Diplomarbeit
(zweiter Teil der Diplomhauptprüfung)

zum Thema

Fallzahlplanung bei Daten mit Survivalendpunkt

im Studiengang Statistik

an der Ludwig-Maximilians-Universität München

Vorgelegt von Cordelia Rogon
am 21.12.2009

Referent:
Prof. Dr. T. Hothorn

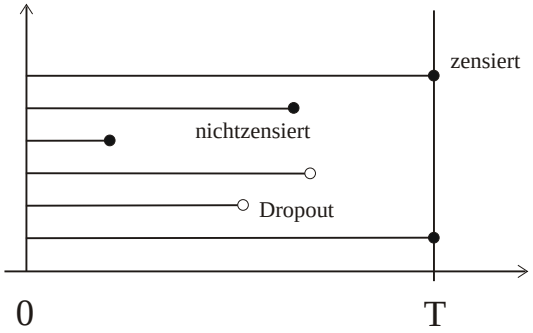
Korreferent:
PD Dr. M. Hennig

Inhaltsverzeichnis

Tabellenverzeichnis

1. Einleitung

In dieser Arbeit wird auf die Fallzahlplanung bei Daten mit Survivalendpunkt eingegangen. Da diese sogenannten Time-to-Event- oder Survivaldaten eine besondere Struktur besitzen, müssen spezielle Methoden zur Analyse, aber auch zur Fallzahlschätzung, angewandt werden. Bei Survivaldaten werden keine stetigen (z.B. Blutdruck) oder binären Endpunkte (z.B. Rückfall / kein Rückfall) betrachtet, sondern Zeitspannen bis zum Eintritt eines wohldefinierten Ereignisses. Dies kann im Rahmen klinischer Studien beispielsweise die bis zum Rückfall oder Tod eines Patienten vergangene Zeit sein. Geprägt durch die medizinische Anwendung wird die Zeitspanne meist als (Über-)Lebenszeit oder Lebensdauer und das Ereignis als Tod bezeichnet. Zeitspannen können jedoch in den verschiedensten Gebieten beobachtet werden, z.B. wenn in der Technik die Lebensdauer eines Geräts, in der Soziologie die Dauer bis zur Heirat oder in der Ökonometrie die Arbeitslosigkeitsdauer gemessen wird.
Bei Daten mit solch einer Struktur können die interessierenden Ereignisse oft nicht direkt beobachtet werden. Es ist beispielsweise nur bekannt, dass der Patient im Beobachtungszeitraum kein Ereignis hatte, unbekannt ist jedoch, ob und wann er das Ereignis nach Studienende hatte. Dies nennt man administrative Zensierung. Eine andere Art der Zensierung entsteht, wenn Patienten frühzeitig aus der Studie ausscheiden, etwa wenn sie nicht mehr zur Behandlung erscheinen. Diese Patienten bezeichnet man als lost to Follow-up oder häufig auch als Dropouts. In der folgenden Abbildung sind der Reihe nach die Überlebenszeiten eines administrativ zensierten Patienten, zweier Patienten, die im Beobachtungszeitraum ein Ereignis haben und damit nicht zensiert sind, zweier Patienten, die aus der Studie ausscheiden, und eines weiteren Patienten, der am Studienende noch nicht das interessierende Ereignis hatte, dargestellt.



Die Anzahl unzensierter Patienten steigt mit zunehmender Studiendauer. Um die Anzahl benötigter Ereignisse zu erreichen, sind klinische Studien deshalb in der Regel sehr lang.
Survivaldaten bestehen aus zwei Variablen pro Patient. Die erste Variable ist entweder der Ereigniszeitpunkt oder der Zensierungszeitpunkt, zu dem der Patient zuletzt beobachtet wurde. Die

zweite Variable ist ein Zensierungsindikator, der angibt, ob der Patient im Beobachtungszeitraum ein Ereignis hatte oder nicht.
In klinischen Studien wird oft eine Standardtherapie mit einer neuen Behandlung verglichen. Dabei wird geprüft, ob die neue Therapie die Überlebensdauer der Patienten signifikant erhöht. Dazu werden die Patienten zufällig einer Kontrollgruppe zugeteilt, die ein Placebo oder die Standardtherapie erhält, oder einer experimentellen Gruppe, die die neue Behandlung erhält. Mittels eines Tests wird untersucht, ob sich die Survivalverteilungen der beiden Behandlungsgruppen signifikant voneinander unterscheiden. Für diesen Vergleich können verschiedene Parameter wie die Länge der Rekrutierungs- und Follow-up-Phase und die Loss-to-Follow-up-, Noncompliance- sowie Rekrutierungsraten berücksichtigt werden.
Im zweiten Kapitel wird eine einheitliche Notation eingeführt, die in der gesamten Arbeit verwendet wird. Außerdem werden Grundlagen wie der Logranktest und das Proportional Hazards Modell von Cox vorgestellt.
Seit den frühen 1980er Jahren wird verstärkt auf dem Gebiet der Fallzahlplanung für Überlebenszeiten geforscht. Die in diesem Zeitraum entwickelten Formeln bilden die Grundlage der meisten später veröffentlichten Methoden. Die grundlegenden Formeln von Freedman (1982), Schoenfeld (1981), Schoenfeld und Richter (1982), Lachin und Foulkes (1986), Rubinstein et *al.* (1981) und Lakatos (1986, 1988) werden im dritten Kapitel angegeben.
Fallzahlplanung bei Daten mit Survivalendpunkt ist ein aktuelles Forschungsgebiet mit zahlreichen Veröffentlichungen. Ein Ziel dieser Arbeit ist es deshalb, im vierten Kapitel einen Überblick über die in den letzten Jahren neu entwickelten Methoden zu geben. Der Überblick von Oellrich et *al.* (1997) [52]) wird auf die folgenden Jahre bis zur Gegenwart fortgeführt und bezüglich Weiterentwicklung der dort vorgestellten Methoden und neuer Ansätze untersucht.
Im fünften Kapitel werden die Programme zur Fallzahlplanung PASS, nQuery, ADDPLAN und East vorgestellt und miteinander verglichen. Es wird untersucht, welche Formeln zur Fallzahlberechnung verwendet werden und ob es Diskrepanzen zwischen den mit verschiedenen Programmen berechneten Fallzahlen gibt. Außerdem wird geprüft, welche der neuen Methoden aus dem vierten Kapitel bereits in den Programmen implementiert wurden.
Im sechsten Kapitel wird ein in R selbstprogrammiertes und validiertes Programm zur Fallzahlberechnung bei Daten mit Survivalendpunkt vorgestellt und die damit berechneten Fallzahlen mit den im fünften Kapitel behandelten Fallzahlen verglichen. Darauf folgt das Schlusskapitel 7 mit Zusammenfassung und Ausblick sowie der Anhang.

2. Fundament

2.1. Notation und Modellierung

In diesem Kapitel wird die im gesamten Dokument verwendete einheitliche Notation eingeführt. Diese unterscheidet sich deshalb in den meisten Fällen von der in den zitierten Artikeln verwendeten Bezeichnungen. Die einheitliche Notation wurde gewählt, um Formeln und Methoden verschiedener Artikel besser miteinander vergleichbar zu machen und Unterschiede zu verdeutlichen. Außerdem wird auf die Modellierung von Daten mit Survivalendpunkt eingegangen.
In einer klinischen Studie werden die Patienten während R Zeitabschnitten rekrutiert. Die Rekrutierung wird auch oft als Eintritt der Patienten in die Studie bezeichnet. An die Rekrutierungsphase schließt die Follow-up-Phase an, in der keine neuen Patienten mehr eingeschlossen werden und ausschließlich die sich schon in der Studie befindenden Patienten weiter beobachtet werden. Die gesamte Studie umfasst T Zeitabschnitten, so dass die Follow-up-Phase aus $F = T - R$ Zeitabschnitten besteht.
Es wird die Fallzahl N für den Vergleich der Survivalverteilungen einer Kontrollgruppe (Gruppe 1) mit der einer experimentellen Gruppe (Gruppe 2) mittels eines Tests mit Signifikanzniveau α und Power $1 - \beta$ bestimmt. Ist $s = 1$, handelt es sich um einen einseitigen Test, bei $s = 2$ um einen zweiseitigen. Der in dieser Situation am häufigsten verwendete Test ist der Logranktest mit der Teststatistik LR. Die Gesamtfallzahl setzt sich aus den Fallzahlen der beiden Behandlungsgruppen N_1 und N_2 zusammen. Bei Methoden, die ein unbalanciertes Design, also ein Design mit unterschiedlich großen Gruppen, zulassen, wird das Verhältnis der Gruppengrößen $r = N_2/N_1$ verwendet. Bei Survivaldaten wird zunächst die Anzahl benötigter Ereignisse d bestimmt. Daraus kann anschließend unter Berücksichtigung verschiedener Annahmen die Fallzahl berechnet werden.
Die Survivalverteilung von Behandlungsgruppe i, $i = 1, 2$, lässt sich durch unterschiedliche Parameter beschreiben. Sei die Lebensdauer X_i eine nichtnegative, stetige Zufallsvariable mit Dichte $f_i(x)$ und Verteilungsfunktion $F_i(x)$. Die Wahrscheinlichkeit, bis zum Zeitpunkt x oder noch länger zu leben, wird durch der Survivalfunktion $S_i(x) := P(X_i \geq x_i) = 1 - F(x_i)$ angegeben. $S_i(T)$ ist die am Studienende erwartete Überlebensrate, $\pi_i(T) = 1 - S_i(T)$ dagegen ist die am Studienende erwartete Ereignisrate. Die mittlere Überlebenszeit, auch Median Survival Time, m_i gibt den Zeitpunkt der Studie an, zu dem bei der Hälfte der Patienten bereits das interessierende Ereignis eingetreten ist. Die Hazardrate $\lambda_i(x)$ gibt das infinitesimale Risiko an, zum Zeitpunkt x ein Ereignis zu haben, wenn man bis dahin überlebt hat.
Diese Parameter stehen bei Annahme einer Exponentialverteilung als Survivalverteilung durch folgende Beziehungen miteinander in Verbindung. Die Wahrscheinlichkeit, bis zum Zeitpunkt x zu überleben beträgt

$$S_i(x) = e^{-\lambda_i x}. \tag{2.1}$$

Die Hazardrate kann durch

$$\lambda_i = \frac{\ln(2)}{m_i} \tag{2.2}$$

aus der mittleren Überlebenszeit berechnet werden. Durch diese beiden Zusammenhänge können bei Vorgabe eines Parameters alle anderen Parameter daraus berechnet werden.
Das Verhältnis der Hazardraten, das Hazardratio, wird mit

$$\theta = \frac{\lambda_2}{\lambda_1} \tag{2.3}$$

angegeben. Dieses Verhältnis kann wegen obiger Zusammenhänge auch durch die anderen Parameter der Survivalverteilungen dargestellt werden:

$$\theta = \frac{\ln(S_2(x))}{\ln(S_1(x))} = \frac{m_1}{m_2} = \frac{\ln(1-\pi_2(x))}{\ln(1-\pi_1(x))}. \tag{2.4}$$

Bei exponentialverteilten Überlebenszeiten ist $\mathrm{E}(X) = 1/\lambda$. Je höher also das konstante infinitesimale Risiko λ ist, desto kürzer ist die erwartete Lebensdauer. Da die Varianz der Lebensdauer $\mathrm{Var}(X) = 1/\lambda^2$ beträgt, streuen die Überlebenszeiten umso mehr, je länger die zu erwartende Lebensdauer ist.
Da klinische Studien in der Regel lange dauern, besteht in gruppensequentiellen oder adaptiven Designs die Möglichkeit, die Daten bereits in Zwischenanalysen zu untersuchen, um die Studie bei Vorliegen eines signifikanten Ergebnisses aus ethischen Gründe frühzeitig zu beenden (siehe Kapitel 4.5). In einem solchen Design besteht die Studie aus I Phasen. Anschließend an jede Phase werden die Daten ausgewertet. Die Endauswertung findet zum Zeitpunkt a_I und die $I-1$ Interims- oder Zwischenanalysen zu den Zeitpunkten $a_1, \ldots, a_{I-1}$ statt. Findet eine Zwischenanalyse erst nach Ablauf der Rekrutierungsphase, also in der Follow-up-Phase, statt, ist nur noch eine Verkürzung der Studiendauer, aber keine Reduktion der Fallzahl mehr möglich.
In Phase i werden pro Behandlungsgruppe N_i Patienten rekrutiert, so dass die insgesamt in der Studie benötigte Fallzahl $2\sum_{i=1}^{I} N_i$ beträgt. Bei der $i$ten Analyse wird eine Teststatistik T_i berechnet und mit einer vorher festgelegten Abbruchgrenze c_i verglichen ($i = 1, \ldots, I$). Die Studie wird im Fall $T_i > c_i$ abgebrochen und andernfalls fortgeführt. Ist T_I immer noch kleiner als c_I, ist der Behandlungseffekt nicht signifikant und die Studie wird beendet. Das Signifikanzniveau α kann beispielsweise mittels einer Alpha Spending Function auf die Zwischenanalyse aufgeteilt werden, so dass das Signifikanzniveau der $i$ten Interimsanalyse α_i beträgt.
Die Notation ist in Angang A in tabellarischer Form zusammengefasst.

2.2. Der Logranktest

Die meisten Methoden zur Fallzahlberechnung bei Daten mit Survivalendpunkt beruhen auf dem Logranktest. Dies ist ein nonparametrischer Test zum Vergleich von zwei Gruppen. Eine möglichst homogene Gruppe von Individuen (z.B. Patienten) wird zufällig in zwei Gruppen aufgeteilt. Gruppe 1 erhält eine Kontrollbehandlung, z.B. Placebo, Gruppe 2 wird mit einer neuen

Therapie behandelt. Seien $\lambda_1(x)$, $\lambda_2(x)$ bzw. $S_1(x)$, $S_2(x)$ die Hazardraten bzw. Survivalfunktionen der Gruppen 1 und 2. Getestet werden soll

$$H_0 : \lambda_1(x) = \lambda_2(x) \qquad \text{für alle } x$$

gegen die Alternativhypothese

$$H_1 : \lambda_1(x) \neq \lambda_2(x)$$

für mindestens ein x. Kann die Nullhypothese H_0 abgelehnt werden, ist ein signifikanter Behandlungsunterschied erkennbar. Um den Logranktest anzuwenden, werden zunächst alle Ereigniszeiten unabhängig von ihrer Gruppenzugehörigkeit gepoolt und aufsteigend sortiert:

$$x_{(1)} < x_{(2)} < \ldots < x_{(r)}.$$

Zu jedem dieser geordneten Zeitpunkte $x_{(j)}$ werden die Ereignisanzahlen d_{1j} aus Gruppe 1 und d_{2j} aus Gruppe 2 bestimmt. Außerdem werden die Anzahlen von Patienten unter Risiko kurz vor $x_{(j)}$ aus Gruppe 1 r_{1j} und aus Gruppe 2 r_{2j} benötigt. Folglich ist $d_j = d_{1j} + d_{2j}$ die Anzahl aller Ereignisse in $x_{(j)}$ und $r_j = r_{1j} + r_{2j}$ die Anzahl aller Individuen unter Risiko kurz vor $x_{(j)}$. Unter H_0 werden alle Individuen zufällig aus einer „Urne“ mit r_j Elementen gezogen mit Label „Ereignis“ bzw. „kein Ereignis“. Dabei tragen d_j Elemente das Label „Ereignis“.
Betrachte nur Gruppe 1. Dann ist d_{1j} gegeben d_j hypergeometrisch verteilt mit Parametern r_j und d_j/r_j, Erwartungswert

$$\mathrm{E}(d_{1j}) = e_{1j} = r_{1j}\frac{d_j}{r_j} \tag{2.5}$$

und Varianz

$$\mathrm{Var}(d_{1j}) = v_{1j} = \frac{r_{1j}r_{2j}d_j(r_j - d_j)}{r_j^2(r_j - 1)} \tag{2.6}$$

Die Idee für die Teststatistik ist der Vergleich der beobachteten (d_{1j}) mit den unter H_0 erwarteten (e_{1j}) Ereignissen für $j = 1, \ldots, r$ bzw. der Vergleich zwischen der beobachteten und erwarteten Gesamtzahl an Ereignissen:

$$U_L = \sum_{j=1}^{r}(d_{1j} - e_{1j}). \tag{2.7}$$

Wegen der bedingten Unabhängigkeit ist

$$\mathrm{Var}(U_L) = \sum_{j=1}^{r} v_{1j} =: V_L. \tag{2.8}$$

Es lässt sich zeigen, dass unter H_0 gilt:

$$\frac{U_L}{\sqrt{V_L}} \overset{a}{\sim} N(0,1).$$

Die Teststatistik des Logranktests schließlich lautet

$$LR = \frac{U_L^2}{V_L} \tag{2.9}$$

und ist unter der Nullhypothese approximativ χ^2-verteilt mit einem Freiheitsgrad.

2.3. Das Proportional-Hazards-Modell von Cox

Viele Fallzahlformeln setzen proportionale Hazardraten voraus. In diesem Abschnitt soll das 1972 von Cox [14] entwickelte Proportional-Hazards-Modell vorgestellt werden. Es hat die Form

$$\lambda(x, z) = \lambda_0(x) \exp(z'\beta) \tag{2.10}$$

mit

$$\lambda(x, z) = \lim_{\Delta x \to 0} \frac{P(x \leq X < x + \Delta x | X \geq x, z)}{\Delta z}. \tag{2.11}$$

Dabei ist $\lambda_0(x)$ eine beliebige, nicht spezifizierte, nicht negative Baseline-Hazardrate. Sie ist identisch für alle Individuen. Der Vektor z besteht aus Kovariablen. Der lineare Prädiktor $z'\beta$ enthält keine Konstante, das heißt $x'\beta = \beta_1 z_1 + \ldots + \beta_p z_p$. Von primärem Interesse ist die Schätzung der Effekte β.
Die charakteristische Eigenschaft dieses Modells ist die Proportionalität der Hazardraten $\lambda(x, z_1)$ und $\lambda(x, z_2)$ für zwei Individuen mit Kovariablen z_1 und z_2. Es gilt

$$\frac{\lambda(x, z_1)}{\lambda(x, z_2)} = \frac{\lambda_0(x)}{\lambda_0(x)} \frac{\exp(z_1'\beta)}{\exp(z_2'\beta)} = \exp((z_1 - z_2)'\beta). \tag{2.12}$$

Das Verhältnis der Hazardraten zweier Individuen ist also zu jedem beliebigen Zeitpunkt gleich. Diese Annahme trifft in der Praxis häufig nicht zu, weshalb Methoden zur Fallzahlschätzung bei nicht proportionalen Hazardraten entwickelt wurden. Einige werden in Kapitel 4.1 vorgestellt.

3. Grundlegende Methoden zur Fallzahlberechnung

In den achtziger Jahren begann die intensive Forschung der Fallzahlplanung bei Survivaldaten. Die Methoden von Freedman (1982), Schoenfeld (1981), Schoenfeld und Richter (1982), Lachin und Foulkes (1986), Rubinstein et. *al.* (1981) und Lakatos (1986, 1988) liegen vielen aktuellen Ansätzen zugrunde und werden in diesem Kapitel eingeführt.

3.1. Freedman

Freedman (1982) [21] setzt proportionale Hazardraten voraus und berechnet die Fallzahl aus dem Erwartungswert und der Varianz der Logrankstatistik. Seien $S_1(T)$ und $S_2(T)$ die Survivalraten am Studienende in der Kontroll- bzw. experimentellen Gruppe und

$$\theta = \frac{\ln(S_2(T))}{\ln(S_1(T))} \tag{3.1}$$

das Hazardratio. Ist N_1 der Anteil der Personen in der Kontrollgruppe, ist das Verhältnis der Gruppengrößen $r = (1 - N_1)/N_1$. Es wird angenommen, dass ein prozentualer Anteil $loss$ lost to Follow-up sein wird. Die Fallzahl für die gesamte Studie berechnet sich zu

$$N = \frac{d(1+r)}{[r(1-S_2(T)) + (1-S_1(T))](1-loss)}, \tag{3.2}$$

wobei die insgesamt in der Studie erwartete Ereignisanzahl d gegeben ist durch

$$d = \frac{(r\theta+1)^2}{r(\theta-1)^2}(z_{1-\alpha/s} + z_{1-\beta})^2. \tag{3.3}$$

s ist dabei 1 für einen einseitigen und 2 für einen zweiseitigen Test. Die Methode von Freedman ist in den Programmen PASS, nQuery und ADDPLAN implementiert.

3.2. Schoenfeld

In der Formel von Schoenfeld (1981) [59] werden ebenfalls das Größenverhältnis r der Gruppen 2 zu 1 und das Hazardratio θ verwendet. Die Dauer von Rekrutierungs- und Follow-up-Phase werden auch hier nicht berücksichtigt. Zunächst wird die Anzahl benötigter Ereignisse

$$d = \frac{(z_{1-\alpha/s} + z_{1-\beta})^2}{r/(1+r)^2(\ln(\theta))^2} \tag{3.4}$$

bestimmt. Die Fallzahl berechnet sich aus der Ereignisanzahl zu

$$N = \frac{d(1+r)}{r(1-S_2(T)) + (1-S_1(T))}, \tag{3.5}$$

wobei $S_1(T)$ bzw. $S_2(T)$ die Survivalraten am Studienende in der Kontroll- bzw. experimentellen Gruppe sind. Diese Methode ist in dem Programm ADDPLAN implementiert.
Schoenfelds Formel liefert in balancierten Designs kleinere benötigte Ereignisanzahlen als Freedmans Formel (Hsieh, 1992 [30]).
In einem weiteren Artikel stellen Schoenfeld und Richter (1982) [61] eine Methode zur Fallzahlberechnung vor, die die Länge von Rekrutierungs- und Follow-up-Phase einbezieht. Bei Annahme exponentialverteilter Überlebenszeiten ist

$$P_j(R) = \frac{1}{R}\int_0^R S_j(R-u)du = \frac{1-\exp(-\log(2)R/m_j)}{\log(2)R/m_j} \tag{3.6}$$

die Wahrscheinlichkeit in Gruppe j, dass ein Patient die Rekrutierungsphase der Länge R überlebt und

$$S_j(F) = \exp(-\log(2)F/m_j) \tag{3.7}$$

die Wahrscheinlichkeit in Gruppe j, dass ein Patient vom Ende der Rekrutierungsphase bis zum Studienende überlebt. Dann setzt sich die Ereignisrate p_j in Gruppe j aus diesen beiden Wahrscheinlichkeiten zusammen zu

$$p_j = 1 - P_j(R)S_j(F). \tag{3.8}$$

Die Ereigniswahrscheinlichkeiten in den beiden Behandlungsgruppen gehen folgendermaßen in die Fallzahlformel ein:

$$N = \frac{2(z_{1-\alpha/s} + z_{1-\beta})^2}{(\ln(\theta))^2}\left(\frac{1}{p_1} + \frac{1}{p_2}\right). \tag{3.9}$$

3.3. Lachin und Foulkes

Mit der Methode von Lachin und Foulkes (1986) [35] kann die Fallzahl für unbalancierte Designs, nicht gleichverteilte Eintrittszeitpunkte der Patienten in die Studie und Loss to Follow-up bei Annahme exponentialverteilter Überlebenszeiten berechet werden. Es wird die Nullhypothese $H_0 : \lambda_1 = \lambda_2$ getestet.
Soll der Einschluss der Patienten in die Studie nicht durch eine Gleichverteilung modelliert werden, kann eine trunkierte Exponentialverteilung mit Parameter γ auf dem Träger $[0; R]$ und Verteilungsfunktion

$$G(x) = \frac{1-e^{-\gamma x}}{1-e^{-\gamma R}}, \qquad 0 \le x \le R, \qquad \gamma \neq 0 \tag{3.10}$$

verwendet werden. Wenn $\gamma > 0$, ist die Rekrutierungsfunktion konvex (schnelle Rekrutierung zu Beginn der Studie), wenn $\gamma < 0$, ist sie konkav (langsamere Rekrutierung). Im Fall von $\gamma = 0$ wird die Gleichverteilung verwendet.
Loss to Follow-up wird durch gruppenspezifische Exponentialverteilungen mit Parametern η_1 und η_2 modelliert.
$\phi(\lambda, \eta, \gamma)$ ist die Komponente der Varianz des Maximumlikelihoodschätzers von λ, die unabhängig von der Fallzahl ist, hängt zusätzlich zu λ von η und γ ab und lautet

$$\phi(\lambda, \eta, \gamma) = \lambda^2 \left\{ \frac{\lambda}{\lambda + \eta} + \frac{\lambda \eta e^{-(\lambda+\eta)T}[1 - e^{(\lambda+\eta-\gamma)R}]}{(1 - e^{-\gamma R})(\lambda + \eta)(\lambda + \eta - \gamma)} \right\}^{-1}. \tag{3.11}$$

Die Fallzahlformel ergibt sich zu

$$N = \frac{(z_{\alpha/s}\sqrt{\phi(\overline{\lambda}, \eta_2, \gamma)Q_2^{-1} + \phi(\overline{\lambda}, \eta_1, \gamma)Q_1^{-1}}) + z_\beta \sqrt{\phi(\lambda_2, \eta_2, \gamma)Q_2^{-1} + \phi(\lambda_1, \eta_1, \gamma)Q_1^{-1}})^2}{(\lambda_2 - \lambda_1)^2}, \tag{3.12}$$

wobei $Q_1 = N_1/N$, $Q_2 = N_2/N$ und $\overline{\lambda} = Q_1\lambda_1 + Q_2\lambda_2$. Die Methode von Lachin und Foulkes ist in dem Programm PASS implementiert.

3.4. Rubinstein et al.

Rubinstein et *al.* (1981) [57] nehmen an, dass die Überlebenszeiten exponentialverteilt sind. Auch die Loss-to-Follow-up-Periode ist exponentialverteilt und es ist möglich, gruppenspezifische Loss-to-Follow-up-Raten η_1 und η_2 anzugeben. Die Fallzahl wird durch

$$N = \left(\frac{z_{1-\alpha/s} + z_{1-\beta}}{\ln(\lambda_2) - \ln(\lambda_1)} \right)^2 \left(\frac{1}{E(P_2)} + \frac{1}{E(P_1)} \right), \tag{3.13}$$

berechnet, wobei

$$E(P_i) = \frac{\lambda_i}{\lambda_i + \eta_i} \left[1 - \frac{e^{-(\lambda_i+\eta_i)(T-R)} - e^{-(\lambda_i+\eta_i)T}}{(\lambda_i + \eta_i)R} \right] \tag{3.14}$$

In dem Programm nQuery kann die Fallzahl basierend auf dieser Methode berechnet werden.

