

Berücksichtigung der Ausgangswerte bei der Effektschätzung / Test im zweiarmigen Studiendesign – ein Simulationsvergleich verschiedener Methoden mit SAS-IML

Vanessa Schaser
Universität Ulm
Schwabstr. 13
89075 Ulm
vanessa.schaser@uni-ulm.de

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

Zusammenfassung

In klinischen Studien liegt das Ziel meist darin, den Therapieeffekt einzelner Behandlungsgruppen miteinander zu vergleichen. Der über die Behandlungszeit auftretende Effekt zwischen den Gruppen kann aber unter anderem an den Ausgangswerten der Patienten liegen. Besteht hier bereits vor/bei Studienbeginn ein Unterschied, kann dieser über die Therapie hinweg bestehen bleiben, größer oder kleiner werden. Dieser mögliche Unterschied in den Ausgangswerten wird in der klinischen Forschung durch Randomisierung reduziert, in Beobachtungsstudien muss immer mit einem solchen Unterschied gerechnet werden. Um vergleichbare Aussagen über die Effekte treffen zu können und verzerrte Studienergebnisse zu vermeiden, sollte eine sogenannte Adjustierung der Baseline-Werte stattfinden. Hierfür werden meist die Werte vor Therapiebeginn und nach Ende der Behandlung eines jeden Patienten gemessen, um einen entsprechenden Effekt der Therapie schätzen zu können. Bisher wurde oft ein unverbundener t-Test über die Differenzen zwischen Ausgangs- und Endwert (Change Score) zwischen den verschiedenen Gruppen durchgeführt. Weitere, in der Literatur angegebene Methoden sind die nicht-Berücksichtigung der Ausgangswerte (Post-Treatment), Varianzanalyse für repeated measurements (rmANOVA) sowie die Kovarianzanalyse (ANCOVA), in der der Baseline-Wert als Kovariable eingesetzt wird. Eine Simulationsuntersuchung mit durch SAS-IML erzeugten Daten soll dies beleuchten. Es lässt sich festhalten, dass die Post Treatment Methode in den meisten Studiensituationen nicht angemessen zu sein scheint. So ist sie in nicht-randomisierten Studien aufgrund des großen Bias zur Change Score Methode auszuschließen. In randomisierten Studien führt die Methode bei geringer Korrelation zu ähnlichen Ergebnissen der Power, p-Werte und Parameterschätzer wie die ANCOVA. Die Change Score Methode scheint in randomisierten Studien vor allem einsetzbar zu sein, wenn eine starke Korrelation der Ausgangs- und Postbehandlungswerte vorhanden ist. In nicht-randomisierten Studien scheint die Change Score Methode die angemessensten Ergebnisse zu erzielen, da diese vorhandene Unterschiede in den Ausgangswerten komplett mit in die Auswertung aufnimmt. Die ANCOVA scheint in randomisierten Studien die Methode zu sein, die präferiert verwendet werden sollte. Da diese im Allgemeinen für jeden Korrelationskoeffizienten große Werte der Power, kleine p-Werte und angemessene Parameterschätzer aufweist. In nichtrandomisierten Studien hingegen zeigt die Methode starke Schwankungen der Werte in Abhängigkeit der Korrelation auf.

Schlüsselwörter: Baseline-Messung, Kovarianzanalyse, Simulation, SAS IML

1 Einleitung

In klinischen Studien liegt das Ziel meist darin, den Therapieeffekt einzelner Behandlungsgruppen miteinander zu vergleichen, um dadurch zukünftig die bestmögliche Therapie herauszufiltern. Die Schwierigkeit den über die Behandlungszeit auftretenden Effekt zwischen den Gruppen zu vergleichen kann unter anderem an den Ausgangswerten der Patienten liegen. Besteht hier bereits vor/bei Studienbeginn ein Unterschied in den Baseline Werten, kann dieser über die Therapie hinweg bestehen bleiben, größer oder kleiner werden. Dieser mögliche Unterschied in den Ausgangswerten wird in der klinischen Forschung durch Randomisierung reduziert, in Beobachtungsstudien muss immer mit einem solchen Unterschied gerechnet werden.

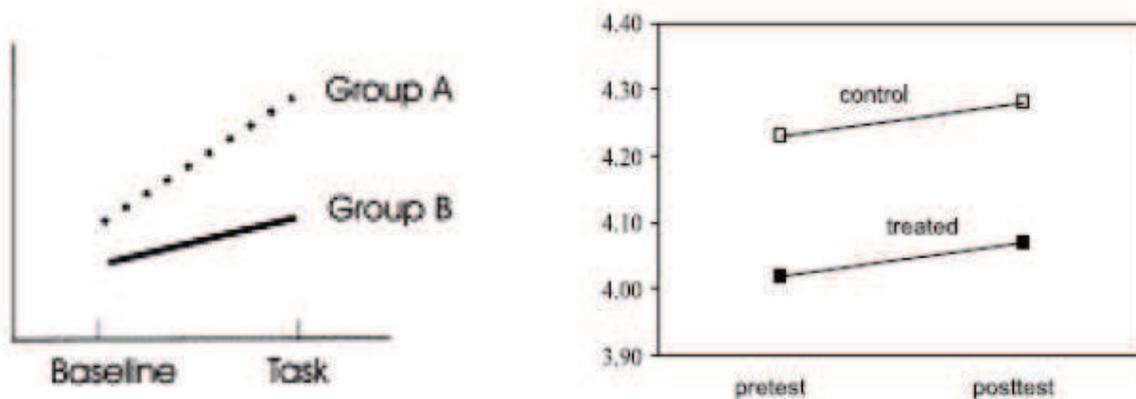


Abbildung 1: Unterschiedliche Verläufe einer Therapie

Betrachtet man in Abbildung 1 lediglich die Werte am zweiten Zeitpunkt, so wird man in beiden Fällen einen Unterschied in den Werten erkennen, obwohl in der rechten Abbildung die Therapieeffekte (gemessen als Veränderung über die Zeit) in beiden Gruppen gleich sind. Hier wird deutlich, dass die Ausgangswerte eine große Rolle bei der Bestimmung des Therapieeffekts und der Beurteilung des Gruppenunterschieds spielen können.

