt. Die RACV lassen sich aber auch für eine Sensitivitätsanalyse nutzen. Dazu werde der optimale gruppensequenzielle Test nach Pocock mit gegebenen Fixkosten von $\nu = 25$ betrachtet. Solange mindestens 107 Teilnehmer pro Behandlungsgruppe für die Studie gewonnen werden können, ist das gruppensequenzielle Design, das eigentlich maximal $4 \cdot 52$ Probanden vorsieht, dem Gauß-Test vorzuziehen. Aufgrund der Höhe der Kosten für eine Auswertung darf jedoch höchstens ein Proband pro Behandlung überzählig sein, um den Vorteil gegenüber dem optimalen Gauß-Test zu bewahren. Kommt es bei einer

Zwischenauswertung zum Studienende, müssen etwaige nicht darin einbezogenen randomisierten Studienteilnehmer in einer weiteren Auswertung an der Hypothesenprüfung partizipiert werden ⟨6⟩. Wassmer bezeichnet dieses Szenario als „random overrunning“ ⟨18c⟩.

4.4 Begutachtung des neuen Studiendesigns anhand der RACV

Nachdem die RACV in Kapitel 4.3 anhand bekannter Entscheidungsregeln als geeignetes Optimierungskriterium befunden werden konnten, werden sie zur Begutachtung des in Kapitel 3 eingeführten SaFDs herangezogen. Die Optimierung selbst eines reduzierten SaFDs ist rechenintensiv. Zum Beispiel kann die numerische Berechnung des Bayes-Risikos eines SaFDs mit 22 verschiedenen möglichen Gesamtfallzahlen und maximal fünf Auswertungen mit nur 21 Stützstellen zwischen [0,3; 0,7] (mit einem Intel Xeon Prozessor E5-2630 mit 2,3 Gigahertz und 64 Gigabyte Random-Access Memory) mehr als 72 Stunden in Anspruch nehmen. Sie muss zur Optimierung der über 400 Variablen mehrfach durchgeführt werden. Hinzu kommen Berechnungen für die Nebenbedingungen sowie gegebenenfalls für Gradienten. Der hohe Rechenaufwand drängt zu Einschränkungen des SaFDs, um es für die Praxis zugänglicher zu machen. Damit gehen zunehmende Zweifel an der Optimalität des optimierten reduzierten SaFDs einher. Nichtsdestotrotz kann es besser als die bestehenden Strategien (aus Kapitel 2.1 und 2.3) sein. Selbst wenn dem nicht so ist, kann das SaFD als Verfahren vorerst nicht verworfen werden. Zukünftig wachsende Rechenkapazitäten ermöglichen SaFDs mit einer höheren Komplexität. Auf diesem Weg nahmen bereits andere statistische Methoden, wie nicht-parametrische Verfahren, erst später in ihrer Popularität zu.

Für einen direkten Vergleich mit dem besten Design aus Kapitel 4.3 werde das Grundgerüst des optimalen gruppensequenziellen Verfahrens nach Pocock in ein SaFD übertragen (vgl. Tabelle 3). Für die Optimierung mit der Software R werde eine Suche mit zufälligen Werten aus dem Intervall $[-3{,}5; 2{,}5]$ mit der mehrfachen Verwendung der Algorithmen Constrained Optimization by Linear Approximation (COBYLA) ⟨14⟩ beziehungsweise Method of Moving Asymptotes (MMA) ⟨17⟩ kombiniert.

