Bachelorarbeit

Zum Thema:

Randomisierung als Grundlage für Inferenz bei Bernoulli-Verteilten primären Endpunkt

Vergleich von Hypothesentests bei Randomisierungsbasierter und Likelihoodbasierter Inferenzstatistik

Hochschule Stralsund
Fachbereich Elektrotechnik und Informatik
Studiengang Medizinisches Informationsmanagement/eHealth

Vorgelegt von: Rika Jurkschat
Matrikelnummer: 16032
Rika.Jurkschat@fh-stralsund.de

Erstgutachter: Prof. Dr. Lieven Kennes
Zweitgutachter: Dr. rer. nat. Paul Wolf

eingereicht am 28.02.2020

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

EBC Efrons Biased Coin Design

PBR Permuted Block Randomisierung

RAR Random allocation Rule

TBD Truncated binomial Design

OR Odds Ratio

1 Einleitung

Randomisierung als Basis für Inferenz? In der heutigen Zeit steht die Randomisierung vor allem für die Erfüllung zweier Rationalen. Zum einen der Unvorhersehbarkeit der Behandlungszuordnung und zum anderen der Schaffung von Gleichheit in den Gruppen bezüglich bekannter und unbekannter Störfaktoren (Bias) [1]. Die dritte Rationale, der Randomisierung als Basis für Inferenz, wenn die Zufallsauswahl nicht aus einem Populationsmodell erfolgt, ist ein unterschätztes Tool für den Nachweis signifikanter Wirksamkeit in randomisierten klinischen Studien. Dabei ist es scheinbar heutzutage in Vergessenheit geraten, oder wird gar nicht erst gelehrt [1]. Die Methoden zur Berechnung von Randomisierungstests entstanden bereits in den 50er Jahren, aber konnten praktisch durch fehlende Rechenleistung nicht genutzt werden [1]. Dabei erscheint es fast paradox, dass in einer Zeit in der die nötige Rechenleistung nicht zur Verfügung stand, ein Bewusstsein für die Randomisierung als Basis für Inferenz vorhanden war, während diese Erkenntnisse in der heutigen Zeit bei Verfügbarkeit der nötigen Rechenleistung, weitestgehend in Vergessenheit geraten sind.

Motivation und Grundlage dieser Arbeit liefern die Annahmen Papers „Randomization - The forgotten component of the Randomized clinical Trial“ von Rosenberger et al. (2018), welche sich folgendermaßen zusammenfassen lassen:

1. Likelihood basierte Tests liefern keine ähnlichen Ergebnisse wie die korrespondierenden Randomisierungstests.
2. Randomisierungsbasierte Inferenz kann für jeden gewünschten Outcome in klinischen Studien genutzt werden.
3. Moderne Computer machen Randomisierungstests berechenbar in Sekunden.
4. Monte Carlo Prozeduren können verwendet werden, um sehr akkurate Randomisierungstest in Sekunden zu berechnen.
5. Randomisierungstest und Permutationstests sind keine Synonyme.

Diese Aussagen führen zu der Fragestellung, wie groß genau der Vorteil von Randomisierungstests im Vergleich zu Standardtests ist. Ziel ist, anhand einer Simulationsstudie konkrete Zahlen für den Unterschied zu qualifizieren und die Aussagen des Papers von Rosenberger et al. (2018) differenziert zu betrachten. Dabei stehen Studiendesigns mit bernoulliverteilten Zufallsvariablen im Fokus.

Als Grundlage für den Vergleich und die Simulationsstudie, wird die Randomisierung näher dargestellt und Unterschiede zwischen den verschiedenen Methoden der Randomisierung sollen verdeutlicht werden. Dazu wird im allgemeinen die Testtheorie dargestellt auf welcher statistische Tests jeglicher Art beruhen. Im Zuge dessen wird der theoretische Unterschied zwischen Randomisierungstests (Randomisierungsbasierte Inferenz) und Standard Hypothesentests (Likelihoodbasierte Inferenz) geklärt. Bei der randomisierungsbasierten Inferenz wird gezeigt, wo der Unterschied zwischen Randomisierungstests und Permutationstests liegt. Praktisch wird das Ganze für klinische Studien dargestellt.

Die Bachelorarbeit wurde am Institut für medizinische Biometrie und Statistik der Universität zu Lübeck (IMBS), unter der Betreuung von Dr. Maren Vens, durchgeführt Das IMBS stellte die Grundlagen und die technischen Mittel für die Arbeit. Die Simulationsstudie wird mithilfe eines Clusters (Großrechners) berechnet, wobei die Programmierung in der Programmiersprache „R“ geschrieben und ausgewertet wird. Die erstellten R-Skripte werden den Gutachtern digital übermittelt und befinden sich auf der CD zu dieser Arbeit.

2 Material und Methoden

2.1 Randomisierung

Randomisierung ist ein Verfahren welches Zufallsmechanismen verwendet, um Probanden (Versuchspersonen) in verschiedene Gruppen einzuteilen. Ziel der zufälligen Verteilung ist eine Eliminierung der Störfaktoren (Bias) bzw. eine gleichmäßige unabhängige Verteilung dieser auf die zu untersuchenden Gruppen, so dass keine Verzerrung der Daten entsteht. Randomisierung bildet somit eine der wichtigsten Grundlagen, um durch einen statistischen Test eine signifikante Aussage über die Grundgesamtheit treffen zu können. In klinischen Studien ist dies vor allem für den sicheren Nachweis der Wirksamkeit, bezogen auf die Grundgesamtheit, ein notwendiges Werkzeug.

2.1.1 Arten und Modelle

Grundlegend werden zwei Arten von Randomisierung unterschieden [2]:

- Random Sampling (zufällige Stichprobenentnahme)
- Random Allocation (zufällige Gruppenzuweisung)

Random Sampling wählt Individuen bzw. Probanden zufällig aus der zu untersuchenden Population. Dabei stellt eine Population die Teilnehmer einer Gruppe. Bei einer Studie mit zwei zu vergleichenden Gruppen, müssen die Teilnehmer der Gruppen dementsprechend aus zwei, durch die Studie definierten, Populationen gezogen werden. Aus dieser Art der Randomisierung ergibt sich das Populationenmodell (Abbildung 2.1). In diesem Modell werden durch das zufällige Ziehen aus den zwei Population zwei unabhängige und identisch verteilte Probandengruppen gezogen. Die Probanden beider Gruppen besitzen die gleiche Verteilung mit unterschiedlichen unbekannten Einflussparametern durch die Behandlung (θ_A und θ_B).

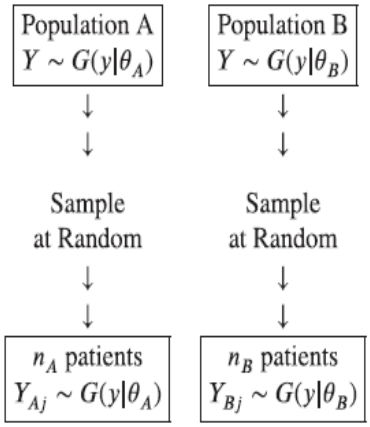

Abbildung 2.1: Populationenmodell [1]

In klinischen Studien ist es meistens nicht möglich aus zwei Populationen Probanden zu ziehen, weil diese nicht existieren. Ein Beispiel dafür die Untersuchung von einem neuen Medikament gegen eine Behandlung mit Placebo. Für ein neues Medikament existiert keine Population, welche es einnimmt. Außerdem gibt es keine Population, welche ein Placebo bekommt gegen eine definierte Krankheit. Aus diesem Grund gibt es zu Beginn einer Studie eine unspezifische Population aus welcher Patienten herausgezogen werden [1]. Dies geschieht undefiniert. Die n Patienten aus der nicht spezifischen Population werden nun mit der Zweiten Art der Randomisierung, Random Allocation, in Behandlungsgruppen verteilt. Dieses Modell wird „Invoked Model“ (Abbildung 2.2) genannt.

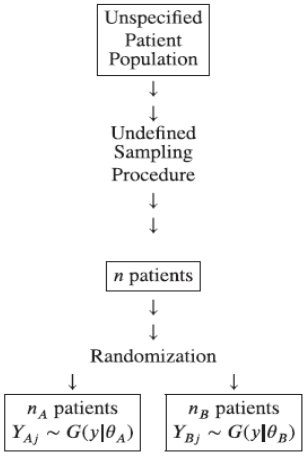

Abbildung 2.2: Invoked Model [1]

Random Allocation teilt die Probanden/Patienten einer Studie durch bestimmte Randomisierungsmethoden ihren Studiengruppen zu. Die Verfahren mit ihren Methoden werden im übernächsten Absatz näher dargestellt. Nach der Randomisierungsprozedur wird im zweiten Modell auch davon ausgegangen, zwei identisch verteilte Patientengruppen mit unterschiedlichen Einflussparametern, zu erhalten. In klinischen Studien sind sehr häufig beide gezeigten Modelle nicht der Realität entsprechend [1]. Durch Mechanismen, wie das Screening, einen Test, ob der Patient die Voraussetzungen für die Studie besitzt, werden Patientengruppen sehr wohl definiert und nicht zufällig aus einer Population gezogen. Aus diesem Sachverhalt ergibt sich das Randomisierungsmodell (Abbildung 2.3).