Um vergleichbare Aussagen über die Effekte treffen zu können und verzerrte Studienergebnisse zu vermeiden, sollte eine sogenannte Adjustierung der Baseline Werte stattfinden. Hierfür werden meist die Werte vor Therapiebeginn und nach Ende der Behandlung eines jeden Patienten gemessen. Der Therapieeffekt ist folglich neben der Behandlungsart auch von den Ausgangswerten abhängig, weshalb es unausweichlich scheint, diese mit in die statistische Auswertung einfließen zu lassen. In der Literatur werden für die Baseline Adjustierung unterschiedliche Methoden aufgeführt [9]. Bisher wurde oft ein t-Test über die Differenzen zwischen Ausgangs- und Endwert zwischen den verschiedenen Gruppen durchgeführt. Es stellt sich jedoch die Frage, ob dies für alle vorliegenden Ausgangssituationen geeignet ist oder für spezielle Studiendesigns oder Eigenschaften der erhobenen Daten eine andere Methode angemessener zu sein scheint. Weitere, in der Literatur angegebene Methoden sind die Varianzanalyse für repeated Measurements (rmANOVA) sowie die Kovarianzanalyse, in der der Baseline-Wert als Kovariable eingesetzt wird [1,2,4,5,10,12].

Für die hier beschriebenen Methodenvergleiche wird ein Zweigruppenvergleich für eine stetige, normalverteilte Zielgröße angenommen. Für Gruppenunterschiede in den Ausgangswerten wird ein Shift im Mittelwert und Streuung festgelegt.

Lord's Paradox

Das von Frederic M. Lord im Jahre 1973 aufgestellte Paradoxon postuliert, dass es bei der Analyse von Behandlungseffekten zu gegensätzlichen Ergebnissen kommen kann, je nachdem welche Methode dafür verwendet wird [2]. Bei Vorliegen einer nicht-randomisierten Studie unterscheiden sich die Ausgangswerte zwischen den Gruppen schon vor Beginn der Studie aufgrund der nicht zufälligen Zuteilung. Neben den Ausgangswerten werden zudem noch Daten an einem zweiten Zeitpunkt, meist nach Ende der Behandlung, gemessen. Die gegensätzlichen Ergebnisse der ANCOVA-Methode, mit den Ausgangswerten als Kovariable, und der Change Score Methode, mit der Differenz von Post- und Ausgangswerten als Zielgröße lassen sich am besten erklären, wenn man den Fall betrachtet, dass es keinen Therapieeffekt gibt. In diesem Fall verändert sich der in den Ausgangswerten bestehende Unterschied zwischen den Gruppen über die Zeit hinweg nicht, es liegt also ein paralleler Verlauf, wie in Abbildung 1, vor. Dies ist der Fall, wenn sich die Mittelwerte durch die jeweilige Behandlung nicht ändern oder der Unterschied trotz Veränderung in den Werten derselbe bleibt, die Gruppen also eine gleich starke Änderung aufweisen. Wendet man auf dieses Szenario eine der zwei Methoden an, so erhält man zwei unterschiedliche Ergebnisse [2,9].

Die Verwendung der Change Score Methode führt zu einem Schätzer des Therapieeffekts der nahezu null ist, wohingegen ANCOVA postuliert, dass ein Therapieeffekt vorliegt. Das Problem hierbei scheint zu sein, dass die ANCOVA Methode einen zuvor bestehenden Unterschied zwischen den Gruppen unterschätzt, da nur ein Anteil dieses Unterschieds im Modell berücksichtigt wird. Im Gegensatz dazu wird bei der Change Score Methode direkt die Differenz des Postbehandlungswertes und des Ausgangswertes betrachtet, was im Falle gleichbleibender Mittelwerte in beiden Gruppen einen 'Null-Effekt' hervorbringt. Dies scheint das intuitiv richtige Ergebnis darzustellen, da kein Unterschied in den Ausgangs- und Postbehandlungswerten besteht und somit auch kein Effekt durch die entsprechende Art der Behandlung erkennbar sein kann [1]. In nicht-randomisierten Studien ergibt sich somit die Frage, welche Methode die korrekten Ergebnisse hervorruft.

2 Studienansätze: randomisierte / nicht-randomisierte Studie

Als Grundlage dienen, wie gerade beschrieben, zweiarmige Studienansätze für eine randomisierte und nicht-randomisierte Zuteilung der Beobachtungen zu den 2 Gruppen (in der Medizin meist Therapiegruppen). Die Patienten werden vor und nach der Therapie zu den Zeitpunkten t_0 (Ausgangswert) und t_1 (Endwert) beobachtet und die Zielgröße erhoben.