Das SaFD in Tabelle 3 erreicht selbst in seiner stark vereinfachten Form ein geringeres Bayes-Risiko als das optimale gruppensequenzielle Design (0,12 < 0,15). Das Bayes-Risiko des optimalen uneingeschränkten SaFDs könnte noch kleiner sein. Ein SaFD ist insbesondere besser als ein gruppensequenzieller Test mit der Option der vorzeitigen Annahme der Nullhypothese. Sonst hätte die Optimierung diese Entscheidungsregel ergeben müssen. Im Gegenteil erreicht keines der in Kapitel 4.3 evaluierten gruppensequenzielles Designs nach Pocock mit einer Ergänzung um die jeweils bezüglich $c_0 \in [-5, 5]$ optimalen Grenzen für eine vorgezogene Annahme der Nullhypothese nach Wang und Tsiatis (für ein Design nach Pocock),

$$c_{n_K, n_{K+1}} = \sqrt{K}\sqrt{\frac{n_K}{2}} - c_0, \qquad K = 1, ..., \kappa - 1,$$

⟨18b⟩ ein Bayes-Risiko unter 0,22 ($\kappa = 2$, $c = 2{,}1584$, $c_{60,120} = \sqrt{1}\sqrt{60/2} - 4{,}7386 \approx 0{,}7386$).

Die Annahmegrenzen von Wang und Tsiatis heben sich mit der Zahl der Auswertungen beziehungsweise mit zunehmender Teilnehmerzahl an. Eine Anhebung ist auch bei einem SaFD zu erwarten. Je mehr Beobachtungen die Schätzung eines kleinen Effekts ergeben, der nicht zur a-priori-Dichte passt, umso näher liegt die Annahme der Nullhypothese. Das SaFD in Tabelle 3 erfüllt die Erwartung nicht. Vermutlich ist die Maximalfallzahl für die kleineren Effekte der a-priori-Dichte zu gering. Zur Bestätigung bräuchte es eine Erweiterung der gegebenen Optionen ◀

Entscheidungskriterien			Entscheidung		
Aktuelle Fallzahl	als finale Fallzahl	Untere Intervallgrenze	Zukünftige Fallzahl	als finale Fallzahl	Testentscheidung
52		$-\infty$	52	$*$	H_0
52		0,3994	156		
52		1,0172	104		
52		2,2849	208	$*$	
52		2,2849	156	$*$	
52		2,2849	104	$*$	
52		2,2849	52	$*$	H_1
104		$-\infty$	104	$*$	H_0
104		-3,0687	156		
104		-1,3010	208	$*$	
104		2,1263	156	$*$	
104		2,2849	104	$*$	H_1
104	$*$	$-\infty$	104	$*$	H_0
104	$*$	2,2849	104	$*$	H_1
156		$-\infty$	156	$*$	H_0
156		1,4800	208	$*$	
156		2,2849	156	$*$	H_1
156	$*$	$-\infty$	156	$*$	H_0
156	$*$	2,2849	156	$*$	H_1
208	$*$	$-\infty$	208	$*$	H_0
208	$*$	2,2849	208	$*$	H_1

Tabelle 3: RACV-optimales flexibles Design für die a-priori-Verteilung $\mathcal{R}[0{,}3; 0{,}7]$ gegeben der Gesamtkostenfunktion $g = \mathbb{E}_\theta(N) + 25\,\mathbb{E}_\theta(K)$ mit vier möglichen finalen Fallzahlen pro Behandlungsgruppe.

um höhere potenzielle Gesamtfallzahlen. Eine zu geringe zulässige Maximalfallzahl ist schon deshalb anzunehmen, weil die maximal mögliche Teilnehmerzahl bei gruppensequenziellen Designs nicht so hoch sein darf, als dass sie weitläufig zum Erreichen einer 100 %-igen Power beansprucht wird. Das Problem gibt es im SaFD nicht unbedingt. Außerdem deuten die Optimierungsergebnisse auf eine zu geringe Maximalfallzahl für das SaFD hin. Ab der ersten Auswertung werden 12,9 % der Pfade bei Gültigkeit von $\theta = 0{,}3$ nicht weiterverfolgt. Damit ist das Erreichen der angestrebten Power von 90 % für $\theta = 0{,}3$ von vornherein ausgeschlossen.