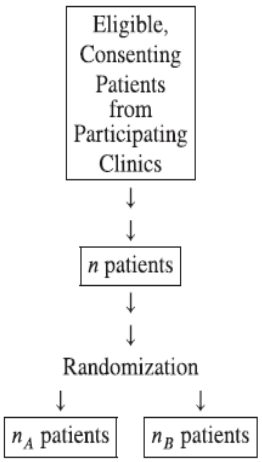

Abbildung 2.3: Randomisierungsmodell [1]

2.1.2 Anforderungen

Das oben genannte Ziel der Randomisierung stellt bestimmte Anforderungen an die Randomisierung und Ihre Verfahren. Nur wenn diese Anforderungen eingehalten werden, sind kausale Aussagen in der Auswertung der erhobenen Daten möglich. Folgende Anforderungen sollte Randomisierung erfüllen[3]:

- Reproduzierbarkeit des Randomisierungsverfahrens
- Nicht-Vorhersehbarkeit
- Zufälligkeit

Die folgenden zwei Beispiele erläutern, wie eine Verletzung der Anforderungen, bei der Randomisierung in einer klinischen Studie, vorliegen kann [3].

Beispiel 1: Die Einteilung der Patienten in die Behandlungsgruppen erfolgt in abwechselnder Reihenfolge. Das heißt der erste Patient ist in Gruppe A, der zweite Patient in Gruppe B, der dritte wieder in Gruppe A usw.
Verletzung: In dieser Randomisierungsmethode ist sowohl die Zufälligkeit als auch die Vorhersehbarkeit verletzt. Die Zuteilung erfolgt nicht zufällig, sondern deterministisch im AB-Schema. Außerdem kann der behandelnde Arzt und das Studienteam stets voraussehen, in welche Gruppe der nächste Patient eingeteilt wird.

Beispiel 2: Bei jedem Patienten der in die Studie aufgenommen wird, wird eine Münze vom Arzt geworfen um die Gruppenzugehörigkeit zuzuteilen.
Verletzung: Dieses Beispiel stellt eine deutliche Verletzung der Reproduzierbarkeit des Randomisierungsverfahrens dar. Hier ist zum einen nicht nachweisbar, ob die Münze vom Arzt wirklich nur einmal geworfen wurde und zum anderen muss vorher geprüft werden ob es sich hier um eine sogenannte faire Münze, also eine Münze mit der Erfolgswahrscheinlichkeit p=1/2, handelt.

2.1.3 Verfahren

Dieser Abschnitt zeigt die drei wichtigsten Verfahren der Randomisierung (Random Allocation) und stellt ihre möglichen Methoden zur Erstellung von Randomisationssequenzen dar. Des Weiteren werden als Basis die zwei möglichen Formen von Randomisationssequenzen dargestellt, welche durch die Methoden erstellbar sind. Die Methoden „Efrons biased Coin Design“ und „Permuted Block Randomisierung“ werden detailliert und beispielhaft als Grundlage für die Simulationsstudie erläutert.

Die Randomisationssequenzen können grundlegend zwei Formen annehmen. Zum einen das „Conditional Setting“, wobei die Gruppengrößen immer ausgeglichen (Verhältnis 1:1) sind und zum anderen das „Unconditional Setting“, wobei alle Aufteilungen in die Gruppen möglich sind. Dabei könnten dementsprechend auch alle Patienten Gruppe A zugeteilt werden. Vorteil des Unconditional Settings ist, dass der Zufall nicht zu Gunsten gleicher Gruppengrößen eingeschränkt werden muss. Es beinhaltet aber auch den Nachteil, dass die Gruppengrößen eine hohe Imbalance beinhalten können und damit keine sichere Aussage über den Unterschied der Gruppen mehr zulassen. Das Conditional Setting hingegen liefert mit den gleichen Gruppengrößen eine sehr gute Grundlage für die statistische Untersuchung der Behandlungsgruppen. Hierbei muss allerdings der Zufall durch Methoden gelenkt werden.

Für die Erstellung von den Randomisierungssequenzen gibt es viele verschiedene Methoden in klinischen Studien, welche sich in vier Grundlegende Verfahren kategorisieren lassen [2,4]:

- Uneingeschränkte Randomisierung
- Balancierte Randomisierung
- Block-Randomisierung
- Adaptive Randomisierung

Die **uneingeschränkte Randomisierung** teilt die Probanden rein zufällig, ohne eine Einschränkung der Behandlungsgruppe, zu. Dieses Verfahren ist mit Wurf einer fairen Münze vergleichbar, wobei nicht auf die Balanciertheit in den Gruppen geachtet wird.

Die **balancierte Randomisierung** teilt die benötigte Gesamtpatientenanzahl gleichmäßig den Gruppen zu und schafft so ein ausgeglichenes Verhältnis (n_A=n_B). Dazu wird eine feste Anzahl der Patienten benötigt, was in klinischen Studien oftmals sehr schwierig sein kann, da einige Probanden zwischenzeitig aus der Studie aussteigen können oder Therapie abbrechen.

Bei der **Block-Randomisierung** wird die benötigte Probandenanzahl in Blöcke gleicher oder unterschiedlicher Länge zerlegt. Jeder Block wird nun einzeln randomisiert, so dass die Randomisierungssequenz am Ende aus einer Zusammensetzung von Blöcken besteht, welche in sich ausgeglichene Gruppengrößen besitzen.

Die **Adaptive Randomisierung** teilt den in die Studie aufgenommen Patienten mit einer bestimmten Wahrscheinlichkeit einer Behandlungsgruppe zu. Diese Wahrscheinlichkeit ist davon abhängig wie viele Patienten sich bereits in den Gruppen befinden, die vorher in die Studie aufgenommen wurden. Somit erfolgt die Zuteilung während der Studie dynamisch. Aus diesem Grund werden balancierte Aufteilungen der Probanden wahrscheinlicher, aber sind nicht garantiert. Außerdem wird durch die geringere Wahrscheinlichkeit für nicht aussagekräftige Randomisationssequenzen, eine sehr große Imbalance (z.B. alle Patienten in Gruppe A) vermieden bzw. sehr unwahrscheinlich. In diesem Verfahren gibt es teilweise auch Methoden, welche eine Unausgeglichenheit der Gruppen nicht zulassen. In der Regel gehören diese Verfahren aber dem Unconditional Setting an.

Die nachfolgende Tabelle (Tabelle 2.1) fasst noch einmal die Formen und Verfahren übersichtlich zusammen und zeigt mögliche Methoden der Randomisierung.

Tabelle 2.1: Gruppierte Methoden der Randomisierung

Unconditional Setting	**Conditional Setting**
Uneingeschränkte Randomisierung: • Münzwurf	Balancierte Randomisierung: • Random Allocation Rule • Truncated Binomial Design
Adaptive Randomisierung: • Efrons Biased Coin Design • Accelerated biased Coin Design	Block-Randomisierung: • Permuted Block Randomisierung • Random Block Randomisierung

Für die Wahl des Verfahrens und der Methode sind verschiedene Faktoren der klinischen Studie entscheidend. Zum Beispiel, ob es sich um eine multizentrische Studie handelt, oder ob zentral bzw. dezentral randomisiert wird. Eine multizentrische Studie ist eine Studie bei der es mehrere medizinische Zentren gibt, welche Probanden in die Studie aufnehmen. Diese Zentren können entweder zentral ihre Probanden den Gruppen zuordnen oder eine übergeordnete Organisation Randomisiert für alle Studienzentren dezentral. Des Weiteren spielt auch die Größe der Studie, bezogen auf die Probandenanzahl, eine wichtige Rolle.

2.1.4 Efrons Biased Coin Design

Efrons biased Coin design stellt eine Adaptive Methode zur Randomisierung des unconditional Settings dar. Dabei wird eine Wahrscheinlichkeit festgelegt, mit welcher ein Proband in eine Gruppe zugeteilt wird. Die Festlegung der Wahrscheinlichkeit hängt von der Anzahl der Probanden in den Gruppen ab, welche bereits aufgenommen und zugeteilt wurden. Das heißt die Gruppe mit weniger Probanden bekommt die höhere Wahrscheinlichkeit. Mathematisch lässt sich das ganze folgendermaßen darstellen [1]:

$$P(A_i) = \begin{cases} 1/2 & \text{für } D_{i-1} = 0 \\ p & \text{für } D_{i-1} < 0 \\ 1-p & \text{für } D_{i-1} > 0 \end{cases}$$

$P(A_i)$ ist hier die Wahrscheinlichkeit, dass der i-te Patient in Gruppe A aufgenommen wird mit $i= 1, \ldots, n$. D_{i-1} beschreibt die Differenz der Anzahlen von den bis Stelle *i-1* bereits eingeteilten Patienten in Gruppe A und B, das heißt, die derzeitige Anzahl in Gruppe A minus der derzeitigen Anzahl an Patienten in Gruppe B. Nach Efron wird die Wahrscheinlichkeit auf $p=2/3$ gesetzt [1]. Schaut man sich die Wahrscheinlichkeiten der Einzelnen möglichen Randomisierungssequenzen an, ist bereits in einem Beispiel mit $n=4$ erkennbar das die Sequenzen mit einer hohen Unausgeglichenheit eine viel kleinere Wahrscheinlichkeit haben. Die Tabelle 2.2 zeigt alle möglichen Gruppenzuteilungen (Permutationen) mit ihren Wahrscheinlichkeiten nach Efrons biased Coin design.

