Auswirkungen von Imbalancen bei der Blockrandomisierung auf die Power am Beispiel des t-Tests

Rainer Muche
Universität Ulm
Schwabstr. 13
89075 Ulm
rainer.muche@uni-ulm.de

Laura Armbrust
Universität Ulm
Schwabstr. 13
89075 Ulm
laura.armbrust@uni-ulm.de

Friederike Rohlmann
Universität Ulm
Schwabstr. 13
89075 Ulm
friederike.rohlmann@uni-ulm

Jens Dreyhaupt
Universität Ulm
Schwabstr. 13
89075 Ulm
jens.dreyhaupt@uni-ulm.de

Zusammenfassung

Die randomisierte kontrollierte klinische Studie wird in der Fachliteratur als „Goldstandard" der klinischen Forschung bezeichnet. Randomisierung bedeutet die zufällige Zuteilung von Patienten zu den einzelnen Studienbehandlungen und zielt damit auf strukturgleiche Behandlungsgruppen hinsichtlich bekannter und unbekannter Störgrößen. Das Problem bei dieser rein zufälligen Zuordnung eines neuen Patienten ist allerdings, dass sowohl gleiche Anzahl der Patienten in den Behandlungsgruppen (balanciertes Design) als auch gleichmäßige Verteilung bekannter wichtiger Störgrößen möglicherweise nicht erreicht werden kann. Deshalb wird häufig die sogenannte stratifizierte Blockrandomisierung eingesetzt. Bei dieser Randomisierung konkurrieren Balance und Unvorhersehbarkeit miteinander, insbesondere wenn die Studienteilnehmer auf viele Schichten (Strata) verteilt werden sollen.

In diesem Zusammenhang kommt der Wahl geeigneter Blocklängen eine große Bedeutung zu. Um bessere Entscheidungen bezüglich der Blocklängen treffen zu können, wurden SAS Makros entwickelt, mit denen es möglich ist, die spezifische stratifizierte Studiensituation mit verschiedenen Blocklängen oder Sets von Blocklängen zu simulieren. Als wichtigste Größe wird dabei die beobachtete Imbalance mit entsprechender Eintrittswahrscheinlichkeit ausgegeben.

Diese Informationen können dazu genutzt werden, die Fallzahlberechnung in der Studie soweit zu korrigieren, dass die vorgegebene gewünschte Power für den balancierten Fall auch bei möglicher Imbalance erreicht wird, denn ungleiche Fallzahlen in den Gruppen reduzieren die Power des Tests. Dazu werden die mittels Simulationsmakros beobachteten Imbalancen in die Fallzahlplanung eingesetzt. Diese Fallzahlplanszenarios werden mit der SAS-Prozedur PROC POWER auf Basis des unverbundenen t-Tests durchgeführt und untersucht. In dem Beitrag werden die notwendigen Schritte sowie der Einsatz der PROC POWER für diese Berechnungen anhand des unverbundenen t-Tests allgemein und exemplarisch an einer konkreten Studienplanung dargestellt.

Schlüsselwörter: Stratifizierte Blockrandomisierung, Balance der Studienbehandlungen, Fallzahlplanung, Power, PROC POWER

1 Einleitung und Problembeschreibung

Die Bestimmung der notwendigen Fallzahl ist ein wichtiger Schritt in der Planung einer klinischen Studie [8]. Von ihr hängen z. B. die Machbarkeit, die Dauer und die Anzahl der notwendigen Studienzentren in der Studie ab. Für die Fallzahlplanung gibt es statistische Vorgehensweisen [2], die u. a. in SAS in der Prozedur PROC POWER [7] implementiert sind. Maximale Power bei fixer Gesamtfallzahl wird erreicht, wenn das Studiendesign balanciert ist, das heißt, dass für jede Behandlungsgruppe einer Studie gleichgroße Fallzahlen vorgegeben werden.

Klinische Studien werden üblicherweise als randomisierte Studien geplant. Die Randomisierung ist das derzeit anerkannteste Mittel, um Vergleichbarkeit der Gruppen hinsichtlich bekannter und unbekannter Störgrößen zu erreichen und unverzerrte Ergebnisse zu bekommen. Allerdings ist die gleichmäßige Verteilung der Störgrößen nicht garantiert, so dass bei bekannten Störfaktoren häufig die sogenannte stratifizierte Blockrandomisierung eingesetzt wird. Durch die Randomisierung innerhalb der Strata (Schichten) der bekannten Störgrößen kann deren Verteilung besser kontrolliert werden. Bei dieser Art der Randomisierung werden insbesondere in offenen Studien [1] möglichst große Blocklängen mit in der Regel variierenden Größen gewählt, um eine Vorhersagbarkeit der nächsten Behandlungszuteilung zu minimieren. Dies häufig genutzte Verfahren garantiert allerdings nicht mehr die Balance der Fallzahlen in den Behandlungsarmen, abhängig von der Anzahl Schichten und den gewählten Blocklängen. Da die Randomisierung in den Strata unabhängig voneinander erfolgt, kann eine Imbalance in der Gesamtpopulation der Studie die Folge sein.

Bei einer eingetretenen Imbalance der Fallzahlen in den Behandlungsgruppen am Ende der Rekrutierung und Randomisierung ist die Power geringer als in der balancierten Situation und damit geringer als in der Fallzahlplanung zur Studie vorgegeben. In diesem Beitrag werden anhand eines Simulationsprogramms (SAS-Makro) [4] die Wahrscheinlichkeiten für Imbalancen quantifiziert und anhand einer sukzessiven Erhöhung der Fallzahlen gezeigt, wann die vorgegebene Power für die Studiensituation mit Imbalance wieder erreicht wird. Somit kann bereits in der Planungsphase der Studie für eine mögliche Imbalance korrigiert werden.