2.1 Randomisierte Studien

In randomisierten Studien wird davon ausgegangen, dass die Gruppen durch die Randomisierung vergleichbar sind. Bei nicht zu kleinen Fallzahlen kann man davon ausgehen, dass die Ausgangswerte in den zu vergleichenden Gruppen zum Zeitpunkt t_0 eine gleiche Verteilung um dasselbe Lagemaß (hier Mittelwert; Shift=0) bei gleicher Standardabweichung haben. Dementsprechend ist zu prüfen, ob eine Korrektur für die Ausgangswerte bei „höchstens“ empirischen Unterschieden der Ausgangswerte notwendig ist und welche Unterschiede sich ergeben. In den Simulationen werden für diese Situation Daten erzeugt mit gleichem Mittelwert und ähnlicher/gleicher Streuung.

2.2 Nicht-randomisierte Studien

In nichtrandomisierten Studien kann man nicht von der gleichen Verteilung der Ausgangswerte ausgehen. Wir untersuchen hier, ob der Shift des Mittelwertes um einen bestimmten Betrag > 0 bei sonst gleicher Verteilung der Werte berücksichtigt werden sollte und wenn ja, welche Methode am geeignetsten scheint.

3 Methoden für die Adjustierung der Ausgangswerte und Beurteilungskriterien

Die Methoden können als Regressionsansatz beschrieben werden.

Y: Endwerte zum Zeitpunkt t_1

X: Ausgangswert zum Zeitpunkt t_0

Gruppe: dichotome Variable der Gruppenzugehörigkeit

3.1 Adjustierungsmethoden

Folgende Adjustierungsmethoden, jeweils als Regressionsmodell beschrieben, werden untersucht:

Keine Berücksichtigung – Post-treatment

Als einfachste Auswertungsmethode könnte man die Ausgangswerte unberücksichtigt lassen. Dann könnte man hier einen unabhängigen t-Test für den Unterschied der Mittelwerte in den beiden Gruppen berechnen oder das einfache Regressionsmodell aufstellen:

$$Y = \text{Gruppe}$$

Unter der Annahme einer randomisierten Studie und vergleichbarer Ausgangswerte könnte dieser Ansatz eventuell Sinn machen.

Analyse der Veränderung – Change-score

Oft wird zur Berücksichtigung der Ausgangswerte der Therapieeffekt definiert durch die Differenz $Y-X$ für jede Beobachtung. Die so um den Ausgangswert korrigierten Endwerte könnten jetzt durch einen t-Test oder durch das Regressionsmodell

$$Y-X = \text{Gruppe}$$

analysiert werden. So gehen die Ausgangswerte in die Analyse ein, der absolute Wert der Ausgangslage allerdings nicht.

Kovarianzanalyse – Baseline-Messung als Kovariable

Alternativ könnte man die Endwerte in ein Regressionsmodell als Zielgröße einsetzen und die Ausgangswerte als zusätzliche Kovariable berücksichtigen:

$$Y = \text{Gruppe} + X$$

Mathematisch ist dieses Modell äquivalent zu dem der Change-Score-Methode, allerdings gibt es einige Literaturhinweise, die Unterschiede in Bezug auf die Ergebnisse beschreiben.

Varianzanalyse mit Repeated Measurements

Man könnte die beiden Zeitpunkte t_0 und t_1 auch als eine (sehr kurze) Zeitreihe auffassen und Methoden für Repeated Measurements (hier ANOVA) ansetzen. Das zugehörige Modell lautet dann (bei T als Zeitpunkt der Messung):

$$Y = \text{Gruppe} + T + \text{Gruppe} * T$$

Dieser Modellansatz wird vor allem dann angewandt, wenn mehr als zwei Zeitpunkte der Datenerhebung vorliegen. In diesen Modellen wird so der Zeitpunkt durch den Index T beschrieben, der neben '0' und '1' auch andere Werte annehmen kann.

3.2 Beurteilungskriterien

Folgende Beurteilungskriterien zur Untersuchung der unterschiedlichen Ergebnisse der verschiedenen Modellansätze wurden gewählt [3]:

Parameterschätzer und Konfidenzintervall

Der Parameterschätzer für den Gruppenunterschied gibt den Mittelwertunterschied in Originaleinheiten wieder. Anhand dieses Ergebnisses kann u.a. die Relevanz des Unterschieds zwischen den Gruppen diskutiert werden. Ein 95%-Konfidenzintervall für diesen Parameter kann auch als Signifikanzprüfung genutzt werden. Deshalb soll als wesentliches Beurteilungskriterium Unterschiede in diesen Größen hergenommen werden.

p-Wert

Der p-Wert des Gruppeneffekts in den verschiedenen Auswertungsmethoden wird als Beurteilungskriterium genutzt. Unterschiede in dieser Größe könnten die Interpretation der Studienergebnisse stark beeinflussen, wenn die p-Werte sehr unterschiedlich sind.

Power

Die Power beschreibt den Fehler 2. Art. Falls man Fallzahl, Mittelwerte, Streuung und Signifikanz festlegt, kann die Power bestimmt werden. Interessant ist zu erkennen, welche Methoden die höchste Power erzeugen.

Bias

Um den Bias zwischen Methoden zu untersuchen, muss eine der Methoden als Standard hergenommen werden. Hier wird die ANCOVA-Methode als Vergleichsstandard hergenommen.

4 Die Simulation

Um die unterschiedlichen Adjustierungsmethoden vergleichen zu können, sollen Daten auf Basis eines realen Beispiels zufällig gezogen werden. Dazu werden im nächsten Abschnitt die grundlegenden Daten des Beispiels hergeleitet und die Simulationsparameter festgelegt. Im Abschnitt 4.2 wird dann der Ansatz der Simulation der Daten mit SAS-IML beschrieben. Die Auswertungsschritte zur Herleitung der Beurteilungskriterien werden anschließend skizziert.