Tabelle 2.2: Wahrscheinlichkeiten beim Efron Biased Coin Design

Sequenz	Wahrscheinlichkeit
A A A A	0.0185185
A A A B	0.0370370
A A B A	0.0370370
A A B B	0.0740741
A B A A	0.0555556
A B A B	0.1111111
A B B A	0.1111111
A B B B	0.0555556
B A A A	0.0555556
B A A B	0.1111111
B A B A	0.1111111
B A B B	0.0555556
B B A A	0.0740741
B B A B	0.0370370
B B B A	0.0370370
B B B B	0.0185185

2.1.5 Permuted Block Randomization

Permuted Block Randomisierung stellt eine Methode zur Block-Randomisierung dar, bei welcher die Blocklänge b festgelegt wird. Die Blockanzahl m ist somit von der gesamten Patientenanzahl n abhängig und lässt sich mathematisch darstellen als:

$$m = n/b$$

Die einzelnen Blöcke werden mit den Methoden Random Allocation Rule (RAR) oder Truncated Binomial Design (TBD) unabhängig voneinander randomisiert. Bei der Methode RAR haben alle möglichen Sequenzen die gleiche Wahrscheinlichkeit, beim TBD nicht. Die Wahrscheinlichkeit p einer kompletten Randomisierungssequenz aus allen Blöcken, für eine Randomisierung mit der Methode RAR, lässt sich wie folgt berechnen:

$$p = \left(\frac{1}{x}\right)^m$$

Mit x, die Anzahl der möglichen Permutationen pro Block. Da alle Permutationen bei RAR die gleiche Wahrscheinlichkeit haben, besitzt eine Blocksequenz die Wahrscheinlichkeit $1/x$. Die Blocklänge kann in Studien individuell nach Probandenanzahl festgelegt werden. Zu beachten ist dabei, dass mit kleineren Blocklängen die Anzahl der Permutationen im Block gesenkt werden und sich die Anzahl der Blöcke erhöht. Soll die Gesamtanzahl möglicher Randomisierungssequenzen x^m möglichst geringgehalten werden, z.B. für die Berechnung von Randomisierungstests, sollte die Blocklänge so kurz wie möglich gehalten werden. Die Kleinste mögliche Blockanzahl ist 1. Für diesen Fall ist zu beachten, dass die Balance zwischen den Gruppen erst am Ende der Studie zu gewährleisten ist. Wird n in mehrere Gruppen aufgeteilt, ist eine balancierte Zwischenauswertung der Studie möglich, wenn eine definierte Anzahl an Blöcken vollständig ist.
Neben einer festen Blocklänge gibt es noch die Möglichkeit variable Blocklängen zu wählen, auch Random Block design genannt. Dabei wird die Patientenanzahl n in Blöcke verschiedener Längen zerlegt. Die Blöcke werden im Anschluss zufällig aneinandergereiht. Vorteil dieser Methode ist die geringe Vorhersehbarkeit der Patienteneinteilung. Bei der Permuted Block Randomisierung, insbesondere mit niedrigen Blocklängen, lassen sich die folgenden Gruppeneinteilungen leichter voraussagen. Beispielsweise bei einem Block der Länge 4, bei dem 2 Patienten bereits in die Gruppe A eingeteilt wurden, ist vorhersagbar, dass die nächsten zwei Patienten in die Gruppe B eingeschlossen werden, wenn die Information über die Blocklänge bekannt ist.

Die Wahrscheinlichkeiten und der Aufbau der Randomisierungssequenzen dienen als Grundlage für Randomisierungstests und werden im Abschnitt „Randomisierungsbasierte Inferenz“ näher betrachtet.

2.2 Testtheorie

Die Testtheorie untersucht und konstruiert statistische Tests und bildet damit ein Teilgebiet der mathematischen Statistik [5]. Statistische Tests dienen der Bildung von Inferenz, also um Schlussfolgerungen über eine gesamte Population zu ziehen, auf der Basis von statistischen Daten einer Stichprobe aus der Population.

Für die Aufstellung statistischer Tests gibt es viele Theorien und Methoden. Das Grundprinzip des Testens ist bei allen Methoden gleich und wird in diesem Abschnitt näher erläutert. Dies dient als Grundlage, die zwei Testtheorien „Randomisierungsbasierte Inferenz“ und „Likelihoodbasierte Inferenz“ näher zu beschreiben und zu differenzieren.
Für die Simulationsstudie wird die Testtheorie mit einer bernoulliverteilten Zufallsvariablen im letzten Punkt näher erläutert.

2.2.1 Statistischer Test

Ein statistischer Test dient dazu, bei einer definierten Irrtumswahrscheinlichkeit, anhand einer Stichprobe eine Aussage über die Grundgesamtheit treffen zu können. Diese Tests werden auch Hypothesentests oder Signifikanztests genannt. Die Stichprobe entstammt der Grundgesamtheit über die eine Aussage getroffen werden soll und ist durch ihre Zufallsvariable X beschrieben. Dabei werden Gültigkeit bzw. Ungültigkeit einer Behauptung überprüft. Diese Behauptung wird Nullhypothese genannt. Das Verfahren liefert für jede Stichprobe ein Ergebnis, ob die Hypothese gestützt, oder verworfen wird [6]. Die Abbildung 2.4 zeigt die Schritte zum Erkenntnisgewinn verdeutlicht.

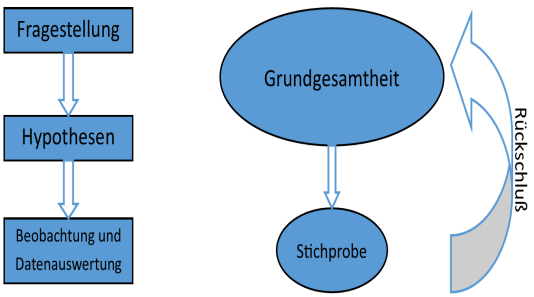

Abbildung 2.4: Prinzip des Signifikanztests [15]

Einzelne Ergebnisse einer Stichprobe sind Realisationen der Zufallsvariablen X

und somit das Ergebnis eines Patienten. Die Stichprobe mit ihren n Ausprägungen von X kann wieder als Zufallsvariable mit einer Verteilung dargestellt werden. Die Zufallsvariable X liefert beim Testprinzip mit ihren möglichen Ausprägungen eine Verteilungsfunktion $FX\ (x/\vartheta)$ [6]. ϑ ist der Parameter, der die Verteilung definiert. Bei parametrischen Tests ist die Verteilung klar und dementsprechend die Parameter. Durch die Verteilung kann die Stichprobe in einer testbaren Funktion dargestellt werden, ein Teststatistikwert berechnet und mit einem Prüfwert verglichen werden. Der Prüfwert wird festgelegt durch die einzuhaltende, vorbestimmte Irrtumswahrscheinlichkeit (α-Fehler), die Verteilung und die Größe der Stichprobe. Die Verteilung der Stichprobe ist abhängig von der Art der Zufallsvariablen bzw. welche Werte sie annehmen kann, zum Beispiel Werte zwischen 0 und 1, ganze Zahlen oder reelle Zahlen. Der Prüfwert gibt an, ab welchem Teststatistikwert die vorher aufgestellte Hypothese sehr unwahrscheinlich ist, also in der Regel einen α-Fehler von 5% aufweist. Bei einer Teststatistik, die größer ist als der Prüfwert, wird die aufgestellte Hypothese verworfen bzw. für ungültig erklärt. In diesem Fall wäre das Ergebnis oder ein noch extremeres, unter der Annahme/Bedingung, dass die aufgestellte Hypothese wahr ist, sehr unwahrscheinlich ($<$5%).

Die Wahl des Tests hängt von verschiedenen Faktoren ab. Es sollten sich bei der Wahl des Tests folgende Fragen beantwortet werden [6]:

- Was ist die Verteilung der zu erhebenden Daten?
- Gibt es zwei unabhängige Stichproben oder eine gebundene?
- Sind die möglichen Ausprägungen der Zufallsvariablen nominal, ordinal oder quantitativ?
- Was soll konkret geprüft werden?

Der letzte Punkt bezieht sich auf die möglichen übergeordneten Kategorien von statistischen Tests und beschreibt, wie die Parameter getestet werden. Die üblichsten Kategorien sind in Tabelle 2.3 mit Beispielen verdeutlicht [7]:

Tabelle 2.3: Kategorien von Hypothesentests

Kategorie	Beispiel
Prüfung auf Mitte	t-Test
Unabhängigkeitstest	χ^2-Test, Fisher´s exakter Test
Anpassungs- oder Verteilungstests	Binomialtest, χ^2-Anpassungstest
Prüfung auf Streuung	Levene-Test
Prüfung von Zusammenhängen	χ^2-Kontingenzanalyse
Nicht parametrische Tests	Wilkoxon-Rangsummen-Test

Allgemein treten folgende Verteilungen von X unter H_0 sehr häufig auf:

- Normalverteilung $X \sim \mathcal{N}\ (\mu,\ \sigma)$
- Bernoulliverteilung $X \sim Ber(p)$
- Binomialverteilung $X \sim Bin(n,k,p)$

2.2.2 Hypothesen

Beim Testen von Hypothesen müssen jeweils zwei Hypothesen aufgestellt werden. Die Nullhypothese und die Alternativhypothese. Dabei wird die Nullhypothese grundlegend angenommen und nur dann verworfen, wenn die Irrtumswahrscheinlichkeit kleiner 5% ist, dass diese Wahr ist. Hypothesen können dabei ein- und zweiseitig aufgestellt werden. Dabei wird je nach Setting der Studie eine individuelle Hypothese aufgestellt, je nachdem was primär gezeigt werden soll. Einseitige Hypothesen sind gerichtete Hypothesen und dürfen nur dann verwendet werden, wenn die Testrichtung gegeben ist (Parameter größer oder kleiner Vergleichswert) und objektiv und sachlich begründet werden kann. Bei einem zweiseitigen Test wird die Irrtumswahrscheinlichkeit symmetrisch auf die Verteilungsfunktion aufgeteilt. Existiert keine Begründung für einen einseitigen Test, sollte der zweiseitige immer bevorzugt werden [8].

2.2.3 Fehler 1. und 2. Art

Die Fehler erster und zweiter Art, auch α- und β-Fehler genannt, werden durch folgende Einflüsse kategorisiert:

- Welche Entscheidung liefert der statistische Test?
- Was ist die Wirklichkeit?