Wären noch kleinere Ausmaße möglich, verschärfte sich das Problem. In der Grenze für die Annahme der Nullhypothese äußert sich demnach unmittelbar die Bedeutung einer korrekt spezifizierten a-priori-Verteilung (vgl. Kapitel 3.3). Die Grenzen müssten maßgeblich von der a-priori-Verteilung abhängen. Die Abhängigkeit ließe sich in zukünftigen Gegenüberstellungen optimaler flexibler Designs verschiedener a-priori-Verteilungen untersuchen. ◀

Je wahrscheinlicher die vorzeitige Annahme der Nullhypothese ist, umso kleiner kann der kritische Wert zur Ausschöpfung des Signifikanzniveaus sein. Das scheint aber nicht der einzige Grund zu sein, warum der kritische Wert im SaFD in Tabelle 3 kleiner als im optimalen gruppensequenziellen Design nach Pocock ist ($2{,}2849 < 2{,}3613$). Wenn dem SaFD die Möglichkeit genommen wird, Auswertungen zu überspringen, wird das Signifikanzniveau nicht mehr eingehalten ($0{,}0273 > 0{,}0250$). Schließlich entfallen durch das Überspringen Gelegenheiten, Ausreißer abzufangen. Das ist ein nachteiliger Aspekt des Überspringens, das zur Kompensation ausreichend niedrige Auswertungszahlen bewirken muss.

Die Wahrscheinlichkeit, (ein oder zwei) Auswertungen zu überspringen, reicht von 38,3 % über eine Halbierung bei circa $\theta = 0{,}4127$ bis hin zu 0,6 %. Damit leistet das Überspringen einen nennenswerten Beitrag zur

Reduktion der durchschnittlichen Auswertungszahl, die in Abbildung 6 auf Seite 47 zu sehen ist. Die Konsequenzen des Überspringens zeigen sich insbesondere darin, dass sich die Wahrscheinlichkeit, alle vier möglichen Auswertungen zu durchlaufen, frühestens ab der achten Nachkommastelle von der Null unterscheidet. Sie beträgt beim optimalen gruppensequenziellen Verfahren bis zu 34,4 %. Die Diskrepanz kann nicht allein durch die vorzeitige Annahme der Nullhypothese erklärt werden. Die Wahrscheinlichkeit hierfür liegt höchstens bei 16,3 % (vgl. „H_0, spontan“ in Abbildung 8).

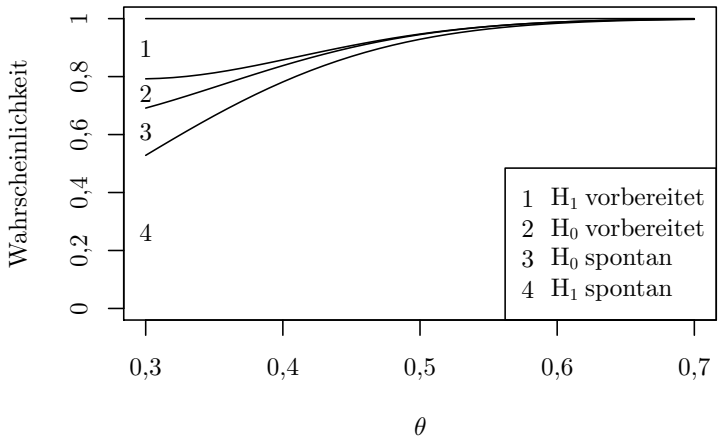

Abbildung 8: Wahrscheinlichkeiten der Studienausgänge, die sich aus einer Testentscheidung (spontan) während einer Fallzahlprüfung oder aus der zuvor durch die Festlegung auf eine endgültige Gesamtfallzahl veranlasste Testung (vorbereitet) ergeben.