3.5. Lakatos

Auch wenn das Hazardratio während der Studie nicht konstant ist, kann die Logrankstatistik verwendet werden. Lakatos (1988) [37] teilt die Studie in Teilintervalle und nimmt einen nichtstationären Markovprozess an. Die Fallzahl wird unter Berücksichtigung von Noncompliance, Dropin und Loss to Follow-up berechnet. Patienten, die die ihnen zugewiesene Behandlung abbrechen, aber trotzdem noch weiter beobachtet werden können, werden für die experimentelle Gruppe als Noncomplier und für die Kontrollgruppe als Dropins bezeichnet. Sie befinden sich

nach dem Wechsel im Compliant-Zustand der jeweils anderen Behandlungsgruppe und nehmen deren Ereignisrate an. Das Markovmodell enthält die vier Zustände „Lost to Follow-up“, „hatte ein Ereignis“, „ist compliant“ und „ist nicht compliant“. Für die experimentelle Gruppe werden die Wahrscheinlichkeiten, sich in diesen vier Zuständen zu befinden, mit $loss_e$, $event_e$, $compl_e = 1 - loss_e - event_e - noncompl$ und $noncompl$ bezeichnet. Für die Kontrollgruppe lauten sie $loss_c$, $event_c$, $compl_c = 1 - loss_c - event_c - dropin$ und $dropin$. Die Studie bestehe aus T Zeiteinheiten. Jede dieser Zeiteinheiten wird wiederum in K gleichlange Intervalle aufgeteilt. Theoretisch können für jedes dieser Intervalle unterschiedliche Ereignis-, Loss-to-Follow-up-, Noncompliance- und Dropinraten spezifiziert werden. Wichtig ist die Annahme konstanter Hazardratios und Verhältnisse der Patienten unter Risiko in jedem Teilintervall.
Die Zustandswahrscheinlichkeiten können für jedes Zeitintervall der Studie i, $i = 1, \ldots, T \cdot K$ unterschiedlich sein. Deshalb werden die Zustandswahrscheinlichkeiten in Form eines Vektors D_{t_i} angegeben. Die Wahrscheinlichkeiten zu den Zeitpunkten $1, \ldots, T$ von der experimentellen in die Kontrollgruppe zu wechseln, werden beispielsweise in einen Vektor

$$noncompl = \begin{bmatrix} noncompl_1 \\ noncompl_2 \\ \ldots \\ noncompl_{T \cdot K} \end{bmatrix}$$

eingegeben.
In diesem grundlegenden Modell wird ein sofortiger Wirkungseintritt der Behandlung und ein simultaner Einschluss der Patienten zu Studienbeginn vorausgesetzt. Die Zustandsvektoren der experimentellen beziehungsweise der Kontrollgruppe werden wie folgt initalisiert:

$$D_{e,0} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \qquad D_{c,0} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Die Übergangsmatrizen $trans_{i,i+1}$ bestehen aus den Wahrscheinlichkeiten $trans_{i,i+1}(j_1, j_2)$, mit denen ein Patient im Zeitintervall $[t_i, t_{i+1}]$ aus Zustand j_1 in Zustand j_2 wechselt:

$$trans_{i,i+1} = \begin{bmatrix} 1 & 0 & loss_{e,i} & loss_{c,i} \\ 0 & 1 & event_{e,i} & event_{c,i} \\ 0 & 0 & compl_{e,i} & dropin_i \\ 0 & 0 & noncompli & compl_{c,i} \end{bmatrix}$$

Wenn $T \cdot K$ das Studienende bezeichnet, ist das Markovmodell für $i \leq T \cdot K$ gleich

$$D_{t_i} = trans_{i-1,i} D_{t_{i-1}}. \tag{3.15}$$

Nun werden $D_e = trans_{T \cdot K-1, T \cdot K} \cdot trans_{T \cdot K-2, T \cdot K-1} \cdot ... \cdot D_{e,0}$ für die experimentelle und $D_c = trans_{T \cdot K-1, T \cdot K} \cdot trans_{T \cdot K-2, T \cdot K-1} \cdot \ldots \cdot D_{c,0}$ für die Kontrollgruppe berechnet. Es resultieren Sequenzen mit $T \cdot K$ Einträgen, die aus Vektoren mit vier Zeilen bestehen. Das Element in der zweiten Zeile des fünften Eintrags in D_e wird beispielsweise mit $D_{e,5,2}$ bezeichnet und gibt

die Ereigniswahrscheinlichkeit für einen Patient aus der experimentellen Gruppe zum Zeitpunkt 5 an. Die Einträge dieser Sequenzen werden im Folgenden benutzt, um die nötige Ereignisanzahl und die Fallzahl für den Logranktest zum Vergleich der beiden Gruppen zu berechnen.
Werden die Ereignis-, Loss-to-Follow-up-, Noncompliance- und Dropinraten ausschließlich pro Zeiteinheit (und nicht pro Intervall) festgelegt und sind die Hazardraten konstant über die Zeit, müssen die Nichtdiagonalelemente b der Übergangsmatrizen durch $1-(1-b)^{1/K}$ ersetzt werden.
Zunächst wird das Verhältnis ϕ der Patienten unter Risiko in der Kontroll- im Vergleich zur experimentellen Gruppe berechnet. Es befinden sich diejenigen Personen unter Risiko, die noch kein Ereignis hatten und nicht Lost to Follow-up sind, also die Summe der Patienten, die entweder compliant sind oder in die andere Behandlungsgruppe wechselten. Das Verhältnis beträgt im ersten Zeitintervall bei gleichgroßen Gruppen 1 und im $(i+1)$ten Zeitintervall

$$\phi_{i+1} = \frac{D_{c,i,3} + D_{c,i,4}}{D_{e,i,3} + D_{e,i,4}}. \tag{3.16}$$

Mit θ_i wird das Hazardratio der experimentellen im Vergleich zur Kontrollgruppe im $i$ten Intervall bezeichnet. Das Hazardratio des ersten Intervalls berechnet sich als

$$\theta_1 = \frac{\ln(1 - event_{c,1})}{\ln(1 - event_{e,1})}. \tag{3.17}$$

Da der Benutzer die Ereignisraten $\pi_i(x)$ in den Vektor $event_c$ eingibt, können daraus die Survivalraten $S_i(x)$ durch $1 - \pi_i(x)$ berechnet werden. Aus einer Survivalrate zum Zeitpunkt x wiederum kann durch die Formel $\lambda_i = -\ln(1 - S_i(x))/x$ die Hazardrate bestimmt werden. Auf diese Weise wird der erste Eintrag von θ berechnet. Danach ändern sich jedoch im Laufe der Studie die Ereignis-, Noncompliance- und Dropinraten. Deshalb muss das Hazardratio unter Berücksichtigung dieser Raten berechnet werden.
Für das $(i+1)$te Intervall wird zunächst der Zuwachs z des Ereignisanteils zwischen der $i$ten und $(i+1)$ten Stufe berechnet. z muss dann durch den Anteil der Patienten unter Risiko im $i$ten Intervall geteilt werden. Dieser Bruch ist die Survivalrate aus der wieder die Hazardrate berechnet werden kann. Der Quotient der Hazardraten der beiden Gruppen ergibt schließlich das Hazardratio θ, dessen $(i+1)$te Komponente

$$\theta_{i+1} = \frac{\frac{-\ln\{1-(D_{c,i+1,2}-D_{c,i,2})\}}{D_{c,i,3}+D_{c,i,4}}}{\frac{-\ln\{1-(D_{e,i+1,2}-D_{e,i,2})\}}{D_{e,i,3}+D_{e,i,4}}} \tag{3.18}$$

lautet. Außerdem werden die Parameter γ und η für das $i$te Intervall als

$$\gamma_i = \frac{\phi_i\theta_i}{1 + \phi_i\theta_i} - \frac{\phi_i}{1 + \phi_i} \tag{3.19}$$

und

$$\eta_i = \frac{\phi_i}{(1 + \phi_i)^2} \tag{3.20}$$

definiert. $\rho_i = d_i/d$ sei der Anteil der im $i$ten Intervall auftretenden Ereignisse an der Gesamtereignisanzahl in der Studie $d = \sum d_i$. Im ersten Intervall ist $\rho_1 = \frac{D_{c,1,2}+D_{e,1,2}}{D_{c,T,2}+D_{e,T,2}}$. In den folgenden Intervallen ist

$$\rho_{i+1} = \frac{D_{c,i+1,2} - D_{c,i,2} + D_{e,i+1,2} - D_e i, 2}{D_c T, 2 + D_{e,T,2}}. \tag{3.21}$$

Es werden also die Zuwächse der Ereignisrate der beiden Gruppen addiert und durch die Summe der Ereignisraten am Studienende geteilt. Die Anzahl insgesamt benötigter Ereignisse wird berechnet durch

$$d = \frac{[z_{1-\alpha/2}(\sum \rho_i \eta_i)^{1/2} + z_{1-\beta}(\sum \rho_i \eta_i)^{1/2}]^2}{(\sum \rho_i \gamma_i)^2} \tag{3.22}$$

und daraus wiederum kann die Fallzahl N durch

$$N = \frac{2d}{D_{c,T,2} + D_{e,T,2}} \tag{3.23}$$

bestimmt werden. Im Nenner stehen dabei die Ereignisraten der beiden Behandlungsgruppen unter Berücksichtigung von Noncompliance, Dropin und Loss to Follow-up. Die Methode von Lakatos ist in dem Programm nQuery und in SIZE von Shih (1995) [66] implementiert.

4. Übersicht über die Literatur von 1997 bis 2009

Oellrich et *al.* (1997) geben einer Überblick über die Literatur zur Fallzahlplanung bei Survivaldaten von 1967 bis 1996. Dieser Überblick wird in diesem Kapitel auf die folgenden Jahre bis zur Gegenwart aktualisiert. Die Artikel wurden durch Referenzen in den Handbüchern der Fallzahlprogramme aus Kapitel 5, Durchsuchung von PubMed nach den Schlagworten „sample size survival“ und Querverweise gefunden.

"*By the years of publication it is seen that the development of sample size methods specialised on survival time data became intensive in the 1980's, but that it is still an ongoing field of research and publication.*" (Oellrich et *al.*, 1997 [52])

4.1. Nichtproportionale Hazardraten

In dem vielen Methoden zugrundeliegenden Artikel von Schoenfeld (1981) [59] spielt die Annahme proportionaler Hazardraten eine zentrale Rolle (siehe Kapitel 2.3). Auch in später erschienenen Artikeln wird diese Annahme häufig vorausgesetzt (siehe beispielsweise Halabi und Singh, 2004 [24]; Schulgen et *al.*, 2005 [62]). Die Annahme ist in der Praxis jedoch nicht immer gerechtfertigt, vor allem bei Studien längerer Dauer. Ist in einer Langzeitstudie der Endpunkt beispielsweise Altersdiabetes, wird die Erkrankungswahrscheinlichkeit mit der Zeit ansteigen, da die Patienten älter werden. So entwickelten unter anderem Heo et *al.* (1998) [28], Ahnn und Anderson (1998) [3], Porcher et *al.* (2002) [54] und Lakatos (2002) [38] Methoden zur Fallzahlschätzung, mit denen auch nicht proportionale Hazardraten berücksichtigt werden können.
Meist wird die Exponentialverteilung zur Modellierung der Überlebenszeiten verwendet. Dabei werden über die Zeit konstante Hazardraten angenommen. Heo et *al.* (1998) [28] erweitern die Fallzahlformel von Schoenfeld und Richter (1982) [61] auf weibullverteilte Überlebenszeiten. Mit der Weibullverteilung können im Studienverlauf zu- oder abnehmende Hazardraten modelliert werden. Sei m_j die mittlere Überlebenszeit in Gruppe j, $j = 1, 2$. Die Survivalverteilung kann als

$$S_j(x) = P(X > x) = \exp\left\{-\ln(2)\left(\frac{x}{m_j}\right)^{\nu}\right\} \tag{4.1}$$

für Gruppe j unter Annahme eines für beide Gruppen gleichen Shapeparameters ν ausgedrückt werden. Dieser Parameter steuert den Verlauf der Survivalverteilung und der Hazardrate. Ist

$\nu > 1$, nimmt die Hazardrate über die Zeit zu, ist $\nu < 1$, nimmt diese im zeitlichen Verlauf ab. Die Hazardrate hat die Form

$$\lambda_j(t) = \frac{\nu \ln(2)}{m_j} \left(\frac{x}{m_j} \right)^{\nu-1}. \tag{4.2}$$

Man erhält die Exponentialverteilung als Spezialfall bei $\nu = 1$, wenn die Hazardrate also konstant ist. Ist die Survivalzeit X weibullverteilt, ist X^ν exponentialverteilt. Bei Annahme einer Weibullverteilung wird die Nullhypothese $H_0 : [m_1/m_2] = 1$ gegen die Alternativhypothese $H_1 : [m_1/m_2]^\nu = \theta^\nu$ getestet. In der Fallzahlformel von Schoenfeld und Richter wird folglich das Hazardratio θ durch θ^ν ersetzt und man erhält als Formel

$$1 - \beta = \Phi \left\{ \frac{N^{1/2} \ln(\theta^\nu)}{\sqrt{p_1^{-1} + p_2^{-1}}} - \Phi^{-1}\{1 - \alpha\} \right\}. \tag{4.3}$$

$p_j = 1 - P_j(R)S_j(F)$ ist wie bei Schoenfeld und Richter die Wahrscheinlichkeit, dass ein Patient in Gruppe j im Laufe der Studie stirbt, und ist abhängig von der Länge der Rekrutierungs- und Follow-up-Phase, R beziehungsweise F. Hier ist

$$P_j(R) = \frac{1}{R} \int_0^R S_j(R - u)du \tag{4.4}$$

die Wahrscheinlichkeit in Gruppe j, dass ein Patient die Rekrutierungsphase überlebt, und

$$S_j(F) = \exp(-\ln(2)F/m_j)^\nu = \exp\left\{ -\ln(2) \left(\frac{F}{m_j} \right)^\nu \right\} \tag{4.5}$$

die Wahrscheinlichkeit in Gruppe j, dass ein Patient vom Ende der Rekrutierungsphase bis zum Studienende überlebt. Die Fallzahl N erhält man durch Umstellen der Gleichung der Power zu

$$N = \frac{(z_{1-\alpha/s} + z_{1-\beta})^2 (p_1^{-1} + p_2^{-1})}{(\nu \ln(\theta))^2}. \tag{4.6}$$

Heo et *al.* berechnen Fallzahlen mit unterschiedlichen Parameterwerten und vergleichen diese miteinander. θ, ν und die Länge der Rekrutierungsphase werden bei festen Werten von Power und mittlerer Überlebenszeit in Gruppe 1 variiert. Zwischen θ beziehungsweise ν und der Fallzahl bestehen jeweils starke positive Abhängigkeiten. Die Fallzahl ändert sich dagegen kaum, wenn die Länge der Rekrutierungsphase variiert wird, besonders bei großen ν. In Programmpaketen zur Fallzahlberechnung, die m_1, m_2, F und R verwenden, kann die Weibullverteilung benutzt werden, indem m_1, m_2 und F mit ν potenziert werden, wenn R null ist. Ist R ungleich null, können diese Größen ebenfalls mit ν potenziert werden. Das ist trotz Verletzung der Annahme einer gleichverteilten Rekrutierungsphase möglich, da die Fallzahl als wenig von der Größe von R abhängig betrachtet wird.
Porcher et *al.* (2002) [54] modellieren Crossover, indem Patienten, die in eine andere Behandlungsgruppe wechseln, die Hazardrate der neuen Gruppe annehmen. Somit ist auch in diesem Modell die Hazardrate eines Patienten nicht konstant über die Zeit (siehe Kapitel 4.3).

Bei Ahnn und Anderson (1998) [3] werden nichtproportionale Hazardraten auf eine andere Art modelliert. Die Studie wird wie bei Lakatos (1988) in einzelne Zeitabschnitte aufgeteilt und es können pro Abschnitt und Gruppe unterschiedliche Hazardraten definiert werden. Außerdem ist der Vergleich von mehr als zwei Behandlungsgruppen möglich (siehe Kapitel 4.4).
Das Markovmodell von Lakatos (2002) [38] ist auf beliebige Survivalverteilungen anwendbar, solange alle Hazardfunktionen durch stückweise lineare Funktionen modelliert werden können (siehe Kapitel 4.5).
Einen ähnlichen Ansatz wie Lakatos (1988) und Ahnn und Anderson (1998) wählen Barthel et *al.* (2006) [6]. Auch sie unterteilen die Studie in gleichlange Zeitabschnitte. Dadurch wird ermöglicht, die Anzahl der Patienten unter Risiko und die Ereignisanzahl für alle Gruppen in jedem Zeitabschnitt einzeln zu untersuchen. Außerdem können durch Modellierung der Survivalverteilungen für jeden einzelnen Zeitabschnitt nichtproportionale Hazardraten berücksichtigt werden. Dadurch ist beispielsweise auch über die Zeit variierender Crossover modellierbar.
Eng und Kosorok (2005) [19] verwenden die Supremum-Logrankstatistik, mit der im Gegensatz zum üblichen Logranktest von Schoenfeld (1983) [60] Unterschiede von Survivalkurven mit nicht proportionalen Hazardraten festgestellt werden können. Wenn sich beispielsweise die beiden Hazardraten überschneiden, die Survivalfunktionen aber geordnet bleiben, zeigt der Supremum-Logranktest im Gegensatz zum Standard-Logranktest einen Unterschied an. Bei Verwendung der Supremum-Logrankstatistik wird eine geringfügig größere Fallzahl benötigt als beim Standard-Logranktest. Man gewinnt jedoch an Power für eine große Anzahl von Alternativhypothesen. Eng und Kosorok schlagen vor, zuerst die Fallzahl mit Schoenfelds Formel 3.5 von 1981 zu berechnen und diese dann für die Supremumversion zu adjustieren. Diese Anpassung erfolgt unabhängig von der in Schoenfelds Formel gewählten Gewichtsfunktion. Die mit dieser Methode berechnete Supremum-Logrankstatistik besitzt im Fall proportionaler Hazardraten dieselbe Power wie die gewichtete Logrankstatistik.
Zhang und Quan (2009) [79] hingegen nehmen das Problem eines verzögerten Wirkungseintritts einer Behandlung zum Anlass, nichtproportionale Hazardraten zu modellieren. Denn tritt der Effekt einer Behandlung nicht unmittelbar nach Verabreichung ein, trifft die Annahme proportionaler Hazardraten nicht zu. Ist beispielsweise aus vorherigen Studien bekannt, dass die Wirkung erst nach mindestens einem Jahr eintritt, kann ein Modell mit einem zweistufigen Hazardratio angewendet werden. Zunächst wäre das Hazardratio zwischen experimenteller und Kontrollgruppe konstant bei eins und würde nach einem Jahr einen konstanten Wert unter eins annehmen, wenn die neue Behandlung wirksamer als die Kontrollbehandlung ist.

4.2. Verteilung der Rekrutierungsphase

Häufig wird angenommen, dass Patienten mit einer konstanten Rate während der Rekrutierungsphase in die Studie eingeschlossen werden. Zur Modellierung wird in diesem Fall eine Gleichverteilung verwendet (siehe Schoenfeld und Richter, 1982 [61]; Heo et *al.*, 1998 [28]; Schulgen et *al.*, 2005 [62]; Desseaux und Porcher, 2007 [16]). Da diese Annahme in der Praxis aber meist nicht zutreffend ist, wurden alternative Methoden entwickelt. So kann beispielsweise modelliert werden, dass ein Großteil der Patienten gleich zu Beginn eingeschlossen wird und die Rekrutierungsrate im Laufe der Studie immer weiter abnimmt.

Lachin und Foulkes (1986) [35] gehörten zu den ersten, die einen nicht gleichverteilten Eintritt der Patienten in die Studie modellierten. Sie benutzten eine trunkierte Exponentialverteilung als Rekrutierungsverteilung. In Yateman und Skene's (1992) [78] Ansatz folgt der Einschluss der Patienten einer stückweise linearen Funktion. Andere Methoden erlauben die Spezifikation eines beliebigen Rekrutierungsmusters (siehe beispielsweise Lakatos, 2002 [38], Chang, 2007 [11]). Denkbar sind außerdem schrittweise Rekrutierungsmuster (siehe Ahnn und Anderson, 1998 [3]; Li und Grambsch, 2006 [45]).
Barthel et *al.* (2006) [6] teilen die Studie in mehrere Zeitabschnitte auf und der Einschluss der Patienten in die Studie erfolgt schrittweise. Die Länge der Abschnitte hängt von den Kapazitäten der Studienzentren und von dem in der Planungsphase vorhandenen Wissen über das Eintrittsverhalten der Patienten ab. Sei $F^R(x)$ die kummulative Verteilungfunktion der Rekrutierungsphase. Dabei ist R die Anzahl der Zeitabschnitte, in denen Patienten rekrutiert werden, $R \leq T$, wobei T die Gesamtanzahl der Zeitabschnitte der Studie ist. Abhängig vom Einschlussmechanismus entspricht $F^R(x)$ beispielsweise der Verteilungsfunktion einer trunkierten Exponentialverteilung oder der einer Gleichverteilung. Die Rekrutierung der Patientenanzahl N basiert dann auf einem Exponential- oder Gleichverteilungsprozess. Außerdem darf $F^R(x)$ ein Punktmaß in 0 besitzen, so dass ein Anteil zwischen 0 und 100% der Patienten schon vor Beginn des ersten Zeitabschnitts in die Studie eingeschlossen werden kann.

4.3. Noncompliance, Crossover und Loss to Follow-up

In der Übersicht von Oellrich et *al.* (1997) werden keine Artikel aufgeführt, die sich Noncompliance, Crossover und Loss to Follow-up befassen.
Patienten, die die ihnen zugewiesene Behandlung nicht bis zum Studienende beibehalten, werden als noncompliant bezeichnet. Dabei kann zwischen Loss to Follow-up (Patienten scheiden aus der Studie aus) und Crossover (Patienten wechseln in eine andere Behandlungsgruppe, werden aber bis zum Ende der Studie weiterverfolgt) unterschieden werden. Crossover ist ein Überbegriff für Dropin und Dropout. Wechselt ein Patient der Kontrollgruppe zur Behandlung der experimentellen Gruppe, nennt man das Dropin, andersherum Dropout. Gründe für den Wechsel können beispielsweise ein Rückfall, kein erkennbarer Behandlungseffekt oder eine starke Verschlechterung der Krankheit sein. Da Crossover in der Praxis häufig vorkommt, wurden in den letzten Jahren einige Methoden zur Fallzahlanpassung entwickelt, da sich ohne die Anpassung die Power möglicherweise stark reduziert.
Ein anderer Aspekt ist die bei vielen Methoden zur Fallzahlberechnung getroffene Annahme, dass Patienten die Behandlung unabhängig von ihrem Risiko für einen Endpunkt nicht nach Anordnung durchführen oder abbrechen. Dies wird als nichtinformative Noncompliance bezeichnet. Allerdings deuten Ergebnisse etlicher veröffentlichter klinischer Studien darauf hin, dass oft eher kränkere Patienten die Studienbehandlung abbrechen als gesündere Patienten (Snapinn et *al.*, 2004 [68]). Hängt das Noncompliance-Verhalten der Patienten von ihrem Risiko für einen Endpunkt ab, wird dies als informative Noncompliance bezeichnet. Etwa ein Viertel aller verordneten Medikamente wird nicht oder nicht so wie vorgesehen eingenommen. Die Kosten der Noncompliance werden von der Bundesvereinigung Deutscher Apothekerverbände für das deutsche Gesundheitswesen auf etwa 10 Milliarden Euro jährlich geschätzt [1].

Bei Intention-to-Treat-Analysen wird die Fallzahl schon in der Designphase angepasst, um die nötige Power auch im Fall von Noncompliance zu erreichen. Die einfache Methode von Lachin und Foulkes (1986) [35] besteht darin, die Fallzahl zu erhöhen, indem sie durch $(1-nc_2-nc_1)^2$ geteilt wird. Dabei sind nc_2 und nc_1 die Anteile von Noncompliern in der experimentellen beziehungsweise in der Kontrollgruppe. Nach einem Behandlungswechsel nehmen Patienten die Hazardrate der Gruppe an, in die sie wechselten, und können nach einem Crossover nicht wieder in ihre ursprüngliche Behandlungsgruppe zurückkehren. Komplexere Methoden passen die Ereignisraten in jeder Behandlungsgruppe bezüglich bestimmter erwarteter Merkmale der Noncompliance an (siehe z.B. Lakatos, 1986 [36]). Lakatos entwickelte 1988 eine allgemeine Methode, bei der Crossover, Loss to Follow-up, Nicht-proportionale Hazardraten und eine nicht konstante Rekrutierungsverteilung berücksichtigt werden können. Im Gegensatz zu der Methode von Lachin und Foulkes (1986) können Patienten nach einem Crossover erneut in ihre ursprüngliche Behandlungsgruppe wechseln.
Die häufig getroffene Annahme, dass Noncomplier in die andere Behandlungsgruppe wechseln und deren Hazardratenrate übernehmen, ist oft unrealistisch. Viele Artikel zeigen, dass Noncomplier meist höhere Risiken für ein Ereignis besitzen als Complier (Coronary-Drug-Project-Research-Group, 1980 [13]; Snappin et *al.*, 2004 [68]). Erst in den letzten Jahren wurden Methoden entwickelt, die dies beachten.
Porcher et *al.* (2002) [54] erweitern in ihrem Artikel die auf dem Logranktest basierende Fallzahlformel von Freedman (1982) auf Berücksichtigung von Crossover und vergleichen ihren Ansatz mit dem von Lakatos (1988). Die Überlebenszeiten in den beiden Gruppen werden durch drei exponentialverteilte Zufallsvariablen X_1, X_2 und X_3 mit den Parametern λ_1, λ_2 und λ_3 modelliert. Dabei ist X_2 die Überlebenszeit in der experimentellen Gruppe. X_1 ist die latente Überlebenszeit und X_3 die Zeit bis zum Rückfall in der Kontrollgruppe. Ein Patient aus der Kontrollgruppe, der einen Rückfall hat, erhält ab diesem Zeitpunkt die experimentelle Behandlung und nimmt damit auch die Hazardfunktion λ_2 der experimentellen Gruppe an. Somit kann die tatsächliche Überlebenszeit $\widetilde{X}_1$ folgendermaßen ausgedrückt werden:

$$\widetilde{X}_1 = X_1 I_{[X_1<X_3]} + (X_3 + X_2^{'}) I_{[X_3<X_1]}. \tag{4.7}$$

Dabei ist $I_{[.]}$ die Indikatorfunktion, die den Wert 1 annimmt, wenn die Bedingung [.] zutrifft und anderenfalls 0 ist. $X_2^{'}$ ist die neue Überlebenszeit der Patienten, die nach einem Rückfall aus der Kontroll- in die experimentelle Gruppe wechselten, und ist mit Parameter λ_2 exponentialverteilt. Dann ist die Survivalfunktion der Kontrollgruppe

$$\begin{aligned} S_{\widetilde{1}}(x) &= 1 - F_{\widetilde{1}}(x) = 1 - P(X_{\widetilde{1}} \leq x) \\ &= \lambda_3 \exp[-\lambda_2 x] + \frac{(\lambda_1 - \lambda_2)\exp[-(\lambda_1+\lambda_3)x]}{\lambda_1 + \lambda_3 - \lambda_2} \end{aligned} \tag{4.8}$$

und die der experimentellen Gruppe ist $S_2(x) = 1 - F_2(x) = 1 - P(X_2 \leq x) = \exp(-\lambda_2 x)$.
Zur Bestimmung der Fallzahl werden sowohl der Logrank- als auch der Gehan-Wilcoxon-Test betrachtet und wie bei Freedman die erwartete Anzahl von Ereignissen verwendet. Die Follow-up-Periode wird in L Intervalle geteilt. Für jedes Intervall l, $l = 1, \ldots, L$, wird die Anzahl der Todesfälle d_l, der Anteil der Patienten unter Risiko kurz vor dem j-ten Ereignis in Gruppe 1 im Verhälnis zu Gruppe 2, ϕ_{lj}, und das Hazardratio θ_{lj} kurz vor den j-ten Ereignis im l-ten

Intervall bestimmt. Sei ω_{lj} die Überlebensrate eines Patienten kurz vor dem $j$ten Ereignis im $k$ten Intervall. Unter Annahme, dass die Varianz der Tarone-Ware-Klasse von Teststatistiken wie bei Freedman (1982) und Lakatos (1988) approximativ 1 ist und $\phi_{lj} \equiv \phi_l$, $\theta_{lj} \equiv \theta_l$ und $\omega_{lj} \equiv \omega_l$ in jedem der L Intervalle gleich sind, erhält man die Gesamtzahl von Ereignissen für den Logranktest d bzw. für den Gehan-Wilcoxon-Test d_W. Die Fallzahlformel, die Crossover berücksichtigt, ist für den Logranktest

$$N = \frac{d}{2 - S_{\tilde{1}}(T) - S_2(T)}, \tag{4.9}$$

und für den Gehan-Wilcoxon-Test

$$N_W = \frac{d_W}{2 - S_{\tilde{1}}(T) - S_2(T)}. \tag{4.10}$$

Die Survivalverteilungen werden hier am Studienende T bestimmt.
Für die korrigierte Fallzahlformel wird zusätzlich zu den auch bei Freedman benötigten Parametern $S_1(T)$, $S_2(T)$ und T, die zu λ_1 und λ_2 führen, nur die Survivalrate S_3 zur Analysezeit T benötigt, da $\lambda_3 = (-lnS_3)/T$. Außerdem muss die Anzahl der Intervalle L festgelegt werden. In Simulationen hat sich gezeigt, dass ein Wert von $L = 50$ in den meisten Fällen zu guten Ergebnissen führt. Die Methode kann auch auf Patienten erweitert werden, die die Behandlung in der experimentellen Gruppe abbrechen und in die Kontrollgruppe wechseln. Außerdem ist es möglich, die Lebensdauer von Patienten, die die Behandlung wechseln, durch eine andere Survivalverteilung zu beschrieben als die der neuen Behandlungsgruppe. Im Fall eines Crossovers von der Kontroll- in die experimentelle Gruppe ist dann der einzige Unterschied, dass $T_2^{'}$ eine exponentialverteilte Zufallsvariable mit beliebigem Parameter $\lambda_2^{'}$ statt λ_2 ist, dem Parameter der Survivalverteilung in der experimentellen Gruppe.
Auch Jiang et *al.* (2004) [32] nehmen an, dass ein Patient nach einem Behandlungswechsel nicht wieder zu seiner ursprünglichen Behandlung zurückkehren kann. Sie verändern die Methode von Lakatos (1988) so, dass die Fallzahlen unter Berücksichtigung informativer Noncompliance berechnet werden können. Patienten, die die ihnen zugeteilte Studienbehandlung abbrechen, werden hier im Follow-up weiterbeobachtet und in eine Intention-to-Treat-Analyse einbezogen. Der Behandlungseffekt kann abgeschwächt sein, da bei den Noncompliern sehr wahrscheinlich kein Effekt beobachtet werden kann.
Die Idee der Methode beruht auf zwei Subpopulationen gleicher Größe. Die eine Subpopulation ist gesünder und hat niedrigere Ereignis- und Abbruchraten, die andere Subpopulation ist kränker mit hohen Ereignis- und Abbruchraten. Die Ereignis- und Abbruchzeiten sind exponentialverteilt. Die ursprüngliche Lakatosmethode verwendet pro Behandlungsgruppe ein Markovmodell. Bei Jiang et *al.* wird für jede der beiden Teilstichproben innerhalb jeder Behandlungsgruppe ein seperates Markovmodell definiert. Mit diesen Modellen werden die Anzahl von Patienten unter Risiko zu Beginn und die Ereignisanzahl innerhalb jedes Intervalls berechnet. Anschließend werden damit die Ergebnisse zusammengefasst und mit der Tarone-Ware-Statistik die Anzahl benötigter Ereignisse bestimmt. Die gesamte Fallzahl N wird mittels der Logrankstatistik berechnet (Lakatos (1988) [37]):

$$N = \frac{[(\sum \rho_i \eta_i)^{1/2}(z_{\alpha/2} + z_\beta)/(\sum \rho_i \gamma_i)]^2}{Q_1\pi_1 + Q_2\pi_2}. \tag{4.11}$$