Beide Fragen beziehen sich auf die Auswahl zwischen der Nullhypothese (H_0) und der Alternativhypothese (H_1). Die folgende Tabelle (Tabelle 2.4 [9]) Zeigt die Einflüsse und ihre Fehler:

Tabelle 2.4: Fehler 1. und 2. Art

		Wirklichkeit	
		H0 ist wahr	H1 ist wahr
Entscheidung des Tests	für H0	Richtig (1-α)	Fehler 2. Art (β)
	für H1	Fehler 1. Art (α)	Richtig (Power = 1-β)

Der **Fehler 1. Art** ist in einer klinischen Studie der wichtigste Fehler. Er zeigt die Wahrscheinlichkeit für falsch positive Ergebnisse, also die Irrtumswahrscheinlichkeit, an. Genauer, die Wahrscheinlichkeit, dass man sich anhand des Tests für H_1 entscheidet aber in Wahrheit H_0 gilt. Dieser Fehler wird auch als Konsumentenrisiko bezeichnet, da eine prozentuale Chance in Höhe von α besteht, dass das Produkt für den Konsumenten in Wahrheit nicht wirkt [9].

Der **Fehler 2. Art**, bzw. β-Fehler, beschreibt die Wahrscheinlichkeit einer falsch negativen Aussage. Das heißt die Wahrscheinlichkeit sich für H_0 zu entscheiden obwohl, H1 in Wahrheit gilt. Dieser Fehler ist vor allem für die Durchführenden einer klinischen Studie von Belang, da sie einen Unterschied und somit H_1 finden möchten. Der Fehler 2. Art wird daher auch Produzentenrisko genannt. Für die Durchführenden einer klinischen Studie ist es wichtig diesen Fehler möglichst gering zu halten. Für die Berechnung wird H_1, also ein Unterschied zwischen den Gruppen, angenommen. Aus dem Fehler 2. Art ergibt sich gleichzeitig die Power eines Tests, beschrieben durch 1-β. Die Power zeigt nun, wie gut der Test den Unterschied der Gruppen in den Daten belegen kann, bzw. wie wahrscheinlich ein positives Testergebnis ist, wenn H_1 wirklich gilt [9]. Die Power wird vor allem in klinischen Studien bei der Fallzahlplanung benötigt. Hierbei darf die Fallzahl n nicht zu niedrig gewählt werden, da der Test sonst nicht ausreichend Power aufbringen kann, aber sie darf auch nicht zu hoch sein da sonst unnötig viele Patienten rekrutiert werden müssen. Der Fehler 2. Art kann dabei durch mehrere Faktoren beeinflusst und gesenkt werden [6]:

- Erhöhung des Stichprobenumfangs n

- Testverfahren (Verwendung des besten statistischen Tests)
- Variabilität in den Daten senken (z.B. Streuung senken)
- Effektstärke erhöhen (Unterschied der Gruppen vergrößern)
- Fehler erster Art erhöhen

Fehler 1. und 2. Art verlaufen entgegengesetzt zueinander. Das heißt, in einer klinischen Studie wird bei Senkung des Fehlers 1. Art der Fehler 2. Art größer und umgekehrt. Nur mit ausreichend großen n lässt sich ein kleines α und β erzwingen.

2.2.4 Statistischer Test bei bernoulliverteilten Zufallsvariablen

Ist die Stichprobe beschrieben durch die Zufallsvariable $X_i \in \{0,1\}$ mit $i=1,..,n$ und der Erfolgswahrscheinlichkeit $P(X = 1) = p$, ist diese bernoulliverteilt mit $X \sim Ber(p)$ [10]. Damit ergibt sich, dass die Stichprobe als Zufallsvariable binomialverteilt ist mit $X \sim Bin(n,k,p)$, mit der Stichprobengröße n, der Anzahl der Erfolge k und der Erfolgswahrscheinlichkeit p, welche der Zufallsvariablen X entsprang. Bei bernoulliverteilter Zufallsvariablen wird meistens auf Unabhängigkeit von zwei solcher Zufallsvariablen getestet. Grundlage für die Berechnung eines statistischen Tests für eine bernoulliverteilten Zufallsvariablen ist die Erstellung einer 2x2 Feldertafel (auch Kontingenztafel) aus der Stichprobe. Diese stellt zusammengefasst die Erfolge und Misserfolge der Stichprobe, aufgeteilt nach den Werten, welche die zwei Zufallsvariablen annehmen können, dar. Beispiele für statistische Tests, anhand von Kontingenztafeln, sind der χ^2-Test oder Fishers exakter Test.

2.3 Likelihoodbasierte Inferenz

2.3.1 Definition eines Likelihood-basierten Tests

Die allgemein bekannten Hypothesentests, wie zum Beispiel der T-test, Chiquadrat-Test, etc., lassen sich auf die Likelihoodfunktion und ihre Methoden, zurückführen. Die Likelihood ist ein Maß, welches die Wahrscheinlichkeit der verschiedenen unbekannten Werte eines Parameters angibt [11]. Dabei besitzt die Stichprobe einer Zufallsvariablen X immer Parameter der zu Grunde liegenden Grundgesamtheit. Die Zufallsvariable weist immer eine Verteilung auf die durch den Einflussparameter bestimmt wird. Sie lässt sich definieren als $FX(x/\vartheta)$ [6]. Diese Verteilungsfunktion entsteht aus einer Dichtefunktion, wobei die Verteilungsfunktion die kumulierten

Wahrscheinlichkeiten und die Dichtefunktion die relativen Häufigkeiten beinhaltet. Aus der Dichtefunktion wird konkret die Likelihoodfunktion gewonnen, welche unter anderem dazu dient, Schätzfunktionen zu konstruieren [11]. Die Schätzfunktion kann dann dazu genutzt werden, anhand von erhobenen Daten einen Schätzwert zu ermitteln. Mit diesem erhält man Informationen über den Parameter der Grundgesamtheit. Um also eine Aussage über die Grundgesamtheit treffen zu können, liefert die Schätzfunktion die Grundlage zur Berechnung von Teststatistiken bei Hypothesentests [12].

Ist die Verteilung der Stichprobe bekannt, sind auch die Parameter definiert. Die Werte die die Parameter annehmen können, werden dann untersucht mit einem zur Verteilung passenden statistischen Test [7]. Im Fall einer bernoulliverteilten Zufallsvariablen ist der Parameter die Erfolgswahrscheinlichkeit p. Bei zwei zu vergleichenden Gruppen (A und B) werden diese beiden Erfolgswahrscheinlichkeiten p_A und p_B nun miteinander auf Unabhängigkeit getestet [13,14].

2.3.2 Unabhängigkeitstests für zwei dichotome Merkmale

Bei Unabhängigkeitstests für zwei dichotome Merkmale stellt die Darstellung dieser in einer 2x2 Feldertafel (Tabelle 2.5), die Grundlage der Berechnung. Dichotom bedeutet, dass die beiden Merkmale bzw. Variablen nur zwei mögliche Ausprägungen haben, also bernoulliverteilt sind. Im Fall einer klinischen Studie wäre dies die Variable X, die den Erfolg (0 oder 1) angibt und die Variable Y, die die Gruppenzugehörigkeit (A oder B) definiert [13,14].

Tabelle 2.5: Kontingenztafel

	Erfolg	**Misserfolg**	**Summe**
Gruppe A	a	b	n_A
Gruppe B	c	d	n_B
Summe	n_1	n_0	n

Der Test untersucht, ob X unabhängig von Y ist. Das heißt, ob ein Erfolg von der Behandlungsgruppe unabhängig ist. Dazu werden die Erfolgswahrscheinlichkeiten pro Gruppe separat betrachtet. Die Erfolgswahrscheinlichkeiten (p_A und p_B) für Gruppe A und B, definiert durch ihre Zufallsvariable, sind gegeben mit:

- $X_A \sim Ber(p_A)$
- $X_B \sim Ber(p_B)$

Für die Wahrscheinlichkeiten p_A und p_B lässt sich ein deskriptiver Schätzer berechnen gegeben durch:

$$\widehat{p_A} = a/n_A$$
$$\widehat{p_B} = c/n_B$$

Die folgenden Hypothesen (zweiseitig) ergeben sich aus diesem Szenario:

$$H_0: p_A = p_B \quad H1: p_A \neq p_B$$

Der kritische Wert (${\chi^2}_{1,1-\alpha}$) zum Vergleich der Teststatistik, wird der χ^2-Verteilungstabelle entnommen. Bei 2x2 Feldertafeln ist die Anzahl der Freiheitsgrade immer 1, wobei immer ein Prüfwert von 3,84 genutzt werden kann. Die χ^2-Verteilung erlaubt durch ihre Herleitung aus der Normalverteilung eine Aussage über die Beschaffenheit eines vermuteten Zusammenhangs [14].

Die folgenden drei Abschnitte beschreiben verschiedene Tests auf Unabhängigkeit und ihre Unterschiede zwischen einander. Grundlegender Unterschied ist die Festsetzung von Werten der 2x2 Feldertafel. Es gibt Kontingenztafeln mit festen Reihensummen, mit zufälligen Randsummen und mit fixen Randsummen.