2 Fallzahlplanung

Zur Überprüfung der Wirksamkeit, Sicherheit und Verträglichkeit von Medikamenten sowie zur Erforschung von Erkrankungen bisher unbekannter Ursache werden Jahr für Jahr weltweit eine große Anzahl von Studien durchgeführt. In der Planungsphase von Studien ist der Fallzahlschätzung eine große Bedeutung beizumessen. Soll z. B. ein klei-

[1] Studien, deren Behandlungen nicht verblindet werden können, so dass Studienleiter und Studienteilnehmer nach Randomisierung wissen, welche Behandlung dem Teilnehmer zufällig zugeteilt wurde.

ner Effekt in der Therapie einer sehr seltenen Erkrankung überprüft werden, würde dafür eine sehr große Fallzahl benötigt werden.

Die Fallzahl sollte hoch genug gewählt werden, um eine verlässliche Antwort auf die vor Beginn der Studie formulierten Fragestellungen zu erhalten. Dennoch darf die Fallzahl nicht beliebig groß gewählt werden, damit ethische Grundsätze in Bezug auf die Studienteilnehmer nicht verletzt und Ressourcen wie Zeit, Personal und Kapital nicht verschwendet werden [2,8] . Ein zu geringer Umfang hingegen könnte zur fälschlichen Verwerfung einer neuen Behandlung führen und damit Therapiechancen vergeben.

Die Fallzahl in einer Studie wird über den statistischen Test bestimmt, der für die Auswertung vorgesehen ist. Die Fragestellung, wie viele Patienten zur Beantwortung einer wissenschaftlichen / klinischen Fragestellung benötigt werden, kann dann (für einfache Testsituationen) folgendermaßen angegangen werden:

1. Als erstes wird der statistische Test ausgewählt, der für die Untersuchung der Fragestellung (Ein-, Zweistichprobensituation, zwei oder mehr Studienbehandlungen) und den Merkmalstyp der gewählten Zielgröße (dichotom, mehrkategorial, ordinal, stetig) eingesetzt werden kann.
2. Der (relevante) Behandlungseffekt, der gefunden bzw. gezeigt werden soll, muss festgelegt werden. Außerdem müssen die Fehlerraten für den Fehler 1. Art (Signifikanzniveau) und 2. Art (β = 1 minus Power) sowie spezifische Parameter für die jeweilige Testsituation festgelegt werden.
3. Die Berechnung ist nun, die Anzahl Patienten zu finden, die den zu zeigenden Effekt bei gegebenen Fehlerwahrscheinlichkeiten 1. und 2. Art gerade als signifikant entdeckt.

Die Größe der Fallzahl für den unverbundenen t-Test wird also im Wesentlichen von vier Faktoren beeinflusst [2]:

- dem zu entdeckenden Effekt (z. B. Unterschied zwischen zwei Therapiegruppen): bei Vergrößerung des erwarteten Effektes reduziert sich die Fallzahl
- dem Fehler 1. Art: mit kleinerer Irrtumswahrscheinlichkeit α vergrößert sich die Fallzahl
- der Power = Wahrscheinlichkeit für ein signifikantes Ergebnis: eine höhere Power (1-β) führt zu einem größeren Stichprobenumfang
- weiterer Parameter, die vom gewählten Test abhängen, z. B. hier die Streuung der Daten: die Fallzahl erhöht sich bei Zunahme der Varianz σ^2

Fallzahlschätzung für den Zweistichproben t-Test

Die Fallzahlabschätzung mit gleicher Fallzahl pro Gruppe (balanciertes Design) für den Zweistichproben-t-Test kann anhand der folgenden Formel nach Bock [2] berechnet werden als:

$$n \approx \frac{2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\left(\left(\mu_A - \mu_B\right)_{REL} / \sigma_V\right)^2}$$

$z_{1-\alpha/2}$: 1-α/2–Quantil der Standardnormalverteilung (zweiseitig, 1-α einseitig)

wobei $(\mu_A-\mu_B)_{REL}$ der zu entdeckende (klinisch) relevante Unterschied zwischen den Mittelwerten der Gruppen A und B ist und σ_v die gemeinsame Standardabweichung. Dieser Quotient im Nenner stellt die relative Effektgröße dar.

Die Fallzahl hängt also von den oben angegebenen Parametern ab. Die folgende Abbildung 1 zeigt den Zusammenhang zwischen Gruppengröße und Effektgröße bei verschiedenen Vorgaben für die Power:

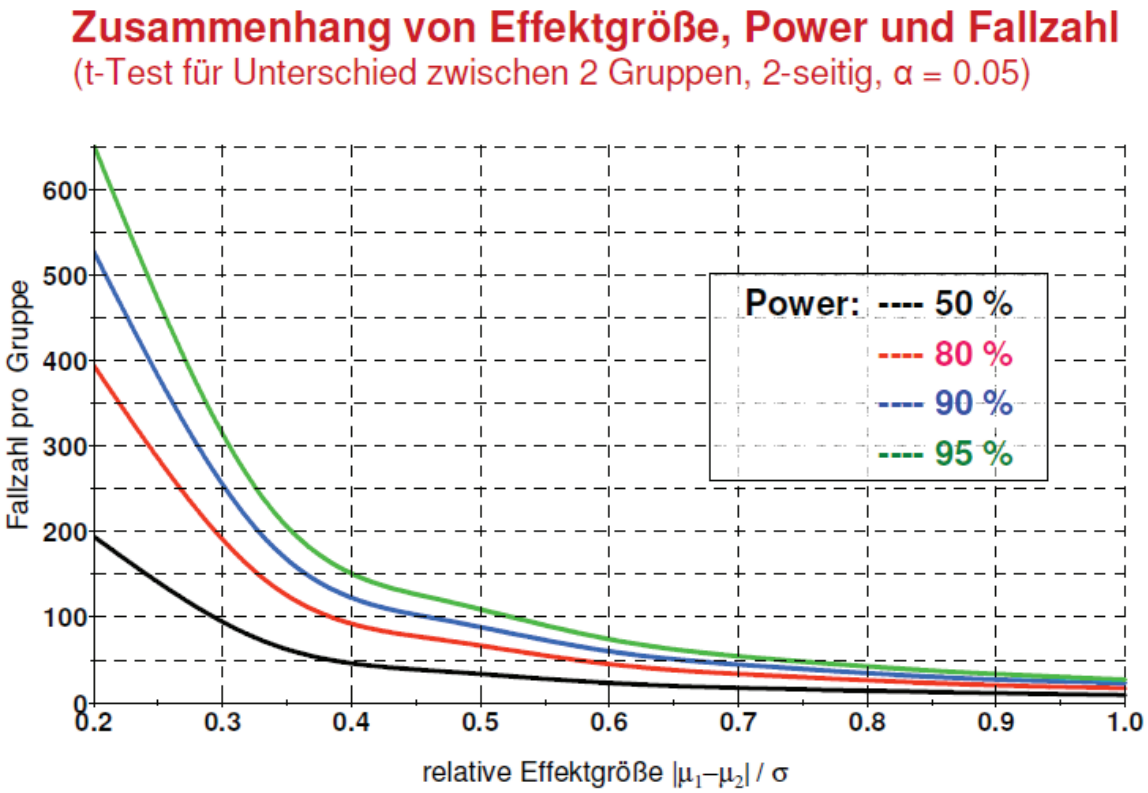


Abbildung 1: Zusammenhang von Effektgröße, Power und Fallzahl im Zweistichproben t-Test

3 Fallzahlplanung mit SAS PROC POWER

Die Prozedur PROC POWER beinhaltet eine Vielzahl an Fallzahlschätzungsmöglichkeiten für die unterschiedlichsten statistischen Testverfahren wie Mittelwertvergleiche, Vergleich von Raten und Wahrscheinlichkeiten, Korrelations- und Regressionsanalysen, Überlebenszeitanalysen, Äquivalenznachweise [3,6,7].
Die Syntax von PROC POWER soll hier vorgestellt werden. Für jedes gewählte Fallzahlplanungsverfahren wird ein spezifisches Statement benötigt, dem Parameter übergeben werden können bzw. müssen.

Hier ein Auszug aus den angebotenen Statements:

```
PROC POWER <options>;
  MULTREG <options>;
  ONECORR <options>;
  ONESAMPLEFREQ <options>;
  ONESAMPLEMEANS <options>;
  ONEWAYANOVA <options>;
  PAIREDFREQ <options>;
  PAIREDMEANS <options>;
  TWOSAMPLEFREQ <options>;
  TWOSAMPLEMEANS <options>;
  TWOSAMPLESURVIVAL <options>;
  PLOT <plot-options> </graph-options>;
```

Nach der Auswahl eines Statements zur Fallzahlschätzung, z. B. TWOSAMPLEMEANS TEST=DIFF für den t-Test auf Unterschied zwischen zwei Gruppen können dort nun alle nötigen Parameter übergeben werden, z. B. unter GROUPMEANS die erwarteten Mittelwerte pro Gruppe oder alternativ unter MEANDIFF der Mittelwertunterschied, unter STDDEV die gemeinsame Standardabweichung, unter POWER die Powerangabe als Wahrscheinlichkeit <1. Für den zu berechnenden Parameter, entweder für die Gesamtfallzahl unter NTOTAL= oder alternativ für die Fallzahl pro Gruppe unter NPERGROUP= wird ein fehlender Wert, d. h. ein Punkt, eingesetzt. Soll bei vorgegebener unterschiedlicher Fallzahl pro Gruppe die Power berechnet werden, so können unter NPERGROUP die entsprechenden Werte und nun für die POWER ein Missing eingetragen werden.

Die Eingabe der Parameter ALPHA= (Default: 0,05) und ein- oder zweiseitiger Test unter SIDES= (Default: 2) ist nur erforderlich, falls ein anderer Wert als der Default benötigt wird. Für den Parameter SIDES= können neben der Voreinstellung 2 für zweiseitiger Test folgende Werte eingetragen werden: 1 für einseitige Fragestellungen, U für den einseitigen Fall, dass der Wert der Alternativhypothese größer als der der Nullhypothese ist oder L für den Fall, dass der Wert der Alternativhypothese kleiner als der der Nullhypothese ist.

Graphiken können optional zusätzlich mit dem PLOT-Statement erstellt werden. Die ausschließliche Erzeugung einer Graphik funktioniert mit dem PLOTONLY-Statement als Option von PROC POWER.

Fallzahlschätzung mit PROC POWER für den t-Test

In der folgenden SAS-Syntax sind einige Optionen für die Fallzahlplanung mit PROC POWER für den Zweistichproben-t-Test angegeben. Die spezifischen Angaben korrespondieren zu dem Beispiel im Kapitel 4.1.

```
PROC POWER;
  TWOSAMPLEMEANS TEST=DIFF
    ALPHA     = 0.05
    SIDES     = 2
    MEANDIFF  = 0.460491818
    STDDEV    = 1
    POWER     = 0.8
    NPERGROUP = .;
RUN;
```

Es werden standardmäßig Gruppengrößen für ein balanciertes Design (NPERGROUP oder NTOTAL) berechnet. Für unterschiedliche Fallzahlen ist statt NPERGROUP die Option GROUPNS anzugeben. Bei Übergabe von 2 Parametern an eine Option sind diese entweder mit dem Verkettungszeichen „|“ (Bsp. 40|60) oder in Klammern durch ein Leerzeichen getrennt (Bsp. (40 60)) anzugeben. Auch hier kann einer von beiden Parametern auf fehlend gesetzt werden, wenn alle anderen nötigen Angaben gemacht wurden.