Dabei ist $\rho_i = d_i/(\sum d_i)$ mit der Ereignisanzahl im $i$ten Intervall d_i, $\eta_i = \varphi_i/(1+\varphi_i)^2$ mit dem Verhältnis φ_i der Patienten unter Risiko im $i$ten Intervall in den beiden Behandlungsgruppen und $\gamma_i = \varphi_i\theta_i/(1+\varphi_i\theta_i) - \varphi_i/(1+\varphi_i)$ mit dem Hazardratio im $i$ten Intervall θ_i. π_1 und π_2 sind die kumulativen Ereignisraten und Q_1 und Q_2 die Anteile der Patienten in der jeweiligen Behandlungsgruppe. ρ_i, η_i und γ_i können aus dem Markovmodell bestimmt werden. Wie bei den meisten Methoden zur Fallzahlberechnung wird hier angenommen, dass Patienten, die die eine Behandlung abbrechen, in die andere Behandlungsgruppe wechseln. Dabei übernehmen sie das Risiko für ein Ereignis der neuen Behandlungsgruppe. Da diese Annahme unrealistisch sein könnte, wird in dem Artikel eine weitere Variante untersucht, bei der sich die Ereignisraten der Noncomplier beider Behandlungsgruppen einem gemeinsamen Wert annähern. Die gemeinsame Ereignisrate kann beipielsweise der Durchschnitt der Ereignisraten in den beiden Behandlungsgruppen sein.
Durch Einführung zusätzlicher Zustände in dem Markovmodell ist eine Berücksichtigung eines verzögerten Wirkungseintritts der Behandlungen und eine verzögerte Wirkungsabnahme nach einem Behandlungsabbruch möglich. Um eine schrittweise Rekrutierung der Patienten zu modellieren, wird angenommen, dass die Studie aus Kohorten mit unterschiedlicher Studiendauer besteht. Für jede Kohorte wird die benötigte Fallzahl mittels der modifizierten Lakatos-Methode berechnet. Dabei basieren Ereignis- und Abbruchraten auf einem Exponentialmodell.
Barthel et *al.* (2006) [6] beschreiben ein System zur Berechnung der Fallzahl bei Survivalstudien, in dem zwei oder mehr Behandlungsgruppen miteinander verglichen werden können, die Rekrutierungsphase nicht gleichverteilt sein muss, auch nicht proportionale Hazardraten möglich sind und Loss to Follow-up und Crossover berücksichtigt werden können. Die Studie wird in mehrere gleichlange Zeitabschnitte unterteilt. So kann die Survivalverteilungen für jeden Zeitabschnitt getrennt modelliert werden, wodurch es möglich ist, über die Zeit variierenden Crossover zu berücksichtigen. Die Patienten können entweder in eine andere Behandlungsgruppe der Studie oder zu einer ganz anderen Behandlung wechseln. Im Gegensatz zu der Methode von Lakatos (1988) dürfen Patienten wie bei Porcher et *al.* (2002) und Jiang et *al.* (2004) nach einem Behandlungswechsel später nicht wieder zu ihrer ursprünglichen Behandlung zurückkehren.
Durch die konservative Annahme, dass kein erneuter Wechsel zur ursprünglichen Behandlung möglich ist, kann die Survivalverteilung $S_k^E(x)$ für die Gruppen $k = 1, 2, \ldots, K$ adjustiert für Crossover berechnet werden. Seien x_E und x_C die Zeiten bis zum Ereignis bzw. bis zum Crossover mit zugehörigen Survivalfunktionen $S_k^E(x)$ und $S_k^C(x)$ und Dichten $f_k^E(x)$ und $f_k^C(x)$. Wenn Crossover zum Zeitpunkt x_C auftritt, ist die bedingte Hazardfunktion

$$\lambda_x(x_E|x_C) = \begin{cases} \lambda_b(x), & x < x_C \\ \\ \lambda_a(x), & x \geq x_C \end{cases}$$

wobei $\lambda_a(x)$ und $\lambda_b(x)$ die jeweiligen Hazardfunktionen vor und nach dem Behandlungswechsel

sind. Dann gilt

$$
\begin{aligned}
S_k^E(x) &= \int_0^\infty P\left\{X \geq x | x_C\right\} f_k^C(x_C) dx_C \\
&= \int_0^\infty \exp\left[-\int_0^x \lambda_x(u|x_C)du\right] f_k^C(x_C)dx_C \\
&= S_0^E(x)S_k^C(x) + \int_0^x \exp\left[-\int_0^{x_C} \lambda_b(u)du - \int_{x_C}^x \lambda_a(u)du\right] f_k^C(x_C)dx_C, \qquad (4.12)
\end{aligned}
$$

wobei $S_0^E(x)$ die Survivalfunktion ohne Berücksichtigung von Crossover ist, das heißt $S_0^E(x) = P\left\{X \geq x | x_C = \infty\right\}$. Die Annahme stückweiser Exponentialverteilungen erleichtert die Berechnung der Integrale. Die für Crossover adjustierte Survivalverteilung $S_k^E(t)$ wird zur Berechnung des Nichtzentralitätsparameters τ benötigt, der in die Formel zur Bestimmung der Fallzahl eingeht (siehe Barthel et *al.* (2006) in Kapitel 4.4).
Außerdem kann bei Barthel et *al.* (2006) [6] Loss to Follow-up berücksichtigt werden. Es wird angenommen, dass die Zeit bis Loss to Follow-up unabhängig von der Survivalzeit ist. Seien $S_k^L(x)$ und $S_k^E(x)$ die Survivalfunktionen der Zeit bis zum Ausscheiden aus der Studie beziehungsweise bis zum Ereignis mit $k = 1, 2, \ldots, K$. $S_k^E(x)$ sei bereits adjustiert für Crossover. Beide Survivalfunktionen können durch stückweise Exponentialverteilungen mit Hazardfunktionen ϵ_{ki} beziehungsweise μ_{ki} approximiert werden, wobei k für die Behandlung und i für den Zeitraum steht und $x \in [0, T]$. Die zu $S_k^L(x)$ gehörige Dichtefunktion für die Zeit bis Loss to Follow-up kann folgendermaßen ausgedrückt werden:

$$
f_k^L(x) = \begin{cases} \mu_{k1} \exp\{-\mu_{k1}x\}, & 0 < x \leq 1 \\ \\ \mu_{ki} \exp\left\{\sum_{j=1}^{i-1}[j(\mu_{k,j+1} - \mu_{kj})] - \mu_{ki}x\right\}, & i-1 < x \leq i, \quad i = 2, \ldots, T. \end{cases}
$$

Die Dichtefunktion für die Zeit bis zum Ereignis wird ähnlich definiert. Sei nun $F^R(s)$ die Rekrutierungsverteilung mit dem Eintrittszeitpunkt s. Dann kann durch $F^R(T-s)$ die Verteilung der Überlebenszeit ausgedrückt werden, wobei $T-R < T-s < T$. Zur Berechnung der Zensierungswahrscheinlichkeit, die für die Fallzahlformel benötigt wird, muss der Anteil der Patienten, die in Behandlungsgruppe k im Laufe der Studie ein Ereignis haben, bestimmt werden.
Der Ansatz von Barthel et *al.* (2006) wurde in dem frei verfügbaren Paket ART für Stata implementiert.
In den bisher betrachteten Artikeln werden zwar über die Zeit variierende Noncompliancerateten betrachtet, jedoch keine über die Zeit variierende Beziehung zwischen Noncompliance und Risiko für den primären Endpunkt. Li und Grambsch (2006) [45] stellen in ihrem Artikel eine auf Lakatos' Markovkettenmodell basierende Methode vor, die diese zeitvariierende Beziehung getrennt für jede Behandlungsgruppe in die Fallzahlberechnung einbezieht. Auch hier wird Noncompliance so definiert, dass Patienten die Studienmedikation dauerhaft abbrechen und nicht

wie bei Lakatos' Methode später wieder zu ihrer ursprünglichen Behandlung zurückkehren dürfen.
Es werden vier Szenarien über die Ereignisrate nach Studienabbruch betrachtet. Szenario 1 ist das aus anderen Artikeln bekannte, in dem Noncomplier das Risiko der anderen Gruppe annehmen. In Szenario 2 dagegen nehmen Noncomplier, unabhängig von ihrer Gruppenzugehörigkeit, nach Studienabbruch eine allgemeine Ereignisrate an. Sowohl in Szenario 3 als auch in Szenario 4 wird eine zeitliche Abhängigkeit zwischen Studienabbruch und Risiko angenommen. In Szenario 3 werden zwei unterschiedliche Ereignisraten für frühe und späte Studienabbrecher definiert. Dabei ist die Ereignisrate derjenigen, die ihre Studienmedikation frühzeitig abbrechen, meist ähnlich der der Kontrollgruppe und die der späten Abbrecher wesentlich größer. Szenario 4 ist eine Kombination der ersten beiden Szenarien. Die frühen Studienabbrecher nehmen die Ereignisrate der anderen Gruppe an, die Späten eine allgemeine hohe Ereignisrate.
Die Erweiterung von Lakatos' Markovkettenmodell besteht in der Einführung zweier Zustände „erhält die zugewiesene Studienbehandlung" und W „hat die zugewiesene Behandlung abgebrochen, hatte aber noch kein Ereignis" an Stelle der beiden Zuständen „erhält die experimentelle Behandlung" und „erhält die Kontrollbehandlung". Unter Szenario 3 und 4 wird wiederum der Zustand W in zwei Zustände aufgespalten, die angeben, ob ein Patient die Studienbehandlung vor oder nach einem bestimmten Cutpoint abgebrochen hat. So können die Übergangsmatrizen entsprechend den unterschiedlichen Beziehungen zwischen früher beziehungsweise später Noncompliance und dem Risiko für ein Ereignis konstruiert werden.
Zur Überprüfung des Zusammenhangs zwischen Noncompliance und Ereignisrisiko werden Fallzahlen für den zweiseitigen Logranktest mit $\alpha = 0.05$, $\beta = 0.15$, 20 Teilintervallen und keinem Loss to Follow-up berechnet. Bei Szenario 4 ist die Fallzahl umso höher, je größer die Ereigniswahrscheinlichkeit der späten Studienabbrecher ist. Außerdem steigt die Fallzahl stark an, wenn der Cutpoint nah am Studienende liegt. Auch unter Szenario 3 wächst die Fallzahl mit der Ereigniswahrscheinlichkeit später Studienabbrecher. Anders als bei Szenario 4 ändert sich die Fallzahl nur wenig, wenn der Cutpoint variiert wird. Die Fallzahl ändert sich also abhängig von der Beziehung zwischen Noncompliance und Risiko für ein Ereignis. Man kann die Methode leicht auf mehrere Cutpoints erweitern, indem man weitere Zustände in das Markovmodell aufnimmt. Auch wenn Loss to Follow-up, schrittweise Rekrutierung, verzögerter Wirkungseintritt und zeitabhängige Ereignis-, Abbruch- und Loss-to-Follow-up-Raten hier nicht behandelt werden, können sie leicht einbezogen werden, da sie schon in der zugrundeliegenden Methode von Lakatos berücksichtigt werden.
Durch Einführung weiterer Zustände in das Markovmodell ist es möglich, auch komplexere zeitliche Zusammenhänge zwischen Noncompliance und Ereignisrisiko zu modellieren.

4.4. Vergleich von mehr als zwei Behandlungen

Fast alle Methoden zur Fallzahlschätzung bei Daten mit Survivalendpunkt können ausschließlich zum Vergleich von nur zwei Behandlungsgruppen verwendet werden. In Oellrich et *al.* (1997) [52] findet sich, dass Ahnn und Anderson (1995) [2] den Ansatz von Schoenfeld (1981) [59] auf den K-Gruppen-Vergleich verallgemeinern. Die Methode ist jedoch eingeschränkt auf den Fall proportionaler Hazardraten und einer einheitlichen Zensierungsverteilung in allen Behand-

lungsgruppen. Außerdem können weder Noncompliance noch Dropin berücksichtigt werden. Die folgenden Artikel erweitern den Ansatz von Lakatos (1988) [37] auf den Vergleich mehrerer, auch teilweise unterschiedlich großer, Gruppen.
Bei Ahnn und Anderson (1998) [3] können wie bei Lakatos (1988) [37] nichtproportionale Hazardraten, zeitabhängige Loss-to-Follow-up-Raten, Noncompliance und Drop-in berücksichtigt werden. Zusätzlich ist der Vergleich von mehr als zwei Behandlungsgruppen möglich. Jeder der N Patienten wird in eine von K gleichgroßen Gruppen randomisiert. Die Patienten wechseln zwischen den Zuständen E (hatte ein Ereignis), L (ist Lost to Follow-up), T_j (erhält Behandlung j) und T_{j_*} (wechselte von Behandlung j zu Behandlung j_* aufgrund von Noncompliance oder Drop-in), $j, j_* = 1, \ldots, K$. Der Vektor D_x gibt die Wahrscheinlichkeiten an, mit denen sich ein Patient zum Zeitpunkt x in den Zuständen $(L, E, T_1, T_2, \ldots, T_K)$ befindet. Gegeben sei die Übergangsmatrix $trans_{i,i+1} = trans_{i,i+1}(j_1, j_2)$ mit der Wahrscheinlichkeit $trans_{i,i+1}(j, j_*)$, im Intervall $[x_i, x_{i+1}]$ von Zustand j zu Zustand j_* zu wechseln. Dann kann D_i rekursiv berechnet werden durch $D_i = T_{i-1,i} D_{i-1}$ für $i \leq n$, wobei x_n das Studienende darstellt. So entsteht für jede Behandlungsgruppe eine Folge von Zustandsverteilungen $\{D_i, i = 0, \ldots, n\}$.
Zur Berechnung der Fallzahl wird eine Teststatistik $Z = U' V^{-1} U$ benötigt. U und V werden aus den Ereignisanzahlen M_{jl} in Gruppe j zum Zeitpunkt x_l, den Personen unter Risiko R_{jl} in Gruppe j kurz vor Zeitpunkt x_l, der Gesamtzahl R_l der Personen unter Risiko kurz vor x_l und einer Menge von Gewichten ω_l berechnet:

$$U_j = \sum_{l=1}^{d} \omega_l \left(M_{jl} - \frac{R_{jl}}{R_l} \right) \tag{4.13}$$

und

$$V_{jq} = \sum_{l=1}^{d} \omega_l^2 \left(\frac{R_{jl}}{R_l} \right) \left(\delta_{jq} - \frac{R_{ql}}{R_l} \right), \tag{4.14}$$

wobei $\delta_{jq} = 0$, wenn $j \neq q$ und $\delta_{jq} = 1$, wenn $j = q$. Daraus werden der Vektor $U = (U_2, \ldots, U_k)'$ und die $(k-1) \times (k-1)$-Matrix

$$V = (V_{jq})_{\substack{j=2,\ldots,k \\ q=2,\ldots,k}}$$

berechnet. Je nach Definition der Gewichte erhält man den Logranktest, den Gehan-Breslow-Test oder einen anderen Test der Klasse von Tarone und Ware (1977) [69]. Die Teststatistik Z ist unter lokalen Alternativen approximativ χ^2-verteilt mit $K-1$ Freiheitsgraden und Nichtzentralitätsparameter τ.
Nach Bestimmung des Nichtzentralitätsparameters unter einer festen lokalen Alternative kann daraus die Fallzahl berechnet werden. Unter Verwendung von Lakatos (1988) und Ahnn und Anderson (1995) erhält man $\tau = E(U') V^{-1} E(U)$. Siehe Ahnn und Anderson (1998) [3] zur Berechnung des Erwartungswertes. Ist die Studie in genügend feine Zeitintervalle unterteilt, können Annahmen getroffen werden, um der Nichtzentralitätsparameter durch

$$\tau = d \Theta' V_*^{-1} \Theta \tag{4.15}$$

berechnen zu können, wobei d die Gesamtzahl der Ereignisse in der Studie ist, $\Theta = (1/d)E(U)$ und $V_* = (1/d)V$. Die zur Berechnung von Θ und V_* benötigten Parameter werden aus den Folgen der Vektoren der Zustandswahrscheinlichkeiten unter dem Markovmodell bestimmt. So kann d aus der Definition von τ und daraus wiederum die Fallzahl

$$N = \frac{Kd}{\sum_{j=1}^{K} P_{E_j}} \tag{4.16}$$

mit der kumulativen Ereignisrate P_{E_j} für Behandlungsgruppe j am Studienende berechnet werden.
Halabi und Singh (2004) [24] erweitern den Ansatz von Ahnn und Anderson (1995) insofern, dass Behandlungsgruppen ungleicher Größe vorliegen dürfen. Voraussetzung sind proportionale Hazardraten und identische Zensierungsverteilungen in allen K Behandlungsgruppen. Die Berücksichtigung von Kovariablen ist möglich, wird hier jedoch nicht behandelt.
Die Logrankstatistik

$$LR = U_0' V_0^{-1} U_0 \tag{4.17}$$

besteht aus der Differenz U_0 der beobachteten abzüglich der erwarteten Ereignisse und der Kovarianzmatrix V_0 von U_0. Die Teststatistik ist unter H_0 asymptotisch χ^2-verteilt mit $K-1$ Freiheitsgraden. Sei das Loghazardratio definiert als $\log\theta_k = \log(\lambda_k/\lambda_1)$, $k = (2,\ldots,K)$.
Konvergiert eine Folge von Alternativhypothesen gegen die Nullhypothese und ist $\log\theta_k$ von der Ordnung $O(N^{-1/2})$, ist die Logrankstatistik unter H_0 nichtzentral χ^2-verteilt mit $K-1$ Freiheitsgraden und Nichtzentralitätsparameter τ. Durch Umstellen der Definition des Nichtzentralitätsparameters erhält man die Formel

$$N = \tau \left\{ (1-c) \left[\sum_{j=2}^{K} P_j(1-P_j)\log\theta_j^2 - 2\sum_{j=2}^{K}\sum_{\substack{l=2 \\ j<l}}^{K} P_j P_l \log\theta_j \log\theta_l \right] \right\}^{-1} \tag{4.18}$$

zur Berechnung der Fallzahl, wobei P_j der Anteil der Patienten ist, die Gruppe j zugeteilt sind, und c der Gesamtanteil von Zensierungen ist.
Auch bei Barthel et *al.* (2006) [6] werden die Survivalverteilungen von K Behandlungsgruppen mittels eines (gewichteten) Logranktests miteinander verglichen. Die Studie wird in gleichlange Zeitabschnitte unterteilt und pro Abschnitt werden die Anzahl der Patienten unter Risiko und die Anzahl der Ereignisse für jede Gruppen einzeln untersucht. Die Hypothese der Gleichheit der K Survivalverteilungen kann durch

$$H_0 : \lambda_1(x) = \lambda_2(x) = \ldots = \lambda_K(x)$$

ausgedrückt werden, wobei $\lambda_k(x)$ die Hazardfunktion zum Zeitpunkt x in Behandlungsgruppe k $(k = 1,\ldots,K)$ ist. Als Alternativhypothese wird angenommen, dass sich mindestens zwei der Hazardraten unterscheiden. Sei das Loghazardratio wie im vorherigen Artikel definiert als $\log\theta_k = \log(\lambda_k/\lambda_1)$, $k = (2,\ldots,K)$.
Der Logranktest zum Vergleich der $k$ten $(k = 2,...K)$ mit der ersten Gruppe basiert auf der Verteilung der Differenz U_k der beobachteten abzüglich der erwarteten Ereignisanzahl unter H_0.

Dabei können Gewichte nach Tarone und Ware (1977) [69] oder Harrington und Fleming (1982) [26] gewählt werden. Die globale Teststatistik Q wird aus dem Vektor $U = (U_2, \ldots, U_K)'$ und der Kovarianzmatrix von U unter H_0, $V(0)$, berechnet:

$$Q = U'V(0)^{-1}U. \tag{4.19}$$

Q ist zentral χ^2-verteilt mit $K-1$ Freiheitsgraden, da die Verteilung von U unter H_0 asymptotisch $N(0, V(0))$ ist. Nun wird die Fallzahl berechnet. Dazu wird wie in den vorigen Artikeln eine Folge lokaler Alternativen zur Nullhypothese betrachtet, wie beispielsweise, dass $\log \theta_k$ von der Ordnung $O(N^{-1/2})$ ist (Schoenfeld, 1981) [59]. Q folgt dann einer nichtzentralen χ^2-Verteilung mit $K-1$ Freiheitsgraden (Cox, 1972) [14] und Nichtzentralitätsparameter τ, aus dessen Definition die Fallzahl berechnet wird. Die Bestimmung von τ ist im Anhang von Barthel et al. (2006) [6] dargestellt. Für den Logranktest ergibt sich

$$N = K\tau\psi \left[\frac{K-1}{K} \sum_{k=2}^{K} (\log \theta_k)^2 - \frac{2}{K} \sum_{k=2}^{K} \sum_{2 \le k \le q}^{K} log\theta_k log\theta_q \right]^{-1}, \tag{4.20}$$

wobei ψ die Wahrscheinlichkeit ist, dass bis zum Ende der Studie keine Zensierung eintritt (Ahnn und Anderson, 1995 [2]).

4.5. Gruppensequentielle Designs

In einem gruppensequentiellen Design werden die Daten nicht erst nach Studienende analysiert, sondern es können schon im Laufe der Studie in sogenannten Interims- oder Zwischenanalysen Teststatistiken berechnet werden. Die Teststatistiken werden mit vorher festgelegten Abbruchgrenzen verglichen. Dadurch bietet sich die Möglichkeit, eine Studie frühzeitig abzubrechen, wenn die Wirksamkeit oder Unwirksamkeit eines Medikaments schon in einer Zwischenanalyse nachweisbar ist. In diesen Fällen ist es unethisch, Patienten aus der Kontrollgruppe das wirksame neue Medikament vorzuenthalten beziehungsweise Patienten aus der experimentellen Gruppe weiter die unwirksamere neue Behandlung zu verabreichen. Mögliche andere Adjustierungen sind eine Neuberechnung der Fallzahl oder die Veränderung der Studiendauer.
Lan und DeMets (1983) [41] entwickelten den Ansatz der Alpha Spending Function, aus der die Abbruchgrenzen für die Zwischenanalysen während der Studie berechnet werden können. Auf diese Weise wird das Problem behoben, dass in der Planungsphase noch nicht bekannt ist, wieviele Ereignisse in den einzelnen Phasen des sequentiellen Designs auftreten werden. Das Signifikanzniveau α wird gemäß der Alpha Spending Function stückweise in jeder Zwischenanalyse verwendet.
Gu und Lai (1999) [23] entwickelten zwei auf Monte Carlo Simulationen basierende Programme für sequentielle Designs. Die Simulationen erweitern das Programm von Halpern und Brown (1993) [25], so dass Zensierungen, Crossover bzw. Noncompliance und unterschiedliche Rekrutierungsraten berücksichtigt werden können, zwischen vier Arten von Abbruchgrenzen gewählt und eine Teststatistik der Betafamilie benutzt werden kann. Die von Self (1991) [63] vorgeschlagene Betafamilie enthält die Teststatistiklasse von Harrington und Fleming (1982) [26].

Spezialfälle sind die Logrank- und die verallgemeinerte Wilcoxonstatistik. Es kann zwischen der Slud-Wei- (1982) [67], der Lan-DeMets- (1983) [41] und der von Gu und Lai in diesem Artikel neu vorgeschlagenen, auf Haybittle (1971) [27] basierenden, Abbruchgrenze und einer vom Benutzer selbst spezifizierten Abbruchgrenze gewählt werden. In einem Beispiel vergleichen Gu und Lai die mit ihrem Simulationsprogramm erhaltene Power des Logranktests zweier Studien mit Loss to Follow-up und Noncompliance mit den Ergebnissen von Lakatos (1988) und erhalten Werte zwischen 0.88 und 0.93.
Lakatos (2002b) [39] geht auf Gu und Lai (1999) ein, da sie in ihrem Artikel Simulationen basierend auf seinem 1998 erschienenen Artikel durchführen. Obwohl die Power bei Lakatos (1988) in allen Fällen ungefähr bei 90% liegt, schwankt sie in den Simulationen von Gu und Lai zwischen 88 und 93%. Dies ist ein überraschendes Resultat für Lakatos, denn in seinem 1992 gemeinsam mit Lan verfassten Artikel [40] wurde die Methode von 1988 in Simulationen für verschiedene Situationen überprüft und die Power lag dabei ausschließlich zwischen 89.1 und 90.9%. Zur weiteren Überprüfung seiner Methode wurde daraufhin erneut eine Simulation mit dem unabhängigen Programm nQuery durchgeführt, wobei die Power wiederum sehr nah bei 90% lag. Durch die wiederholte Überprüfung zeigt Lakatos, dass erstens das Simulationsprogramm von Lakatos und Lan korrekt funktioniert und dass zweitens die Methode von Lakatos zur Fallzahlberechnung korrekte Ergebnisse liefert.
Für Lakatos (2002a) ist der Begriff der „Informationszeit“ (Lan und Zucker,1993 [42]) wichtig, um sequentielle Designs bei Survivaldaten anzuwenden. Information wird als Kehrwert der Varianz des aktuellen Schätzwerts des Parameters definiert, der die Effizienz der neuen Behandlung misst. Bei Survivaldaten ist die Information

$$\left(\frac{1}{r_2}+\frac{1}{r_1}\right)^{-1}\frac{d}{(r_2+r_1)}, \tag{4.21}$$

wobei d die Gesamtzahl der Ereignisse und r_2 und r_1 die Personen unter Risiko in der experimentellen und der Kontrollgruppe sind (Lan und Zucker, 1993 [42]). Die Information verhält sich proportional zur Ereignisanzahl und ist somit eine komplexe Funktion der Survivalrate, des Rekrutierungsmusters und anderer Faktoren. Als Informationszeit wird der Anteil der aktuellen Information an derjenigen Information bezeichnet, die bei nicht vorgezogener Beendigung der Studie vorhanden wäre.
Früher mussten in gruppensequentiellen Designs im Voraus Zeitpunkte für die Zwischenanalysen und zugehörige kritische Werte festgelegt werden. Diese Zeitpunkte wurden in Form von Informationszeiten und nicht Kalenderzeiten angegeben. Lan und DeMets (1983) [41] entwickelten einen Ansatz, bei dem die Interimsanalysen zu Kalenderzeiten stattfinden und nicht wenn eine bestimmte Anzahl an Ereignissen aufgetreten ist. Sie definieren eine Alpha Spending Function $\alpha^*(u)$, die angibt, wieviel des globalen Signifikanzniveaus α zu den einzelnen Zwischenanalysen verbraucht wird. Wenn zu den Kalenderzeiten $a_1,\ldots,a_I$ Analysen durchgeführt werden, ist mit den zugehörigen Informationszeiten $u_1,\ldots,u_I$ das in der i-ten Interimsanalyse verbrauchte α gleich

$$\alpha_i=\alpha^*(u_i)-\alpha^*(u_{i-1}). \tag{4.22}$$

Die zu $a_1,\ldots,a_I$ gehörigen Informationszeiten $u_1,\ldots,u_I$ können mittels des Markovmodels von Lakatos (1986) [36] berechnet werden. Mit dem Markovmodell werden der Mittelwert und

die Varianz der Logrankstatistik bei jeder Interimsanalyse bestimmt. Es wird die Nullhypothese $H_0 : (1 - S_1(x)) = (1 - S_2(x))$ gegen die Alternative $H_1 : (1 - S_1(x)) \neq (1 - S_2(x))$ getestet, wobei $S_1(x)$ und $S_2(x)$ die Survivalfunktionen der beiden Behandlungsgruppen sind. Die gewichtete Logrankstatistik der $i$ten Zwischenanalyse wird nach Schoenfeld (1981) berechnet als LR^i_ω, die der $(i-1)$ten ist LR^{i-1}_ω. Dann seien die Zuwächse der gewichteten Logrankstatistik der einzelnen Phasen

$$\mu^i = E(LR^i_\omega) - E(LR^{i-1}_\omega). \tag{4.23}$$

R_i sei der Anteil der bei der $i$ten Interimsanalyse rekrutierten Patienten. q_2 sei die kumulative Wahrscheinlichkeit, der experimentellen Gruppe zugeteilt zu werden, und p^i_2 die kumulative Wahrscheinlichkeit, bei der $i$ten Interimsanalyse in dieser Gruppe ein Ereignis zu haben. Entsprechend sind q_1 und p^i_1 für die Kontrollgruppe definiert.
Außerdem sei

$$\gamma^i_h = \frac{\varphi^i_h \theta^i_h}{1 + \varphi^i_h \theta^i_h} - \frac{\varphi^i_h}{1 + \varphi^i_h}. \tag{4.24}$$

Dabei ist φ^i_h das Verhältnis der Patienten unter Risiko und θ^i_h das Verhältnis der Hazardraten in den beiden Behandlungsgruppen kurz vor dem $h$ten Ereignis.
Mit diesen aus den Markovmodell berechenbaren Parametern lässt sich die Fallzahl N aus der folgenen Gleichung bestimmen:

$$\begin{aligned} \frac{\mu^i}{N} = {} & (R_i(q_2 p^i_2 + q_1 p^i_1)) \sum_{l=1}^{a_i} \gamma^i_l \\ & - (R_{i-1}(q_2 p^{i-1}_2 + q_1 p^{i-1}_1)) \sum_{h=1}^{a_{i-1}} \gamma^{i-1}_h \end{aligned} \tag{4.25}$$

4.6. Adaptive Designs

In klinischen Studien wird gewöhnlich eine im Voraus festgelegte Patientenanzahl rekrutiert. Die Bestimmung dieser Fallzahl hängt von Parametern wie dem erwarteten Behandlungseffekt, der angenommenen Variabilität des Endpunkts und den a priori festgelegten Fehlerraten erster und zweiter Art ab. Während der letzten beiden Jahrzehnte wurden gruppensequentielle Designs weitverbreitet genutzt. Sie ermöglichen einen vorgezogenen Studienabbruch, wenn in einer Zwischenanylse ein begründeter Hinweis auf die Effizienz oder Ineffizienz der neuen Behandlung vorliegt, während immer noch die Fehlerraten erster und zweiter Art kontrolliert werden. Allerdings beruhen solche Designs noch auf der vorherigen Festlegung der oben erwähnten Schlüsselparameter. Folglich kann eine Fehlspezifikation dieser Parameter zu unterpowerten Studien oder unnötig hohen Fallzahlen führen.
Um diesem Problem zu begegnen, wurden in der Folge adaptive Designs entwickelt (Bauer und Köhne, 1994 [8]). Sie ermöglichen Designmodifikationen bei Interimsanalysen, während die globalen Fehlerraten kontrolliert werden. Häufige Designänderungen oder -anpassungen sind eine Neuschätzung der Fallzahl, die Wahl einer passenderen Teststatistik, das Einfügen oder Überspringen von Interimsanalysen, und das Verwerfen oder Hinzufügen von Behandlungsgruppen.