2.3.3 χ^2-Test

Der χ^2-Test entstammt dem binomialen Z-Test. Dieser bringt eine Teststatistik hervor welche approximativ standardnormalverteilt ist. Die Testsstatistik des χ^2-Tests ist daher die quadrierte Teststatistik des binomialen Z-Tests [15]. Da die χ^2-Verteilung auch der Normalverteilung entspringt, durch die Summierung der quadrierten Zufallsvariablen $X1^2 + \ldots + Xn^2$, kann diese zur Berechnung des Prüfwertes herangezogen werden [16]. Der Chiquadrat-Test bezieht sich auf eine Kontingenztafel bei der die Reihensummen fest sind. Der Teststatistikwert unter H0 wird wie folgt berechnet [14]:

$$T = n * \frac{(a*d - b*c)^2}{(a+b)*(c+d)*(a+c)*(b+d)} \sim \chi_1^2$$

Der p-Wert, welcher mit dem Signifikanzniveau verglichen wird ergibt sich aus [13]:

$$p = \frac{1}{2} * 10^{\frac{-\chi^2}{3,84}}$$

Der Test ist nur anzuwenden für Fallzahlen größer als 60, da die Erwartungswerte für *a, b, c* und *d* jeweils größer als 5 sein müssen, weil die Daten ansonsten nicht approximativ χ^2-Verteilt sind. Der Erwartungswert für die einzelnen Felder lässt sich berechnen aus der Zeilensumme*Spaltensumme/n. Ist die Voraussetzung nicht erfüllt, wird Fisher´s exakter Test empfohlen [13].

2.3.4 Fisher's exakter Test

Fishers exakter Test ist ein Test auf Unabhängigkeit, welcher von einer Kontingenztafel mit festen Randsummen ausgeht. Daraus ergibt sich, dass der Wert für a, in der Kontingenztafel, die Verteilung und somit die Teststatistik stellt. Fisher zeigte, das a hypergeometrisch verteilt ist mit $a \sim Hyp\ (n,\ n_A,\ n_1)$. In einer Kontingenztafel mit festen Randsummen lässt sich die Verteilungsfunktion folgendermaßen darstellen [16]:

$$F(k) = P(a \leq k)$$

wobei

$$P(a = k) = \frac{\binom{n_A}{k} * \binom{n_B}{n_1 - k}}{\binom{n}{n_1}}\ mit\ 0 \leq k \leq min(n_A, n_1)$$

die Wahrscheinlichkeiten aller möglichen Werte für a angibt. Für diese Funktion lässt sich die Verteilung von a darstellen, indem alle möglichen Werte für a, was k entspricht, eingesetzt werden. Anhand der Werte und dem festgelegten Signifikanzniveau, lassen sich eine Ober- und Untergrenze für a berechnen, ab welchen der Test signifikant wird. H_0 wird dementsprechend verworfen, wenn [16]:

$$H_0\ wird\ verworfen,\ falls \begin{cases} F(a) < \frac{\alpha}{2} & \text{für } a \leq \frac{k}{2} \\ 1 - F(a) < \frac{\alpha}{2} & \text{für } a > \frac{k}{2} \end{cases}$$

Bei dieser Berechnung und einer kleinen Fallzahl, entsteht der Nachteil, dass das α-Niveau nicht richtig ausgeschöpft werden kann. Das bedeutet, dass die Werte der Ober- und Untergrenze für *F(a)* an einem Punkt liegen, welcher deutlich geringer ist als $\alpha/2$ bzw. 1- $\alpha/2$, aber dennoch der erstmögliche Wert für k ist, der das festgelegte α-Niveau einhält. Das α-Niveau gibt an, ab welchem p-Wert die Nullhypothese verworfen wird. Dieser Verlust des α-Niveaus wird durch das Neyman-Pearson Lemma beschrieben und senkt gleichzeitig die Power des Tests. Je größer demnach die Stichprobe ist, desto genauer werden die Ober- und Untergrenzen für a.

Für Fisher´s exakten Test ist auch die Berechnung des p-Wertes möglich, welcher ohne die genaue Berechnung der Ober- und Untergrenzen für a möglich ist. Der p-Wert ergibt sich aus *P(a=k)* mittels Umformung. Er ist konkret definiert durch [16]:

$$p = \frac{(a+b)! * (c+d)! * (a+c)! * (b+d)!}{a!b!c!d!n!}$$

In klinischen Studien kommt es nahezu gar nicht vor, dass die Randsummen von Kontingenztafeln fest sind und das Ergebnis nur von dem Erfolg der Gruppe A abhängt. Dabei sind die Erfolge von Gruppe A und B unabhängig und die Randsumme der Erfolge entsteht aus den einzelnen Erfolgen und kann daher niemals fest sein. Die Reihensummen hingegen, sind durch die Studie sehr wohl definiert durch die Fallzahlplanung und die Randomisierungsmethode. Ein Modell, welches die Exaktheit des Fisher Tests anwendet bei einer Kontingenztafel, ist der Fisher-Boschloo-Test [15].

2.3.5 Fisher-Boschloo Test

Der Fisher-Boschlootest für ein binomiales Modell geht, anders als der exakte Test nach Fisher, nicht von festen Randsummen aus. Er projiziert Fishers exakten Test auf ein Modell mit festen Reihen- oder Spaltensummen, wodurch er besser geeignet bzw. designet ist für den unconditional Fall. Er beschreibt demnach ein binomiales Modell in dem der Test von den Erfolgswahrscheinlichkeiten der Behandlungsgruppen abhängen.

Der Test nutzt den *p-Wert* des exakten Tests nach Fisher als Ordnungsfunktion. Dabei verwendet er ein erhöhtes α-Level, wodurch gleichzeitig die Power erhöht wird, da die Wahrscheinlichkeit geringer ist, die Nullhypothese Abzulehnen. Das er-

höhte Alpha Niveau α^* wird anhand gegebener Tabellen definiert durch n_A, n_B und α. Dieser Wert stellt den Prüfwert bzw. kritischen Wert und wird mit dem Teststatistikwert vom Fishers exakten Test verglichen. Der Teststatistikwert, von Fishers exakten Test, ist sein p-Wert.

Der Test wurde von R.D. Boschloo im Paper "Raised conditional level of significance for the 2×2-table when testing the equality of two probabilities"dargestellt, und kann sowohl für ein binomiales als auch für ein multinormales Modell (zufällige Randsummen in der Kontingenztafel) angewendet werden.

2.4 Randomisierung als Basis für Inferenz

Randomisierungsbasierte Inferenzstatistik stellt Verfahren zur Berechnung von p-Werten, genau wie die likelihoodbasierte Inferenzstatistik. Die Basis für die Berechnung der p-Werte ist in Randomisierungs- und Permutationstests gegeben durch die Wahrscheinlichkeiten der möglichen Permutationen und die Annahme der Austauschbarkeit der Zufallsstichprobe. Austauschbarkeit ist dann vorhanden, wenn die Beobachtungen einer Stichprobe unabhängig sind und die gleiche Verteilung haben oder wenn die Stichprobe aus einer endlichen Grundgesamtheit stammt [6]. Die genaue Berechnung hängt von der Verteilung der Zufallsvariablen X ab. Dabei kann die Teststatistik bei normalverteilten Daten beispielsweise über die Differenzen und die von bernoulliverteilten Zufallsvariablen über den χ^2-, Fischer- oder Boschlootest [1] berechnet werden.

Dieser Abschnitt erläutert die Unterschiede und die Berechnung der beiden Tests und zeigt wie auch bei hohen Fallzahlen die Methoden mit Hilfe von Monte Carlo Prozeduren angewandt werden können.

2.4.1 Randomisierungstest vs. Permutationstest

Randomisierungstests und Permutationstests werden in der Literatur häufig als Synonym verwendet [1]. Bei der Sichtung der Literatur ist dargestellt, dass der Unterschied zwischen Randomisierungstest und Permutationstest in den Wahrscheinlichkeiten der Permutationen liegt [1]. Dabei handelt es sich um alle möglichen Einteilungen in die Behandlungsgruppen mit ihren einzelnen Wahrscheinlichkeiten. Bei Permutationsverfahren haben alle Permutationen dieselbe Wahrscheinlichkeit und folgen dementsprechend derselben Likelihood. In Randomisationssequenzen ist die

Wahrscheinlichkeit einer Permutation abhängig von der Randomisationsmethode. Dabei können die Permutationen unterschiedliche Wahrscheinlichkeiten haben bei festem n_A, n_B bzw n. Somit folgen diese nicht derselben Likelihood. Daraus folgt, dass die Daten *(X1,..,Xn)*, in den unterschiedlichen Reihenfolgen durch die verschiedenen Permutationen, auch unterschiedlichen Likelihoods folgen und dementsprechend die Austauschbarkeit der Daten nicht gegeben ist. Aus diesem Grund kann es wichtig sein die Wahrscheinlichkeiten von Permutationen mittels Randomisierungsstest mit einzubeziehen [1].

2.4.2 Testprinzipien

Beide Testarten sind sogenannte „Resampling“ Verfahren und können komplett oder mit Hilfe von Monte Carlo-Prozeduren berechnet werden [6]. Das bedeutet, dass zu einer festen Stichprobe mit dem dazugehörigen Outcome, die Gruppenzuweisung immer wieder geändert wird bzw. die Gruppenzuweisung per Randomisierung oder Permutation M-Mal neu generiert wird [1]. Der Outcome sind die Ergebnisse aller Patienten einer Stichprobe. Dabei wird zu jeder neu generierten Gruppenzuweisung(Permutation) R die Wahrscheinlichkeit p_R des Auftretens, definiert durch Randomisierungs- oder Permutationsmethode, berechnet. Die Teststatistik (T_R) wird dann für jede Permutation mit dem festen Outcome berechnet. Den Vergleichswert für die Testentscheidung liefert die Teststatisik (T_0) des festen Outcomes mit ihrer originalen Gruppenzuweisung der klinischen Studie.