Demgegenüber enden mehr als die Hälfte der Studien mit einer spontanen Ablehnung der Nullhypothese. Damit fällt die Entscheidung für eine abschließende Aufnahme zusätzlicher Studienteilnehmer zu höchstens 30,9 %. Sie wird bei einer bedingten Erfolgswahrscheinlichkeit von mindestens 59,6 % getroffen. Das ist insofern beachtlich, als dass die Entscheidung für eine geplante endgültige Gesamtfallzahl unter anderem bei negativen Behandlungseffekten ($\theta = \sqrt{\frac{2}{104}} Z_{104} > -0{,}1804$) und nicht den

wenigsten Beobachtungen getroffen wird (vgl. Tabelle 3). Abbildung 9 zeigt die variierenden bedingten Wahrscheinlichkeiten je nach Situation. Eine allgemeine Beziehung, wie konstante bedingte Wahrscheinlichkeiten, scheint daraus nicht ersichtlich zu sein.

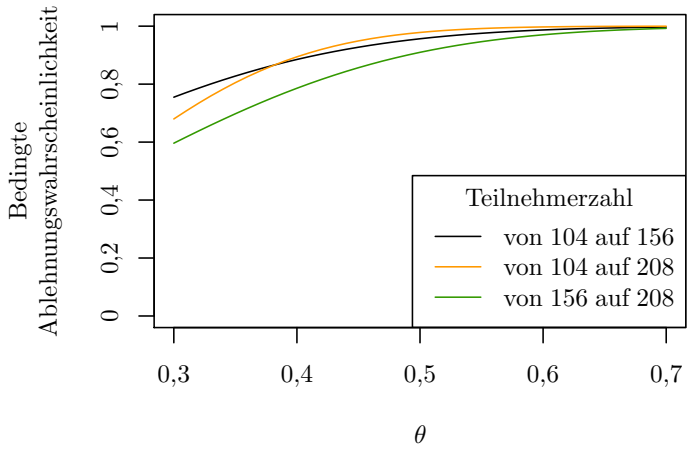

Abbildung 9: Bedingte Wahrscheinlichkeit, dass mit der Entscheidung für eine finale Auswertung die Ablehnung der Nullhypothese gelingt.

Die Entscheidung für eine letzte Fallzahlerhöhung sollte eine Annäherung an die geforderte Power von 90 % durch das Unterbinden des Strebens nach einer Power von 100 % bewirken. Die Steigung der in Abbildung 4 auf Seite 46 gezeigten Power ist aber ausnahmslos höher als beim optimalen gruppensequenziellen Test nach Pocock. Am größten sind die Unterschiede kurz nachdem das SaFD die geforderte Power übertrifft. Das belegt aber nicht die Nutzlosigkeit der Planung einer endgültigen Gesamtfallzahl. Die höhere Steigung kann auf die vorzeitige Annahme der Nullhypothese zurückgeführt werden. Der verhaltene Ressourceneinsatz reduziert die Ablehnungswahrscheinlichkeit für kleinere Behandlungseffekte. Auf der anderen Seite erhöht der kleinere kritische Wert die Power auch für weniger große Effekte. Dagegen kommt die Option einer abschließenden Aufnahme zusätzlicher Studienteilnehmer anscheinend nicht an.

Eine Erklärung ist die geringe Auswahl von Gesamtfallzahlen. Die Power des Designs nach Hartung demonstriert, dass schon eine Auswertung mit vielen Optionen für eine sichtbare Annäherung an die geforderte Power ausreichen kann (vgl. Abbildung 4 auf Seite 46). Wenn die Sprünge zwischen den möglichen Teilnehmerzahlen so hoch sind, dass die nächste Auswertung durch eine hohe Fallzahl zwangsläufig die letzte Auswertung sein wird, erübrigt sich die Differenzierung zwischen der Entscheidung für die zukünftige endgültige Gesamtfallzahl oder der zukünftig spontanen Entscheidung für ein Rekrutierungsende. Daher überrascht die seltene Inanspruchnahme der Entscheidung für eine größere endgültige Fallzahl nicht (vgl. Abbildung 8). Die höheren Wahrscheinlichkeiten bei kleineren Behandlungseffekten sind darauf zurückzuführen, dass die Entscheidung für die zu geringe und daher zu oft gefragte Maximalfallzahl zwangsläufig endgültig ist. Mit einer größeren Auswahl endgültiger Fallzahlen ◀ sollte sich daher ein anderes Bild als in Abbildung 8 ergeben. Der Anstoß müsste von einer geringeren durchschnittlichen Fallzahl für größere Behandlungseffekte ausgehen. Dieser Vorteil kann sich im optimierten erweiterten SaFD aber auch an anderer Stelle, wie dem Erreichen der angestrebten Power für kleinere Behandlungseffekte, zeigen.