Die Ausgabe der Power, die für die im Weiteren gezeigten Analysen notwendig ist, ist allerdings auf 3 Nachkommastellen gerundet. Dies ist für die nachfolgenden Untersuchungen bzgl. Power nicht ausreichend genau. Durch Ausgabe des Prozedur-Outputs mit ODS und Formatierung in PROC PRINT kann dies Problem umgangen werden:

```
ODS TRACE ON;
ODS OUTPUT OUTPUT = test1;
PROC POWER;
  TWOSAMPLEMEANS TEST=diff
  ...;
RUN;
ODS TRACE OFF;

PROC PRINT DATA=test1;
  FORMAT Alpha     8.4
         Power    12.10
         MeanDiff 12.9;
  VAR _NUMERIC_;
RUN;
```

4 Balance in der stratifizierten Blockrandomisierung

Die randomisierte kontrollierte klinische Studie wird in der Fachliteratur als „Goldstandard“ der klinischen Forschung bezeichnet. Randomisierung bedeutet zufällige Zuteilung von Patienten zu den verschiedenen Studienbehandlungen mit dem Ziel, die Störgrößen gleichmäßig auf die Behandlungsgruppen zu verteilen. Eine uneingeschränkte

Randomisierung bedeutet vollkommen zufällige Zuteilung der Patienten zu den Behandlungen. Es gilt als das beste Verfahren hinsichtlich Unvorhersehbarkeit und Vermeidung von systematischen Fehlern. Das Problem bei diesem Vorgehen ist, dass trotz vorgegebener gleicher Eintrittswahrscheinlichkeit für jede Studienbehandlung insbesondere bei kleineren Studien eine gleiche Anzahl der Patienten in den Behandlungsgruppen (Balance) oft nicht erreicht wird. Deshalb kommt häufig die Blockrandomisierung zum Einsatz. Der Zufall wird eingeschränkt, da pro Block vorgegeben wird, dass den Gruppen gleich viele Patienten zugeteilt werden. Ist z. B. bei zwei möglichen Studienbehandlungen die Anzahl in einer der Gruppen erreicht, so ergeben sich die letzten Zuteilungen zur anderen Gruppe in einem Block deterministisch. Der Balance-Vorteil wird also durch die Reduzierung der Unvorhersagbarkeit der nächsten Behandlung(en) (Concealment) eingeschränkt.

Um das Concealment bei der Blockrandomisierung zu verbessern, können abhängig von der Fallzahl unterschiedliche Blockgrößen gewählt und zufällig aneinandergefügt werden (permutierte Blockrandomisierung). In klinischen Studien erfolgt die Randomisierung häufig zusätzlich stratifiziert, d. h. es gibt separate geblockte Randomisierungslisten für jede Ausprägung eines prognostischen Faktors (z. B. männlich/weiblich) bzw. Ausprägungskombination bei mehreren Faktoren (z. B. Zentrum1-männlich/Zentrum1-weiblich/ Zentrum2-männlich …). Das Verfahren wird dann stratifizierte Blockrandomisierung genannt. In multizentrischen Studien wird dieses Verfahren mit Schichtung nach Zentrum und ggf. weiteren Faktoren häufig eingesetzt.

Die wichtigsten zu wählenden Parameter bei der Planung einer solchen stratifizierten, ggf. permutierten Blockrandomisierung sind demnach die Blocklängen je Stratum. Dabei ist zu beachten, dass sich mit steigender Anzahl Strata die Wahrscheinlichkeit der Imbalance durch viele angebrochene Blöcke am Ende der Rekrutierung erhöht, da die Zuteilung der Therapie in den Strata unabhängig von denen der anderen Strata erfolgt.

In [4] wurde ein SAS-Makro vorgestellt, mit dem man die Gesamtbalance in einer stratifizierten Blockrandomisierung simulieren und untersuchen kann. Somit lässt sich für eine konkrete Studienplanung untersuchen, wie sich bei gegebener Fallzahl und verschiedener aufgetretener Imbalancen die Power des Tests verringert.

4.1 Angaben zum Studienbeispiel

Zu Demonstrationszwecken wird ein einfaches Studienbeispiel gewählt mit einer geplanten Fallzahl von 150 Teilnehmern und zwei Stratifizierungsmerkmalen Zentrum und Geschlecht mit je zwei Ausprägungen, so dass sich also insgesamt „nur“ vier Strata (Schichten) ergeben. Die Studienbehandlungen A und B sollen im Verhältnis 1:1 (also balanciert) zufällig zugeteilt werden. Die Simulierung der Studienrekrutierung wird 1000mal wiederholt. Der in dieser Studie erwartete Anteil Studienteilnehmer pro Zentrum und Geschlecht ist in Tabelle 2 dargestellt.

Tabelle 1: Angaben zum Studienbeispiel

Stratifizierungsvariablen	Ausprägungen	Erwarteter Anteil Studienteilnehmer (%)
Zentren	Zentrum 1	70%
	Zentrum 2	30%
Geschlecht	männlich	40%
	weiblich	60%

Aus den erwarteten Anteilen pro Stratifizierungsmerkmal werden die erwarteten Anteile pro Stratum (unter Annahme der Unabhängigkeit) als Produkt aus den Wahrscheinlichkeiten ihrer jeweiligen Ausprägungen errechnet (Bsp. Stratum 1: 0.7*0.4=0.28), siehe Tabelle 2. Auf Grundlage der resultierenden absoluten Fallzahlen werden pro Schicht geeignet scheinende Blocklängen festgelegt.

Tabelle 2: Erwartete Fallzahlen in den Strata und gewählte Blocklängen

Stratum Nr.	Stratum-Bezeichnung	Erwarteter Anteil in %	Erwartete absolute Fallzahlen	Gewählte (permutierte) Blocklängen
1	Zentrum 1 / männlich	28%	42	6 / 8
2	Zentrum 1 / weiblich	42%	63	6 / 8
3	Zentrum 2 / männlich	12%	18	4 / 6
4	Zentrum 2 / weiblich	18%	27	4 / 6
Gesamt		100%	150	

4.2 Ergebnis der Simulation zum Balanceverhalten im Studienbeispiel

Nach 1000 Simulationen ergibt sich die in Tabelle 3 dargestellte Verteilung für Behandlung A. Wie erwartet, liegt der Modalwert bei N=75, also der Hälfte der geplanten Studienteilnehmerzahl.