Ein großer Vorteil adaptiver Designs ist, dass diese Änderungen nicht im Vorhinein festgelegt werden müssen (Desseaux und Porcher, 2007 [16]).
Lawrence (2002) [43] beispielsweise behandelt die Änderung der Teststatistik in einem adaptiven Design. Wenn bei einer Zwischenanalyse festgestellt wird, dass eine andere Teststatistik mehr Power für die Studie liefern würde, kann zukünftig eine andere Teststatistik aus der Klasse von Fleming und Harrington verwendet werden. Die Methode kann auf Daten verschiedener Endpunkte angewendet werden, auch auf Survivaldaten.
Bisher werden adaptive Designs allerdings fast ausschließlich zur Fallzahladjustierung verwendet. Andere mögliche Designmodifikationen werden in der medizinischen Praxis erst selten angewendet. Die meisten Artikel über die Nutzung adaptiver Designs in konkreten Studien stammen aus Deutschland, da dort früh und umfassend auf diesem Gebiet geforscht wurde (Bauer und Einfalt, 2006 [7]).
Aufgrund des in den letzten Jahren entstandenen starken Interesses für adaptive Designs veranstalteten die FDA und die Harvard-MIT Division of Health Sciences and Technology im Oktober 2004 gemeinsam einen Workshop mit dem Titel „Adaptive Clinical Trial Designs: Ready for Prime Time?“ (Ellenberg, 2006 [18]). In den Workshops wurden vor allem adaptive Designs behandelt, die eine Neuschätzung der Fallzahl auf der Basis von Interimsergebnissen erlauben. Ein Hauptthema war die potentielle Effizienz der Designs, die geringer ist als die traditioneller sequentieller Designs.

"*For any adaptive design, one can always construct a standard group-sequential test based on the sequential likelihood ratio test statistic that, for any parameter value in the space of alternatives, will reject the null hypothesis earlier with higher probability, and, for any parameter value not in the space of alternatives, will accept the null hypothesis earlier with higher probability.*"
(Tsiatis und Mehta, 2003 [71])

Befürworter wenden ein, dass adaptive Designs diesen Effizienzverlust jedoch durch ihre große Flexibilität ausgleichen.
Da bereits regelmäßig Interimsanalysen zur Überprüfung der Effizienz und Sicherheit der neuen Behandlung durchgeführt werden, bietet es sich an, die Zwischenanalysen zusätzlich zur Anpassung des Studiendesigns zu nutzen. Statistiker sollten jedoch darauf achten und hinweisen, dass die Anpassung der Fallzahl in Interimsanalysen nicht die sorgfältige Fallzahlplanung ersetzt (D'Agostino, 2006 [15]). Ein Schwachpunkt adaptiver Designs ist die potentielle Entblindung der Studiendaten durch Fallzahladjustierung bei Interimsanalysen (Ellenberg, 2006 [18]). Eine Reduktion der Fallzahl könnte zu einem erhöhten Crossover führen, eine Erhöhung der Fallzahl hingegen scheint anzuzeigen, dass der beobachtete Behandlungseffekt kleiner ist als ursprünglich angenommen (D'Agostino, 2006 [15]).
Zudem besteht die Komplexität adaptiver Designs einerseits in dem Problem multiplen Testens, das durch die wiederholte Analyse der Daten entsteht und eine Adjustierung der p-Werte notwendig macht. Außerdem trifft bei adaptiven Designs die Annahme unabhängiger Beobachtungen nicht mehr zu, da beispielsweise die Daten von Patienten, die vor der ersten Zwischenauswertung in die Studie eingeschlossen wurden, in mehreren Zwischenauswertungen verwendet werden. Ein großer Kritikpunkt an adaptiven Designs ist, dass im Fall von Designmodifikationen an Stelle gebräuchlicher Teststatistiken nicht standardisierte Prüfgrößen angewendet werden

müssen (Bauer und Einfalt, 2006 [7]).
Hinter den in den letzten 20 Jahren entwickelten Methoden adaptiver Designs stehen zwei unterschiedliche Konzepte. Einerseits gibt es Kombinationstests, die stufenweise Teststatistiken verwenden, welche mit einer vordefinierten Kombinationsfunktion verknüpft werden (Bauer und Köhne, 1994 [8]). Anderseits besagt das Prinzip der Conditional Power, dass, solange der bedingte Fehler des neuen Designs nicht den bedingten Fehler des im Voraus geplanten Designs überschreitet, jede Art von Designmodifikation zu jedem Zeitpunkt einer Studie durchgeführt werden kann (Proschan und Hunsberger, 1995 [56]; Müller und Schäfer, 2001 [49], 2004 [51]).
In den letzten Jahren wurde eine Vielzahl von Artikeln im Bereich adaptiver Designs veröffentlicht. Allerdings beschäftigen sich die meisten davon mit adaptiven Designs bei normalverteilten Daten. In dem Artikel von Oellrich et *al.* (1997) [52] gibt es nur einen referenzierten Artikel zu diesem Thema. Im Folgenden werden Konzepte für adaptive Designs bei Daten mit Survivalendpunkt vorgestellt.
Schäfer und Müller (2001) [58] verallgemeinern die Idee von Proschan und Hunsberger (1995) zur Berechnung der bedingten Fehlerrate erster Art bei einer geplanten oder sogar ungeplanten Interimsanalyse. Sie entwickeln eine Methode, mit der in Studien mit zensierten Survivaldaten Interimsanalysen zu beliebigen Zeitpunkten durchgeführt werden können, ohne die Fehlerrate erster Art zu erhöhen. Dazu wird der beobachtete Behandlungsunterschied betrachtet, um entscheiden zu können, ob die Studie abgebrochen werden soll oder Änderungen am Studiendesign nötig sind.
Wird beispielsweise in einer ungeplanten Zwischenanalyse ein großer Behandlungseffekt beobachtet, muss bei Verwendung einer Alpha Spending Function das Signifikanzniveau α künftig schneller abgegeben werden. Die Folge kann eine unkontrollierte Erhöhung der Fehlerrate erster Art sein. Fast alle bisher zur Kontrolle dieser Fehlerrate vorgeschlagenen Methoden setzen vor Studienbeginn festgelegte Zeitpunkte für die Interimsanalysen voraus. Mit der Methode von Müller und Schäfer (2001) [49] kann ein ungeplanter zusätzlicher Blick in die Daten geworfen werden, ein sogenannter Interim Look, wobei eine Ablehung der Nullhypothese nicht möglich ist und somit kein Teil von α verbraucht wird. Stattdessen können jedoch Designmodifikationen durchgeführt werden. Der Blick in die Daten kann zu einem beliebigen Zeitpunkt während der Studie stattfinden, auch zeitgleich mit einer geplanten Interimsanalyse, oder nachdem bereits andere Zwischenanalysen durchgeführt wurden.
Es wird die Nullhypothese $H_0 : \log\theta = 0$ gegen die Alternativhypothesen $H_1^+ : \log\theta > 0$ und $H_1^- : \log\theta < 0$ getestet, mit dem Loghazardratio der Kontrollgruppe gegen die experimentelle Gruppe $\log\theta = \log(\lambda_1/\lambda_2)$. Die bis zum Interim Look beobachteten Ereignisse werden mit d_0 bezeichnet. Die laut Studienprotokoll geplanten Zwischenauswertungen *nach* dem Interim Look sollen nach der Beobachtung von $d_1, d_2, \ldots, d_{m-1}$ und die finale Analyse nach d_m Ereignissen stattfinden mit $d_0 < d_1 < d_2 < \ldots < d_m$. Bei dem Interim Look wird zunächst die Logranktestatistik $LR(d_0)$ mit dem Wert τ berechnet. Seien R^+ und R^- die Ereignisse, dass H_0 zugunsten H_1^+ beziehungsweise H_1^- bei einer der folgenden Interimsanalysen oder der finalen Analyse abgelehnt wird. Ein zentraler Punkt des Ansatzes von Schäfer und Müller ist die Berechnung der bedingten Wahrscheinlichkeiten dieser Ereignisse, gegeben $LR(d_0) = \tau$ unter der Nullhypothese

$$\epsilon^+(\tau, d_0) = P(R^+ | LR(d_0) = \tau; \log\theta = 0) \tag{4.26}$$

und

$$\epsilon^-(\tau, d_0) = P(R^-|LR(d_0) = \tau; \log\theta = 0). \tag{4.27}$$

Ein großer Vorteil der Methode ist, dass nach Bestimmung von $\epsilon^+(\tau, d_0)$ und $\epsilon^-(\tau, d_0)$ alle bis dahin erhobenen Daten zur Modifikation des Studiendesigns verwendet werden können. Möglich ist nun ein beliebiges anderes Design zum Testen der oben genannten Hypothesen mit Fehlerraten erster Art gleich $\epsilon^+(\tau, d_0)$ und $\epsilon^-(\tau, d_0)$. Siehe Schäfer und Müller (2001) [58] zur genauen Berechnung von $\epsilon^+(\tau, d_0)$ und $\epsilon^-(\tau, d_0)$ im Fall einer gruppensequentiellen Studie mit dem Faltungssatz von Armitage et *al.* (1969) [5]. Nach dem Interim Look können mittels des beobacheten Hazardratios die bei der finalen Analyse oder bei folgenden Interimsanalysen insgesamt zusätzlich benötigten Ereignisanzahlen ΔK neu berechnet werden. Im Fall keiner weiteren Zwischenanalysen ergibt sich:

$$\Delta K = \frac{4[\Phi^{-1}(1-\beta) + \Phi^{-1}(1-\epsilon^+(\tau, d_0))]^2}{(\log\theta^*)^2}, \tag{4.28}$$

wobei $\log\theta^*$ der Wert des Loghazardratios ist, den man in der verbleibenden Studienzeit nachweisen möchte. Im Fall weiterer Interimsanalysen muss ΔK mit einem Faktor $f > 1$ multipliziert werden, der sich aus den Grenzen des gruppensequentiellen Designs ergibt. Außerdem können mit Hilfe der geschätzten Survivalverteilungen die Fallzahl und die zur Beobachtung der benötigten Ereignisanzahl nötige Länge der Follow-up-Phase neu geschätzt werden. Zur Bestimmung der Fallzahl muss die bis zum Zeitpunkt t zusätzlich benötigte Ereignisanzahl $\Delta d(x)$ berechnet werden. Sei $\widehat{S}(x)$ die bei dem Interim Look geschätzte Survivalfunktion und $\widehat{F}(x) = 1 - \widehat{S}(x)$. Mit der Dauer von Beginn der Studie bis zum Interim Look B, der Zeit e_i zwischen Beginn der Studie und dem Einschluss des $i$ten Patienten vor dem Interim Look, den Personen unter Risiko bei dem Interim Look r, der Rekrutierungszeit nach dem Interim Look $\tilde{R}$, der in dieser Zeit rekrutierten Personen n und der seit dem Interim Look vergangenen Zeit x kann $\Delta d(x)$ folgendermaßen berechnet werden

$$\Delta d(x) = \sum_{i \in r} \widehat{F}(x + B - e_i) + \frac{n}{\tilde{R}} \int_0^{min(\tilde{R}, x)} \widehat{F}(x - s)ds. \tag{4.29}$$

Sei F die Dauer der Follow-up-Periode. Durch Gleichsetzen von ΔK mit $\Delta d(F + \tilde{R})$ kann die zusätzlich benötigte Fallzahl bestimmt werden. Die Power $1 - \beta$ ist hier ab den Zeitpunkt der Interimsanalyse eine Conditional Power (siehe Kapitel 4.7).
Eine Erweiterung dieses Ansatzes besteht in der bisher noch nicht möglichen Berechnung von p-Werten und Konfidenzintervallen für den Behandlungseffekt.
In einem „Letter to the editor“ nehmen Bauer und Posch (2004) [9] Bezug auf Schäfer und Müller (2001) [58] und geben zu bedenken, dass es nicht möglich ist, die gesamten zum Zeitpunkt einer Zwischenanalyse verfügbaren Informationen für Designänderungen zu benutzen, ohne die Fehlerrate erster Art zu beeinträchtigen. Dieses Problem trete außerdem nur bei Survivaldaten auf. Müller und Schäfer (2004) ([50]) widersprechen dem in einer „Authors' reply“ und legen dar, dass dieses Problem in jeder Studie mit langer Follow-up-Phase auftreten kann. Werden zur

Berechnung der folgenden Stufe des Designs ausschließlich Informationen von Patienten verwendet, die nach der vorherigen Interimsanalyse in die Studie eingeschlossen wurden, wird das Problem auch bei Survivaldaten nicht auftreten. In dem Ansatz von Schäfer und Müller wird das Prinzip bedingter Ablehnwahrscheinlichkeiten verwendet. Dieses Prinzip sagt aus, dass auf alle für die Designanpassung benutzten Informationen bedingt werden muss und dass die bedingte Wahrscheinlichkeit, die Nullhypothese abzulehnen, gegeben diese Informationen, unter dem alten und neuen Studiendesign gleich sein muss. Auf diese Weise kann die gesamte zum Zeitpunkt der Interimsanalyse verfügbare Information zur Designänderung benutzt werden, ohne die Fehlerrate erster Art zu beeinträchtigen.
Der Ansatz von Wassmer (2006) [72] ist im Fallzahlplanungungsprogramm ADDPLAN implementiert. Zur Designanpassung einer Studie mit Survivalendpunkt wird der Logranktest zum Vergleich zweier Survivalverteilungen und die inverse Normalmethode verwendet. Die Methode nutzt die Tatsache, dass zwar die Logrankstatistiken der einzelnen Phasen abhängig, die Zuwächse der Teststatistiken jedoch unabhängig und normalverteilt sind (siehe Jones und Whitehead, 1979 [33]; Selke und Siegmund, 1982 [64]; Tsiatis, 1981 [70]; Olschewski und Schumacher, 1986 [53]). Unter Verwendung der Unabhängigkeit der Zuwächse wird die Logrankstatistik leicht verändert, um das Design abhängig von den Beobachtungen anpassen zu können.
Bei gruppensequentiellen Designs wird die Teststatistik, hier der Logranktest, für jede Phase der Studie einzeln berechnet. In Phase i, $i = 1, \ldots, I$, wird die Ereignisanzahl d_i beobachtet und es befinden sich in Behandlungsgruppe 1 r_{1ik} und in Gruppe 2 r_{2ik} Patienten unter Risiko, wenn das $k$te Ereignis auftritt. Die Logrankstatistik in Phase i lautet dann

$$LR_i = \frac{\sum_{k=1}^{d_i} \left(I_{2ik} - \frac{r_{2ik}}{(r_{1ik}+r_{2ik})} \right)}{\sqrt{\sum_{k=1}^{d_i} \frac{r_{1ik} r_{2ik}}{(r_{1ik}+r_{2ik})^2}}}, \qquad i = 1, \ldots, I, \tag{4.30}$$

wobei I_{2ik} eine Indikatorvariable ist, die angibt, ob das Ereignis in Gruppe 2 ($I_{2ik} = 1$) oder in Gruppe 1 ($I_{2ik} = 0$) auftrat. Für festes d_i ist LR_i approximativ normalverteilt mit Varianz 1 und bekanntem Erwartungswert (Schoenfeld, 1981)[59]. Getestet wird die Nullhypothese H_0 : $S_1(x) = S_2(x)$ für alle x gegen die Alternative, dass für mindestens ein x die Gleichheit nicht gilt.
Im ersten Schritt des Designs wird wie in einem einstufigen Design die zur Erreichung der Power $1 - \beta$ benötigte Ereignisanzahl d, gegeben ein relevantes Hazardratio θ^*, bestimmt. d kann entweder nach Schoenfeld (1981) oder nach Freedman (1982) bestimmt werden. Die gesamte Fallzahl N wird durch Division von d durch einen Faktor ψ erhalten. Eine einfache Version dieses Faktors ist

$$\psi = \frac{\pi_1 + r\pi_2}{1 + r}, \tag{4.31}$$

wobei π_1 und π_2 die Ereignisraten in Gruppe 1 bzw. 2 sind und $r = N_2/N_1$ das Verhältnis der Gruppengrößen angibt. Kompliziertere Versionen von ψ erlauben die Berücksichtigung von schrittweisen Rekrutierungsmustern.

In jeder Stufe wird eine standardisierte Teststatistik berechnet, die auf der Kombination der stufenweisen p-Werte p_i, $i = 1, \ldots, I$, basiert. Bei der inversen Normalmethode wird die standardisierte Teststatistik Z_i mit der Kombinationsfunktion $\Phi^{-1}(1 - p_i)$ und vorher festgelegten Gewichten $\omega_i > 0$ gebildet. Der entscheidende Punkt ist, dass die Gewichte vorher festgelegt und auch nicht mehr verändert werden. Unter H_0 sind die Kombinationsfunktionen $\Phi^{-1}(1-p_i)$ unabhängig und standardnormalverteilt, wenn die p-Werte gleichverteilt sind. In einem gruppensequentiellen Design, bei dem der Logranktest verwendet wird, lauten die approximativ normalverteilten Zuwächse

$$Z_i^* = \frac{\sqrt{d_i}LR_i - \sqrt{d_{i-1}}LR_{i-1}}{\sqrt{d_i - d_{i-1}}}, \qquad i = 2, \ldots, I \tag{4.32}$$

(siehe Shen und Cai, 2003 [65]). Die standardisierte Teststatistik wird in diesem Fall wie folgt angegeben

$$Z_i^* = \left(\sum_{\tilde{i}=1}^{i} \omega_{\tilde{i}}^2 \right)^{-1/2} \sum_{\tilde{i}=1}^{i} \omega_{\tilde{i}} Z_i^* \tag{4.33}$$

und die Gewichte sind als $\omega_1 = \sqrt{\xi}_1$, $\omega_i = \sqrt{\xi_i - \xi_{i-1}}$, $i = 2, \ldots, I$, definiert, wobei ξ_i die geplante kumulierte Ereignisanzahl in Phase i ist. Anschließend wird angegeben, wie Konfidenzintervalle für das Hazardratio berechnet werden können.
Designänderungen sollten nur basierend auf den stufenweisen Logrankstatistiken und auf den bis dahin beobachteten Ereignissen durchgeführt werden, um die Fehlerrate erster Art einzuhalten. Im Folgenden werden drei Möglichkeiten der Fallzahlneuschätzung vorgestellt.
Die Fallzahl kann entweder basierend auf Conditional Power (siehe Kapitel 4.7), auf der angestrebten Breite des Konfidenzintervalls für das Hazardratio oder auf der Berechnung der bedingten Fehlerrate erster Art bestimmt werden. Letzterer Ansatz ist der flexibelste, da zusätzlich zur Neuschätzung der Fallzahl die Anzahl sowie die Abbruchgrenzen der folgenden Zwischenanalysen geändert werden können. Die „conditional Type I error function“ gibt die bedingte Wahrscheinlichkeit an, H_0 fälschlicherweise in einer der folgenden Phase abzulehnen.
Die Arbeit von Chang (2007) [11] basiert auf Bauer und Köhne (1994) [8] und benutzt Fisher's Kombination unabhängiger p-Werte für Teilstichproben aus unterschiedlichen Phasen. Deren Methode ist jedoch in der Praxis bezüglich der Abbruchgrenzen der Prozedur zu unflexibel. Um die praktische Anwendbarbeit zu verbessern, schlägt Chang hier eine adaptive Designmethode vor, die eine Linearkombination der p-Werte aus Teilstichproben der aktuellen und vorherigen Stufen benutzt.
Bei jeder der I Interimsanalysen wird ein Hypothesentest durchgeführt. Abhängig von dem Ergebnis des Tests kann die Studie wegen sehr großer Effizienz oder zu geringer Wirkung der neuen Behandlung abgebrochen, die Fallzahl neu berechnet und die Randomisierung geändert werden.
Getestet wird die globale Hypothese

$$H_0 : H_{01} \cap ... \cap H_{0I},$$

wobei H_{0i} $(i = 1, \ldots, I)$ die Nullhypothese der $i$ten Interimsanalyse ist. Alle Teilhypothesen testen denselben Endpunkt der Studie und basieren auf einer Teilstichprobe der jeweiligen Stufe. Im Folgenden wird ohne Beschränkung der Allgemeinheit angenommen, dass die Überlegenheit der Studienmedikation überprüft werden soll, also

$$H_{0i} : \eta_{i1} \geq \eta_{i2} \qquad \text{versus} \qquad H_{1i} : \eta_{i1} < \eta_{i2},$$

wobei η_{i1} und η_{i2} die Responsevariablen der beiden Gruppen in der $i$ten Stufe sind.
Wenn der Response in beiden Behandlungsgruppen gleich ist, ist der p-Wert p_i in der $i$ten Phase häufig unter H_0 gleichverteilt auf $[0, 1]$ (Bauer und Köhne (1994) [8]). Daraus konstruieren Bauer und Köhne (1994) die folgende Teststatistik

$$T_k = \sum_{i=1}^{k} \omega_{ki} p_i, \qquad k = 1, \ldots, I \tag{4.34}$$

mit $\omega_{ki} > 0$ als Gewichte der p-Werte in der Linearkombination.
Ein Abbruch der Studie erfolgt bei großer Effizienz der neuen Behandlung, wenn $T_k \leq \alpha_k$, oder wegen zu geringer Wirkung der neuen Behandlung, wenn $T_k \geq \beta_k$. Andernfalls wird die Studie fortgesetzt. α_k, β_k und T_k sind monoton in k wachsende Funktionen, $\alpha_k < \beta_k$ $(k = 1, \ldots, I-1)$ und $\alpha_I = \beta_I$.
Die Wahrscheinlichkeit, die $k$te Phase zu erreichen, ist

$$\begin{aligned} \psi_k(t) &= Pr(T_k < t, \alpha_1 < T_1 < \beta_1, \ldots, \alpha_{k-1} < T_{k-1} < \beta_{k-1}) \\ &= \int_{\alpha_1}^{\beta_1} \ldots \int_{\alpha_{k-1}}^{\beta_{k-1}} \int_{-\infty}^{t} f_{T_1 \ldots T_k}(t_1, \ldots, t_k) dt_k dt_{k-1} ... dt_1, \end{aligned} \tag{4.35}$$

wobei $t \geq 0$, T_i $(i = 1, \ldots, k)$ die Teststatistik der i-ten Phase und $f_{T_1 \ldots T_k}$ die Dichte der gemeinsamen Verteilung der Teststatistiken ist.
Die globale Fehlerrate erster Art

$$\alpha = \sum_{k=1}^{I} e_k \tag{4.36}$$

setzt sich aus den Fehlerraten $e_k = \psi_k(\alpha_k)$ der einzelnen Stufen zusammen.
Wenn für die Teststatistik in der k-ten Stufe gilt $T_k = t = \alpha_k$, ist der p-Wert gleich dem bis dahin verbrauchten α, nämlich $\sum_{i=1}^{k} e_i$. Der zu einer in der k-ten Stufe beobachteten Teststatistik $T_k = t$ gehörige adjustierte p-Wert ist

$$p(t; k) = \sum_{i=1}^{k-1} e_i + \psi_k(t), \qquad k = 1, \ldots, I. \tag{4.37}$$

Je früher die Nullhypothese aufgrund des adjustierten p-Werts abgelehnt wird, desto mehr spricht gegen sie, da noch nicht so viel α verbraucht wurde. Die p-Werte können mit einer beliebigen

Teststatistik berechnet werden, die Hauptsache ist, dass sie unabhängig sein müssen. Nun können Grenzen für den Studienabbruch aufgrund von großer Effizienz oder zu geringer Wirksamkeit der Studienbehandlung berechnet werden.
Bei der Fallzahlneuschätzung wird die adjustierte Fallzahl

$$N = \min\left(N_{max}, \max\left(N_{min}, \text{sign}(E_0 E)\left|\frac{E_0}{E}\right|^c N_0\right)\right) \tag{4.38}$$

berechnet.
Dabei ist N_{max} die höchste (aus finanziellen oder anderen Gründen) vertretbare Fallzahl, N_{min} die Fallzahl für die Interimsanalyse, E_0 die vor Studienbeginn geschätzte Effektgröße, E die beobachtete Effektgröße, N_0 die vor Studienbeginn durch ein klassisches Design geschätzte Fallzahl und c eine Konstante. Aufgrund von Simulationsergebnissen empfiehlt der Autor einen Wert von $c = 2$, der aber auch erst nach der ersten Interimsanalyse festgelegt werden kann, da die stufenweisen p-Werte ungeachtet der Art der Fallzahladjustierung unabhängig sind. Die Funktion $\text{sign}(E_0 E)$ in der Formel führt dazu, dass keine Fallzahladjustierung erfolgt, wenn E und E_0 unterschiedliche Vorzeichen besitzen. Die beobachtete Effektgröße erhält man aus der standardisierten Differenz der geschätzten Responsevariable in den Behandlungsgruppen, das heißt

$$E = \frac{\widehat{\eta}_{i2} - \widehat{\eta}_{i1}}{\widehat{\sigma}_i}, \tag{4.39}$$

wobei sich die Varianzschätzung $\widehat{\sigma}_i^2$ für große Stichenprobenumfänge bei einem Survivalendpunkt ergibt durch

$$\widehat{\sigma}_i^2 = \overline{\eta}_i^2\left[1 - \frac{e^{\overline{\eta}_i T_0} - 1}{T_0 \overline{\eta}_i e^{\overline{\eta}_i T_s}}\right]^{-1}. \tag{4.40}$$