Bei einem **kompletten Test** ist M immer die Anzahl aller möglichen Permutationen, nach dem conditional oder uncontitional Fall. Dabei wird zu jeder Permutation mit dem festen Outcome die Teststatistik berechnet. Der p-Wert ist dann die Summe aller Wahrscheinlichkeiten jener Permutationen, die eine Teststatistik hervorbringen, welche größer sind als T_0. Mathematisch lässt sich diese Berechnung wie folgt darstellen:

$$p = \sum_{i=1}^{n=A_x} p_{xi} * d$$

Dabei ist $d \in [0,1]$ wobei 0 bedeutet, dass $T_R \leq T_0$ und 1 bedeutet $T_R > T_0$. Der Wert p_{xi} ist die Wahrscheinlichkeit einer einzelnen Permutation mit $x \in$ (c,u). Die Anzahl der Möglichkeiten A_x für die Permutationen ergeben sich ausfolgenden Gleichungen für den conditional Fall:

$$A_c = \binom{n_A + n_B}{n_A}$$

Und folgende Berechnung für den Unconditional Fall:

$$A_u = 2^{n_A + n_B}$$

Der Zähler i nimmt dementsprechend die Werte *1,...,Ax* an. Daraus ergeben sich beispielhaft folgende Permutationen bei einer Fallzahl *n=4* (Abbildung 2.6) und dem conditional Setting.

Tabelle 2.6: Permutationen bei n=4, conditional Fall

x
A A A A
A A A B
A A B A
A A B B
A B A A
A B A B
A B B A
A B B B
B A A A
B A A B
B A B A
B A B B
B B A A
B B A B
B B B A
B B B B

Beim Einsetzen von möglichen Fallzahlen in die Gleichungen für Ac und Au, welche in klinischen Studien üblich sind, wird deutlich, dass die Anzahl der möglichen Permutationen in einem kompletten Test, sich sehr schnell erhöht und demnach viel

Rechenleistung benötigt. Die Tabelle 2.7 zeigt den starken Anstieg der Anzahlen der möglichen Permutationen beispielhaft.

Tabelle 2.7: Anzahlen möglicher Permuationen

n	Au	Ac
10	1.024000e+03	2.520000e+02
50	1.125900e+15	1.264106e+14
100	1.267651e+30	1.008913e+29

Für klinische Studien sind dies jedoch sehr kleine Fallzahlen, welche aber bereits zeigen, dass für die Berechnung aller Wahrscheinlichkeiten viel Rechenleistung notwendig wird. Um diese Randomisierungs- und Permutationstests dennoch rechentechnisch effizient nutzen zu können, kann ein Monte-Carlo Verfahren angewandt werden.

Bei einem **Monte-Carlo-Verfahren** wird M festgelegt, da eine komplette Berechnung auf Grund des hohen Ax nicht möglich wäre, oder nur mit sehr viel Rechenleistung. Die Festlegung von M liegt im Bereich zwischen 10000 und 15000. So viele Gruppensequenzen werden zufällig generiert und dazu die Teststatistik berechnet. Die Auswertung bzw. Zusammenfassung zu einem p-Wert ist in zwei Arten möglich. Einmal als Anteil der M generierten Permutationen (Proportions-Prinzip) oder mit Hilfe der Summe der$M$-Sequenzen (Summen-Prinzip).
Das **Proportions-Prinzip** zählt, bei wie vielen Permutationen eine Teststatistik T_R größer als T_0 entstanden ist. Diese Anzahl a wird dann durch M geteilt. Der p-Wert ergibt sich demnach aus

$$p = a/M$$

Dieses Prinzip ist die klassische Monte-Carlo Prozedur. Die Gewichtung durch die Wahrscheinlichkeiten der Permutationen sind in diesem Prinzip durch das häufigere Auftreten höherwahrscheinlicher Sequenzen gegeben. Nach Rosenberger et. al. (2018) wird Anhand dieser Methode ein konsistenter Schätzer des p-Wertes berechnet [1].
Das **Summenprinzip** bezieht die genauen Wahrscheinlichkeiten der Permutatio-

nen mit ein und projiziert sie auf die Summe der Wahrscheinlichkeiten aller M-Permutationen. Die folgende Formel verdeutlicht die Berechnung des p-Wertes:

$$p = \frac{\sum_{i=1}^{n=m} p_{xi} * d}{\sum_{i=1}^{n=m} p_{xi}}$$

Die Berechnung der Wahrscheinlichkeiten des conditional Falls sind bei einigen Randomisierungsmethoden für das Monte Carlo-Prinzip nicht einfach zu realisieren und müssen theoretisch berechnet werden. Ein Beispiel dafür ist die Efrons biased Coin Methode für ausgeglichene Gruppengrößen. Dabei werden im kompletten Randomisierungsprinzip die Wahrscheinlichkeiten der conditinal Permutationen durch die Summe aller conditional Permutationen geteilt. Durch dieses Verfahren erhält man die Wahrscheinlichkeiten der Sequenzen für den conditional Fall. In großen Fallzahlen müssten für diese Methode immer alle Permutationswahrscheinlichkeiten berechnet werden, was die Monte Carlo Prozedur eigentlich unterbinden soll. Für diesen Fall ist die Formel der theoretischen Berechnung gegeben durch [1] :

$$\varnothing(A_j) \begin{cases} P(A_j) * \frac{P(N_A(n)=n_A|N_A(j)=m_{j-1}+1)}{P(N_A(n)=n_A|N_A(j-1)=m_{j-1})} & \text{für } 1 < j \leq n \\ \frac{P(N_A(n)=n_A|N_A(1)=1)}{2P(N_A(n)=n_A)} & \text{für } j = 1 \end{cases}$$

2.4.3 Teststatistik bei bernoulliverteilten ZufallsVariablen

Bei der Berechnung der Teststatistik einer bernoulliverteilten Zufallsvariablen müssen für jede genutzte Permutation mit der festen Outcomesequenz zuerst die 2x2-Feldertafeln generiert werden. Anhand der 2x2-Feldertafel können die Teststatistiken berechnet werden, beispielsweise über den χ^2-, Fisher-, oder Boschloo-Test. Besonders wichtig ist, dass bei der Berechnung des p-Wertes dieselbe Methode für die Berechnung von T_0 und die Berechnung aller T_R konstant genutzt wird.

2.5 Simulationsstudie

2.5.1 Definition einer Simulationsstudie

Eine Simulationsstudie ist eine Studie mit einer konkreten Fragestellung zu der die Daten nach bestimmten Parametern selbst generiert werden. Das heißt verschiedene Outcomes der Studie werden definiert und die Daten dazu mit Zufallsmethoden

erstellt. Im Gegensatz zu richtigen klinischen Studien ist die Realität durch die erstellten Daten bekannt. Durch das Wissen über die „Wahrheit“ lassen sich gezielt Methoden und Tests auf richtige Ergebnisse prüfen und vergleichen. Um im Durschnitt das eingestellte Ergebnis zu erhalten wird die Studie m-Mal durchgeführt. Es werden dementsprechend m Stichproben generiert die im Durchschnitt der eingestellten Wahrheit entsprechen.
Im Falle der nachfolgend beschriebenen Simulationsstudie werden die Hypothesentests auf ihre Ausschöpfung des α- und β-Niveaus geprüft. Dabei wird gezeigt, welches Testverfahren die genausten Ergebnisse liefert hinsichtlich des Fehlers 1. Und 2. Art. Die Datenerzeugung und Berechnung der statistischen Tests wurden über einen Cluster (Großrechner) realisiert.

2.5.2 Setting

Diese Simulationsstudie beschreibt eine klinische Studie in der zwei Behandlungsgruppen, Gruppe A und Gruppe B, bezüglich ihres Erfolges der Behandlung, miteinander verglichen werden. Die zwei Variablen „Gruppe“ und „Erfolg“ sind dichotome Merkmale und werden auf Unabhängigkeit getestet. Im Fokus steht demnach die Frage, ob die Behandlungsgruppe vom Erfolg unabhängig ist. Dazu werden die zwei Variablen folgendermaßen definiert:

- Behandlungsgruppe: $X \sim Ber(p)$, für X $\in$ A,B
- Outcome: $Y_i \sim Ber(p_i)$, für Y $\in$ 0,1, i $\in$ A,B

Die Variablen lassen sich in einer Kontingenztafel darstellen. Da die Gruppen unabhängig voneinander sind und somit unabhängige Outcomes haben, folgt jede Gruppe einer Erfolgswahrscheinlichkeit p_i. mit i $\in$ A,B, welche anteilig auf die Gruppen je nach Setting verteilt wird. Dabei bildet jede Gruppe eine Zufallsvariable Y_i, wobei

$$p_i = P(Y_i = 1)$$

die Erfolgswahrscheinlichkeit definiert.

Für die Simulation der Daten nach Hypothesen müssen die Ergebnisse für Y_A und Y_B so eingestellt werden, das

- H_0: $p_A = p_B$

- H_1: $p_A \neq p_B$

Für die Abbildung der Hypothesen in einer Kontingenztafel werden sogenannte Odds Ratios (OR) verwendet. Das OR gibt die Stärke eines Zusammenhangs zwischen den Behandlungsgruppen an. Ein OR von 1 stellt den Fall H_0 dar. Ein OR $\neq 0$ stellt H_1 dar. Je höher das OR, desto größer ist der Unterschied und desto leichter ist es, diesen zu zeigen. Das OR ist in einer Kontingenztafel mit Anzahlen folgendermaßen definiert:

$$OR = \frac{a * d}{b * c}$$

2.5.2.1 Simulationsarameter

Die Datensätze der Simulationsstudie werden durch die Parameter n der Stichprobengröße (mit den Gruppengrößen n_A und n_B), p der Erfolgswahrscheinlichkeit, OR dem Odds-Ratio und m die Anzahl der simulierten Studien bestimmt. Aus diesen Angaben werden verschiedene Datensätze generiert, die entweder H_0 darstellen oder H_1. Folgende Einstellungen wurden berechnet:

- n $\in$ {40,100,200}
- p $\in$ {0.05, 0.1}
- OR $\in$ {1,3,5}
- m = 10.000

Zur Randomisierung der$n$ Patienten wurden die Methoden Efrons biased Coin Design und Permuted Block Randomisierung mit fester Blocklänge von 4 verwendet. Dadurch entstehen pro Einstellung, definiert durch die Randomisierungsmethode,n, p und OR, ein Datensatz mit 10.000 Stichproben/Experimenten. Für die gesamte Studie wurden 36 Datensätze generiert. Alle Daten Sätze mit einem OR von 1 stellen H_0 dar und prüfen den Fehler 1. Art. Die Daten Sätze mit einem OR>1 prüfen die Power (1-β). Die EBC Methode stellt dabei den Unconditional und die PBR Methode den conditional Fall dar. Außerdem besitzen im Fall EBC die Randomisierungssequenzen alle unterschiedliche Wahrscheinlichkeiten und im Fall PBR sind alle Sequenzen gleichwahrscheinlich.