Tabelle 3: Häufigkeiten der Behandlung A im Gesamtkollektiv von 150 Studienteilnehmern und 1000 Simulationen

Behandlung A	Frequency	Percent	Cumulative Frequency	Cumulative Percent
72	9	0.90	9	0.90
73	63	6.30	72	7.20
74	243	24.30	315	31.50
75	381	38.10	696	69.60
76	238	23.80	934	93.40
77	60	6.00	994	99.40
78	6	0.60	1000	100.00

Aus diesem Beispiel ergeben sich folgende Häufigkeiten für Imbalancen bezogen auf das Gesamtkollektiv (Tabelle 4). Unter der Annahme, dass die tatsächlichen Fallzahlen in den Schichten zufällig mit einer Standardabweichung von 5 um die erwarteten Häufigkeiten streuen, ist also bei Wahl der oben angegebenen Blockgrößen die Wahrscheinlichkeit für eine Imbalance ≥ 6 etwa 1.5%. Das bedeutet, dass bei 1000 Simulationen maximal eine Imbalance von 72 zu 78 Patienten beobachtet wird.

Tabelle 4: Imbalance im Gesamtkollektiv von 150 Studienteilnehmern und 1000 Simulationen

Imbalance	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	381	38.10	381	38.10
2	481	48.10	862	86.20
4	123	12.30	985	98.50
6	15	1.50	1000	100.00

Es zeigt sich, dass Imbalancen von 8 bis maximal möglicher Imbalance von 14 bei 1000 Simulationen nicht vorgekommen, also ziemlich unwahrscheinlich sind. Die maximale Imbalance würde sich ergeben, wenn bei Rekrutierungsende nach 150 Studienteilnehmern in jedem Stratum der letzte Block immer der Größte und je zur Hälfte mit immer der gleichen Therapie angebrochen wäre. Für das Beispiel ergäbe sich ein Maximum von 4+4+3+3=14; das hätte bedeutet, dass 82 Teilnehmer zur einen und 68 Teilnehmer zur anderen Studienbehandlung zugewiesen worden wären.

5 Fallzahlbestimmung bei erwarteter Imbalance

Folgende Schritte sind durchzuführen, um die notwendige Fallzahl unter Vorliegen einer möglichen Imbalance durch die Nutzung der stratifizierten Blockrandomisierung abschätzen zu können:

1. Berechnung der Fallzahl für balancierte Gruppen
2. Simulation des Imbalance-Verhaltens bei gegebenen Strata und Blocklängen
3. Berechnung der Power für beobachtete Imbalancen (aus Simulation)
4. Schrittweise gleiche Erhöhung der Fallzahl in den Gruppen, bis die unter 1. verwendete vorgegebene Power wieder erreicht wird[2]

Im folgenden Kapitel 5.1 wird für das oben eingeführte Beispiel die für Imbalancen angepasste Fallzahl bestimmt. In Kapitel 5.2 wird dann für andere Studienumfänge systematisch und tabellarisch die Powerentwicklung bei zunehmender Imbalance gezeigt.

[2] Die Überlegungen beruhen auf der Annahme, dass sich maximal beobachtete Imbalancen bei zusätzlicher Rekrutierung anhand bestehender Randomisierungslisten nicht weiter vergrößern.

5.1 Durchführung im Beispiel

Die Effektstärke (hier Differenz der Gruppenmittelwerte) von 0.460491818 (s. a. Syntaxbeispiel in Kapitel 3) ist so gewählt, dass für den Vergleich zweier Gruppen 75 Patienten pro Gruppe benötigt werden (Schritt 1) und wurde als Ergebnis aus folgendem Aufruf erhalten:

```
PROC POWER;
  TWOSAMPLEMEANS TEST=DIFF
    ALPHA     = 0.05
    SIDES     = 2
    MEANDIFF  = .
    STDDEV    = 1
    POWER     = 0.8
    GROUPNS   = 75 | 75;
RUN;
```

In der in Kapitel 4.2 aufgeführten Simulation wird mit 1.5%iger Wahrscheinlichkeit eine maximale Imbalance von 6 beobachtet (Schritt 2). So kann als Nächstes (Schritt 3) die Power für den Vergleich mit den entsprechend veränderten Gruppengrößen berechnet werden:

```
PROC POWER;
  TWOSAMPLEMEANS TEST=DIFF
    MEANDIFF = 0.460491818
    STDDEV   = 1
    POWER    = .
    GROUPNS  = 72 | 78;
RUN;
```

Obs	Index	Mean Diff	Std Dev	N1	N2	Null Diff	Alpha	Power
1	1	0.460491818	1	72	78	0	0.0500	0.7993717220

Abbildung 2: Power für den Zweistichproben t-Test bei einer Imbalance in den Fallzahlen pro Gruppe von 72 zu 78 Patienten

Die resultierende Power (siehe Abb. 2) ist nur unwesentlich kleiner als die geforderte Power von 80%. Für praktische Zwecke wäre sie wohl ausreichend. Als Schritt 4 (s. oben) kann nun die Fallzahl pro Gruppe schrittweise um jeweils 1 erhöht werden. Die folgende Abbildung 3 zeigt, dass schon bei Erhöhung auf 73:79 Patienten die geforderte Power von 80% erreicht ist. Sollte dies nicht der Fall sein, würde die Fallzahl weiter sukzessive pro Gruppe um 1 erhöht werden, bis 80% erreicht sind.