Dabei ist $\overline{\eta}_i = (\widehat{\eta}_{i1} + \widehat{\eta}_{i2})/2$.
Auch in dem Artikel von Desseaux und Porcher (2007) [16] geht es um adaptive Designänderungen bei einer Zwischenanalyse in einem zweistufigen Design. Bei mehrstufigen Designs werden stufenweise Teststatistiken oder p-Werte kombiniert. Hier wird der Ansatz von Bauer und Köhne (1994) auf Survivaldaten erweitert.
Die Nullhypothese $H_0 : \theta = 0$ wird gegen die Alternative $H_1 : \theta > 0$ getestet, wobei θ im Fall von Survivaldaten das Hazardratio ist. Bauer und Köhne (1994) [8] beschreiben den zweistufigen Kombinationstest folgendermaßen. Die erste Stufe des Designs besteht in der Beobachtung von N_1 Patienten, deren Daten bei einer Zwischenanalyse ausgewertet werden, woraus ein p-Wert p_1 berechnet wird. Anschließend kann abhängig von p_1 und zwei Entscheidungsgrenzen α_0 und α_1 entweder H_0 abgelehnt, H_0 angenommen oder die Studie in einer zweiten Stufe fortgeführt werden. Während der zweiten Stufe werden gegebenenfalls weitere N_2 Patienten beobachtet und am Ende ein weiterer p-Wert p_2 berechnet. Aufgrund einer von p_1 und p_2 abhängigen Kombinationsfunktion $C(p_1, p_2)$ wird H_0 abgelehnt, wenn der Wert der Kombinationsfunktion kleiner als ein aus α_2 berechneter kritischer Wert ist. Hier wird Fishers Produktkriterium $C(p_1, p_2) = p_1 \cdot p_2$ verwendet. α_0, α_1 und α_2 werden so bestimmt, dass die Gesamtfehlerrate erster Art α eingehalten wird. Mit der Indikatorvariable I für die Behandlung (z.B. $I = 0$ für

die Kontrollgruppe und $I = 1$ für die experimentelle Gruppe) kann die Hazardfunktion bedingt auf I für $x \geq 0$ durch $\lambda(x|I) = \lambda_0(x)\exp(-\theta I)$ ausgedrückt werden. Die Daten sind hier in der zweiten Stufe nicht unabhängig, da sowohl Daten von Patienten untersucht werden, die in der zweiten Stufe rekrutiert werden, als auch von Patienten, die die erste Stufe überlebten. p-Werte aus stufenweisen Logrankstatistiken können folglich nicht benutzt werden. Dennoch ist die Berechnung unabhängiger p-Werte mittels des Logranktests möglich.
Nach k beobachteten Ereignissen lautet die Logrankstatistik

$$T(k) = \sum_{l \in D_k} (1 - I_l) - \frac{r_1(k)}{r_2(k) + r_1(k)} \tag{4.41}$$

mit dem Patientenindex l, der Menge D_k der Indizes aller Patienten, die vor dem $k$ten Ereignis der Studie selbst ein Ereignis hatten, dem Indikator I_l für die Behandlungsgruppe des $l$ten Patienten und den Patientenanzahlen unter Risiko $r_2(k)$ und $r_1(k)$ zum Zeitpunkt des $l$ten Ereignisses in der experimentellen beziehungsweise Kontrollgruppe.
Sei k_1 die Ereignisanzahl in der ersten Phase und k_2 die Ereignisanzahl am Ende der Studie, wenn $N_1 + N_2$ Patienten beobachtet wurden. Die gemeinsame Verteilung der Zufallsvariablen $T(k_1)$ und $T(k_2)$ kann für große Fallzahlen unter Annahme kontanter Hazardraten durch eine bivariate Normalverteilung approximiert werden. Wenn in der ersten Stufe T_1 zur Berechnung von p_1 und am Ende der Studie $T'(k_2) = T(k_2) - T(k_1)$ zur Berechnung von p_2 verwendet werden, folgt die Unabhängigkeit der p-Werte. Folglich kann der zweistufige Kombinationstest angewendet werden.
Zunächst wird die Anzahl in der ersten Stufe benötigter Ereignisse bestimmt (siehe Posch und Bauer, 2000 [55]). Um die auf frühen Studienabbruch bedingte Power zu kontrollieren, muss

$$\frac{P_{H_1}(Z_1 < z_{1-\alpha_0})}{P_{H_1}(Z_1 < z_{1-\alpha_0}) + P_{H_1}(Z_1 > z_{1-\alpha_1})} = \beta \tag{4.42}$$

gelten. Dabei ist $Z_1 = T(k_1)/\sqrt{k_1}$ unter der Alternative normalverteilt mit Erwartungswert $\sqrt{k_1}\theta/2$ und Varianz 1. Um die Ereignisanzahl $k_1(\alpha_0, \alpha_1)$ zu erhalten, wird die eindeutige Lösung der Gleichung

$$\frac{1-\beta}{\beta}\Phi_{(\sqrt{k_1}\theta/2,1)}(z_{1-\alpha_0}) = 1 - \Phi_{(\sqrt{k_1}\theta/2,1)}(z_{1-\alpha_1}) \tag{4.43}$$

für alle $\alpha_1 \in [c_{\alpha_2}/(1-\beta), \alpha]$, $\alpha_0 \in [\alpha_1, 1-\alpha_1]$ und festes $\beta < \frac{1}{2}$ bestimmt. Den kritischen Wert c_{α_2} erhält man dabei aus α_2. Daraus wird anschließend die Fallzahl

$$N_1(\alpha_0, \alpha_1) = \frac{k_1(\alpha_0, \alpha_1)R_1}{\int_0^{R_1} \overline{S}(x_1 - u)du} \tag{4.44}$$

für die erste Stufe berechnet, wobei $\overline{S}$ die gepoolte Survivalverteilungsfunktion über die Behandlungsgruppen ist, der Eintritt der Patienten während der Rekrutierungsphase $[0, R_1]$ als gleichverteilt angenommen wird und die Daten zum Zeitpunkt $x_1 (x_1 \geq R_1)$ analysiert werden. Zur Vereinfachung wird vorausgesetzt, dass es in der ersten Stufe keine Follow-up-Phase

gibt und die Daten am Ende der Rekrutierungsphase ausgewertet werden, also dass $R_1 = x_1$. Falls das Kriterium $p_1 \geq \alpha_1$ erfüllt ist, geht man in die zweite Stufe über. Aufgrund der Interimsergebnisse können die erwarteten Ereignisraten und der Wert des Log-Hazardratio auf θ^* geändert werden. Bei Verwendung von Fisher's Produktkriterium wird H_0 nach der zweiten Stufe abgelehnt, wenn $p_2 < c_{\alpha_2}/p_1$, was als Signifikanzniveau $\tilde{\alpha}_2(p_1)$ der zweiten Stufe angesehen werden kann. Jetzt kann daraus die Fallzahl N_2 mit Standardmethoden berechnet werden. Dazu wird wieder zunächst die Anzahl in der zweiten Stufe benötigter Ereignisse

$$\Delta k = \frac{4([z_{1-\tilde{\alpha}_2(p_1)} + z_{1-\beta}]^+)^2}{\theta^{*2}} \tag{4.45}$$

berechnet. $[]^+$ bezeichnet dabei den Positivteil des Arguments. Patienten, die in der zweiten Stufe ein Ereignis haben, wurden entweder schon in der ersten oder erst in der zweiten Stufe rekrutiert. r_1 sei die Menge der Patienten, die in der ersten Stufe rekrutiert wurden und zum Zeitpunkt der Interimsanalyse a_1 noch unter Risiko stehen und x_i die Zeit zwischen Studenbeginn und Einschluss des i-ten Patienten. Sei τ die Zeit seit der Interimsanalyse und R_2 die Dauer der Rekrutierunsphase in der zweiten Stufe. Dann ist

$$\Delta k(\tau) = \sum_{i \in r_1} \frac{\overline{F}(\tau + a_1 - x_i) - \overline{F}(a_1 - x_i)}{1 - \overline{F}(a_1 - x_i)} + \frac{N_2}{R_2} \int_0^{min(R_2,\tau)} \overline{F}(\tau - s)ds \tag{4.46}$$

die Anzahl der erwarteten zusätzlichen Ereignisse zum Zeitpunkt τ. Aus dem Zusammenhang dieser beiden Ereignisanzahlen $\Delta k(R_2 + F_2) = \Delta k$, und der Follow-up-Dauer F_2 des als letztes in der zweiten Phase eingeschlossenen Patienten, kann schließlich die Fallzahl N_2 bestimmt werden. Bei der Interimsanalyse ist eine Anpassung der Fallzahl möglich, wenn die Ereignisraten bis dahin höher oder niedriger als erwartet sind. Dazu wird die gepoolte Survivalverteilung $\overline{F}$ aus den bis a_1 vorliegenden Daten geschätzt und mit der Methode von Whitehead et *al.* (2001) [75] extrapoliert. Außerdem können auch $\alpha_0, \alpha_1, \alpha_2$ und das Log-Hazardratio angepasst werden. Hier wird eine gleichverteilte Rekrutierungsphase und keine Drop-outs vorausgesetzt. Daher empfehlen die Autoren, die Formeln von Lachin und Foulkes (1986) [35] zu verwenden, um die Patientenanzahl aus der Ereignisanzahl zu bestimmen, wenn komplexere Annahmen getroffen werden sollen.

4.7. Conditional Power Berechnungen

Unter Conditional oder bedingter Power versteht man die Wahrscheinlichkeit, die Nullhypothese bei der Endauswertung abzulehnen, gegeben die bis dahin beobachteten Daten. Dieser Wert kann zu der Entscheidung beitragen, die Studie abzubrechen oder fortzuführen.
Betensky (1997) [10] gibt einen Literaturüberblick der Jahre vor 1997 zur Conditional Power Berechnung und zeigt Alternativen auf, wenn es um den frühen Abbruch einer Studie geht.
Cook (2003) [12] zeigt, wie man mit dem Markovkettenmodell von Lakatos (1986, 1988) Korrekturen während einer Studie vornimmt. Dabei können willkürliche Rekrutierungsmuster, Hazardraten und andere Designparameter in das Modell aufgenommen werden. Eine mögliche Korrektur kann der Abbruch der Studie aufgrund sehr starker oder sehr schwacher Wirkung

der neuen Behandlung sein. Die Analyse einer solchen Fragestellung erfolgt entweder auf Basis von bedingter oder von prädiktiver Power. Dazu müssen die Daten entblindet werden. Andere Designmodifikationen wie eine Änderung der Fallzahl oder der Länge der Follow-up-Phase dagegen können auch mit verblindeten Daten durchgeführt werden, wenn die Daten nicht nach Gruppen getrennt analysiert werden. Ein mögliches Problem gewöhnlich verwendeter Methoden besteht darin, dass während der Studie verfügbare Daten nicht zur Umgestaltung verwendet werden können, da beispielsweise die Patienten unregelmäßig rekrutiert werden oder die beobachteten Ereignisraten irregulär sind. Die Tests in diesem Artikel gehören zur Fleming-Harrington-Familie (1991) [20] gewichteter Logranktests. Im Falle von Crossover wird eine Intention-to-treat-Analyse durchgeführt, das heißt, dass Patienten entsprechend ihrer ursprünglich zugeteilten Behandlung analysiert werden. Es wird der Fall zweier Behandlungsgruppen betrachtet. Lakatos' Markovkettenmodell besitzt die vier Zustände Lost to Follow-up (L), hatte ein Ereignis (E), erhält die experimentelle Behandlung (A_E) oder erhält die Kontrollbehandlung (A_C). Wenn ein Patient einen der Zustände A_E oder A_C verlässt, nimmt er die Übergangsraten des jeweils anderen Zustands an. Im Gegensatz zu Lakatos (1988) wird hier kein verzögerter Wirkungseintritt betrachtet.
Sei LR_{ij} die Logrankstatistik, wenn die Rekrutierung bis zum Zeitpunkt τ_j stattfand und die bisherige Studiendauer x_i ist, $i = 1, \ldots, I$, $j = 1, \ldots, J$. $\tau_J = R$ ist der Zeitpunkt, zu dem die Rekrutierung abgeschlossen ist, und $x_I = T$ ist das Studienende. Die Conditional Power hängt von dem bedingten Erwartungswert und von der bedingten Varianz der Logrankstatistik ab, gegeben die bis dahin beobachteten Daten. Der bedingte Erwartungswert ist $\mathrm{E}(LR_{IJ}|LR_{ij}) = LR_{ij}+\mathrm{E}(LR_{IJ}-LR_{ij})$, wobei der Erwartungswert der Logrankstatistik E_{ij} in Gleichung (2) im Artikel angegeben ist. Die bedingte Varianz wird angegeben als $\mathrm{Var}(LR_{IJ}|LR_{ij}) = V_{IJ} - V_{ij}$, wobei die Varianz der Logrankstatistik V_{ij} in Gleichung (3) des Artikels angegeben ist. Da ein geringer Unterschied zwischen der tatsächlich beobachteten Varianz der Logrankstatistik, bezeichnet mit $\widetilde{V}_{ij}$, und V_{ij} bestehen kann, wird bei der Konstruktion der finalen Teststatistik $\widetilde{V}_{IJ} = V_{IJ} - V_{ij} + \widetilde{V}_{ij}$ als Varianz der Logrankstatistik LR_{IJ} verwendet. Ist $c > 0$ der kritische Wert des Tests auf Überlegenheit, ergibt sich die Conditional Power zu

$$\begin{aligned} P(LR_{IJ} > c\widetilde{V}_{IJ}^{1/2}) &= P(LR_{IJ} - LR_{ij} - (E_{IJ} - E_{ij}) > cV_{IJ}^{1/2} - LR_{ij} - (E_{IJ} - E_{ij})) \\ &\approx 1 - \Phi\left(\frac{cV_{IJ}^{1/2} - LR_{ij} - (E_{IJ} - E_{ij})}{(V_{IJ} - V_{ij})^{1/2}}\right) \end{aligned} \tag{4.47}$$

Nach Chang (2007) [11] kann die Fallzahl sowohl wie in Kapitel 4.6 durch eine Formel, in die das Verhältnis der erwarteten und der beobachteten Effektgröße eingeht, als auch basierend auf Conditional Power während einer Studie adjustiert werden. Seien p_k, z_k und α_k jeweils der p-Wert, der normalverteilte z-Score und das Signifikanzniveau der Teilstichprobe der $k$ten Stufe eines adaptiven Designs. H_0 wird abgelehnt, falls $z_2 \geq B(\alpha_2, p_1)$. $B(\alpha_2, p_1)$ ist das Abbruchkriterium in der zweiten Stufe des Designs und hängt von der Art der Kombination der stufenweisen p-Werte ab. Daraus ergibt sich die Conditional Power der zweiten Stufe als

$$P_c(p_1, \delta) = 1 - \Phi\left(B(\alpha_2, p_1) - \frac{\delta}{\sigma}\sqrt{\frac{N_2}{2}}\right), \qquad \alpha_1 < p_1 \leq \beta_1, \tag{4.48}$$

wobei δ den Behandlungsunterschied angibt und σ sich aus der Varianz des Behandlungsunterschieds in der zweiten Stufe ergibt.
Für gegebene Conditional Power $P_c(p_1, \delta)$ berechnet sich die adjustierte Fallzahl für die zweite Stufe durch Umstellen nach N_2 als

$$N_2 = \left[\frac{\sqrt{2}\sigma}{\delta} (B(\alpha_2, p_1) - \Phi^{-1}(1 - P_c(p_1, \delta))) \right] \tag{4.49}$$

4.8. Zensierungen

In den meisten klinischen Studien werden nicht alle Patienten solange beobachtet, bis sie ein Ereignis haben. Stattdessen wird die Studie beendet, sobald die vorher festgelegten Rekrutierungs- und Follow-up-Phasen verstrichen sind. Patienten, die am Studienende noch nicht das interessierende Ereignis hatten, werden als rechtszensiert bezeichnet. Alternativen zu dieser häufig vorkommenden Zensierungsart werden in diesem Kapitel behandelt.
Bei der Methode von Xiong et *al.* (2003) [77] wird die Studie nicht nach Ablauf der im Vorhinein festgelegten Studiendauer beendet, sondern sobald ein vorher spezifizierter Patientenanteil ein Ereignis hatte. Beträgt dieser Anteil beispielsweise 80% und die Fallzahl 100, werden die Patienten solange beobachtet, bis 80 Ereignisse aufgetreten sind. Bei auf diese Weise zensierten Daten ist die Studiendauer somit unbekannt.
Xiong et *al.* vergleichen nicht die Survivalkurven selbst miteinander, sondern deren Mittelwerte. Es wird sowohl die Verteilungsfamilie proportionaler Hazardraten als auch die location-scale Familie logtransformierter Überlebenszeiten betrachtet. Der Logranktest kann für beide Verteilungsfamilien berechnet werden, wobei die meisten Methoden zur Fallzahlberechnung auf der Familie proportionaler Hazardraten basieren.
Seien X_1 und X_2 die Überlebenszeiten in der Kontroll- bzw. Behandlungsgruppe. $Y_i = \ln X_i$, $i = 1, 2$, besitze die Dichte einer location-scale Familie mit Erwartungswert $\mathrm{E}(Y_i) = \mu_i$ und Varianz $\mathrm{Var}(Y_i) = \sigma_i^2$:

$$\frac{1}{\sigma_i} g\left(\frac{y - \mu_i}{\sigma_i}\right), \qquad -\infty < y < \infty, \tag{4.50}$$

wobei $g(s) > 0$ eine differenzierbare positive Funktion ist. Sei außerdem $\sigma_1 = \sigma_2 = \sigma > 0$. Da $X_i = \exp Y_i$ den Erwartungswert

$$E(X_i) = \int_{-\infty}^{\infty} e^y \frac{1}{\sigma} g\left(\frac{y - \mu_i}{\sigma}\right) dy = e^{\mu_i} \int_{-\infty}^{\infty} e^{\sigma s} g(s) ds \tag{4.51}$$

für $i = 1, 2$ besitzt, ist

$$\frac{E(X_2)}{E(X_1)} = e^{\mu_2 - \mu_1}. \tag{4.52}$$

Somit muss nur die Differenz und nicht das Verhältnis der Mittelwerte der beiden Gruppen getestet werden. Dazu wird angenommen, dass aus den Verteilungen von X_1 und X_2 zwei entsprechende Stichproben der Größe N_1 bzw. N_2 gezogen werden. Aufgrund der besonderen Zensierungsart wird jeweils nur ein festgelegter Anteil der Stichproben beobachtet. Dieser Anteil darf sich zwischen beiden Gruppen unterscheiden.
Die Nullhypothese $H_0 : \mu_1 = \mu_2$ soll gegen die Alternative $H_1 : \mu_1 \neq \mu_2$ zu einem asymptotischen Signifikanzlevel α $(0 < \alpha < 1)$ getestet werden. Aus der Definition der zu erreichenden Power $1 - \beta$ bei einem nachzuweisenden Unterschied von $d = \mu_2 - \mu_1$ kann die Fallzahl N_1 berechnet werden. Durch Multiplikation von N_1 mit einem vorher festgelegten Faktor r, der das Verhältnis der Stichprobenumfänge der beiden Gruppen angibt, folgt sofort die Fallzahl N_2.
Es ist auch möglich, die Nullhypothese $H_0 : \mu_1 = \mu_2$ gegen die einseitige Alternative $H_1 : \mu_2 > \mu_1$ zu testen, wobei eine positive Differenz der Mittelwerte $M = \mu_2 - \mu_1$ nachgewiesen werden kann.
Die Weibullverteilung ist die einzige Verteilung, die sowohl zur location-scale Familie logtransformierter Überlebenszeiten als auch zur Familie proportionaler Hazardsraten gehört. So können einige Zusammenhänge der Fallzahlbestimmung in dem Artikel von Xiong et *al.* mit den auf dem Logranktest basierenden Methoden aufgezeigt werden. Ist $\sigma = 1$ ergibt sich aus der Weibullverteilung die Exponentialverteilung. In diesem Fall ist aus Gleichung 4.52 ersichtlich, dass M gleich dem natürlichen Logarithmus des Hazardratios zwischen der Kontroll- und der Behandlungsgruppe ist. Die Fallzahlformel für den einseitigen Test ist identisch mit der Formel von Rubinstein et *al.* (1981) [57], wenn $N_1 = N_2$, die Patienten gleichzeitig in die Studie eintreten und kein Loss to Follow-up vorhanden ist. Dieselbe Fallzahlformel wird auch von Lakatos und Lan (1992) [40] basierend auf Rubinstein et *al.* (1981) berechnet.
Williamson et *al.* (2009) [76] entwickeln eine Methode zur Power- und Fallzahlberechnung bei einer anderen Zensierungsart. Current Status Daten sind ein Spezialfall intervallzensierter Daten. Für den Ereigniszeitpunkt E ist nur bekannt, ob er vor oder nach einer zufälligen Untersuchungszeit U liegt. Die Daten sind von der Form $(0, U]$ oder $[U, \infty)$ und werden häufig als intervallzensierte Daten erster Art bezeichnet. Bisher gibt es wenig Literatur zur Fallzahlplanung bei Current Status Daten. Diese Daten enthalten jedoch weniger Information und die entsprechenden Tests besitzen weniger Power als bei rechtszensierten Daten (Lin et *al.* (1998) [46]).
Die Vorteile parametrischer Modellierung sind die einfachere Modellspezifikation gegenüber nicht- oder semiparametrischer Modellierung und eine größere Flexibilität, da Kovariablen miteinbezogen werden können. Zur Powerberechnung wird hier ein Weibullmodell für die Survivalverteilung verwendet. Außerdem wird das Hazardratio $e^{-\beta}$ festgelegt.
Da zur Powerberechnung der Waldtest verwendet wird, wird zunächst die Likelihood für Current Status Data

$$L(c_i, \delta_i; \rho, k) = \prod_{i=1}^{N} S(c_i; \rho, k)^{1-\delta_i}(1 - S(c_i; \rho, k))^{\delta_i} \tag{4.53}$$

aufgestellt. Dabei ist N die Fallzahl, C_i die logtransformierte Zensierungszeit des i-ten Individuums $(i = 1, \ldots, N)$, $\delta_i = I(\log E_i \leq c_i)$ der Zensierungsindikator mit der logtransformierten Ereigniszeit $\log E_i$ des i-ten Individuums, ρ ein Spaltenvektor mit Scale- oder Shapeparametern

und k ein Spaltenvekor mit Regressionskoeffizienten. Durch Nullsetzen der Scorefunktion erhält man die Parameterschätzer für $\eta = [\rho', k']'$.
Werden zwei Gruppen miteinander verglichen, enthält das Modell eine binäre Kovariable $x = 0, 1$ für die Gruppenzugehörigkeit. Es wird angenommen, dass die Survivalzeiten einer Weibullverteilung mit Intercept Δ und Shapeparameter ν folgen und dass die Zensierungsverteilung in beiden Gruppen gleich ist.
Zum Test der Nullhypothese $H_0 : k = 0$ gegen die Alternative $H_1 : k \neq 0$ wird der Waldtest mit der Teststatistik

$$W = \frac{\widehat{k}^2}{\widehat{var}(\widehat{k})}, \tag{4.54}$$

verwendet, wobei $\widehat{k}$ der Maximumlikelihoodschätzer für k ist. Da W unter der Alternativhypothese nichtzentral χ^2-verteilt ist mit Nichtzentralitätsparameter $\tau = W$, kann die Power für den Waldtest folgendermaßen bestimmt werden:

$$1 - \beta = \int_{\chi^2_{1-\alpha}(1)}^{\infty} \frac{e^{-(u^2+\tau^2)/2} u\tau}{\sqrt{\tau u}} B_{-\frac{1}{2}}(\tau u) du, \tag{4.55}$$

wobei α und β die Fehlerraten 1. beziehungsweise 2.Art sind, $\chi^2_{1-\alpha}(1)$ der kritische Wert der zentralen χ^2_1-Verteilung und $B_a(b)$ eine geänderte Besselfunktion erster Art ist.
Die Berechnung der Varianz von $\widehat{k}$ erfordert die Bestimmung der erwarteten Fisherinformationsmatrix

$$E\left[-\sum_{i=1}^{N} \frac{\partial^2}{\partial \eta^2} \log L(c_i, \delta_i; \eta)\right] = -\sum_{i=1}^{N} \int_0^{\infty} \frac{\partial^2}{\partial \eta^2} \log L(c_i, \delta_i; \eta) f_{\delta|c}(c_i) f_c(c_i) dc \tag{4.56}$$

unter Verwendung der bedingten Verteilungsfunktion $f_{\delta|c}(c_i) = P(\delta_i = 1|C_i = c_i) = 1 - S(c_i)$. Dazu muss eine Verteilung für den natürlichen Logarithmus der Zensierungsvariable C festgelegt werden. Bei Annahme einer während der gesamten Studie konstanten Zensierungsverteilung wird eine Exponentialverteilung mit Parameter ϕ verwendet. Eine steigende oder fallende Zensierungsrate wird durch eine Weibullverteilung modelliert. Wird angenommen, dass die Zensierungszeiten in ungefähr gleichen Intervallen während der Studie auftreten, kann eine Gleichverteilung gewählt werden.
Bei Annahme einer Exponentialverteilung als Zensierungsverteilung ist

$$f_c(c_i) = \phi e^{c_i} \exp(-\phi e^{c_i}). \tag{4.57}$$

Um ϕ zu bestimmen, muss die folgende Formel für den in der Studie erwarteten Ereignisanteil

$$\int_{-\infty}^{\infty} \left[\frac{N_1}{N_1 + N_2}(1 - S_1(c_i)) + \frac{N_2}{N_1 + N_2}(1 - S_2(c_i))\right] f_c(c_i) dc \tag{4.58}$$

nach ϕ umgestellt werden. Dabei sind N_j und $S_j(x)$ die Fallzahl und Survivalfunktion von Gruppe j, $j = 1, 2$.
Um die Fallzahl zu bestimmen, wird ein parametrisches Weibullmodell für Current Status Daten gefittet. Die Parameterschätzer für den Intercept, die Kovariable und den Shapeparameter werden basierend auf früheren Studien festgelegt. Für die Zensierungszeiten wird beispielsweise eine Exponentialverteilung mit Parameter ϕ angenommen. ϕ wird so bestimmt, dass man einen bestimmen Ereignisanteil erhält. Die Fallzahl kann so berechnet werden, dass eine bestimmte Power erreicht wird. Das Design muss nicht balanciert sein und es ist möglich, die Methode auf den Fall von mehr als zwei Behandlungsgruppen zu erweitern.

4.9. Sonstiges

Einige Artikel lassen sich in keine der bisherigen Kategorien einteilen und werden im Folgenden kurz wiedergegeben.
Die Anzahl der Patienten, die behandelt werden müssen, um ein zusätzliches Ereignis zu verhindern (number needed to treat; NNT), ist bei klinischen Studien mit binärem Endpunkt ein weitverbreitet benutztes Maß für den Behandlungseffekt. Altmann und Andersen (1999) [4] zeigen, wie die NNT für Studien mit Survivalendpunkt berechnet werden kann.
Hsieh und Lavori (2000) [31] geben eine Formel zur Berechnung der benötigten Ereignisanzahl für ein Proportional-Hazards-Regressionsmodell mit nicht binären Kovariablen an. Zusätzlich zu der Vorraussetzung proportionaler Hazardraten müssen keine weiteren Annahmen über die Verteilung der Überlebenszeit und der Prädiktorvariablen getroffen werden.
Bei mehreren möglichen Ereignissen wird ein Ereignis E als primärer Endpunkt definiert und alle anderen Ereignisse als „Competing Risks“ aufgefasst. Beipielsweise in der von Fresenius Biotech durchgeführten Finke-Studie werden Competing Risks analysiert. In der Studie wird die Effizienz eines neuen Arzneimittel überprüft, das nach Knochenmark- und Stammzelltransplantation dazu beiträgt, die Abwehrreaktion des Immunsystems gegen das Transplantat (Graft-versus-Host-Disease, GvHD) zu unterdrücken. Der primäre Endpunkt ist das Auftreten einer GvHD von Grad III-IV oder der Tod des Patienten. Competing Risks sind unter anderem die Zeitspanne, bis eine GvHD eintritt, die Häufigkeit und Schwere von Infektionen und die Überlebenszeit der Patienten.
Ein Grund, Competing Risks zu betrachten, ist die Annahme, dass die Studienbehandlung nur den Hazard des primären Endpunkt beeinflusst, nicht aber die der anderen Endpunkte. Außerdem können mögliche Einflüsse der Studienbehandlung auf die Competing Risks untersucht werden. Der Artikel von Schulgen et *al.* (2005) [62] behandelt die Fallzahlplanung für zwei Behandlungsgruppen bei Survivaldaten mit Competing Risks. Das Auftreten der verschiedenen Ereignisse wird in einem Multi-State-Modell dargestellt. Es wird angenommen, dass die Rekrutierungszeitpunkte einer Gleichverteilung folgen und weder Noncompliance noch Loss to Follow-up auftreten. Im ersten Schritt der Fallzahlberechnung wird die Anzahl benötigter Ereignisse nach Schoenfeld (1981, 1983) berechnet. Zur Bestimmung der Fallzahl werden im zweiten Schritt die ereignisspezifischen Intensitäten, die Länge der Rekrutierungs- und Follow-up-Phasen und die geschätzte Rekrutierungsrate festgelegt. Dabei ist die benötigte Patientenanzahl $N = d/\Psi$, wobei d die Anzahl benötigter Ereignisse und Ψ die Wahrscheinlichkeit ist,

einen primären Endpunkt zu beobachten. Diese Wahrscheinlichkeit lässt sich mit der Länge der Rekrutierungsphase R und der Länge der Follow-up-Phase F durch

$$\Psi = \frac{\lambda_1 + \lambda_2}{\lambda}\left[1 - \frac{\exp(-\lambda F) - \exp(-\lambda(R+F))}{\lambda R}\right] \qquad \text{mit } \lambda = \lambda_1 + \lambda_2 + \lambda_3 \qquad (4.59)$$

berechnen, wobei $\lambda_i(t)$, $i = 1, 2, 3$, die Hazardfunktion für den Wechsel von Status 0 zu Status i zum Zeitpunkt t bezeichnet.

5. Programmvergleich zur Fallzahlberechnung

In diesem Kapitel werden die kommerziellen Programmpakete nQuery, PASS, ADDPLAN und East vorgestellt. Zunächst werden die Programmeigenschaften beschrieben und die zugrundeliegenden Formeln angegeben. Es wird untersucht, ob in den Programmen neu entwickelte Methoden aus Kapitel 4 zur Berücksichtigung von beispielsweise nichtproportionalen Hazardraten oder von mehr als zwei Behandlungsgruppen implementiert sind. Anschließend werden mit den unterschiedichen Programmen berechnete Fallzahlen für verschiedene Szenarien miteinander verglichen.