Zusammenfassend lässt sich festhalten, dass ein SaFD sogar in stark reduzierter Form einem gruppensequenziellen Test nach Pocock überlegen sein kann. Darin sind Abweichungen vom Studienplan eingeschlossen. Im untersuchten Fall müssen nur mindestens 94 statt 107 Probanden beobachtet werden, um den RACV-optimalen Gauß-Test zu übertreffen. Die Vorteile gegenüber dem gruppensequenziellen Test nach Pocock liegen hier hauptsächlich in einer geringeren durchschnittlichen Auswertungszahl, indem Zwischenauswertungen übersprungen werden können. Das hat eine Reduktion der Studiendauer zur Folge, falls die Sprünge nicht unangemessen weit sind. Weitere potenzielle Stärken können sich höchstens bei komplexeren SaFDs durchsetzen. Mit der Unterbindung

des Strebens nach einer 100 %-igen Power könnten Probanden bei Behandlungseffekten mit ausreichender Power eingespart werden.

Insgesamt kann das SaFD als vielversprechendes Vorgehen bei ungewissen Fallzahlen bewertet werden. Demzufolge sollte es bei der Studienplanung in möglichst umfangreicher Form in Erwägung gezogen werden. Das beinhaltet die Gegenüberstellung mit alternativen Verfahren, wofür sich die RACV in Kapitel 4.3 empfohlen haben.

5 Resümee der Beiträge der Arbeit

Die vorliegende Arbeit hinterfragte einleitend das konventionelle und die publizierten alternativen Vorgehen bei der Bestimmung der Fallzahl für eine empirische Studie an Lebewesen. Die dargelegten Einwände bezüglich der Verfahren begründeten den Bedarf, ein anderes Vorgehen zu entwickeln. Die Dringlichkeit verdeutlichte die Einordnung der Bedeutung angemessener Fallzahlen in der medizinischen Forschung. Aufgrund der interdisziplinären Zusammenarbeit bei Studienplänen dieses Forschungsgebiets wurde auf mathematische Formulierungen möglichst verzichtet, um eine vielseitige Diskussion zu eröffnen.

In der Auseinandersetzung mit den Schwierigkeiten bei der Fallzahlberechnung wurde ein Ansatz zur Lösung des Problems identifiziert. Anhand von inhaltlichen Überlegungen wurde eine Struktur für Entscheidungsprozesse zur Fallzahlbestimmung entwickelt, die unter anderem das bislang etablierte Vorgehen als Spezialfall prinzipiell beinhaltet.

In diesem Zusammenhang wurde der Bedarf für ein erweitertes Verständnis einer optimalen Entscheidungsregel aufgezeigt, um die allgemeine Struktur dahingehend auszugestalten. Demzufolge wurde ein Optimierungskriterium zur Beurteilung von Verfahren zur Fallzahlbestimmung erarbeitet.

Ein zusätzlicher Beitrag der Arbeit liegt in dem Vorschlag eines erwartungstreuen Stichprobenschätzers, der sich auch wegen seiner einfachen Bestimmung durch einen breiten Anwendungsbereich auszeichnet.

Die genannten Entwicklungen regen weitere Forschungstätigkeiten an, die mitunter an entsprechender Stelle aufgezeigt wurden (◀).