Obs	Index	Mean Diff	Std Dev	N1	N2	Null Diff	Alpha	Power
1	1	0.460491818	1	73	79	0	0.0500	0.8046322459

Abbildung 3: Power für den Zweistichproben t-Test bei einer Imbalance in den Fallzahlen pro Gruppe von 73 zu 79 Patienten

Somit würden bei einer erwarteten maximalen Imbalance von 6 Patienten 73 + 79 = 152 Patienten ausreichen, um eine Power von mindestens 80% zu erreichen. Im Studienplan könnte man daher für jede Gruppe 76 Patienten vorsehen. Die Fallzahl wäre somit gegenüber einer balancierten Studiensituation um 1.3% von 150 auf 152 Patienten erhöht.

Setzt man die maximal mögliche Imbalance von 14 ein, also 68 und 82 Patienten pro Gruppe, so reduziert sich die geplante Power von 80% auf 79.7%. Auch hier würde ein weiterer Patient pro Gruppe reichen, um wieder die 80% zu erreichen. Selbst wenn beide zusätzliche Patienten der gleichen Gruppe angehören, werden jedes Mal die 80% überschritten.

5.2 Korrekturvorschläge für verschiedene Studienumfänge

Ähnliche Ergebnisse erhält man auch für kleine (2 x 20 Patienten) und große (2 x 250 Patienten) Studien, wie die beiden nächsten Tabellen zeigen [1]. Dabei ist hier die Imbalance als Imbalance pro Gruppe angegeben, d. h. eine Imbalance von 4 in Tabelle 5 entspricht Gruppengrößen von 16 und 24 Patienten für N=40.

In Tabelle 5 zeigt sich, dass eine relativ große Imbalance von ±8 pro Gruppe mit einer Erhöhung von 3 Patienten pro Gruppe auf N=46 in Bezug auf die Power korrigiert werden kann. Diese Studie sollte demnach dann mit 2 x 23 Patienten geplant werden, damit bei einer Imbalance von 8 Patienten pro Gruppe (15 vs. 31 Patienten) die gewünschte Power von 80% erhalten bleibt.

Die Ergebnisse für eine Studie mit einem Umfang von 2 x 250 Patienten gibt Tabelle 6 wieder. Auch hier sind nur geringe Fallzahlerhöhungen notwendig, um die Power bezüglich Imbalance korrigieren zu können: 1 Patient pro Gruppe auf N=502 für eine Imbalance von ±15 Pat./Gruppe (235 vs. 265); 2 Patienten pro Gruppe auf N=504 bei Imbalance von ±22 Pat./Gruppe (228 vs. 272) und 3 Patienten pro Gruppe auf N=506 bei Imbalance von ±27 Pat./Gruppe (223 vs. 277).

Tabelle 5: Power in Abhängigkeit von Imbalancen (±) pro Gruppe bei 2 x 20 Patienten sowie schrittweiser Erhöhung der Fallzahlen

Imbalance	$N = 40$	$N = 42$	$N = 44$	$N = 46$
0	0.8000245	0.8198105	0.8378750	0.8543319
1	0.7990427	0.8189534	0.8371271	0.8536794
2	0.7960725	0.8163611	0.8348652	0.8517063
3	0.7910388	0.8119692	0.8310342	0.8483657
4	0.7838140	0.8056685	0.8255408	0.8435774
5	0.7742148	0.7973017	0.8182501	0.8372262
6	0.7619970	0.7866591	0.8089823	0.8291581
7	0.7468499	0.7734731	0.7975075	0.8191757
8	0.7283901	0.7574120	0.7835397	0.8070333
9	0.7061547	0.7380732	0.7667305	0.7924304
10	0.6795969	0.7149769	0.7466620	0.7750051

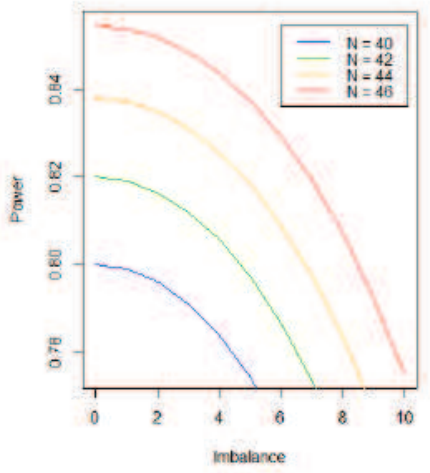
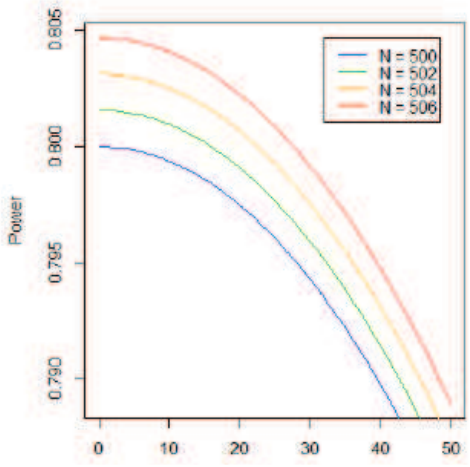


Tabelle 6: Power in Abhängigkeit von Imbalancen (±) pro Gruppe bei 2 x 250 Patienten sowie schrittweiser Erhöhung der Fallzahlen