5.1. Programmeigenschaften und zugrundeliegende Formeln

5.1.1. PASS

PASS 2008 [29] ist ein menügesteuertes Programm zur Fallzahlberechnung bei verschiedenen Datentypen. Für Daten mit Survivalendpunkt kann die Fallzahl für den Logranktest nach Freedman, Lachin und Foulkes oder Lakatos, zum Nachweis der Unterlegenheit einer Behandlung oder für gruppensequentielle Designs berechnet werden.
Voraussetzung für die Verwendung der Fallzahlformel für den Logranktest nach Freedman (1982) [21] sind proportionale Hazardraten. Die verwendeten Formeln stammen aus Machin et *al.* (1997) [47]. Sei θ das Verhältnis der Hazardraten von experimenteller zur Kontrollgruppe. $S_1(T)$ und $S_2(T)$ geben die Anteile der Patienten an, die in der Kontroll- und experimentellen Gruppe die gesamte Studiendauer überleben. Durch den Logranktest wird die Gleichheit der Hazardraten in experimenteller und Kontrollgruppe ($H_0 : \lambda_1 = \lambda_2$) überprüft. Zur Berechnung der Power wird im Benutzerhandbuch [29] die Formel

$$z_{1-\beta} = \frac{|\theta - 1|\sqrt{N(1 - loss)r[(1 - S_1(T)) + r(1 - S_2(T))]/(1 + r)}}{(1 + r)\theta} - z_{1-\alpha/s} \qquad (5.1)$$

angegeben, wobei s gleich 1 für einen einseitigen und gleich 2 für einen zweiseitigen Test ist. $loss$ ist der Anteil der Patienten, die aus der Studie ausgeschieden sind, und r gibt das Verhältnis der Umfänge von experimenteller zur Kontrollgruppe an. Durch Umstellen erhält man die in Kapitel 3.1 angegebene Fallzahlformel von Freedman 3.2 für den Fall ungleicher Gruppengrößen und Berücksichtigung von Loss to Follow-up:

$$N = \frac{d(r + 1)}{r[(1 - S_1(T)) + r(1 - S_2(T))](1 - loss)} \qquad (5.2)$$

mit

$$d = \frac{(r\theta + 1)^2}{(\theta - 1)^2}(z_{1-\alpha/s} + z_{1-\beta})^2. \tag{5.3}$$

Die Berücksichtigung von Loss to Follow-up erfogt hier auf denkbar einfache Art, indem die Fallzahl durch $1 - loss$ geteilt wird.
Alternativ kann der Logranktest nach Lachin und Foulkes (1986) [35] verwendet werden. Auch hier werden proportionale Hazardraten vorausgesetzt. Zusätzlich können die Länge der Rekrutierungsphase R und Follow-up-Phase F in die Berechnung miteinbezogen werden. Die gesamte Studiendauer wird mit $T = R + F$ bezeichnet. Zusätzlich zur Länge der Rekrutierungsphase wird die Zeit, bis 50% der Patienten rekrutiert werden, eingegeben. PASS legt eine trunkierte Exponentialverteilung zugrunde, deren Parameter γ aus den beiden Benutzerangaben bestimmt wird. Losses to Follow-up werden durch eine Exponentialverteilung modelliert. Der Benutzer kann gruppenspezifische Anteile $P_1(x)$ und $P_2(x)$ eingeben, die lost to Follow-up sein werden. PASS berechnet daraus mit der Formel

$$\eta_i = -\frac{\ln(1 - P_i(x))}{x} \tag{5.4}$$

die Parameter η_1 und η_2 der zugrundeliegenden Verteilung.. Unter Annahme exponentialverteilter Survivalverteilungen mit Hazardraten λ_1 und λ_2 wird im Benutzerhandbuch einer Version der Formel von Lachin und Foulkes' (1986) zur Fallzahlberechnung folgendermaßen angegeben:

$$N = \frac{z_\alpha^2 \phi(\overline{\lambda}) \left(\frac{1}{Q_1} + \frac{1}{Q_2}\right) + z_\beta^2 \left(\frac{\phi(\lambda_1)}{Q_1} + \frac{\phi(\lambda_2)}{Q_2}\right)}{(\lambda_1 - \lambda_2)^2}, \tag{5.5}$$

wobei $Q_i = N_i/N$, $i = 1, 2$, der Patientenanteil in Gruppe i ist und $\overline{\lambda} = Q_1\lambda_1 + Q_2\lambda_2$ ist. $\phi(\lambda, \eta, \gamma)$ berechnet sich aus

$$\phi(\lambda, \eta, \gamma) = \lambda^2 \left(\frac{\lambda}{\lambda + \eta} + \frac{\lambda\gamma e^{-(\lambda+\eta)T}[1 - e^{(\lambda+\eta-\gamma)R}]}{(1 - e^{-\gamma R})(\lambda + \eta)(\lambda + \eta - \gamma)} \right)^{-1}, \tag{5.6}$$

wobei η der Anteil der aus der Studie ausgetretenen Patienten ist. Die Formeln stimmen mit den in Kapitel 3.3 angegebenen überein, wenn $\eta_1 = \eta_2 = \eta$ in beiden Gruppen einheitlich ist. Die im Benutzerhandbuch angegebenen Formeln können jedoch nicht den Berechnungen in PASS zugrundeliegen, da es möglich ist, gruppenspezifische Loss-to-Follow-up-Raten einzugeben. Korrekt wäre eine Angabe der Formeln aus Kapitel 3.3.
Eine weitere Alternative ist die Berechnung des Logranktests nach Lakatos (1988) [37] basierend auf einem Markovmodell. Diese Methode erlaubt die Berücksichtigung vieler praxisrelevanter Parameter. Für jede Gruppe können eigene Crossover- und Loss-to-Follow-up-Raten eingegeben werden. Das Größenverhältnis der Gruppen kann variiert werden. Die Länge der Rekrutierungsphase und entweder die Gleichverteilung oder ein benutzerspezifisches Rekrutierungsmuster kann festgelegt werden. Aus der gesamten Studiendauer wird die Länge der Follow-up-Phase bestimmt. Die Hazardraten müssen im Gegensatz zu den beiden vorigen Prozeduren nicht proportional sein und können vom Benutzer über die Zeit variierend eingegeben werden.

Der Nutzer kann zwischen vier Alternativen zur Spezifikation der Effektgröße wählen. Er kann entweder die Hazardraten, die mittleren Überlebenszeiten, die Survivalrate zu einem bestimmten Zeitpunkt oder die Ereignisraten angeben. Siehe Kapitel 3.5 dieser Arbeit für eine Darstellung des Markovmodells von Lakatos (1988).
Nachstehende Tabelle aus dem Benutzerhandbuch von PASS [29] liefert einen Vergleich der drei Logrankprozeduren: Bei allen drei Methoden kann entweder eine Power- oder eine Fall-

	Algorithm		
Feature/Capability	**Simple (Freedman)**	**Advanced (Lachin)**	**Markov Process (Lakatos)**
Test Statistic	Logrank statistic	Mean hazard difference	Logrank statistic
Hazard Ratio	Constant	Constant	Any pattern including time-dependent
Basic Time Distribution	Constant hazard ratio	Constant hazard ratio (exponential)	Any distribution
Loss to Follow Up Parameters	Yes	Yes	Yes
Accrual Parameters	No	Yes	Yes
Drop In Parameters	No	No	Yes
Noncompliance Parameters	No	No	Yes
Duration Parameters	No	Yes	Yes
Input Hazard Ratios	No	No	Yes
Input Median Survival Times	No	No	Yes
Input Proportion Surviving	Yes	Yes	Yes
Input Mortality Rates	No	No	Yes

Tabelle 5.1.: Vergleich der Prozeduren in PASS.

zahlberechnung durchgeführt werden. Dabei wird die insgesamt benötigte Fallzahl angegeben. Der Output besteht aus einer Tabelle der numerischen Ergebnisse der Berechnungen, Definitionen der verwendeten Parameter, ausformulierten Zusammenfassungen der Ergebnisse und vom Benutzer anforderbaren Graphiken inklusive Beschreibung.

5.1.2. nQuery

nQuery Advisor 7 [17] enthält drei verschiedene Methoden zur Fallzahlberechnung für Daten mit Survivalendpunkt.
Mit der Prozedur STT0 wird die Fallzahl für den Logranktest nach Freedman (1982) [21] unter Annahme einer festen Beobachtungszeit für alle Patienten und eines konstanten Hazardratios berechnet:

$$N = \frac{(z_{1-\alpha/s} + z_{1-\beta})^2(\theta+1)^2}{(2-\pi_1-\pi_2)(\theta-1)^2}. \tag{5.7}$$

Dabei sind $\pi_1(x)$ und $\pi_2(x)$ die Anteile Überlebender in Gruppe 1 bzw. 2 zum Zeitpunkt x, $\theta = \ln(\pi_1(x))/\ln(\pi_2(x)) = \ln(S_2(x))/\ln(S_1(x))$ das Hazardratio. s ist wiederum gleich 1

für einen einseitigen und gleich 2 für einen zweiseitigen Test. Durch Ersetzen von $\pi_i(x)$ durch $1 - S_i(x)$ erhält man die in Kapitel 3.1 angegebene Fallzahlformel. Im Gegensatz zu dieser Formel kann hier kein Loss to Follow-up berücksichtigt werden. Erwartet man, dass der Anteil $loss$ bis zum Ende der Studie ausscheiden wird, können die mit STT0 berechneten Fallzahlen einfach manuell durch $(1 - loss)$ geteilt werden, um für Loss to Follow-up zu adjustieren.
Die Prozedur SST1 berechnet die Fallzahl für einen Test, der auf exponentialverteilten Survivalverteilungen basiert. Aufgrund dieser Annahme liefert die Prozedur im Allgemeinen kleinere Fallzahlen als der Logranktest. Zusätzlich zur Fallzahl pro Gruppe wird die Anzahl insgesamt benötigter Ereignisse zur Erreichung der angegebenen Power berechnet.
Eine Erweiterung von STT1 ist die Prozedur STT2, bei der zusätzlich die Länge der Rekrutierungsphase und der gesamten Studie sowie eine Dropoutrate spezifiziert werden können. Es wird angenommen, dass die Dropouts mit Parameter d exponentialverteilt sind. Um Verwechslungen zu vermeiden, wird in diesem Abschnitt der Dropoutparameter mit d bezeichnet und die Ereignisanzahl dafür mit E. Mit der Tabelle „Conversion to alternate rates“ kann der Dropoutparameter d aus der Angabe des Patientenanteils bestimmt werden, der sich nach einer bestimmten Zeit noch in der Studie befindet. Diese Berechnung muss der Benutzer allerdings selbst durchführen.
Die Formeln für sowohl STT1 als auch STT2 basieren laut Benutzerhandbuch auf Lakatos und Lan (1992) [40], stammen jedoch ursprünglich von Rubinstein et *al.* (1981) [57]. Die Fallzahl wird wie folgt berechnet:

$$N = \left(\frac{z_{1-\alpha/s} + z_{1-\beta}}{\ln(\lambda_2) - \ln(\lambda_1)} \right)^2 \left(\frac{1}{E(P_2)} + \frac{1}{E(P_1)} \right), \tag{5.8}$$

wobei $E(P_i) = \frac{\lambda_i}{\lambda_i + d} \left[1 - \frac{e^{-(\lambda_i+d)(T-R)} - e^{-(\lambda_i+d)T}}{(\lambda_i+d)R} \right]$, R die Dauer der Rekrutierungsphase, T die gesamte Studiendauer, λ_i die Hazardrate in Gruppe i, $i = 1, 2$ ist.
Für alle bisher behandelten Methoden kann die Anzahl benötgter Ereignisse E approximativ mit folgender Formel berechnet werden:

$$E = \frac{4(z_{\alpha/2} + z_\beta)^2}{[\ln(\theta^*)]^2}. \tag{5.9}$$

Die Methode STT3 ist am flexibelsten, da der Benutzer die Studie in Teilintervalle aufteilen und für jedes Intervall Überlebens- , Hazard-, Rekrutierungs- und Dropoutraten festlegen kann. Allerdings ist ausschließlich eine Power- und keine Fallzahlberechnung möglich. Die Anteile bis zum Ende des jeweiligen Intervalls Überlebender und die Hazardraten können für beide Gruppen unterschiedlich sein, das Rekrutierungsmuster und die Dropoutrate müssen dagegen für das gesamte Patientenkollektiv zutreffen. Die Dropoutrate der exponentialverteilten Dropoutverteilung muss wieder vom Benutzer manuell berechnet werden. Es können auch unterschiedliche Gruppengrößen betrachtet werden. Die zugrundeliegenden Methoden oder Formeln werden jedoch nicht im Manual angegeben. Es wird auf Lee (1980) [44] verwiesen. Dort wird angegeben, wie die Teststatistik für den Cox-Mantel-Test gebildet wird. In nQuery werden zwei Behandlungsgruppen unter Verwendung der Benutzereingaben simuliert. Die Power erhält man durch Division der Anzahl der Simulationen, in denen die Teststatistik ein signifikantes Ergebnis liefert durch die Gesamtanzahl der durchgeführten Simulationen.

Es wird kein automatischer Output generiert. Die Tabelle mit den Eingaben und Ergebnissen kann jedoch in ein PDF-Dokument gedruckt werden. Dabei ist auch eine Anzeige der Referenzen möglich. Der Benutzer kann zudem interaktive Grafiken und Summary Statements erstellen.

5.1.3. ADDPLAN

ADDPLAN ist ein Programmpaket zur Planung, Simulation und Analyse klinischer Studien, das eine auf der Kombination von Teststatistiken basierende adaptive Methode verwendet. Sei λ_i die Hazardrate für Behandlungsgruppe i, $i = 1, 2$. Ein Vergleich von mehr als zwei Gruppen ist nicht möglich. ADDPLAN schätzt die Fallzahl für den Logranktest mit der Hypothese $H_0 : \theta = \lambda_2/\lambda_1 = 1$ unter der Annahme proportionaler Hazardraten. Die Anzahl benötigter Ereignisse kann entweder nach Freedman (Formel 3.2) oder nach Schoenfeld (Formel 3.4) berechnet werden. Bei balancierten Designs erhält man mit Schoenfelds Formel kleinere Anzahlen benötigter Ereignisse (Hsieh (1992)[30]).
Der Nutzer kann wählen, ob die Fallzahl entweder auf Basis von Ereignisraten oder Annahmen über Rekrutierungs- und Follow-up-Phase berechnet werden soll. Werden nur Ereignisraten $\pi_i(T) = 1 - S_i(T)$ angegeben, wird die Fallzahl folgendermaßen berechnet:

$$N = \frac{d(1+r)}{\pi_1 + r\pi_2}. \tag{5.10}$$

Werden die Länge der Rekrutierungsphase R und der Follow-up-Phase F spezifiziert, kann die Fallzahl der Kontrollgruppe durch

$$N_1 = \frac{d(1+r)}{\psi_{\lambda_1}(R+F) + r\psi_{\lambda_2}(R+F)} \tag{5.11}$$

bestimmt werden, wobei

$$\psi_{\lambda_i}(R+F) = 1 - e^{-\lambda_i(R+F)}\frac{e^{\lambda_i R} - 1}{\lambda_i R}. \tag{5.12}$$

Die Fallzahl der experimentellen Gruppe ergibt sich durch den Zusammenhang $r = N_2/N_1$ zu $N_2 = rN_1$.
Zusätzlich müssen die maximale Anzahl der Stufen, das globale Signifikanzniveau α, die Bestimmungsmethode der Entscheidungsgrenzen, die Grenze für einen Abbruch wegen Vergeblichkeit, die Informationsraten, die Ereignisraten, Hazardraten oder mittleren Überlebenszeiten, die Power $1 - \beta$ und das Verhältnis der Gruppengrößen eingegeben werden. Die Entscheidungsgrenzen können beispielsweise nach Pocock, O'Bien-Fleming oder einer selbstspezifizierten Alpha Spending Function gewählt werden. Als Informationsraten werden die kummulierten Anteile der Patienten bezeichnet, die in den einzelnen Stufen des Designs rekrutiert sind. Erfolgt der Einschluss der Patienten beispielsweise gleichmäßig in drei Zeiteinheiten, müssen die Werte 1/3, 2/3, 1 eingegeben werden.
Für jede Stufe des Designs wird eine eigene Logranktestatistik berechnet. Da die Folge dieser Statistiken unabhängige und normalverteilte Zuwächse besitzt (Sellke und Siegmund (1982) [64]), können gruppensequentielle Testdesigns angewendet werden.

Alternativ zur Fallzahlschätzung kann auch eine Powerberechnung durchgeführt werden. Außerdem sind Nichtunterlegenheits- (Noninferiority-) und Äquivalenztests möglich. In diesen Fällen wird Schoenfelds Formel, adaptiert für diese Situation, zur Berechnung der benötigten Ereignisse verwendet. Bei einem Nichtunterlegenheitstest muss ein sogenannter Noninferiority Margin δ_0 gewählt werden, der angibt, in welchem Maße man ein schlechteres Abschneiden der experimentellen Gruppe toleriert. Es wird die Nullhypothese $H_0 : \theta = \delta_0$ getestet, wobei $\delta_0 \neq 1$. Falls $\theta - \delta_0 > 0$, lautet die Alternativhypothese $H_1 : \theta > \delta_0$. Ist der Behandlungseffekt $\theta - \delta_0$ negativ, ist die Alternative $H_1 : \theta < \delta_0$. Schoenfelds Formel zur Berechnung der Ereignisanzahl angepasst an dieses Design ist

$$d = \frac{(z_{1-\alpha/s} + z_{1-\beta})^2}{r/(1+r)^2(\ln(\theta^*) - \ln(\delta_0))^2}, \tag{5.13}$$

wobei θ^* das erwartete Hazardratio ist. Auf den Äquivalenztest zweier Behandlungen wird hier nicht genauer eingegangen.
Die Neuschätzung der Fallzahl basierend auf Conditional Power wird in Kapitel 7.3 des ADDPLAN Benutzerhandbuchs beschrieben (ADDPLAN User's Guide (2007) [73]).

5.1.4. East

In East 5 können mit dem Modul EastSurv Überlegenheits- und Nichtunterlegenheitstests mit jeweils sowohl festem als auch variablem Follow-up durchgeführt werden. Ist dieses Modul nicht verfügbar, können Zweistichproben Überlegenheits- und Unterlegenheitstests durchgeführt und geplant werden. Im Gegensatz zu EastSurv müssen bei diesen grundlegenden Prozeduren die Hazardraten proportional und die Rekrutierungsphase gleichverteilt sein und es können keine Dropouts berücksichtigt werden. Im Folgenden wird die Vorgehensweise bei Noninferioritytests ohne Verwendung von East Surv beschrieben.
Unter Nichtunterlegenheit wird verstanden, dass die experimentelle Therapie nicht schlechter als die Kontrollbehandlung ist. Ein solcher Vergleich ist dann sinnvoll, wenn die neue Behandlung ähnlich wirksam ist wie die Kontrolle und zusätzliche Vorteile besitzt wie eine leichtere Darreichungsform, geringere Kosten oder wenn sie weniger Nebenwirkungen verursacht. Die Survivalfunktionen der beiden Behandlungsgruppen werden als exponentialverteilt angenommen. Hier wird der Fall betrachtet, dass das Auftreten eines Ereignises negativ ist, wie beispielsweise ein Rückfall. Der gegenteilige Fall, wenn kleine Ereigniszeiten bevorzugt werden wie im Fall von beispielsweise einer Heilung, ist ähnlich.
Überprüft werden soll, ob die mittlere Überlebenszeit in der experimentellen Gruppe nicht kleiner ist als in der Kontrollgruppe. Die Nullhypothese

$$H_0 : \theta = \frac{m_1}{m_2} = \frac{\lambda_2}{\lambda_1} \geq \delta_0$$

wird gegen die einseitige Alternativhypothese

$$H_1 : \theta < \delta_0$$

getestet. $\delta_0 (\geq 1)$ ist ein Noninferiority Margin, der so definiert wird, dass die experimentelle Behandlung als nicht schlechter angesehen wird als die Kontrollbehandlung, wenn $\theta \leq \delta_0$. Die Berechnung der Fallzahl ist in Kapitel A.5.2 des Benutzerhandbuchs [48] angegeben.

5.2. PASS und nQuery im Vergleich

Dieser Abschnitt dient dem Vergleich der Eigenschaften von East und nQuery und von mit beiden Programmen berechneten Fallzahlen für verschiedene Szenarien.
nQuery ist auf die Spezifikation einer einheitlichen Dropoutrate für beide Behandlungsgruppen festgelegt. Außerdem kann in nQuery kein Crossover berücksichtigt werden. In PASS dagegen ist die Annahme gruppenspezifischer Dropout- und Crossoverraten möglich. Die Beachtung dieser Parameter ist wichtig, um eine Studie realistisch planen zu können, da Dropout und Crossover in der Praxis häufig vorkommen.
Beim Vergleich muss die unterschiedliche Definition des Hazardratios in PASS und nQuery beachtet werden. In PASS ist $\theta = \lambda_2/\lambda_1$ wie in der Notation dieser Diplomarbeit, in nQuery ist es andersherum als $\theta^* = \lambda_1/\lambda_2$ definiert. Ein Nachteil von nQuery ist, dass keine Fallzahl-, sondern nur eine Powerberechnung durchgeführt werden kann. Bei PASS kann der Benutzer zwischen der Berechnung beider Parameter wählen. In nQuery wird die Fallzahl pro Gruppe berechnet, was den Programmvergleich erschwert, da bei der Powerberechnung nur gerade Gesamtfallzahlen betrachtet werden können. PASS berechnet Gesamtfallzahlen, die auch ungerade sein können.
In PASS ist es bei allen Prozeduren möglich, unterschiedliche Gruppengrößen festzulegen. In nQuery besteht diese Möglichkeit nur für die Prozedur STT3. Adaptive Designs und Conditional-Power-Berechnungen sind weder in PASS noch in nQuery implementiert.
Im Folgenden werden mit PASS und nQuery unter bestimmten Annahmen berechnete Fallzahlen miteinander verglichen. In verschiedenen Szenarien wird ein zweiseitiger Test mit Signifikanzniveau $\alpha = 0.05$ und Power $1 - \beta = 0.9$ zum Vergleich zweier Behandlungsgruppen gleicher Größe durchgeführt. Aufgrund der unterschiedlichen Definitionen des Hazardratios in nQuery und PASS wird im Folgenden die Kontrollgruppe in PASS weiterhin als Gruppe 1 und die experimentelle Gruppe als Gruppe 2 bezeichnet. In nQuery wird die Zuordnung umgekehrt. Unter Annahme exponentialverteilter Überlebenszeiten beträgt die mittlere Überlebenszeit in der Kontrollgruppe eine Zeiteinheit. Der Einschluss der Patienten folgt einer Gleichverteilung und findet während der ersten zwei Zeiteinheiten einer gesamten Studiendauer von vier Zeiteinheiten statt. Zum Vergleich der Programme werden sowohl konstante als auch über die Zeit variierende Hazardratios betrachtet. Werden nicht konstante Hazardraten untersucht, wird das Hazardratio θ_1 nach zwei Zeiteinheiten zu θ_2 geändert. Dabei werden für jeden θ_1-Wert (0.1 bis 0.9 in Schritten von 0.1) sowohl der Fall $\theta_1 = \theta_2$ als auch der Wechsel zu vier weiteren θ_2-Werten ($0.1 \leq \theta_2 \leq 0.9$) betrachtet. Außerdem wird für jede Kombination der Hazardratios der Anteil der Patienten, die vor Studienende aus der Studie ausscheiden, von 0.0, 0.05, 0.1 bis 0.6 in Schritten von 0.1 variiert. Da angenommen wird, dass die Follow-up-Periode exponentialverteilt ist, müssen diese Anteile in nQuery manuell in die Parameter einer Exponentialverteilung konvertiert werden. Dies führt zu Parameterwerten von 0.0, 0.0128, 0.0263, 0.0558, 0.0892, 0.1277, 0.1722 und 0.2291. In PASS kann der Anteil direkt eingegeben werden. Wird die Fallzahl mit der Methode von Lakatos berechnet, müssen die Loss-to-Follow-up-Anteile pro Zeiteinheit eingegeben werden. Dies ergibt Werte von 0.0, 0.0125, 0.025, 0.05, 0.075, 0.1, 0.125 und 0.15. Da es in nQuery im Gegensatz zu PASS nicht möglich ist, gruppenspezifische Dropoutraten festzulegen, konnte dieser Fall nicht untersucht werden. Zudem können nichtproportionale Hazardraten in PASS nur mit der Methode nach Lakatos und in nQuery nur mit der Prozedur STT3 betrachtet

werden. Für die Fälle konstanter Hazardraten wurden die Fallzahlen zusätzlich in PASS für die Methoden nach Lachin und Foulkes und nach Freedman und in nQuery für die Prozeduren STT0 und STT2 berechnet. Insgesamt wurden auf diese Weise 1008 Fallzahlen berechnet, miteinander verglichen und in Tabelle nQuery_PASS_R_PH.xls auf beiligender CD gespeichert.

5.2.1. Proportionale Hazardraten

In Tabelle 5.2.1 sind die mit PASS nach Lakatos berechneten Fallzahlen abzüglich der mit der Prozedur STT3 in nQuery berechneten dargestellt. Dabei wurden zur besser Vergleichbarkeit mit PASS berechnete ungerade Werte auf die nächstgrößere gerade Zahl aufgerundet. In Klammern sind die prozentualen Abweichungen bezogen auf n_{nQuery} angegeben. Eine Zahl von -1.5 bedeutet dann, dass die mit nQuery berechnete Fallzahl 1.5 Prozent größer ist als die mit PASS berechnete. Der Fallzahlvergleich bei Annahme proportionaler Hazards liefert folgende Ergeb-

Dropout-rate	**Hazardratio**								
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0	0 *(0%)*	-2 *(-6.7%)*	0 *(0%)*	0 *(0%)*	0 *(0%)*	0 *(0%)*	+2 *(0.5%)*	+18 *(1.8%)*	+4 *(0.1%)*
0.05	0 *(0%)*	0 *(0%)*	-2 *(-4.2%)*	0 *(0%)*	0 *(0%)*	+2 *(1.0%)*	+4 *(1.0%)*	+16 *(1.6%)*	+4 *(0.1%)*
0.1	-2 *(-10.0%)*	0 *(0%)*	-2 *(-4.2%)*	0 *(0%)*	+2 *(1.7%)*	0 *(0%)*	+2 *(0.5%)*	+4 *(0.4%)*	+8 *(0.2%)*
0.2	-2 *(-10%)*	-2 *(-6.3%)*	0 *(0%)*	0 *(0%)*	+2 *(1.6%)*	0 *(0%)*	-2 *(-0.5%)*	+20 *(1.9%)*	0 *(0%)*
0.3	-2 *(-10.0%)*	0 *(0%)*	-4 *(-7.4%)*	-2 *(-2.5%)*	-2 *(-1.5%)*	+2 *(0.9%)*	-12 *(-2.6%)*	-20 *(-1.8%)*	-4 *(-0.1%)*
0.4	-2 *(-9.1%)*	-2 *(-5.9%)*	0 *(0%)*	-2 *(-2.4%)*	-2 *(-1.5%)*	-4 *(-1.7%)*	-10 *(-2.1%)*	-22 *(-1.9%)*	-72 *(-1.4%)*
0.5	-5 *(-20.8%)*	-3 *(-8.3%)*	-4 *(-7.1%)*	-4 *(-4.5%)*	-6 *(-4.2%)*	-10 *(-4.0%)*	-20 *(-4.1%)*	-34 *(-2.8%)*	-172 *(-3.3%)*
0.6	-4 *(-16.7%)*	-6 *(-15.0%)*	-6 *(-10.0%)*	-6 *(-6.5%)*	-8 *(-5.4%)*	-18 *(-6.8%)*	-38 *(-7.3%)*	-60 *(-4.8%)*	-274 *(-5.0%)*

Tabelle 5.2.: Vergleich der mit nQuery und PASS berechneten Fallzahlen bei proportionalen Hazardraten.

nisse. Im Allgemeinen stimmen die Fallzahlen gut überein. Bei großen Dropoutraten sind die prozentualen Unterschiede jedoch hoher als bei kleinen. Bei einer Dropoutrate von mindestens 0.2 ist in den meisten Fällen die mit nQuery berechnete Fallzahl n_{nQuery} größer als die mit PASS berechnete n_{PASS}. Ist die Dropoutrate geringer, ist bei Annahme kleiner Hazardratios n_{nQuery} ebenfalls größer als n_{PASS} und bei Annahme von Hazardratios nahe an 1 ist n_{nQuery} kleiner als n_{PASS}. Betrachtet man die absoluten Fallzahlunterschiede, sind diese für hohe Hazardratios größer als für kleine. Die prozentualen Unterschiede verhalten sich jedoch umgekehrt, da

bei Annahme eines Hazardratios nahe an 1 hohe Fallzahlen benötigt werden. Die Methode von Freedman ist als einzige sowohl in PASS als auch in nQuery implementiert. Werden die eben beschriebenen Szenarien mit beiden Programmen nach Freedman berechnet, erhält man in den meisten Fällen übereinstimmende Ergebnisse. Die Fallzahlen unterscheiden sich höchstens um 5, wobei die Fallzahlen in diesen Fällen sehr hoch sind und der Unterschied damit minimal ist. Alle Fallzahlen sind in der Exceltabelle nQuery_PASS_R_PH.xls gespeichert. Zusätzlich befinden sich die Outputs von nQuery und PASS für die Berechnungen auf der beiliegenden CD.

5.2.2. Nichtproportionale Hazardraten

In Tabelle 5.3 sind die durchschnittlichen absoluten Abweichungen der Fallzahlen zwischen nQuery und PASS für jeden Wert θ_1 des Hazardratio in der ersten Studienhälfte getrennt dargestellt. Wie oben erwähnt, wird jeder θ_1-Wert mit fünf θ_2-Werten kombiniert, wodurch ein Wechsel des Hazardratios in der Studienmitte modelliert wird. Siehe Tabelle nQuery_PASS_R_nonPH.xls auf beiliegender CD, in der die einzelnen Kombinationen der Werte von θ_1 mit den Werten von θ_2 und die resultierenden Fallzahlen dargestellt sind. In Klammern sind die prozentualen durchschnittlichen Abweichungen bezogen auf die mit nQuery berechneten Fallzahlen angegeben.

	Hazardratio								
Dropout-rate	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0	2.4 (*8.4%*)	1.6 (*4.5%*)	1.2 (*2.1%*)	0.4 (*0.6%*)	0.4 (*0.2%*)	0.8 (*0.4%*)	1.6 (*0.6%*)	11.6 (1.5%)	2.4 (*0.1%*)
0.05	2.0 (*3.2%*)	1.2 (*2.0%*)	1.2 (*0.8%*)	0.8 (*0.6%*)	0.8 (*1.1%*)	2.0 (*1.1%*)	3.2 (*1.3%*)	10.0 (*1.3%*)	18.4 (*1.3%*)
0.1	2.8 (*10.4%*)	1.2 (*2.7%*)	1.2 (*2.3%*)	0.8 (*1.1%*)	0.8 (*0.7%*)	0.8 (*0.3%*)	2.0 (*0.7%*)	6.0 (*0.8%*)	16.8 (*0.9%*)
0.2	2.0 (*7.2%*)	1.6 (*4.3%*)	1.2 (*2.1%*)	0.8 (*1.0%*)	1.6 (*1.1%*)	1.6 (*0.8%*)	2.8 (*0.9%*)	10.0 (*1.4%*)	8.4 (*0.8%*)
0.3	2.8 (*9.0%*)	1.6 (*4.0%*)	2.0 (*3.6%*)	2.0 (*2.4%*)	2.0 (*1.8%*)	2.0 (*1.1%*)	12.0 (*3.3%*)	18.0 (*2.5%*)	32.4 (*2.2%*)
0.4	2.4 (*8.5%*)	2.8 (*6.7%*)	2.0 (*3.5%*)	2.4 (*2.7%*)	2.0 (*1.6%*)	6.4 (*3.1%*)	13.6 (*3.8%*)	26.0 (*3.4%*)	83.6 (*4.2%*)
0.5	4.0 (*13.8%*)	2.8 (*6.8%*)	3.6 (*5.8%*)	4.0 (*4.6%*)	7.2 (*5.3%*)	10.0 (*4.6%*)	20.4 (*5.2%*)	38.4 (*4.7%*)	178 (*7.5%*)
0.6	4.4 (*14.6%*)	5.2 (*11.7%*)	5.6 (*8.7%*)	7.2 (*7.7%*)	9.2 (*6.4%*)	18.8 (*7.9%*)	36.8 (*8.6%*)	72.0 (*8.0%*)	285.6 (*10.9%*)

Tabelle 5.3.: Vergleich der mit nQuery und PASS berechneten Fallzahlen bei nicht proportionalen Hazardraten.