4.1 Ausgangslage / -parameter

Die Simulation wird auf Grundlage einer realen Studie aufgesetzt, in der in zwei Gruppen mit je 13 Beobachtungen (Teil einer Cross-Over-Studie) zwei nichtinvasive Beatmungsformen (nCPAP, nHFOV) verglichen werden [6]. Die Zielgröße ist der pCO₂-Wert (in mmHg). Der Gesamtmittelwert und die Standardabweichung der Ausgangswerte über alle 26 Messungen stellen die Grundlage für den Simulationsansatz dar. Bei randomisierten Studien wird angenommen, dass beide Gruppen gleiche Ausgangswerte und Streuung haben, bei nicht-randomisierten Studien sollen die Ausgangswerte sich um 10% bzw. 20%-Punkte unterscheiden.

The MEANS Procedure

gruppe	N Obs	Variable	Label	Mean	Std Dev
nCPAP	13	bl	Ausgangswert pCO_2 (mmHg)	50.00	4.04
		end	Postwert pCO_2 (mmHg)	49.00	8.06
nHFOV	13	bl	Ausgangswert pCO_2 (mmHg)	49.85	8.59
		end	Postwert pCO_2 (mmHg)	47.69	9.49

Abbildung 2: Deskriptive Angaben zum Studienbeispiel

Obs	gruppe	NObs	Variable	Label	Mean	Std	Var
1	total	26	bl	Ausgangswert pCO_2 (mmHg)	49.93	6.32	39.94

Abbildung 3: Deskriptive Angaben der gemittelten Ausgangswerte

Weiterhin eingestellt und untersucht werden die Parameter Fallzahl, Effektstärke und Korrelation.

Fallzahlen: Es werden die Effekte der unterschiedlichen Adjustierung der Ausgangswerte bei verschieden großen Studienumfängen untersucht. Gruppengrößen von jeweils 10, 50, 250 und 1000 Patienten werden betrachtet.

Effektstärke: Untersucht wird außerdem der Effekt der unterschiedlichen Effekte zwischen den Gruppen anhand Cohen's Effektstärke. Kleine (0.2), mittlere (0.5) und große Effektstärke (0.8) werden eingesetzt.

Korrelation: Als weitere Einflussgröße auf die unterschiedliche Adjustierung wird in der Literatur die Korrelation zwischen Ausgangs- und Endwert beschrieben. Um dies untersuchen zu können werden die Korrelationsstärken 0.0, 0.2, 0.5 und 0.8 angenommen.

Aus diesen Angaben lassen sich dann die Kovarianzen für die Verteilungsannahmen der Simulation der Daten bestimmen. Die Kovarianzmatrix nCPAP für das IML Programm (s. Abbildung 4) bei $d=0.5$ wird folgendermaßen festgelegt:

39.94 = Quadrat der SD der zusammengefassten Daten (6.32)

25.47 = $\text{corr} \times \text{SD}(\text{zusammengefasste Daten}) \times \text{SD}(\text{nCPAP_end})$

bei $r=0.5$: $0.5 \times 6.32 \times 8.06 = 25.47$

64.69 = Quadrat der SD bei nCPAP_end (8.06)

Wie in Tabelle 1 und Abbildung 4 im SAS-Programm zu sehen sind diese explizit anzugeben. Eine Berechnung in einem Data-Step in SAS wäre möglich gewesen. Aber da kein allgemeingültiges Simulationsprogramm angestrebt wurde ist dies hier unterblieben. Tabelle 1 gibt so alle Parameter für die Simulationsansätze wieder.

Tabelle 1: Ansätze der Simulation

Anzahl Patienten/Gruppe	N = 10, 50, 250, 1000
Effektgröße (Cohen)	d = 0.2, 0.5, 0.8
Korrelation	r = 0, 0.2, 0.5, 0.8
Baselineunterschiede (nicht-randomisierte Studie)	10%, 20%
Wiederholungen	jedes Szenario 1000 Mal
Ausgangswert nCPAP in jeder Studiensituation	49.93
Ausgangswert nHFOV randomisierte Studie	49.93
Ausgangswert nHFOV bei 10% Abweichung	54.92
Ausgangswert nHFOV bei 20% Abweichung	59.92
Postwerte nHFOV in randomisierter Studie	47.56, 45.41, 43.25
Postwerte nHFOV bei 10% Abweichung	52.5, 50.3, 48.1
Postwerte nHFOV bei 20% Abweichung	57.45, 55.21, 52.96
Kovarianzen nCPAP randomisierte Studie & nicht-randomisierte Studie	0, 10.19, 25.47, 40.75
Kovarianzen nHFOV randomisierte Studie	0, 10.19, 25.47, 40.75
Kovarianzen nHFOV 10% Abweichung	0, 10.68, 26.71, 42.74
Kovarianzen nHFOV 20% Abweichung	0, 11.16, 27.9, 44.64

4.2 Simulation mit SAS-IML

Nach Festlegung der Fallzahl und der Simulationsanzahl in den Makrovariablen &N bzw. &NumSamples wird dann in PROC IML [7] der Mittelwertvektor und die Kovarianzmatrix angegeben. Mit der Funktion RandNormal werden normalverteilte Zufallszahlen erzeugt, die die gewünschten Eigenschaften besitzen. In Abbildung 4 wird hier die Erzeugung von 1000 Wiederholungen von Studiendaten mit 10 Beobachtungen für die Kontrollgruppe in einer randomisierten Studie dargestellt [11].

```
%LET N=10;
%LET NumSamples=1000;
PROC IML;
  Mean={49.93, 49};
  Cov= {39.94 25.47,
        25.47 64.96};
  CALL randseed (2321);
  X=RandNormal(&N*&NumSamples, Mean, Cov);
  X1=X[,1];
  Y1=X[,2];
  ID=REPEAT(T(1:&NumSamples),1,&N);
  SampleId=colvec(ID);
  pid=repeat(T(1:&N),&NumSamples,1);
  treat=j(&Numsamples,&N,0);
  c = "x1" : "x2";
  PRINT (X[1:5,]) [label="First 5 obs in control group"];
  CREATE X VAR{ pid, SampleId,treat, X1, Y1}; APPEND;CLOSE X;
```