QUELLENVERZEICHNIS

⟨1⟩ Brannath, Posch und Bauer (2002). Recursive combination tests, *Journal of the American Statistical Association* 97 (457), Seiten 236-244.

⟨2⟩ Burman und Sonesson (2006). Are flexible designs sound?, *Biometrics* 62 (6), Seiten 664-683.

⟨3⟩ Chang (2008). *Adaptive design theory and implementation using SAS and R*, Chapman & Hall,

⟨3a⟩ Seiten 51 ff,

⟨3b⟩ Seite 348.

⟨4⟩ Chow und Chang (2008). Adaptive design methods in clinical trials - a review, *Orphanet Journal of Rare Diseases* 3 (11).

⟨5⟩ Chow, Shao und Wang (2008). *Sample size calculations in clinical research*, Chapman & Hall,

⟨5a⟩ Seite 8,

⟨5b⟩ Seite 16,

⟨5c⟩ Seiten 57-58,

⟨5d⟩ Seiten 60, 119 f.

⟨6⟩ European Medicines Agency (2007). *Reflection paper on methodological issues in conformatory clinical trials with flexible design and analysis plan* (CHMP/EWP/2459/02).

⟨7⟩ Food and Drug Administration (2010). *Guidance for industry - adaptive design clinical trials for drugs and biologics.*

⟨8⟩ Hartung (2006). Flexible designs by adaptive plans of generalized Pocock and O'Brien-Fleming-type and by self-designing clinical trials, *Biometrical Journal* 48 (4), Seiten 521-536.

⟨9⟩ Jennison und Turnbull (2006). Adaptive and nonadaptive group sequential tests, *Biometrika* 93 (1), Seiten 1-21.

⟨10⟩ Liu, Hall, Yu und Wu (2006). Estimation following a group sequential test for distributions in the one-parameter exponential family, *Statistica Sinica* 16 (1), Seiten 165-181.

⟨11⟩ Liu, Zhu und Cui (2008). Evaluating the adaptive performance of flexible sample size designs with treatment difference in an interval, *Statistics in Medicine* 27 (4), Seiten 584-596.

⟨12⟩ Müller und Schäfer (2004). A general statistical principle for changing a design any time during the course of a trial, *Statistics in Medicine* 23 (16), Seiten 2497-2508.

⟨13⟩ Pocock (1977). Group sequential methods in the design and analysis of clinical trials, *Biometrica* 64 (2), Seiten 191-199.

⟨14⟩ Powell (1994). A direct search optimization method that models the objective and constraint functions by linear interpolation, *Advances in Optimization and Numerical Analysis*, Kluwer Academic Publishers, Seiten 51-67.

⟨15⟩ Shen und Fisher (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis, *Biometrics* 55 (1), Seiten 190-197.

⟨16⟩ Stansen (2010). *Flexible Designs in klinischen Prüfungen mit binärer und ordinaler Zielvariable*, Dissertation, Seite 46.

⟨17⟩ Svanberg (2002). A class of globally convergent optimization methods based on conservative convex separable approximations, *SIAM Journal on Optimization* 12 (2), Seiten 555-573.

⟨18⟩ Wassmer (1999). *Statistische Testverfahren für gruppensequentielle und adaptive Pläne in klinischen Studien: Theoretische Konzepte und deren praktische Umsetzung mit SAS*, Mönch,

⟨18a⟩ Seiten 1 ff.,

⟨18b⟩ Seiten 30 ff.,

⟨18c⟩ Seite 61.

⟨19⟩ Whitehead (1986). On the bias of maximum likelihood estimation following a sequential test, *Biometrika* 73 (3), Seiten 573-581.

⟨20⟩ Wu und Cui (2012). Group sequential and discretized sample size re-estimation designs: a comparison of flexibility, *Statistics in Medicine* 31 (24), Seiten 2844-2857.

⟨21⟩ Zhang, Cui und Yang (2016). Optimal flexible sample size design with robust power, *Statistics in Medicine* 35 (19), Seiten 3385-3396.