Imbalance	$N = 500$	$N = 502$	$N = 504$	$N = 506$
0	0.8000000	0.8015694	0.8031283	0.8046766
1	0.7999937	0.8015632	0.8031221	0.8046706
2	0.7999749	0.8015446	0.8031037	0.8046523
3	0.7999435	0.8015135	0.8030730	0.8046219
4	0.7998995	0.8014701	0.8030300	0.8045794
5	0.7998430	0.8014142	0.8029747	0.8045247
6	0.7997740	0.8013458	0.8029071	0.8044578
7	0.7996923	0.8012650	0.8028271	0.8043787
8	0.7995980	0.8011718	0.8027349	0.8042874
9	0.7994912	0.8010660	0.8026303	0.8041839
10	0.7993717	0.8009478	0.8025133	0.8040682
11	0.7992395	0.8008171	0.8023840	0.8039403
12	0.7990947	0.8006738	0.8022422	0.8038000
13	0.7989371	0.8005179	0.8020880	0.8036475
14	0.7987669	0.8003495	0.8019214	0.8034826
15	0.7985839	0.8001685	0.8017423	0.8033054
16	0.7983881	0.7999748	0.8015507	0.8031159
17	0.7981796	0.7997684	0.8013465	0.8029139
18	0.7979581	0.7995494	0.8011298	0.8026995
19	0.7977238	0.7993176	0.8009005	0.8024726
20	0.7974766	0.7990730	0.8006585	0.8022332
21	0.7972164	0.7988156	0.8004039	0.8019813
22	0.7969433	0.7985453	0.8001365	0.8017168
23	0.7966570	0.7982622	0.7998564	0.8014397
24	0.7963577	0.7979661	0.7995634	0.8011499
25	0.7960453	0.7976570	0.7992577	0.8008473
26	0.7957197	0.7973348	0.7989390	0.8005321
27	0.7953808	0.7969996	0.7986073	0.8002040
28	0.7950286	0.7966512	0.7982626	0.7998630
29	0.7946631	0.7962896	0.7979049	0.7995091
30	0.7942842	0.7959147	0.7975341	0.7991422
31	0.7938918	0.7955265	0.7971500	0.7987623



6 Fallzahlkorrektur beim Chi-Quadrat- und Log-Rank-Test

Für das in Kapitel 4 vorgestellte stratifizierte Studienbeispiel mit einer Gesamtfallzahl von 150 Patienten (2 x 75 Patienten pro Gruppe) wird im Folgenden jeweils eine Beispielrechnung für eine Fallzahlkorrektur auf Basis des Chi-Quadrat-Tests für eine dichotome Zielgröße und des Log-Rank-Tests für Überlebenszeiten dargestellt, da diese Verfahren neben dem t-Test häufig eingesetzt werden.

6.1 Fallzahlkorrektur beim Chi-Quadrat-Test

Bei der Fallzahlplanung auf Basis des Chi-Quadrat-Tests will man einen signifikanten Unterschied zwischen den Erfolgsraten in zwei Therapiegruppen bei vorgegebenem Signifikanzniveau (hier 5%) und Power (hier 80%) zeigen. Ein Unterschied zwischen der Rate von 35% in Gruppe A und 15.2% in Gruppe B ergibt bei Balance genau eine Fallzahl von 75 Patienten pro Gruppe mit folgender Syntax in PROC POWER [7]:

```
PROC POWER;
  TWOSAMPLEFREQ TEST=PCHI[3]
    ALPHA               = 0.05
    SIDES               = 2
    GROUPPROPORTIONS    = (0.35 0.152)
    NULLPROPORTIONDIFF  = 0
    POWER               = 0.8
    NPERGROUP           = .;
RUN;
```

Bei einer erwarteten maximalen Imbalance von 6, wie in Kapitel 4.2 gezeigt, liegt hier die Power immer noch über 80%, so dass keine Korrektur der Fallzahl zur Berücksichtigung der Imbalance notwendig ist. Untersucht man die Power bei maximaler Imbalance von 14 Patienten (siehe Kapitel 4.2, hier 82 vs. 68 Patienten), so ergibt sich eine Power von 0.798.

Bei der Untersuchung der Power ist zu beachten, dass der Chi-Quadrat-Test nicht symmetrisch ist, d. h. es ist nicht egal, welcher Rate welche Fallzahl zugeordnet wird. Wird die größere Fallzahl zur größeren Rate von 35% zugeordnet, ist die Power kleiner als bei entgegengesetzter Zuordnung. Diese Situation ist in den ersten beiden Zeilen von Abbildung 4 dargestellt. Um eine Powerkorrektur für die ungünstigere Situation in Zeile 2 (82 vs. 68) zu erzielen, reicht die Erhöhung um einen Patienten pro Gruppe (siehe Zeile 3).

[3] Mit PCHI = Pearson's chi-square test und den Parametern: GROUPPROPORTIONS = Erfolgs- (oder Misserfolgs-) Rate in den zu vergleichenden Gruppen und NULLPROPORTIONDIFF = Unterschied zwischen den Gruppen unter der Hypothese H_0 (Voreinstellung: 0, also kein Grp.-Unterschied)

Obs	Index	Null Proportion Diff	Proportion1	Proportion2	N1	N2	Alpha	Power
1	1	0	0.35	0.152	68	82	0.0500	0.8053340672

Obs	Index	Null Proportion Diff	Proportion1	Proportion2	N1	N2	Alpha	Power
1	1	0	0.35	0.152	82	68	0.0500	0.7976940605

Obs	Index	Null Proportion Diff	Proportion1	Proportion2	N1	N2	Alpha	Power
1	1	0	0.35	0.152	83	69	0.0500	0.8032157736

Abbildung 4: Power für den Chi-Quadrat-Test bei einer Imbalance pro Gruppe von 68 vs. 82 und 82 vs. 68 Patienten sowie Erhöhung der Fallzahlen auf 83 vs. 69 Patienten

6.2 Fallzahlkorrektur beim Log-Rank-Test

Derselbe Ansatz wie beim Chi-Quadrat-Test wird auch für den Log-Rank-Test genutzt. Dazu wurden die entsprechenden Parameter so gewählt, dass sich für das balancierte Studiendesign eine Fallzahl von 150 Patienten ergibt. Mit der folgenden Syntax [7] kann dies nachvollzogen werden:

```
PROC POWER;
  TWOSAMPLESURVIVAL TEST=logrank[4]
    ALPHA             = 0.05
    SIDES             = 2
    GROUPSURVIVAL     = "Grp A" | "Grp B"
    CURVE ("Grp A")   = (60):(0.2)
    CURVE ("Grp B")   = (60):(0.0755)
    ACCRUALTIME       = 24
    FOLLOWUPTIME      = 84
    POWER             = 0.8
    NPERGROUP         = .;
RUN;
```