Abbildung 4: Ansatz der Simulation mit SAS-IML (Mittelwertvektor und Kovarianzmatrix ist einzugeben je nach Ausgangswerten)

In dem Programmabschnitt in Abbildung 4 werden die Ausgangswerte X1 und Endwerte Y1 in der Datei X für die Kontrollgruppe erzeugt. Entsprechend werden die Daten der zweiten Behandlungsgruppe (X1, Y1) in der Datei Y erzeugt. Die beiden Gruppen werden anhand der Variablen „treat“ auseinandergelassen. Anschließend werden die beiden Dateien dann im Data-Step mit dem SET-Befehl zusammengesetzt (siehe Abbildung 5).

```
/*in einem Datensatz zusammenfuehren*/
DATA XY;
SET X Y;
changescore= Y1- X1;
percentchange=((Y1-X1)/X1)*100;
run;
```

Abbildung 5: Zusammenfügen der Dateien im Data-Step

Die Auswertungen der Regressionsanalysestrategien (siehe Abschnitt 3.1) werden dann mit PROC GLM [8] durchgeführt und die Beurteilungsparameter wie Parameterschätzer, Power und Bias ausgelesen.

```
/*Change Score Methode*/
PROC GLM DATA=XY;
BY SampleId;
CLASS treat;
MODEL changescore=treat/SOLUTION;
ODS OUTPUT 'Solution'=change;
run;
```

Abbildung 6: Auswertung der Change-Score-Methode mit PROC GLM

```
PROC GLM DATA=XY;
BY SampleId;
CLASS treat;
MODEL Y1=X1 treat/SOLUTION;
ODS OUTPUT 'Solution'=solAncova;
run;
```

Abbildung 7: Auswertung der ANCOVA-Methode mit PROC GLM

Über die 1000 Wiederholungen werden die über ODS ausgelesenen Parameter wie die Parameterschätzer oder der p-Wert dann gemittelt und im Ergebnisteil im Kapitel 5 angegeben. Für die Berechnung des Bias werden die Ergebnisse der Change-Score- und ANCOVA-Methode direkt miteinander verglichen und der Unterschied bestimmt. Eine genauere Beschreibung der Berechnung ist detailliert in [9] dargestellt.

5 Ergebnisse

Im Folgenden wird etwas ausführlicher das Ergebnis bzgl. randomisierter Studien für die Change-Score – und ANCOVA-Methode für eine mittlere Effektsize ($d=0.5$) dargestellt. Die Ergebnisse für nicht-randomisierte Studien werden aus Platzgründen nur zusammenfassend dargestellt. Es wird auf [9] verwiesen.

5.1 Randomisierte Studien

Die beispielhafte Studiensituation in randomisierten Studien ist in Abbildung 8 dargestellt. Es wird von gleichen mittleren Ausgangswerten ausgegangen.

Die Power in Abhängigkeit der Fallzahl und Korrelation zwischen den Ausgangs- und Endwerten im Vergleich der Adjustierungsmethoden zeigt Abbildung 9. Anschließend werden in Tabelle 2 die Ergebnisse genauer dargestellt.

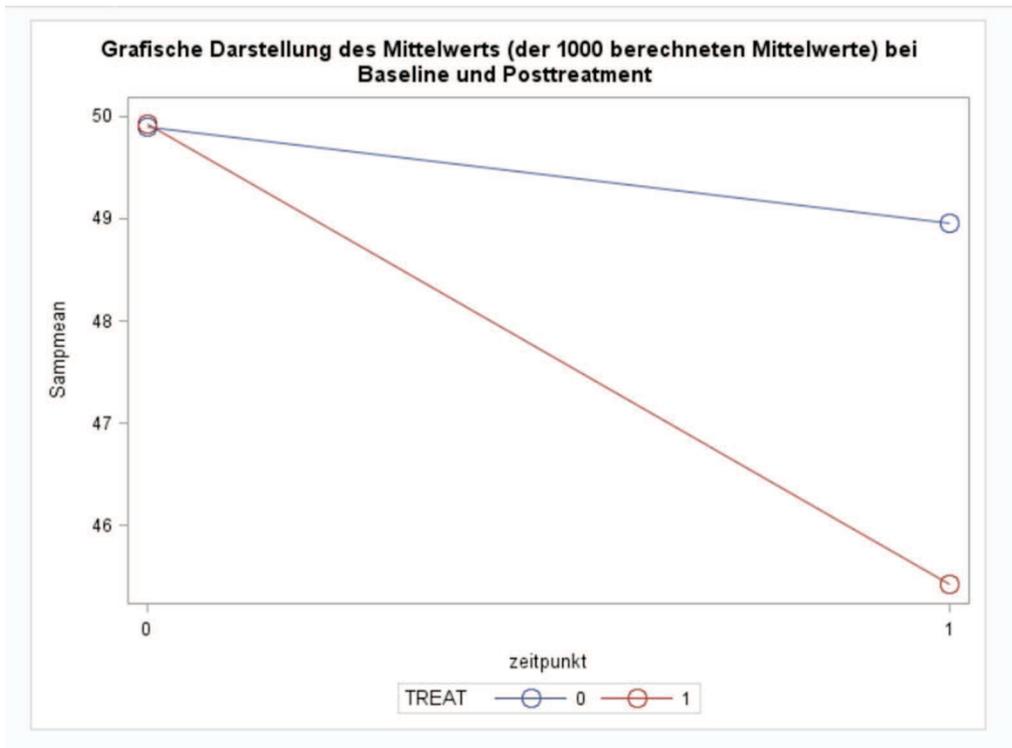


Abbildung 8: Verlauf der Mittelwerte in beiden Gruppen bei randomisierten Studien