Bei Annahme geringer Dropoutraten und mittlerer bis großer Hazardratios stimmen die Fallzahlen der beiden Programme weitestgehend überein. Im Allgemeinen sind die durchschnittlichen

Unterschiede bei einer festen Dropoutrate bei sehr großen und besonders bei sehr kleinen θ_1-Werten am höchsten. Für Hazardratios θ_1 im mittleren Bereich unterscheiden sich die Fallzahlen bei mäßigen Dropoutraten kaum. Betrachtet man dagegen variierende Dropoutraten bei festen Hazardratios θ_1, wird deutlich, dass die Unterschiede bei hohen Dropoutraten am größten sind. Auch hier sind alle Fallzahlen in einer Exceltabelle nQuery_PASS_R_nonPH.xls gespeichert und die Outputs von nQuery und PASS für die Berechnungen befinden auf der beiliegenden CD.

5.3. ADDPLAN und East im Vergleich

Um Unterschiede von mit East und ADDPLAN berechneten Fallzahlen zu untersuchen, werden verschiedene Szenarien in zwei-, drei- und vierstufigen Designs betrachtet. In allen Fällen wird ein Noninferioritytest mit Signifikanzniveau $\alpha = 0.05$ und Power $1 - \beta = 0.9$ durchgeführt. Die mittlere Überlebenszeit beträgt in der Kontrollgruppe 10 Zeiteinheiten. Von Interesse sind Steigerungen der mittleren Überlebenszeiten in der experimentellen Gruppe um 5, 10, 15 und 20 %. Dies resultiert in entsprechende Hazardratios von $\theta_1 = 0.9524$, $\theta_2 = 0.9091$, $\theta_3 = 0.8696$ und $\theta_4 = 0.8333$. Als Noninferioritymargins werden Werte von $\delta_0 = 1.1$ und $\delta_0 = 1.2$ betrachtet. Alle Fallzahlen werden sowohl mit Entscheidungsgrenzen nach Pocock als auch nach O'Brien-Fleming berechnet.
Außerdem wird eine Rekrutierungsphase der Länge 2 und eine gesamte Studiendauer von 4 Zeiteinheiten angenommen. In ADDPLAN können diese Werte direkt eingegeben werden. In East dagegen ist eine direkte Eingabe der Dauern der Studienphasen nicht möglich. Stattdessen wird die Patientenanzahl, die in einer Zeiteinheit rekrutiert werden kann, spezifiziert. Dies erfordert eine komplizierte Eingabe und Variation der Parameterwerte in East, um die beiden Programme vergleichbar zu machen. Dazu wurde die Berechnung zunächst mit der Standardeinstellung von 8 Patienten pro Zeiteinheit durchgeführt. Anschließend wurde dieser Wert so variiert, dass die Eingabe einer Rekrutierungsdauer von 2 Zeiteinheiten möglich war. Nun wurde die Rekrutierungsrate unter Annahme einer Rekrutierungsphase von 2 Zeiteinheiten so variiert, dass als maximale Studiendauer 4 Zeiteinheiten berechnet wurden.
In Tabelle 5.4 sind die absoluten und prozentualen Unterschiede der mit East und ADDPLAN berechneten Fallzahlen dargestellt. In East werden die Fallzahlen durch Multiplikation der Länge der Rekrutierungsphase mit der Patientenanzahl berechnet, die während einer Zeiteinheit rekrutiert werden kann. Da hier die Rekrutierungsphase die Länge zwei besitzt, berechnet East nur gerade Fallzahlen. Deshalb wurden zur besseren Vergleichbarkeit mit ADDPLAN berechnete ungerade Werte auf die nächstgrößere gerade Zahl aufgerundet. Die absoluten Werte sind die Differenz der mit East abzüglich der mit ADDPLAN berechneten Fallzahlen. Dieser Unterschied wird durch den mit ADDPLAN berechneten Wert geteilt und mit 100 multipliziert, um den prozentualen Unterschied zu erhalten, der in Klammern angegeben ist.
Auch wenn die absoluten Unterschiede der Fallzahlen in einigen Fällen groß erscheinen, muss man zur Interpretation der Tabelle berücksichtigen, dass für die untersuchten Szenarien Fallzahlen von mindestens $N = 1512$ berechnet wurden. Deshalb sollten vor allem die proentualen Abweichungen betrachtet werden, die in keinem Fall über 1.5% liegen und damit in allen Fällen sehr gering sind.
In dem zweistufigen Design mit nach Pocock bestimmten Entscheidungsgrenzen ist die mit

I	δ_0	Entscheidungs-grenze	$\theta_1 = 0.9524$	$\theta_2 = 0.9091$	$\theta_3 = 0.8696$	$\theta_4 = 0.8333$
2	$\delta_0 = 1.1$	PK	44 *(0.4%)*	34 *(0.6%)*	34 *(0.9%)*	30 *(1.1%)*
		OBF	-62 *(-0.7%)*	-24 *(-0.5%)*	-10 *(-0.3%)*	-2 *(-0.1%)*
	$\delta_0 = 1.2$	PK	16 *(0.4%)*	16 *(0.6%)*	18 *(0.9%)*	18 *(1.1%)*
		OBF	-24 *(-0.7%)*	-14 *(-0.6%)*	-6 *(-0.3%)*	0 *(-0.0%)*
3	$\delta_0 = 1.1$	PK	-52 *(-0.5%)*	-22 *(-0.4%)*	-4 *(-0.1%)*	2 *(0.1%)*
		OBF	-54 *(-0.6%)*	-22 *(-0.4%)*	-6 *(-0.2%)*	0 *(0.0%)*
	$\delta_0 = 1.2$	PK	-22 *(-0.5%)*	-10 *(-0.3%)*	-4 *(-0.2%)*	2 *(0.1%)*
		OBF	-22 *(-0.6%)*	-12 *(-0.5%)*	-4 *(-0.2%)*	0 *(0.0%)*
4	$\delta_0 = 1.1$	PK	-158 *(-1.5%)*	-82 *(-1.3%)*	-46 *(-1.1%)*	-26 *(-0.8%)*
		OBF	-46 *(-0.5%)*	-18 *(-0.3%)*	-4 *(-0.1%)*	2 *(0.1%)*
	$\delta_0 = 1.2$	PK	-62 *(-1.5%)*	-38 *(-1.3%)*	-26 *(-1.2%)*	-16 *(-0.9%)*
		OBF	-18 *(-0.5%)*	-10 *(-0.4%)*	-4 *(-0.2%)*	0 *(0.0%)*

Tabelle 5.4.: Vergleich der mit East und ADDPLAN berechneten Fallzahlen.

ADDPLAN berechnete Fallzahl $n_{ADDPLAN}$ geringer als die mit East berechnete n_{East}, in Designs mit drei und vier Stufen ist $n_{ADDPLAN}$ größer als n_{East}. Werden die Entscheidungsgrenzen nach O'Brien-Fleming bestimmt, ist sowohl in dem zwei-, als auch in den drei- und vierstufigen Designs $n_{ADDPLAN}$ größer als n_{East}.
Für Szenarien mit Noninferioritymargins von $\delta_0 = 1.1$ sind die Fallzahlen stets größer als in mit Noninferioritymargins von $\delta_0 = 1.2$. Auch wenn in ersterem Fall die absoluten Fallzahlunterschiede zwischen den beiden Programmen größer sind als in zweiterem Fall, unterscheiden sich die prozentualen Unterschiede nicht oder nur sehr gering.
Unabhängig von der Wahl der Entscheidungsgrenzen und des Noninferioritymargins sind die prozentualen Unterschiede umso höher, je geringer, das heißt je weiter von 1 entfernt, das Hazardratio ist. Eine Ausnahme davon bilden die Fälle, in denen in einem zweistufigen Design die Fallzahlen mit Entscheidungsgrenzen nach Pocock berechnet werden.
Wassmer und Vandemeulebroecke (2006) [74] geben einen Überblick über Programme, mit denen Fallzahlen für sequentielle und adaptive Designs berechnet werden können. Auch sie kommen zu dem Schluss, dass ADDPLAN und East als gleichwertig angesehen werden können.
In den Tabellen 5.5 und 5.6 und sind die prozentualer Zuwächse der Fallzahlen bei Verwendung eines drei- bzw. vierstufigen anstelle eines zweistufigen Designs dargestellt. Dazu wurden die für obigen Vergleich mit ADDPLAN und East berechneten Fallzahlen verwendet. Zur besseren Vergleichbarkeit der Programme wurden erneut die mit East berechneten Fallzahlen auf den nächst höheren geraden Wert aufgerundet. Tabelle 5.5 enthält die Zuwächse bei mit nach Pocock bestimmten Entscheidungsgrenzen, Tabelle 5.6 die Zuwächse bei mit nach O'Brien-Fleming bestimmten Entscheidungsgrenzen.
Die Gegenüberstellung zeigt, dass der Anstieg der Fallzahlen mit zunehmender Anzahl der Interimsanalysen höher ist, wenn die Entscheidungsgrenzen nach Pocock anstelle von O'Brien-Fleming bestimmt werden. Verwendet man nach Pocock berechnete Entscheidungsgrenzen, ist der Zuwachs der Fallzahl bei mehr als zwei Interimsanalysen höher, wenn ADDPLAN zur Be-

	δ_0	$\theta_1 = 0.9524$	$\theta_2 = 0.9091$	$\theta_3 = 0.8696$	$\theta_4 = 0.8333$	durchschnittlicher Zuwachs
3 statt 2 Interimsanalysen	1.1	5.04 / 4.06	5.04 / 4.06	5.07 / 4.06	5.05 / 4.02	5.05 / 4.05
	1.2	5.04 / 4.05	5.02 / 4.05	5.09 / 4.00	5.08 / 4.07	5.06 / 4.04
4 statt 2 Interimsanalysen	1.1	8.32 / 6.27	8.30 / 6.27	8.34 / 6.24	8.27 / 6.24	8.31 / 6.26
	1.2	8.28 / 6.25	8.29 / 6.29	8.36 / 6.19	8.34 / 6.22	8.32 / 6.24

Tabelle 5.5.: Prozentuale Zuwächse der mit ADDPLAN/East berechneten Fallzahlen bei Verwendung eines drei- bzw. vierstufigen anstelle eines zweistufigen Designs mit nach Pocock berechneten Entscheidungsgrenzen.

	δ_0	$\theta_1 = 0.9524$	$\theta_2 = 0.9091$	$\theta_3 = 0.8696$	$\theta_4 = 0.8333$	durchschnittlicher Zuwachs
3 statt 2 Interimsanalysen	1.1	1.05 / 1.15	1.05 / 1.13	1.01 / 1.12	1.07 / 1.15	1.05 / 1.14
	1.2	1.07 / 1.13	1.03 / 1.12	1.05 / 1.16	1.06 / 1.06	1.05 / 1.12
4 statt 2 Interimsanalysen	1.1	1.75 / 1.94	1.76 / 1.92	1.73 / 1.91	1.76 / 1.92	1.75 / 1.92
	1.2	1.74 / 1.93	1.75 / 1.92	1.79 / 1.90	1.85 / 1.85	1.78 / 1.90

Tabelle 5.6.: Prozentuale Zuwächse der mit ADDPLAN/East berechneten Fallzahlen bei Verwendung eines drei- bzw. vierstufigen anstelle eines zweistufigen Designs mit nach O'Brien-Fleming berechneten Entscheidungsgrenzen.

rechnung benutzt wird, als wenn East verwendet wird. Liegen dagegen nach O'Brien-Fleming berechnete Entscheidungsgrenzen zugrunde, sind umgekehrt die Zuwächse bei den mit East berechneten Fallzahlen höher als bei den mit ADDPLAN berechneten. Die Unterschiede der Zuwächse sind in diesem Fall jedoch geringer als bei nach Pocock berechneten Entscheidungsgrenzen. Die Wahl eines Noninferioritymargins von $\delta_0 = 1.1$ im Vergleich zu $\delta_0 = 1.2$ hat sozusagen keinen Einfluss auf die Zuwächse. Bei diesen Vergleichen muss jedoch die begrenzte Anzahl der untersuchten Fälle beachtet werden. Zur Verallgemeinerung dieser Ergbnisse müssen weitere Szenarien betrachtet werden.
In der Datei ADDPLAN_East.xls auf beiliegender CD sind zusätzlich zu den Fallzahlen die Ereignisanzahlen und weitere Werte für die untersuchten Szenarien gespeichert. Im unteren Bereich befinden sich die Berechnungen der prozentualen Zuwächse der Fallzahlen, die den Tabellen 5.5 und 5.6 zugrundeliegen. Um die Berechnungen nachvollziehen zu können, befinden sich die Outputs der Programme ebenfalls auf der CD.

6. Programmierung und Validierung eines R-Programms

Ein Ziel dieser Diplomarbeit ist die Programmierung eines Programms in R zur Fallzahlberechnung bei Daten mit Survivalendpunkt. In diesem Kapitel wird die Implementierung der grundlegenden Methoden von Freedman, Schoenfeld, Lachin und Foulkes und Rubinstein und einer Erweiterung von Heo et *al.* beschrieben. Außerdem wird das Programm durch Vergleiche mit Beispielen aus der Litertur und den Szenarien aus Kapitel 5.2 validiert. In allen R-Programmen werden statt der Indizes $_1$ bzw. $_2$ die Indizes $_c$ bzw. $_e$ zur Markierung der Zugehörigkeit zur Kontroll- bzw. experimentellen Gruppe verwendet.

6.1. Freedman

Die Fallzahlberechnung mit Formel 3.2 von Freedman wurde zunächst in der R-Funktion `Freedman.r` implementiert und anschließend in die Funktion `samplezize.r` in die Methode „freedman“ eingebunden. Bei der Berechnung der Szenarios aus Kapitel 5.2 erhält man mit den R-Funktionen dieselben Fallzahlen wie mit der Prozedur PASS Freedman (siehe Tabelle nQuery_PASS_R_PH.xls auf beiliegender CD).

6.2. Schoenfeld

In dem Programm `Schoenfeld.r` sind die Formeln 3.4 und 3.5 aus Kapitel 3.2 implementiert. Sollen auch die Länge von Rekrutierungs- und Follow-up-Phase berücksichtigt werden, kann das Programm `Schoenfeld1982.r` verwendet werden, in dem die Formeln 3.6 bis 3.9 aus Kapitel 3.2 umgesetzt sind. Beide Funktionen sind die in die R-Funktion `samplesize.r` in den Methoden „schoenfeld“ und „schoenfeld1982“ eingebunden. Ein Vergleich mit nQuery und PASS ist nicht möglich, da diese Formeln dort nicht implementiert sind. Siehe Kapitel 6.5 für die Möglichkeit, mit der Methode von Schoenfeld die Weibullverteilung statt der Exponentialverteilung zur Modellierung der Überlebenszeiten zu verwenden.

6.3. Lachin und Foulkes

Im Folgenden werden die Funktionen `lachin_rd_loss_nonuniform.r` und `lachin_rd_loss_nonuniform_anteil.r` schrittweise aufgebaut, indem immer mehr Annahmen getroffen werden. Die Funktionen der Zwischenschritte können durch konkrete Zahlenbeispiele im Artikel von Lachin und Foulkes (1986) auf ihre Richtigkeit hin überprüft werden und wurden hier ausschließlich zur Validierung des Programms implementiert.

Nur die finale Funktion `lachin_rd_loss_nonuniform_anteil.r` ist in das Programm `samplesize.r` eingebunden. Zu den beiden finalen Funktionen gibt es im Artikel von Lachin und Foulkes (1986) keine Zahlenbeispiele. Deshalb wurden einige Szenarien zum Vergleich von `lachin_rd_loss_nonuniform_anteil.r` und PASS berechnet und miteinander verglichen. Die Formel von Lachin und Foulkes ist bereits in der R-Funktion `nSurvival.r` im Paket gsDesign implementiert, allerdings muss dort direkt der Parameter der trunkierten Exponentialverteilung, die als Rekrutierungsverteilung verwendet wird, eingegeben werden. Die Spezifikation eines Anteils der Rekrutierungsphase, der vergangen sein wird, bis 50% der Patienten in die Studie eingeschlossen sein werden wie in PASS ist nicht möglich.
Unter Verwendung der Differenz der Hazardraten $\lambda_2 - \lambda_1$ wird von Lachin und Foulkes die Fallzahlformel

$$N = \frac{(z_{\alpha/s}\sqrt{\phi(\overline{\lambda})(Q_2^{-1} + Q_1^{-1})} + z_\beta\sqrt{\phi(\lambda_2)Q_2^{-1} + phi(\lambda_1)Q_1^{-1}})^2}{(\lambda_2 - \lambda_1)^2} \tag{6.1}$$

angegeben, wobei $Q_1 = N_1/N$, $Q_2 = N_2/N$, $\overline{\lambda} = Q_1\lambda_1 + Q_2\lambda_2$ und $\phi(\lambda)$ die Komponente der Varianz des Maximumlikelihoodschätzers von λ ist, die unabhängig von der Fallzahl ist. Alternativ geben sie die Formel von George und Desu (1974) [22] an, die das Hazardatio λ_1/λ_2 verwenden:

$$N = \frac{\left(z_{\alpha/s}\sqrt{\frac{\phi(\overline{\lambda})}{\overline{\lambda}^2}(Q_2^{-1} + Q_2^{-1})} + z_\beta\sqrt{\frac{\phi(\lambda_2)}{\lambda_2^2 Q_2} + \frac{\phi(\lambda_1)}{\lambda_1^2 Q_1}}\right)^2}{(\ln(\lambda_1/\lambda_2))^2}. \tag{6.2}$$

Bei Annahme einer gleichverteilten Rekrutierungsphase und ausschließlich administrativer Zensierungen zeigte Lachin (1981) [34], dass

$$\phi(\lambda) = \lambda^2\left[1 - \frac{e^{-\lambda(T-R)} - e^{-\lambda T}}{\lambda R}\right]^{-1}. \tag{6.3}$$

In diesem Fall ist die unter der Alternativhypothese erwartete Ereignisanzahl

$$d = \frac{N_1\lambda_1^2}{\phi(\lambda_1)} + \frac{N_2\lambda_2^2}{\phi(\lambda_2)}. \tag{6.4}$$

Die Formeln 6.1 und 6.2 wurden in den Funktionen `lachin_rr.r` und `lachin_rd.r` implementiert. Zur Validierung von `lachin_rd.r` wurde ein Beispiel von Lachin (1981) verwendet. Ein Unterschied zwischen den Hazardraten der beiden Behandlungsgruppen von $\lambda_1 = 0.3$ und $\lambda_2 = 0.2$ soll bei einem einseitigen Signifikanzniveau von $\alpha = 0.05$ und Power $1 - \beta = 0.9$ in einer fünfjährigen Studie mit dreijähriger Rekrutierungsphase und balanciertem Design entdeckt werden. Lachin (1981) berechnet die Fallzahl 378 und die unter H_1 erwartete Ereignisanzahl 215. Die Funktion `lachin_rr.r` berechnet die Fallzahl 376.1823 und die unter H_1 erwartete Ereignisanzahl 213.3074. Beide Ergebisse von `lachin_rd.r` liegen aufgerundet eins unter den im Artikel angegebenen, stimmen jedoch mit den mit der R-Funktion nSurvival berechneten Werten überein.

Soll ein nicht gleichverteilter Einschluss der Patienten in die Studie modelliert werden, kann eine trunkierte Exponentialverteilung mit Parameter γ verwendet werden. Deren Verteilungsfunktion ist in Gleichung 3.10 angegeben. Die in der Fallzahlformel von Lachin und Foulkes verwendete Funktion $\phi(\lambda)$ lautet in diesem Fall

$$\phi(\lambda) = \lambda^2 \left\{ 1 + \frac{\gamma e^{-\lambda T}[1 - e^{(\lambda-\gamma)R}]}{(1 - e^{-\gamma R})(\lambda - \gamma)} \right\}^{-1} . \tag{6.5}$$

Diese Methode ist in der R-Funktion `lachin_rd_nonuniform_gamma.r` implementiert. Wie in PASS soll es möglich sein, nicht γ direkt einzugeben, sondern den prozentualen Anteil A_{prop} der Rekrutierungsphase, der vergangen ist, bis 50% der Patienten in die Studie eingeschlossen wurden. Durch eine iterative Suche wird der zu diesem Anteil gehörige Wert von gamma bestimmt. In R wird dazu die Funktion uniroot() verwendet, die die Nullstelle der Gleichung

$$1 - \frac{\exp(-\gamma A_{prop} R))}{(1 - \exp(-\gamma R))} - 0.5 = 0 \tag{6.6}$$

iterativ bestimmt. Die Methode ist in der R-Funktion `lachin_rd_nonuniform_anteil.r` implementiert.
In Tabelle 6.1 werden die mit den R-Funktionen `lachin_rd_nonuniform_anteil.r` und `lachin_rd_nonuniform_gamma.r` und mit PASS berechneten Fallzahlen den Ergebnissen in Tabelle 1 in Lachin und Foulkes (1986) [35] gegenübergestellt. In wenigen Fällen

γ	**Anteil**	**mit R berechnete Fallzahl**	**mit PASS berechnete Fallzahl**	**Fallzahl aus Lachin und Foulkes (1986)**
0.0	0.5	377	377	378
-0.5	0.5619	404	404	404
-1.0	0.6201	430	430	430
-1.5	0.6722	452	452	452
-2.0	0.7169	468	468	468
-2.5	0.7543	480	480	480
-3.0	0.7851	489	489	490
-3.5	0.8105	496	496	496
-4.0	0.8313	502	502	502
-4.5	0.8484	506	506	506
-5.0	0.8627	510	510	510
-5.5	0.8747	513	513	512
-6.0	0.8849	515	515	516

Tabelle 6.1.: Vergleich der mit der R-Funktion `lachin_rd_nonuniform_gamma.r` und PASS berechneten Fallzahlen mit Lachin und Foulkes (1986) [35].

ist die mit R berechnete Fallzahl um eins kleiner als die im Artikel angegebene bei sonst völliger

Übereinstimmung. Die mit PASS berechneten Fallzahlen stimmen mit den mit R berechneten überein. Auch die Berechnung der selben Szenarios mit der R-Funktion nSurvival liefert exakt die gleichen Fallzahlen wie die selbst programmierte Funktion.
Loss to Follow-up wird durch gruppenspezifische Exponentialverteilungen mit Parametern η_2 und η_1 modelliert. Außerdem wird zunächst eine gleichverteilte Rekrutierungsphase angenommen. Im Vergleich zu `lachin_rd.r` unterscheiden sich die Definitonen von $\phi()$ und N und η_1 und η_2 werden zusätzlich berücksichtigt. Die Funktion $\phi()$ wird in der Fallzahlformel durch

$$\phi(\lambda, \eta) = \lambda^2 \left\{ \frac{\lambda}{\eta + \lambda} \left[1 - \frac{e^{-(T-R)(\eta+\lambda)} - e^{-T(\eta+\lambda)}}{R(\eta + \lambda)} \right] \right\}^{-1} \tag{6.7}$$

ersetzt. Da $\phi()$ jetzt zusätzlich zu λ von η abhängt, ergibt sich die neue Fallzahlformel

$$N = \frac{(z_{\alpha/s}\sqrt{\phi(\overline{\lambda}, \eta_2)Q_2^{-1} + \phi(\overline{\lambda}, \eta_1)Q_1^{-1}}) + z_\beta\sqrt{\phi(\lambda_2, \eta_2)Q_2^{-1} + phi(\lambda_1, \eta_1)Q_1^{-1}})^2}{(\lambda_2 - \lambda_1)^2}. \tag{6.8}$$

Diese Formeln sind in der R-Funktion `lachin_rd_loss.r` implementiert. In Tabelle 6.2 werden die mit dieser Funktion und und mit PASS berechneten Fallzahlen den Werten in Tabelle 3b in Lachin und Foulkes (1986) [35] gegenübergestellt. In einigen Fällen sind sowohl die

$\eta_1 \rightarrow$ $\eta_2 \downarrow$	0	.05	.10	.15	.20
0	377	393	409	427	444
	377	**393**	**409**	**427**	**444**
	378	*394*	*410*	*428*	*444*
.05	390	406	423	440	457
	390	**406**	**423**	**440**	**457**
	390	*406*	*424*	*440*	*458*
.10	403	420	436	453	471
	403	**419**	**436**	**453**	**471**
	404	*420*	*436*	*454*	*472*
.15	417	434	450	467	485
	417	**434**	**450**	**468**	**485**
	418	*434*	*450*	*468*	*486*
.20	432	448	465	482	500
	432	**448**	**465**	**482**	**500**
	432	*448*	*466*	*482*	*500*

Tabelle 6.2.: Vergleich der mit der R-Funktion `lachin_rd_loss.r` und **PASS** berechneten Fallzahlen mit *Lachin und Foulkes (1986)* [35].

mit R als auch mit PASS berechneten ungeraden Fallzahlen um eins kleiner als die im Artikel angegebenen bei ansonsten völliger Übereinstimmung. Das mag daran liegen, dass im Artikel nur gerade Fallzahlen angegeben werden, da es sich um zwei gleichgroße Gruppen handeln soll.

PASS und das R-Programm berechnen jedoch auch ungerade Fallzahlen. Auffällig ist, dass bei Vertauschung der Werte von η_1 und η_2 keine übereinstimmenden Fallzahlen berechnet werden, sondern dass die Dropoutraten tatsächlich als gruppenspezifisch angesehen werden müssen.
Der Vergleich mit der R-Funktion `nSurvival.r` kann nur eingeschränkt erfolgen, da die Eingabe gruppenspezifischen Dropoutraten bei `nSurvival.r` nicht möglich ist. Folglich werden nur die Fälle $\eta_1 = \eta_2$ betrachtet und die mit `lachin_rd_loss.r` beziehungsweise mit `nSurvival` berechneten Fallzahlen miteinander verglichen. Beide Funktionen liefern exakt die selben Ergebnisse. Ersetzt man in `nSurvival.r` η durch Durchschnittswerte von η_1 und η_2, führt das nicht zu den gleichen Ergebnissen wie die Berücksichtigung gruppenspezifischer Werte in `lachin_rd_loss.r`, da η_1 und η_2 nicht austauschbar sind.
Soll Loss to Follow-up und zusätzlich auch nicht gleichverteilter Eintritt der Patienten in die Studie berücksichtigt werden, muss die Definition von ϕ geändert werden zu

$$\phi(\lambda, \eta, \gamma) = \lambda^2 \left(\frac{\lambda}{\eta + \lambda} + \frac{\lambda\gamma e^{-(\lambda+\eta)T}[1 - e^{(\lambda+\eta-\gamma)R}]}{(1 - e^{-\gamma R})(\lambda + \eta)(\lambda + \eta - \gamma)} \right)^{-1}. \tag{6.9}$$

Die Fallzahlformel ist in Gleichung 3.12 angegeben.
Diese ist in der R-Funktion `lachin_rd_loss_nonuniform.r` implementiert.
Es ist möglich, wie in PASS statt η einen Anteil FU einzugeben, der bis zum Studienende Lost to Follow-up ist, da dieser für den Benutzer leichter zu spezifizieren ist. Dann kann η mit $\eta = -\ln(1-FU)/T$ daraus berechnet werden. Zusätzlich soll wieder nicht der Parameter γ der Rekrutierungsverteilung, sondern der Anteil A_{prop} spezifiziert werden können. In der Funktion `lachin_rd_loss_nonuniform_anteil.r` ist die benutzerfreundliche Eingabe dieser beiden Parameter möglich.
Für diese finalen Methoden werden keine konkreten Fallzahlen im Artikel von Lachin und Foulkes angegeben. Aus diesem Grund werden die mit `lachin_rd_loss_nonuniform.r`, mit `lachin_rd_loss_nonuniform_anteil.r`, mit `nSurvival.r`und mit PASS für bestimmte Szenarien berechneten Fallzahlen einander gegenübergestellt. Es wird wieder das Beispiel aus Lachin (1981) [34] verwendet (siehe oben), wobei γ von -6 bis 6 in 1er-Schritten und η bzw. η_1 und η_2 von 0 bis 2 in 0.05er-Schritten variiert werden.
Die mit `lachin_rd_loss_nonuniform.r` und `nSurvival.r` berechneten Fallzahlen sind in allen Fällen exakt gleich. Die mit `lachin_rd_loss_nonuniform_anteil.r` berechneten Fallzahlen stimmen bis mindestens zur zweiten Nachkommastelle völlig mit den anderen Fallzahlen überein. In drei der 65 Fälle weicht die mit PASS berechnete Fallzahl um 1 von den mit den drei R-Funktionen berechneten ab, was jedoch einer Abweichung von den mit R-berechneten Fallzahlen von höchstens 0.25% entspricht.
Die Funktion `lachin_rd_loss_nonuniform_anteil.r` ist eingebunden in das Programm `samplesize.r`, so dass alle oben behandelten Teilschritte enthalten sind, wenn als Methode „lachin“ gewählt wird. Zur Validierung von `samplesize.r` wurden die Szenarios aus Kapitel 5.2 berechnet. In allen Fällen stimmen die Fallzahlen mit den mit der PASS-Prozedur berechneten überein (siehe Tabelle nQuery_PASS_R_PH.xls auf beiligender CD).