[4] Mit den Parametern: GROUPSURVIVAL = Grp.-Bezeichnungen, CURVE = t_i:p_i = Angabe Überlebenswahrscheinlichkeit p_i zum Zeitpunkt t_i, ACCRUALTIME = Rekrutierungsdauer, FOLLOWUPTIME = Beobachtungsdauer [6]

Untersucht man auch hier wieder, wie sich die maximale Imbalance von 14 auswirkt, so ergibt sich für 68 vs. 82 eine Power von 0.797 und für 82 vs. 68 eine Power von 0.800. Mit einer Erhöhung um jeweils einen Patienten pro Gruppe wird dann im ungünstigeren ersten Fall wieder die gewünschte Power von mindestens 80% erreicht (s. Abbildung 5).

Obs	Index	Accrual Time	Follow-up Time	Alpha	N1	N2	N Sub Interval	Loss Exp Hazard 1	Loss Exp Hazard 2	Power
1	1	24	84	0.0500	68	82	12	0	0	0.7973784987

Obs	Index	Accrual Time	Follow-up Time	Alpha	N1	N2	N Sub Interval	Loss Exp Hazard 1	Loss Exp Hazard 2	Power
1	1	24	84	0.0500	69	83	12	0	0	0.8026772505

Abbildung 5: Power für den Log-Rank-Test bei einer Imbalance von 68 vs. 82 Patienten pro Gruppe und Erhöhung der Fallzahlen auf 69 vs. 83 Patienten

7 Diskussion und Ausblick

Bei der in klinischen Studien häufig genutzten stratifizierten Blockrandomisierung wird zur Verhinderung einer potentiellen Vorhersage der nächsten Therapiezuteilung insbesondere in offenen Studien eine möglichst große Blocklänge gewählt. Nachteil dieses Vorgehens vor allem bei vielen Strata ist eine mögliche Imbalance der Fallzahlen in den Therapiegruppen am Ende der Rekrutierung. Diese geht mit einer Reduktion der Power einher. Um diesem möglichen Verlust entgegenzuwirken, kann man die Fallzahl geeignet moderat erhöhen und diesen Effekt kontrollieren.

In diesem Beitrag zeigen wir einen Weg auf, den Einfluss der Imbalance zu quantifizieren. Grundlage dafür ist das in [4] vorgestellte Simulationsprogramm, welches die Größenordnung und Wahrscheinlichkeit für Imbalancen in konkreten Studienplänen abschätzt. Auf Basis dieser simulierten Ergebnisse kann dann mit Fallzahlplanungssoftware, hier PROC POWER in SAS, durch Einsetzen unterschiedlicher Fallzahlen in den Behandlungsgruppen die Auswirkung auf die Power berechnet werden. Durch sukzessive Erhöhung der ungleichen Fallzahlen in den Gruppen kann dann die für die gewünschte Power (in der Regel 80 oder 90%) notwendige Fallzahl ermittelt werden.

In den Untersuchungen zum Zweistichproben-t-Test sowie in den Beispielen zum Chi-Quadrat- und Log-Rank-Test bei vier Strata zeigt sich, dass auch eine größere Imbalance nur wenig Einfluss auf die Power hat, so dass für diese Situationen eine kleine Erhöhung der Fallzahl ausreichen wird, einen möglichen Powerverlust zu korrigieren. Systematische Untersuchungen der Auswirkungen von Imbalancen für diese und weitere Studien-Settings sind notwendig und stehen noch aus (viele Strata, weitere Testverfahren usw.). Eine mögliche Automatisierung mittels SAS-Makros wäre ebenfalls denkbar.

Da eine pauschale Erhöhung der Teilnehmerzahlen [5] nicht mehr akzeptabel ist, kann mit den aufgeführten Untersuchungen eine gemäßigte Fallzahlerhöhung für potentielle Imbalancen begründet werden.

Literatur

[1] L. Armbrust: Auswirkung von Imbalancen auf die Power am Beispiel des zweiseitigen t-Test. Unveröffentlichter Praktikumsbericht, Uni Ulm, 2015

[2] J. Bock: Bestimmung des Stichprobenumfangs. Oldenbourg-Verlag, München, 1998

[3] K. Häußler, R. Muche: Fallzahlplanung mit der Statistiksoftware SAS: Möglichkeiten und Limitierungen. In: Hilgers, Heussen, Herff, Ortseifen (Hrsg.): Proceedings der 12. KSFE-Tagung, Shaker-Verlag, Aachen, 2008, S. 49-67

[4] L. Hupperz, F. Rohlmann, B. Einsiedler, R. Muche: Untersuchungen zum Balanceverhalten der stratifizierten Blockrandomisierung – eine Lösung mit SAS-Makros. In: Muche, Minkenberg (Hrsg.): Proceedings der 17. KSFE-Tagung, Shaker-Verlag, Aachen, 2013, S. 249-259

[5] ICH-E9 Statistical Principles for Clinical Trials http://www.ich.org/products/guidelines/efficacy/efficacy-single/article/statistical-principles-for-clinical-trials.html

[6] R. Minkenberg: Power- und Fallzahlanalyse mit SAS 9. In: Rödel, Bödeker (Hrsg.): Proceedings der 9. KSFE-Tagung, Shaker-Verlag, Aachen, 2005, S. 245-277

[7] The POWER Procedure: Syntax:: SAS/STAT 9.3 User`s Guide : https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_power_sect001.htm

[8] M. Schumacher, G. Schulgen: Methodik klinischer Studien. Springer-Verlag, Heidelberg, 2008