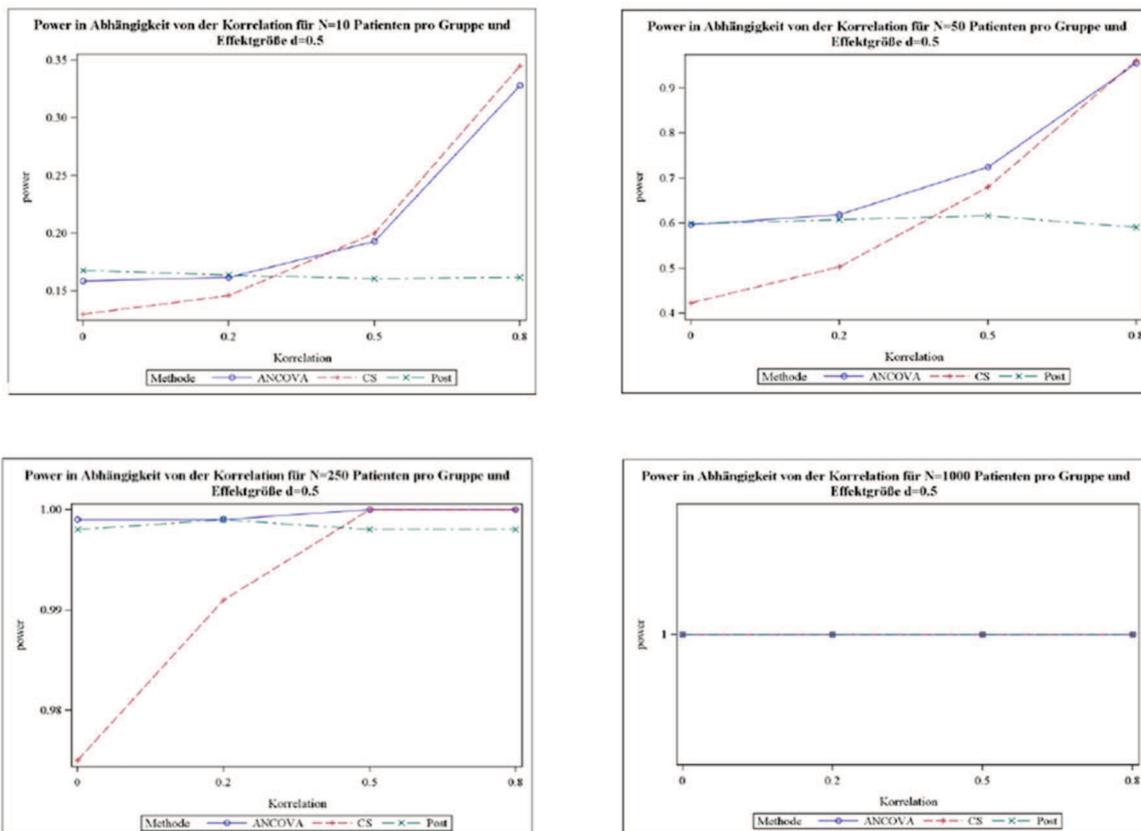


Abbildung 9: Verlauf der Power in Abhängigkeit der Korrelation: randomisierte Studie, mittlere Effektstärke ($d=0.5$)

Tabelle 2: Ergebnisse Simulationen: randomisierte Studien, mittlere Effektstärke (0.5)
Ansätze der Simulation (hier Vergleich ANCOVA und Change-Score (CS))