6.4. Rubinstein et al.

Im Benutzerhandbuch von nQuery wird angegeben, dass die Prozedur STT2 auf Lakatos und Lan(1992) [40] basiert. Dieser Artikel ist allerdings ein Überblick über schon veröffentlichte Methoden und die STT2 zugrundeliegende Formel stammt ursprünglich von Rubinstein et *al.* (1981) [57]. Obwohl die Formel gruppenspezifische Loss-to-Follow-up-Raten enthält, ist es in nQuery wie in Kapitel 5.1.2 beschrieben nur möglich, eine einheitliche Loss-to-Follow-up-Rate für beide Gruppen zu spezifizieren. Zudem könnte es für den Benutzer schwierig sein, die Parameter η_1 und η_2 der Exponentialverteilungen direkt anzugeben. Deshalb kann in der R-Funktion `Rubinstein.r` der Anteil $loss_1$ der Patienten, der in der Kontrollgruppe zu einem bestimmten Zeitpunkt t_{loss} lost to Follow-up sein wird, spezifiziert werden. Daraus wird durch die Formel

$$\eta_1 = -\frac{\ln(1 - loss_1)}{t_{loss}} \tag{6.10}$$

der Parameter der Exponentialverteilung berechnet. Auf dieselbe Art wird aus der Eingabe eines Anteils $loss_2$ der Parameter η_2 bestimmt. Die Fallzahl wird durch die Gleichungen 3.13 und 3.14 berechnet. Auch `Rubinstein.r` ist in die Gesamtfunktion `samplesize.r` in der Methode „rubinstein“ eingebunden. Berechnung der Szenarien aus Kapitel 5.2 zeigt, dass die mit dem R-Programm nach Rubinstein et *al.* berechneten Fallzahlen mit den mit der entsprechenden Prozedur STT2 in nQuery übereinstimmen. Eine Ausnahme bilden 11 Szenarien, in denen das Hazardratio nahe an 1 liegt, und bei denen sich die Fallzahlen um 2 unterscheiden. In Relation zu der Größe der mit nQuery berechneten Fallzahlen gesetzt, beträgt dieser Unterschied jedoch höchstens 0.17% (siehe Tabelle nQuery_PASS_R_PH.xls auf beiligender CD).

6.5. Heo et al.

In Kapitel 4.1 wird die Erweiterung der Methode von Schoenfeld und Richter (1982) [61] auf weibullverteilte Überlebenszeiten durch Heo et *al.* (1998) [28] beschrieben. Auf diese Weise können nichtproportionale Hazardraten berücksichtigt werden. Die Survivalverteilung von Gruppe i ist in Formel (4.1) und die Berechnung der Fallzahl in Formel (4.4) angegeben. Die Formeln sind in der R-Funktion `Heo.r` sowie, bei Wahl der Methode „heo“, in der Funktion `samplesize.r` implementiert.
In dem Artikel von Heo et *al.* werden Fallzahlen für verschiedene Szenarien angegeben. In allen Fällen beträgt das Signifikanzniveau $\alpha = 0.05$, die Power $1-\beta = 0.85$, die mittlere Überlebenszeit in der Kontrollgruppe $m_c = 1$, die Länge der Rekrutierungsphase $R = 1$ bei einer gesamten Studiendauer von $T = 4$. Der Parameter der Weibullverteilung ν wird zwischen 1 (entspricht der Exponentialverteilung), 3 und 5 variiert. Das Hazardratio $\theta = m_c/m_e$ nimmt Werte zwischen 1.0 und 1.5 in Schritten von 0.05 an. In Tabelle 6.3 sind jeweils in der oberen Zeile eines Tabelleneintrags die mit der R-Funktion berechneten Fallzahlen dargestellt, darunter befinden sich die entsprechenden Werte aus dem Artikel von Heo et *al.*. In einigen Fällen unterscheiden sich die Fallzahlen um 1 bei sonstiger völliger Übereinstimmung.
Möchte man in einer Funktion, die eine Exponentialverteilung zugrundelegt, stattdessen eine Weibullverteilung zur Modellierung der Überlebenszeiten verwenden, schlagen Heo et *al.* vor,

$\theta \rightarrow$ $\nu \downarrow$	1.05	1.10	1.15	1.20	1.25	1.30	1.35	1.40	1.45	1.50
1	13358 *13358*	3523 *3522*	1649 *1648*	976 *976*	656 *656*	478 *478*	368 *368*	295 *294*	243 *242*	206 *206*
3	1343 *1342*	352 *352*	164 *164*	97 *96*	65 *64*	47 *46*	36 *36*	29 *28*	24 *24*	20 *20*
5	484 *484*	127 *126*	59 *58*	35 *34*	24 *24*	17 *16*	13 *12*	11 *10*	9 *8*	7 *6*

Tabelle 6.3.: Vergleich der Fallzahlen der Methode „heo“ der R-Funktionen `samplesize.r` mit *Heo et al. (1998)* [28].

die Eingabeparameter m_c, m_e und F mit dem Parameter der Weibullverteilung ν zu potenzieren. Zur Überprüfung dieses Vorgehens wurden obige Szenarien ebenfalls mit der Methode „schoenfeld1982“ der R-Funktion `samplesize.r` nach Potenzieren der entsprechenden Eingabeparameter berechnet. Es wurden exakt dieselben Ergebnisse erhalten wie mit der Methode „heo“. Die Berechnungen sind in dem R-Programm `Heo.r` unter der eigentlichen Funktion dargestellt, um sie nachvollziehen zu können.

6.6. Gesamtfunktion

Mit der Gesamtfunktion `samplesize.r` ist die Berechnung der Fallzahl nach Freedman, Schoenfeld, Schoenfeld und Richter, Lachin und Foulkes, Rubinstein und Heo et *al.* wie in den vorherigen Abschnitten beschrieben möglich. Die Eingabe der Parameter für die verschiedenen Methoden ist in dem R-Programm selbst dokumentiert. In der Gesamtfunktion ist eine Eingabe von zwei beliebigen Parametern möglich, die die Survivalfunktionen der beiden Gruppen beschreiben. Der Benutzer kann entweder die Hazardraten λ_1 und λ_2, die mittleren Überlebenszeiten m_1 und m_2, die Survivalraten zu einem beliebigen Zeitpunkt S_1 und S_2, λ_1 und das Hazardratio θ, λ_2 und θ, m_1 und θ, m_2 und θ, S_1 und θ, S_2 und θ, λ_1 und m_2, λ_1 und S_2, λ_2 und m_1, λ_2 und S_1, m_1 und S_2 oder m_2 und S_1 eingeben. Innerhalb der Funktion werden aus der Eingabe die benötigten Parameter berechnet und zusätzlich zu den eingegebenen Werten werden alle andere Parameterwerte im Output ausgegeben. Dabei werden die Survivalraten am Studienende angegeben. Die Validierung der einzelnen Methoden wurden bereits in den vorherigen Abschnitten beschrieben. Alle R-Funktionen sind auf beiliegender CD gespeichert.

7. Zusammenfassung und Ausblick

Die Fallzahlplanung bei Survivaldaten ist ein aktuelles Forschungsgebiet, in dem es viele Veröffentlichungen gibt. Der Literaturüberblick in Kapitel 4 zeigt die Vielfältigkeit der Artikel und verdeutlicht die ständige Weiterentwicklungen von grundlegenden Methoden. Um die Fallzahl möglichst realistisch zu schätzen, können beispielsweise die Verteilung der Eintrittszeiten der Patienten in die Studie und gruppenspezifische Noncompliance- und Loss-to-Follow-up-Raten berücksichtigt werden. Auch ein Vergleich von mehr als zwei Behandlungsgruppen ist möglich. Speziell im Bereich adaptiver Designs ist weitere Forschung nötig, um Methoden für normalverteilte und binäre Endpunkte auf Überlebenszeiten zu erweitern oder weitere spezielle Möglichkeiten für diese komplexeren Endpunkte zu entwickeln.
Es wurden die kommerziellen Programmpaketen zur Fallzahlberechnung zugrundeliegenden Methoden untersucht. Bei gleichen Annahmen berechnen die Programme abhängig von der verwendeten Prozedur nicht immer übereinstimmende Werte. Die Erklärung für die Differenzen sind die unterschiedlichen zugrundeliegenden Formeln und Methoden, die nur Näherungen sind. Einzig die Methode von Freedman ist sowohl in PASS als auch in nQuery verfügbar und die mit den entsprechenden Prozeduren berechneten Fallzahlen stimmen in den meisten Fällen überein. Alle anderen Prozeduren der beiden Programme basieren auf unterschiedlichen Methoden und sind somit nicht direkt vergleichbar.
Für verschiedene Szenarien mit ADDPLAN und East berechnete Fallzahlen unterscheiden sich um höchstens 1.5%. Dabei wurden allerdings ausschließlich Fallzahlen für Noninferioritytests berechnet. Wassmer(2006) [74], Mitinhaber von ADDPLAN, sieht East jedoch allgemein als gleichwertig mit ADDPLAN an.
Interessiert man sich für den Einfluss der Anzahl von Interimsnalysen auf die Fallzahl, müssen weitere Szenarien berechnet werden, wie sie in Kapitel 5.3 beschrieben sind. In den betrachteten Szenarien ist der Anstieg der Fallzahlen mit zunehmender Anzahl der Interimsanalysen höher, wenn die Entscheidungsgrenzen nach Pocock anstelle von O'Brien-Fleming bestimmt werden. Die Wahl des Noninferioritymargins hat sozusagen keinen Einfluss auf den Zuwachs der Fallzahlen. Durch Berechnung weiterer Szenarien und Vergleich der Fallzahlzuwächse können diese Zusammenhänge eventuell verallgemeinert werden. Als Basis für die Berechnung der Zuwächse sollten dabei auch die Fallzahlen von entsprechenden einstufigen Designs verwendet werden.
Ein Ziel dieser Diplomarbeit war die Programmierung von Methoden zur Fallzahlberechnung in R. Die Programme wurden dazu genutzt, mit kommerziellen Programmen berechnete Fallzahlen zu überprüfen. Die Methoden von Freedman (1982), Schoenfeld (1981), Schoenfeld und Richter (1982), Lachin und Foulkes (1986), Rubinstein (1981) und Heo et *al.* (1998) wurden in der R-Funktion `samplesize.r` implementiert, die sich auf beiliegender CD befindet. Die Vergleiche mit entsprechenden Prozeduren in nQuery und PASS zeigen, dass das R-Programm als gleichwertig betrachtet werden kann, da dieselben Fallzahlen berechnet werden.
Eine Erweiterung der R-Programme ist auf verschiedene Arten möglich. Es könnte die Möglich-

keit geschaffen werden, Ereignisraten anstelle von Survivalraten einzugeben, da diese Parameter in der Praxis ebenfalls sehr geläufig sind. Außerdem könnten alle Formeln zur Fallzahlberechnung nach der Power umgestellt werden, so dass entweder eine Fallzahl- oder eine Powerberechnung durchgeführt werden kann. Auch die Implementierung weiterer Ansätze wie die Methode von Ahnn und Anderson (1995) [2] oder Porcher et *al.* (200) (siehe Kapitel 4.3) sind denkbar. Das R-Programm könnte auch in ein SAS-Makro überführt werden. Bei der Literaturrecherche fiel auf, dass einige Autoren ihrer Artikel SAS-Programme anhängten, in denen ihre Methoden implementiert sind. Dazu sei auf die beiden Artikel von Lakatos ([36], [37]) und SIZE von Shih [66] verwiesen. Jiang et *al.* (2004) geben eine Internetseite an, von der ein SAS-Programm heruntergeladen werden kann, in dem ihre modifizierte Lakatos-Methode implementiert ist.
Auch bei den kommerziellen Programmen werden schrittweise neue Methoden und benutzerfreundlichere Eingabemethoden implementiert. So besteht beispielsweise eine Neuerung von nQuery 7 im Vergleich zu nQuery 5 in der Möglichkeit der Spezifikation eines nicht gleichverteilten Rekrutierungsmusters und von nicht proportionalen Hazardraten. Als Erweiterung von PASS 2004 wurden erst in der aktuellen Version von 2008 zusätzliche Möglichkeiten zur Berücksichtigung von nicht proportionale Hazardraten und Crossover implementiert. Daraus wird ersichtlich, dass die Programme ständig weiter entwickelt werden, um immer mehr wichtige Einflussfaktoren auf Fallzahl und Power berücksichtigen zu können. Angesichts der Vielzahl aktuell veröffentlichter Artikel im Bereich Fallzahlplanung bei Daten mit Survivalendpunkt wird eine Verbesserung und Erweiterung der Programme erwartet.

Danksagung

Mein Dank gilt besonders Herrn Prof. Dr. T. Hothorn und Herrn PD Dr. M. Hennig für die engagierte Betreuung und Unterstützung sowie die vielseitigen Ratschläge und Verbesserungsvorschläge.
Herrn PD Dr. M. Hennig danke ich zudem für die Bereitstellung eines Arbeitsplatzes mit Laptop, an dem ein Großteil dieser Diplomarbeit verfasst wurde, sowie der Programme zur Fallzahlberechnung, die in dieser Arbeit verwendet wurden.
Auch dem gesamten Team Biostatistik/Datenmanagement gilt mein herzlicher Dank für die freundliche Arbeitsatmosphäre und die Anregungen für meine Arbeit.
Bei Herrn Dr. M. Baehr und Herrn A. Bayerstadler möchte ich mich für das Korrekturlesen dieser Arbeit und für die hilfreichen Anmerkungen bedanken.
Nicht zuletzt möchte ich meinen Eltern danken, die mir durch ihre fortwährende Unterstützung das Studium und diese Arbeit ermöglichten und sie mit Anteilnahme verfolgt haben.

A. Übersicht Notation

Dies ist ein kurzer Überblick über die verwendete Notation. Details lese man in Kapitel 2.1 nach.

α	Signifikanzniveau; Fehler erster Art
β	Fehler zweiter Art
N	Fallzahl
N_1	Fallzahl in der Kontrollgruppe
N_2	Fallzahl in der experimentellen Gruppe
r	Verhältnis der Gruppenumfänge N_2/N_1
d	Anzahl benötigter Ereignisse
R	Länge der Rekrutierungsphase
F	Länge der Follow-up-Phase
T	Länge der gesamten Studie
s	Indikator, ob es sich um einen ein- oder zweiseitigen Test handelt
LR	Teststatistik des Logranktests
X_i	Lebensdauer in Gruppe i
$f_i(x)$	Dichtefunktion der Lebensdauer in Gruppe i zum Zeitpunkt x
$F_i(x)$	Verteilungsfunktion der Lebensdauer in Gruppe i zum Zeitpunkt x
$S_i(x)$	Survivalfunktion in Gruppe i zum Zeitpunkt x
$\pi_i(x)$	Ereignisrate in Gruppe i zum Zeitpunkt x
m_i	mittlere Überlebenszeit in Gruppe i
$\lambda_i(x)$	Hazardrate in Gruppe i zum Zeitpunkt x
θ	Hazardratio $= \lambda_2/\lambda_1$
I	Anzahl der Phase in einem gruppensequentiellen oder adaptiven Design
a_i	Zeitpunkt der $i$ten Interimsanalyse
T_i	bei der $i$ten Interimsanalyse berechnete Teststatistik
c_i	Abbruchgrenze für die Teststatistik bei der $i$ten Interimsanalyse
α_i	Signifikanzniveau bei der $i$ten Interimsanalyse

B. Anlage: CD

Auf der beigelegten CD sind die in Kapitel 6 referenzierten R-Funktionen in Tinn-R-Dateien (Endung .r) gespeichert. Außerdem enthält die CD Dateien, anhand derer die in Kapitel 5 erläuterten Fallzahlberechnungen nachvollzogen werden können. Diese sind in Ordnern mit dem jeweiligen Namen des zur Berechnung verwendeten Programms abgelegt. Die Ordnerstruktur wird in dem Worddokument Beschreibung.doc erklärt.

Literaturverzeichnis

[1] ABDA: *Internet: http://www.abda.de/1232.html [Stand: 12.12.2009]*. Bundesverband Deutscher Apothekerverbände, 2007.

[2] AHNN S., ANDERSON S.J.: *Sample Size Determination for Comparing more than Two Survival Distributions*. Statistics in Medicine, 14:2273–2282, 1995.

[3] AHNN S., ANDERSON S.J.: *Sample Size Determination in Complex Clinical Trials Comparing More Than Two Groups for Survival Endpoints*. Statistics in Medicine, 17:2525–2534, 1998.

[4] ALTMANN D.G., ANDERSEN P.K.: *Calculating the number needed to treat for trials where outcome is time to an event*. British Medical Journal, 319:1492–95, 1999.

[5] ARMITAGE P., MCPHERSON C.K., ROWE B.C.: *Repeated signifcance tests on accumulating data*. Journal of the Royal Statistical Society, 132:235–244, 1969.

[6] BARTHEL F.M.-S., BABIKER A., ROYSTON P. ET AL.: *Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over*. Statistics in Medicine, 25:2521–2542, 2006.

[7] BAUER P., EINFALT J.: *Application of Adaptive Designs - a Review*. Biometrical Journal, 48:493–506, 2006.

[8] BAUER P., KÖHNE K.: *Evaluation of Experiments with Adaptive Interim Analyses*. Biometrics, 50:1029–1041, 1994.

[9] BAUER P., POSCH M.: *Letter to the editor: Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections*. Statistics in Medicine, 23:1333–1335, 2004.

[10] BETENSKY, R.A.: *Early Stopping to Accept H0 Based on Conditional Power: Approximations and Comparisons*. Biometrics, 53:794–806, 1997.

[11] CHANG, M.: *Adaptive design method based on sum of p-values*. Statistics in Medicine, 26:2772–2784, 2007.

[12] COOK, T.D.: *Methods for mid-course corrections in clinical trials with survival outcomes*. Statistics in Medicine, 22:3431–3447, 2003.

[13] CORONARY-DRUG-PROJECT-RESEARCH-GROUP: *Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project*. New England Journal of Medicine, 303:1038–1041, 1980.

[14] COX, D.: *Regression models and life tables (with discussion)*. Journal of the Royal Statistical Society, Series B, 34:187–220, 1972.

[15] D'AGOSTINO, R.B.: *Discussion*. Statistics in Medicine, 25:3302–3304, 2006.

[16] DESSEAUX K., PORCHER R.: *Flexible two-stage design with sample size reassessment for survival trials*. Statistics in Medicine, 26:5002–5013, 2007.

[17] ELASHOFF, J.D.: *nQuery Advisor Version 7.0*. Help File, 2007.

[18] ELLENBERG S.S., GOLUB H., METHA C. ET AL.: *Preface*. Statistics in Medicine, 25:3229–3230, 2006.

[19] ENG K.H., KOSOROK M.R.: *A Sample Size Formula for the Supremum Log-Rank Statistic*. Biometrics, 61:86–91, 2005.

[20] FLEMING T.R., HARRINGTON D.P.: *Counting Process and Survival Analysis*. Whiley: New York, 1991.

[21] FREEDMAN, L.S.: *Tables of the Number of Patients Required in Clinical Trials Using the Log Rank Test*. Statistics in Medicine, 1:121–129, 1982.

[22] GEORGE S.L., DESU M.M.: *Planning the size and duration of a clinical trial studying the time to some critical event*. Journal of Cronic Diseases, 27:15–24, 1974.

[23] GU M., LAI T.L.: *Determination of Power and Sample Size in the Design of Clinical Trials with Failure-Time Endpoints and Interim Analyses*. Controlled Clinical Trials, 20:423–438, 1999.

[24] HALABI S., SINGH B.: *Sample size determination for comparing several survival curves with unequal allocations*. Statistics in Medicine, 23:1793–1815, 2004.

[25] HALPERN J., BROWN B.W.J.: *A computer program for designing clinical trials with arbitrary survival curves and group sequential testing*. Controlled Clinical Trials, 14:109–122, 1993.

[26] HARRINGTON D.P., FLEMING T.R.: *A class of rank test procedures for censored survival data*. Biometrika, 69:553–566, 1982.

[27] HAYBITTLE, J.L.: *Repeated assessments of results in clinical trials of cancer treatment*. The British Journal of Radiology, 44:793–797, 1971.

[28] HEO M., FAITH M.S., ALLISON D.B.: *Power and sample size for survival analysis under the Weibull distribution when the whole lifespan is of interest*. Mechanisms of Ageing and Development, 102:45–53, 1998.

[29] HINTZE, J.: *PASS 2008*. NCSS, LCC: Kaysville, Utah, 2008.

[30] HSIEH, F. Y.: *Comparing sample size formulae for trials with unbalanced allocation using the logrank test*. Statistics in Medicine, 11:1091–1098, 1992.

[31] HSIEH F.Y., LAVORI P.W.: *Sample-Size Calculations for the Cox Proportional Hazards Regression Model with Nonbinary Covariates*. Controlled Clinical Trials, 21:552–560, 2000.

[32] JIANG Q, SNAPINN S, IGLEWICZ B.: *Calculation of sample size in survival trials: the impact of informative noncompliance*. Biometrics, 60:800–806, 2004.

[33] JONES D., WHITEHEAD J.: *Sequential forms of the logrank and modified Wilcoxon tests for censored data*. Biometrika, 66:105–133, 1979.

[34] LACHIN, J.M.: *Introduction to Sample Size Determination and Power Analysis for Clinical Trials*. Controlled Clinical Trials, 2:93–113, 1981.

[35] LACHIN J.M., FOULKES M.A.: *Evaluation of Sample Size and Power for Analyses of Survival with Allowance for Nonuniform Patient Entry, Losses to Follow-up, Noncompliance, and Stratification*. Biometrics, 42:507–519, 1986.

[36] LAKATOS, E.: *Sample sizes for clinical trials with time-dependent rates of losses and noncompliance*. Controlled Clinical Trails, 7:189–199, 1986.

[37] LAKATOS, E.: *Sample Sizes Based on the Log-Rank Statistic in Complex Clinical Trials*. Biometrics, 44:229–241, 1988.

[38] LAKATOS, E.: *Designing complex group sequential survival trials*. Statistics in Medicine, 21:1969–1989, 2002.

[39] LAKATOS, E.: *Letters to the Editor*. Controlled Clinical Trials, 23:182–183, 2002.

[40] LAKATOS E., LAN K.K.G.: *A comparison of sample size methods for the logrank statistic*. Statistics in Medicine, 11:179–191, 1992.

[41] LAN K.K.G., DEMETS D.L.: *Discrete sequential boundaries for clinical trials*. Biometrika, 70:659–663, 1983.

[42] LAN K.K.G., ZUCKER D.M.: *Sequential monitoring of clinical trials: the role of information and Brownian motion*. Statistics in Medicine, 12:753–765, 1993.

[43] LAWRENCE, J.: *Strategies for changing the test statistic during a clinical trial*. Journal of Biopharmaceutical Statistics, 12:193–205, 2002.

[44] LEE, E.T.: *Statistical Methods for Survival Data Analysis*. Lifetime Learning Publications: Belmont, California, 1980.

[45] LI B., GRAMBSCH P.: *Sample size calculation in survival trials accounting for time-varying relationship between noncompliance and risk of outcome event*. Clinical Trials, 3:349–359, 2006.

[46] LIN D.Y., OAKES D., YING Z.: *Additive hazards regression with current status data*. Biometrika, 85:289–298, 1998.

[47] MACHIN D., CAMPBELL M., FAYERS P. ET AL.: *Sample Size Tables for Clinical Studies*. Blackwell Science, 2. Auflage, 1997.

[48] MEHTA, C.: *East 4*. Software Documentation, 2007.

[49] MÜLLER H.H., SCHÄFER H.: *Adaptive group sequential designs for clinical trials: combining the advantage of adaptive and classical group sequential approaches*. Biometrics, 57:886–891, 2001.

[50] MÜLLER H.H., SCHÄFER H.: *Authors' Reply*. Statistics in Medicine, 23:1334–1335, 2004.

[51] MÜLLER H.H., SCHÄFER H.: *A general statistical principle for changing a design any time during the course of a trial*. Statistics in Medicine, 23:2497–2508, 2004.

[52] OELLRICH S., FREISCHLÄGER F., BENNER A. ET AL.: *Sample Size Determination on Survival Data - A Review*. Informatik, Biometrie und Epidemiologie in Medizin und Biologie, 28:64–85, 1997.

[53] OLSCHEWSKI M., SCHUMACHER M.: *Sequential analysis of survival times in clinical trials*. Biometrical Journal, 28:273–293, 1986.

[54] PORCHER R., LÉVY V., CHEVRET S.: *Sample size corrections for treatment crossovers in randomized clinical trials with a survival endpoint*. Controlled Clinical Trials, 23:650–661, 2002.

[55] POSCH M., BAUER P.: *Interim analysis and sample size reassessment*. Biometrics, 56:1170–1176, 2000.

[56] PROSCHAN M.A., HUNSBERGER S.A.: *Designed extension of studies based on conditional power*. Biometrics, 51:1315–1324, 1995.

[57] RUBENSTEIN L.V., GAIL M.H., SANTNER T.J.: *Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation*. Journal of Chronical Diseases, 34:469–479, 1981.

[58] SCHÄFER H., MÜLLER H.-H.: *Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections*. Statistics in Medicine, 20:3741–3751, 2001.

[59] SCHOENFELD, D.A.: *The asymptotic properties of nonparametric tests for comparing survival distributions*. Biometrika, 68:316–319, 1981.

[60] SCHOENFELD, D.A.: *Sample-Size Formula for the Proportional-Hazards Regression Model*. Biometrics, 39:499–503, 1983.

[61] SCHOENFELD D.A., RICHTER J.R.: *Nomograms for Calculating the Number of Patients Needed for a Clinical Trial With Survival as an Endpoint*. Biometrics, 38:163–170, 1982.

[62] SCHULGEN G., OLSCHEWSKI M., KRANE V. ET AL.: *Sample sizes for clinical trials with time-to-event endpoints and competing risks*. Contemporary Clinical Trials, 26:386–396, 2005.

[63] SELF, S.G.: *An adaptive weighted logrank test with application to cancer prevention and screening trials*. Biometrics, 47:975–986, 1991.

[64] SELLKE T., SIEGMUND S.: *Sequential analysis of the proportional hazards model*. Biometrika, 70:315–326, 1982.

[65] SHEN Y., CAI J.: *Sample size re-estimation for clinical trials with censored survival data*. Journal of the American Statistical Association, 98:418–426, 2003.

[66] SHIH, J.H.: *Sample Size Calculation for Complex Clinical Trials with Survival Endpoints*. Controlled Clinical Trials, 16:395–407, 1995.

[67] SLUD E., WEI L.J.: *Two-sample repeated significance tests based on the modified Wilcoxon statistics*. Journal of the American Statistical Association, 77:862–868, 1982.

[68] SNAPINN S. M., JIANG Q., IGLEWICZ B.: *Informative noncompliance in endpoint trials*. Current Controlled Trials in Cardiovascular Medicine, 5:5, 2004.

[69] TARONE R.E., WARE J.: *On distribution-free tests for equality of survival distributions*. Biometrika, 64:156–160, 1977.

[70] TSIATIS, A.A.: *The asymptotic joint distribution of the efficient scores test for the proprtional hazards model calculated over time*. Biometrika, 68:311–315, 1981.

[71] TSIATIS A.A., MEHTA C.: *On the inefficiency of adaptive design for monitoring clinical trials*. Biometrika, 90:367–378, 2003.

[72] WASSMER, G.: *Planning and Analyzing Adaptive Group Sequential Survival Trials*. Biometrical Journal, 48:714–729, 2006.

[73] WASSMER G., EISEBITT R.: *ADDPLAN Adaptive Designs - Plans and Analyses. Release 4*. Software Documentation. http://www.addplan.com, 2007.

[74] WASSMER G., VANDEMEULEBROECKE M.: *A Bief Review on Software Developements for Group Sequential and Adaptive Designs*. Biometrical Journal, 48:732–737, 2006.

[75] WHITEHEAD J., WHITEHEAD A., TODD S. ET AL.: *Mid-trial design reviews for sequential clinical trials*. Statistics in Medicine, 20:165–176, 2001.

[76] WILLIAMSON J.M., LIN H.M., KIM H.Y.: *Power and sample size calculations for current status survival analysis*. Statistics in Medicine, 28:1999–2011, 2009.

[77] XIONG C., YAN Y., JI M.: *Sample sizes for comparing means of two lifetime distributions with type II censored data: application in an aging intervention study*. Controlled Clinical Trials, 24:283–293, 2003.

[78] YATEMAN N.A., SKENE A.M.: *Sample Sizes for Proprortional Hazards Survibal Studies with Arbitrary Patient Entry and Loss to Follow-up Distributions*. Statistics in Medicine, 11:1103–1113, 1992.

[79] ZHANG D., QUAN H.: *Power and sample size calculation for log-rank test with a time lag in treatment effect*. Statistics in Medicine, 28:864–879, 2009.

Hiermit versichere ich, dass ich diese Diplomarbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

München, den 21.12.2009