2.5.2.2 Berechnung der Erfolgswahrscheinlichkeiten über das OR

Die Aufteilung der Erfolgswahrscheinlichkeit p anhand der Odds-Ratios, auf die Gruppenerfolgswahrscheinlichkeiten p_A und p_B, um H_0 oder H_1 zu erzeugen, erfolgt über die Kontingenztafel mit relativen Häufigkeiten (Tabelle 2.8):

Tabelle 2.8: Kontingenztafel relative Häufigkeiten

	Erfolg	**Misserfolg**	**Summe**
Gruppe A	h_a	h_b	h_{n_A}
Gruppe B	h_c	h_d	h_{n_B}
Summe	p	1-p	1

Daraus ergeben sich die Erfolgswahrscheinlichkeiten:

$$p_A = \frac{h_a}{h_{n_A}},\ bei\ 0 < h_a < p$$
$$p_B = \frac{p - h_a}{h_{n_B}},\ bei\ 0 < h_a < p$$

Die Lage der Daten nach dem OR ist demnach abhängig von a. Die Berechnung von a wird über die Formel des OR und der Erfolgswahrscheinlichkeit p hergeleitet. Dabei entsteht eine quadratische Funktion, welche zwei Werte für a berechnet.

$$a = min(a_{1,2})$$

mit

$$a_{1,2} = -\frac{s}{2} \pm \sqrt{\frac{s^2}{2} - q}$$

Der im Betrag größere Wert für a kann ausgeschlossen werden, da dieser größer als die Erfolgswahrscheinlichkeit wird und somit nicht im möglichen Bereich für a liegt. Die Variablen s und q müssen nach der Erfolgswahrscheinlichkeit p folgendermaßen berechnet werden (Tabelle 2.5.2.2):

Tabelle 2.9: Formeln zur Berechnung von q und s

	p=0.05	**p=0.1**
s	$\frac{-0.55*OR-0.45}{(OR-1)}$	$\frac{-0.6*OR-0.4}{(OR-1)}$
q	$\frac{2.5*OR}{(OR-1)}$	$\frac{5*OR}{(OR-1)}$

2.5.2.3 Clusterfunktionen

Pro Randomisierungsmethode gibt es ein Cluster Skript in der Programmiersprache R. Ein Skript beinhaltet zwei Funktionen, die nacheinander vom Cluster ausgeführt werden. Die erste Funktion generiert je eine Stichprobe und berechnet im Anschluss mit der zweiten Methode die p-Werte der Hypothesentests. Die folgenden Tests wurden an den Daten durchgeführt:

- Likelihood-Verfahren: χ^2-Test,Fischers-Test, Fisher-Boschloo-Test
- Monte Carlo Standard: χ^2-Test,Fischers-Test, Fisher-Boschloo-Test
- Monte Carlo Summenmethode: χ^2-Test,Fischers-Test, Fisher-Boschloo-Test

Alle Hypothesentests wurden pro Datensatz berechnet. Ein Datensatz beinhaltet 10.000 Stichproben eines bestimmten Settings. Zu jeder der 10.000 Stichproben sind die neun p-Werte der oben beschriebenen Tests berechnet, damit ist die Vergleichbarkeit zwischen den Tests gewährleistet. Außerdem wurden die Teststatistiken der Likelihood-Verfahren als T_0 für die Randomisierungstests verwendet. T_0 ist in der Studie der p-Wert. Für die Korrektheit und Vergleichbarkeit hat das keine Relevanz da der p-Wert aus der Test Statistik berechnet wird.
Die Monte-Carlo Summenmethode wird im Zuge dieser Simulationsstudie erstellt, da sie das Standardverfahren von Monte-Carlo mit den exakten Wahrscheinlichkeiten der Randomisierungssequenz verwendet. Sie stellt somit eine Mischung aus der Monte-Carlo oder kompletten Randomisierung dar.

2.5.2.4 Clusterergebnisse

Die Ergebnisdateien der Clusterberechnung werden im R-Skript „Cluster-Auswertung.R“ ausgewertet und zum Alpha und Betafehler zusammengefasst. Dabei werden die 10.000 p-Werte pro Test und Setting auf Signifikanz getestet und der Fehler erster oder zweiter Art als Anteil der 10.000 ausgegeben. Im Fall OR= 1 umfasst der Alphafehler prozentual alle Stichproben, die ein signifikantes Ergebnis hervorbringen,

obwohl kein Unterschied zwischen den Gruppen vorliegt. Bei der Simulation von H_1 bzw. OR $\in \{3,5\}$ wird die Power, durch den Anteil der Stichproben, die ein richtig signifikantes Ergebnis geliefert haben bestimmt. Dabei ist die Power beschrieben durch 1-β. Das R-Skript befindet sich auszugsweise im Anhang.

2.5.3 Bezug zu klinischen Studien

In der Praxis stellt diese Studie eine klinische Studie der Phase III, nach der FDA („Food and Drug“ Administration), dar. Sie wird durchgeführt, um einen signifikanten Wirksamkeitsnachweis neuer Medikamente zu zeigen. Dabei dienen diese Studien der Markzulassung der Medikamente. Für die WHO („World Health Organisation“) steht der Alphafehler, für die Sponsoren der Studie die Power im Vordergrund. Deshalb sind für diese Art klinischer Studien Patientenanzahlen im Bereich von 200 bis 10.000 Patienten, je nach OR und Erfolgswahrscheinlichkeit, üblich. Um den Fehler erste und zweiter Art bestmöglich auszuschöpfen sollte dieser Bereich nicht unterschritten werden.
Die Simulationsstudie stellt demnach sehr kleine Fallzahlen bis hin zum unteren Bereich klinischer Studien dar. Grund dafür, sind die Berechnungszeiten bei hohen Fallzahlen.

3 Ergebnisse der Simulationsstudie

Der folgende Abschnitt beschreibt deskriptiv die aus den Ergebnistabellen gewonnenen Daten. Damit legt dieser Abschnitt die Grundlage für die Interpretation und das Fazit der Arbeit.

Dabei stellen die hier gezeigten Tabellen die Ausschöpfung des α-Niveaus und der Power (1-β) dar. Der Test ist dementsprechend der beste, welcher beim α-Niveau der 5 % am dichtesten kommt und beim den geringsten Betafehler (größte Power) erzeugt. Ziel der Auswertung ist der Vergleich der Hypothesentests bezogen auf den Fehler erster Art und der Power. Dabei werden die Tests bei verschiedenen Einstellparametern miteinander verglichen und beschrieben. Für jede der folgenden Einstellungen wird veranschaulicht, welches der allgemein beste Test und welches der beste Test einer Gruppe ist. Die Gruppen sind definiert durch die Testverfahren: Likelihood-Verfahren, Monte-Carlo Prozedur und Monte-Carlo Summenverfahren. Der beste Test einer gesamten Einstellung (Tabellenzeile) ist rosa hinterlegt. Der beste Test einer Gruppe trägt die Schriftfarbe dunkelrot. Die Werte der Tabelle sind in den darauffolgenden Grafiken anschaulich dargestellt.

Für die deskriptive Beschreibung wurden die Fallzahlen und die Erfolgswahrscheinlichkeiten Kategorisiert. Eine n von 40 beschreibt eine kleine, ein n von100 eine mittlere und ein n von 200 eine große Fallzahl. Die untere Erfolgswahrscheinlichkeit liegt bei 0.05% und die obere bei 0.1%.

3.1 p-Werte der Datensätze

Eine Besonderheit die in die den Daten aufgetreten ist, ist das ein paar Stichproben für bestimmte Tests keinen p-Wert besitzen, aber dennoch in den Anteil der 10.000 Stichproben für die Fehlerarten mit einfließen müssen. Dafür wurde anhand der 2x2 Feldertafeln geprüft, wann der p-Wert das Ergebnis „NaN“ (Not a Number) ausgibt. Das Problem entsteht vorwiegend bei und vor allen durch die kleinen Fallzahlen, niedrigen Erfolgswahrscheinlichkeiten und dem Chiquadrat-Test. Ist der

p-Wert einer Stichprobe nicht berechenbar, gab es in beiden Behandlungsgruppen keinen Erfolg. Das heißt die Spaltensumme n_1 ist 0. Im Fall des Chiquadrat-Tests bedeutet das bei der Teststatistik, das durch 0 geteilt wird, was mathematisch nicht möglich ist. Da bei diesem Szenario in beiden Gruppen kein Erfolg vorhanden war ist dementsprechend die Erfolgswahrscheinlichkeiten bei beiden gleich. Somit kann dieser Fall den nicht signifikanten Stichproben zugeordnet werden.