Obs	Methode	effectsize	PatientenAnzahl	Korrelation	ParameterEstimate	CI_Estimate	Pvalue	CI_Pvalue	power
1	ANCOVA	0.5	10/10	0.0	3.7555486	(3.53;3.98)	0.3674	(0.3493;0.3855)	0.159
2	ANCOVA		10/10	0.2	3.7522020	(3.53;3.98)	0.3627	(0.3447;0.3808)	0.162
3	ANCOVA		10/10	0.5	3.7333657	(3.54;3.93)	0.3328	(0.3149;0.3507)	0.193
4	ANCOVA		10/10	0.8	3.6893248	(3.55;3.83)	0.2187	(0.2030;0.2344)	0.328
5	ANCOVA	0.5	50/50	0.0	3.6014625	(3.5;3.7)	0.1088	(0.0972;0.1203)	0.597
6	ANCOVA		50/50	0.2	3.6012307	(3.5;3.7)	0.1024	(0.0912;0.1136)	0.619
7	ANCOVA		50/50	0.5	3.5999265	(3.51;3.69)	0.0668	(0.0581;0.0754)	0.724
8	ANCOVA		50/50	0.8	3.5968772	(3.54;3.66)	0.0089	(0.0067;0.0111)	0.955
9	ANCOVA	0.5	250/250	0.0	3.5386526	(3.49;3.58)	0.0005	(0.0003;0.0007)	0.999
10	ANCOVA		250/250	0.2	3.5396906	(3.5;3.58)	0.0004	(0.0002;0.0006)	0.999
11	ANCOVA		250/250	0.5	3.5455330	(3.51;3.58)	<.0001	(<.0001;<.0001)	1.000
12	ANCOVA		250/250	0.8	3.5591929	(3.53;3.59)	<.0001	(<.0001;<.0001)	1.000
13	ANCOVA	0.5	1000/1000	0.0	3.5869762	(3.57;3.61)	<.0001	(<.0001;<.0001)	1.000
14	ANCOVA		1000/1000	0.2	3.5870373	(3.57;3.61)	<.0001	(<.0001;<.0001)	1.000
15	ANCOVA		1000/1000	0.5	3.5873814	(3.57;3.61)	<.0001	(<.0001;<.0001)	1.000
16	ANCOVA		1000/1000	0.8	3.5881858	(3.58;3.6)	<.0001	(<.0001;<.0001)	1.000
17	CS	0.5	10/10	0.0	3.7627416	(3.48;4.04)	0.4098	(0.3914;0.4283)	0.130
18	CS		10/10	0.2	3.7574913	(3.51;4.01)	0.3926	(0.3740;0.4111)	0.146
19	CS		10/10	0.5	3.7358248	(3.53;3.94)	0.3395	(0.3216;0.3574)	0.200
20	CS		10/10	0.8	3.6889980	(3.56;3.82)	0.2078	(0.1926;0.2230)	0.345
21	CS	0.5	50/50	0.0	3.6747972	(3.55;3.8)	0.1869	(0.1720;0.2019)	0.423
22	CS		50/50	0.2	3.6561190	(3.54;3.77)	0.1484	(0.1351;0.1617)	0.503
23	CS		50/50	0.5	3.6270604	(3.54;3.72)	0.0766	(0.0674;0.0858)	0.680
24	CS		50/50	0.8	3.5960858	(3.54;3.66)	0.0086	(0.0064;0.0107)	0.961
25	CS	0.5	250/250	0.0	3.5678828	(3.51;3.62)	0.0061	(0.0042;0.0080)	0.975
26	CS		250/250	0.2	3.5617236	(3.51;3.61)	0.0024	(0.0014;0.0034)	0.991
27	CS		250/250	0.5	3.5566813	(3.52;3.6)	0.0002	(<.0001;0.0003)	1.000
28	CS		250/250	0.8	3.5592878	(3.53;3.59)	<.0001	(<.0001;<.0001)	1.000
29	CS	0.5	1000/1000	0.0	3.5978699	(3.57;3.63)	<.0001	(<.0001;<.0001)	1.000
30	CS		1000/1000	0.2	3.5951006	(3.57;3.62)	<.0001	(<.0001;<.0001)	1.000
31	CS		1000/1000	0.5	3.5912187	(3.57;3.61)	<.0001	(<.0001;<.0001)	1.000
32	CS		1000/1000	0.8	3.5878301	(3.58;3.6)	<.0001	(<.0001;<.0001)	1.000

Man erkennt in der Abbildung 9 und vor allem in der Tabelle 2, dass die Korrelation zwischen Ausgangs- und Endwerten eine große Rolle spielt. Die ANCOVA-Methode und die Change-Score Methode sind bei hohen Korrelationen vergleichbar. Wenn die beiden Werte zu den zwei Zeitpunkten nur eine geringe Korrelation zeigen oder unabhängig sind, unterscheiden sich die Ergebnisse der beiden Methoden allerdings. Die Nutzung der Kovarianzanalyse zeigt hier höhere Power-Werte und im Vergleich geringeren Bias (nicht gezeigt).

5.2 Nicht-randomisierte Studien

Bei nichtrandomisierten Studien (Unterschiede von 10%- bzw. 20%-Punkten im mittleren Ausgangswert) verhalten sich die Ergebnisse anders. Bei der Power ist die Change-Score-Methode besser und die ANCOVA-Methode zeigt Schwankungen u.a. in der Power bei geringen Korrelationen zwischen den Ausgangs- und Endwerten. Insgesamt über alle Ergebnisparameter erscheint die Nutzung der Change-Score-Methode über alle Kombinationen als angezeigt.

5.3 Ergebnisse des Studienbeispiels

Die Werte des Studienbeispiels (vgl. Kapitel 4.1), die für den Simulationsansatz gedient haben, sollen nun bezüglich des Therapieeffekts ausgewertet werden. Dafür wurde auf die Daten ebenfalls die jeweilige statistische Adjustierungsmethode mittels SAS angewandt. Dies führte zu den folgenden Ergebnissen in Abbildung 10.

Obs	Methode	PatientenAnzahl	Korrelation	Estimate	CI_Est	Pvalue
1	ANCOVA	13/13	0.79	1.147205234	(-3.28;5.57)	0.5968
2	CS	13/13	0.79	1.153846154	(-3.17;5.48)	0.5872
3	Post	13/13	0.79	1.307692308	(-5.82;8.44)	0.7083

Abbildung 10: Ergebnisse des Studienbeispiels

Es handelt sich bei der Studie um ein Beispiel mit einer geringen Anzahl an Patienten pro Gruppe und starker Korrelation. Des Weiteren unterscheiden sich die Ausgangswerte in beiden Gruppen kaum, weshalb eine randomisierte Studie vorliegt. Die p-Werte sind alle relativ groß, es kann also kein Unterschied zwischen den Behandlungsgruppen nachgewiesen werden. Es kommt jedoch zu leicht unterschiedlichen Ergebnissen, je nachdem, welche Methode verwendet wird.

6 Zusammenfassung

In dieser Arbeit wurde untersucht, welchen Einfluss die Ausgangswerte und der unterschiedliche Umgang mit diesen auf das Ergebnis haben. Hierfür wurden eine stetige Zielgröße, zwei Behandlungsgruppen und die Erhebung der Daten an zwei Zeitpunkten betrachtet. Die Change Score Methode, die ANCOVA und die Post Treatment Methode, die unter anderem die in der Literatur genannten Methoden zur Adjustierung von Ausgangswerten darstellen, wurden durch die Ergebnisse der durchgeführten Simulation verglichen. Durch den Einsatz einer Simulation unter Hilfe von SAS-IML konnten somit sowohl eine randomisierte, als auch nicht-randomisierte Studiensituationen auf Basis realer Werte eines Studienbeispiels ausgeführt werden.

Es wurde jeweils der Einfluss der Korrelation, der Effektgröße und der Fallzahl betrachtet, indem die Werte dieser Größen variierten. Abschließend lässt sich festhalten, dass die Post Treatment Methode in den meisten Studiensituationen aufgrund der Vernachlässigung von Ausgangswerten und damit der Korrelation zwischen Ausgangs- und Postbehandlungswerten nicht angemessen zu sein scheint. So ist sie in nicht-randomisierten Studien aufgrund des großen Bias zur Change Score Methode auszuschließen. In randomisierten Studien führt die Methode bei geringer Korrelation zu ähnlichen Ergebnissen der Power, p-Werte und Parameterschätzer wie die ANCOVA. Somit besteht die theoretische Möglichkeit, in einer solchen Studiensituation die Post Treatment Methode zu verwenden. Die Change Score Methode scheint in randomisierten Studien vor allem zielführend zu sein, wenn eine starke Korrelation der Ausgangs- und Postbehandlungswerte vorhanden ist. In solchen Fällen sind die Werte der Power, p-Werte und Parameterschätzer ähnlich zu den mittels ANCOVA berechneten. Folglich ist auch der Bias der Change Score Methode in einer solchen Situation gering.

In nicht-randomisierten Studien scheint die Change Score Methode die angemessensten Ergebnisse zu erzielen, da diese vorhandene Unterschiede in den Ausgangswerten komplett mit in die Auswertung aufnimmt. Die ANCOVA scheint in randomisierten Studien die Methode zu sein, die präferiert verwendet werden sollte, da sie im Allgemeinen für jeden Korrelationskoeffizienten große Werte der Power, kleine p-Werte und angemessene Parameterschätzer aufweist. Bei Verwendung dieser Methode in Studienprotokollen muss die Korrelation vor Studiendurchführung nicht bekannt sein. In nicht-randomisierten Studien hingegen zeigt die Methode starke Schwankungen der Werte in Abhängigkeit der Korrelation auf. Die Richtung des Schätzers und somit des Effekts variiert, je nachdem welche Korrelation vorliegt, was zu stark abweichenden Parameterschätzern im Vergleich zur Change Score Methode führen kann. Aus diesem Grund weist die ANCOVA teilweise große Werte für den Bias auf und ist somit in nicht-randomisierten Studien, in denen sich die Mittelwerte vor/bei Studienbeginn unterscheiden nicht die bevorzugte anwendbare Methode.

Mit den Ergebnissen der Simulation lässt sich zusammenfassend sagen, dass in randomisierten Studien der Einsatz einer ANCOVA und in nicht-randomisierten Studien der Einsatz der Change Score Methode zu empfehlen ist. Alles in allem spielt die Korrelation bei der Wahl der Methode eine große Rolle und sollte deshalb betrachtet werden.

Eine genaue Dokumentation der Ergebnisse [9] kann von den Autoren angefordert werden.

Literatur

- [1] P.D. Allison: Change scores as dependent variables in regression analysis. *Sociological Methodology* (1990) 93-114
- [2] G.van Breukelen: ANCOVA versus Change from Baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology* 59(2006) 920-925
- [3] B.E. Egbevale et al.: Bias, precision and statistical power of ancova in the analysis of randomized trials with baseline imbalance: a simulation study. *BMC Med Res Method* (2014), 49
- [4] G.F. Liu et al.: Should baseline be a covariate or dependent variable in analysis of change from baseline in clinical trials? *Stat Med* (2009), 2509-2530
- [5] J. Jamieson: Dealing with baseline differences: two principles and two dilemmas. *Int. J. Psychophysiology* 31(1999) 155-161
- [6] D. Klotz et al.: Non-invasive high frequency oscillatory ventilation in preterm infants: a randomised controlled cross-over trial. *BMJ – Archives of Diseases in Childhood, Fetal and Neonatal Edition* (2017) Epub before Print
- [7] SAS/IML 13.1 User's Guide: <https://support.sas.com/documentation/cdl/en/imlug/66845/HTML/default/viewer.htm#titlepage.htm> (9.2.2018)
- [8] SAS/STAT 9.3 User's Guide PROC GLM: https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_glm_sect008.htm (9.2.2018)
- [9] V. Schaser: Berücksichtigung von Ausgangswerten beim Therapievergleich in klinischen Studien. Bachelorarbeit, Studiengang Mathematische Biometrie, Universität Ulm (2017)
- [10] S. Senn: Change from baseline and analysis of covariance revisited. *Statistics in Medicine* 25(2006) 4334-4344
- [11] R. Wicklin: *Simulating data with SAS*. SAS Institute, Cary, NC (2013)
- [12] S. Zhang, J. Paul, M. Nantha-Aree et al.: Empirical comparison of four baseline adjustment methods in analysis of continuous outcomes in randomized controlled trials. *Clinical Epidemiology* 6(2014) 227-235