3.2 Efrons Biased Coin Design

3.2.1 Fehler 1. Art

Tabelle 3.1: Alphafehler EBC bei n=40

		Likelihood			Monte-Carlo			Monte-Carlo-Summe		
OR	p	Chi	Fisher	Boschloo	MCChi	MCFisher	MCBoschloo	MCSChi	MCSFisher	MCSBoschloo
1	0.05	0.0006	0.0030	0.0034	0.1496	0.0207	0.0200	0.1655	0.0435	0.0363
1	0.10	0.0067	0.0157	0.0196	0.0416	0.03	0.0294	0.0482	0.041	0.0361

Tabelle 3.2: Alphafehler EBC bei n=100

		Likelihood			Monte-Carlo			Monte-Carlo-Summe		
OR	p	Chi	Fisher	Boschloo	MCChi	MCFisher	MCBoschloo	MCSChi	MCSFisher	MCSBoschloo
1	0.05	0.0076	0.0112	0.0273	0.0386	0.0365	0.0319	0.0481	0.0424	0.0417
1	0.10	0.0178	0.0230	0.0374	0.0379	0.0399	0.0370	0.0409	0.0415	0.0413

Tabelle 3.3: Alphafehler EBC bei n=200

		Likelihood			Monte-Carlo			Monte-Carlo-Summe		
OR	p	Chi	Fisher	Boschloo	MCChi	MCFisher	MCBoschloo	MCSChi	MCSFisher	MCSBoschloo
1	0.05	0.0175	0.0239	0.0297	0.0381	0.0386	0.0382	0.0428	0.0445	0.0427
1	0.10	0.0271	0.0333	0.0411	0.0426	0.0437	0.0422	0.0470	0.0482	0.0471

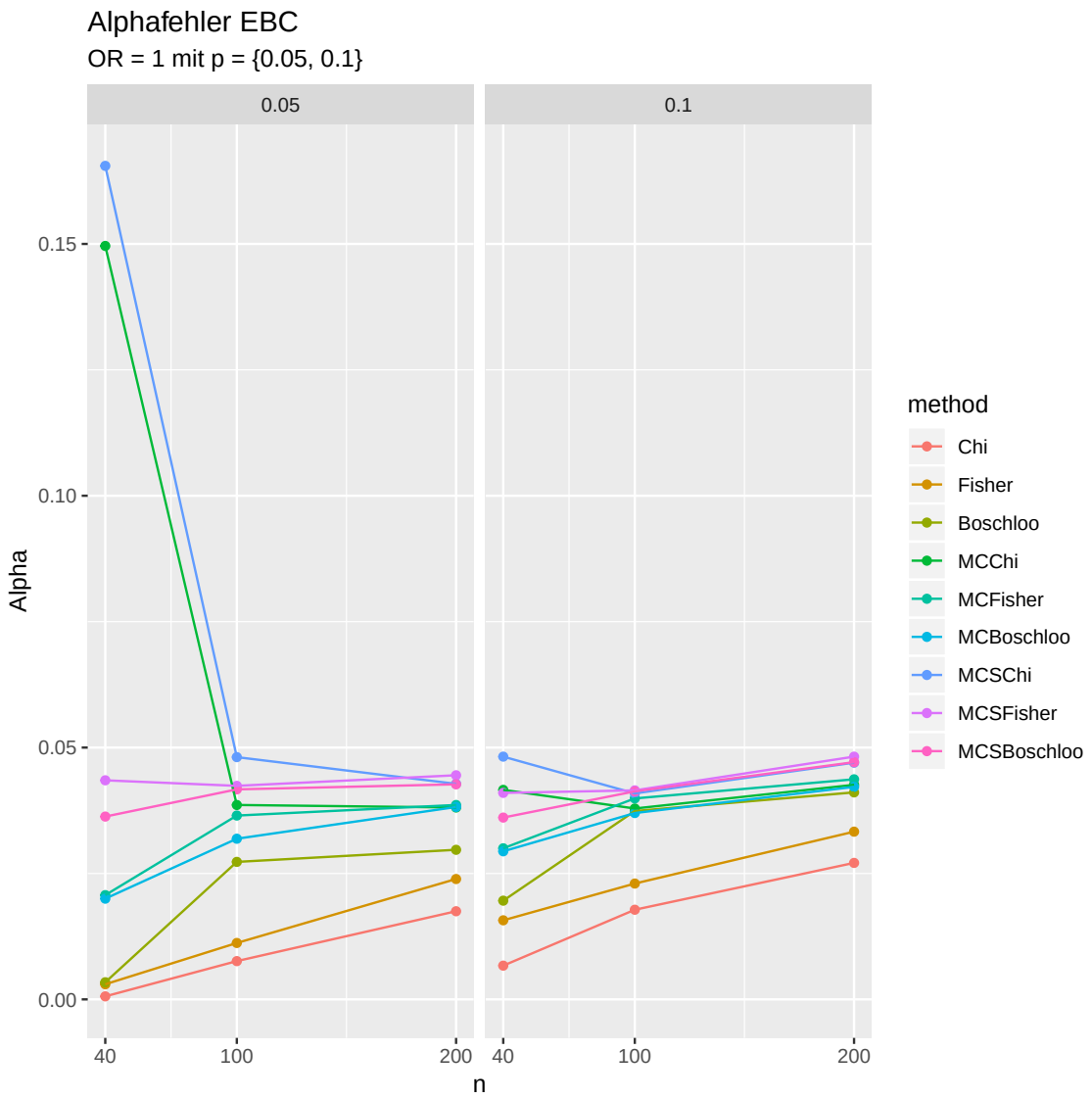

Abbildung 3.1: Alphafehler EBC

Die Tabellen 3.1 bis 3.3 zeigen die Ergebnisse des Alphaniveaus nach Fallzahl für die Efrons Biased Coin Methode zu den zwei verschiedenen Erfolgswahrscheinlichkeiten. Die Abbildung 3.1 beruht auf den Daten der Tabellen 3.1 bis 3.3. Beim ganzheitlichen Vergleich aller Methoden ist zu erkennen, dass die Methoden MCChi und MCSChi das Alphaniveau bei einer Fallzahl von 40 und einer Erfolgswahrscheinlichkeit von 0,05 nicht einhalten. Ab einer Erfolgswahrscheinlichkeit von 0,1 oder einer Fallzahl von 100, halten alle Tests das Alphaniveau ein. Mit steigender Erfolgswahrscheinlichkeit, können alle Tests das Niveau besser ausschöpfen.

Die likelihodbasierten Tests schneiden in allen Einstellungen am schlechtesten ab. Vor allem bei kleiner Fallzahl läuft das Alphaniveau gegen 0. Eine steigende Fallzahl lässt diese Tests genauer werden. Deutlich erkennbar ist, dass von den likelihoodbasierten Tests der Fisher-Boschloo-Test in allen Einstellungen am besten abschneidet und der Chiquadrat-Test das Niveau am wenigsten ausschöpfen kann.
Bei den Randomisierungstests ist erkennbar, dass mit steigender Fallzahl die Streuung des Alpha-Niveaus sinkt. Das heißt, der Unterschied zwischen den Ergebnissen dieser Tests wird sehr gering. In kleinen Fallzahlen und der unteren Erfolgswahrscheinlichkeit schneiden MCSFisher und MCSBoschloo am besten ab. Bei einer Kleinen Fallzahl und der oberen Erfolgswahrscheinlichkeit schneidet der Test MCSChi wiederrum am besten ab. Der Vergleich der Verfahren Monte Carlo und Monte Carlo Summe zeigt einen leichten Vorteil der Summenverfahren, welcher bei steigender Fallzahl und Erfolgswahrscheinlichkeit aber marginaler wird.

3.2.2 Fehler 2. Art

Tabelle 3.4: Power EBC bei n=40

		Likelihood			Monte-Carlo			Monte-Carlo-Summe		
OR	p	Chi	Fisher	Boschloo	MCChi	MCFisher	MCBoschloo	MCSChi	MCSFisher	MCSBoschloo
3	0.05	0.0030	0.0124	0.0135	0.1694	0.0398	0.0391	0.1904	0.0621	0.0613
3	0.10	0.0321	0.0701	0.0815	0.1086	0.0969	0.0936	0.1182	0.1052	0.1029
5	0.05	0.0055	0.0224	0.024	0.1806	0.0535	0.0527	0.2028	0.0753	0.0762
5	0.10	0.0635	0.1313	0.1494	0.1797	0.1702	0.1654	0.1884	0.1769	0.1742

Tabelle 3.5: Power EBC bei n=100

		Likelihood			Monte-Carlo			Monte-Carlo-Summe		
OR	p	Chi	Fisher	Boschloo	MCChi	MCFisher	MCBoschloo	MCSChi	MCSFisher	MCSBoschloo
3	0.05	0.0554	0.0719	0.1305	0.136	0.1299	0.1308	0.1432	0.1363	0.1375
3	0.10	0.2124	0.2368	0.2976	0.2919	0.2944	0.2901	0.2889	0.2897	0.2893
5	0.05	0.1123	0.1388	0.2376	0.2347	0.2279	0.2297	0.2413	0.2343	0.2370
5	0.10	0.4125	0.4378	0.5219	0.4982	0.5019	0.4953	0.491	0.492	0.4917

Tabelle 3.6: Power EBC bei n=200

		Likelihood			Monte-Carlo			Monte-Carlo-Summe		
OR	p	Chi	Fisher	Boschloo	MCChi	MCFisher	MCBoschloo	MCSChi	MCSFisher	MCSBoschloo
3	0.05	0.2116	0.2418	0.27	0.2905	0.2918	0.29	0.2893	0.292	0.2887
3	0.10	0.4789	0.5011	0.5292	0.5246	0.5271	0.5243	0.5026	0.5038	0.5031
5	0.05	0.419	0.4474	0.4828	0.4972	0.498	0.4965	0.4750	0.4776	0.4748
5	0.10	0.7271	0.7017	0.6977	0.7395	0.7393	0.7397	0.6667	0.6678	0.6664

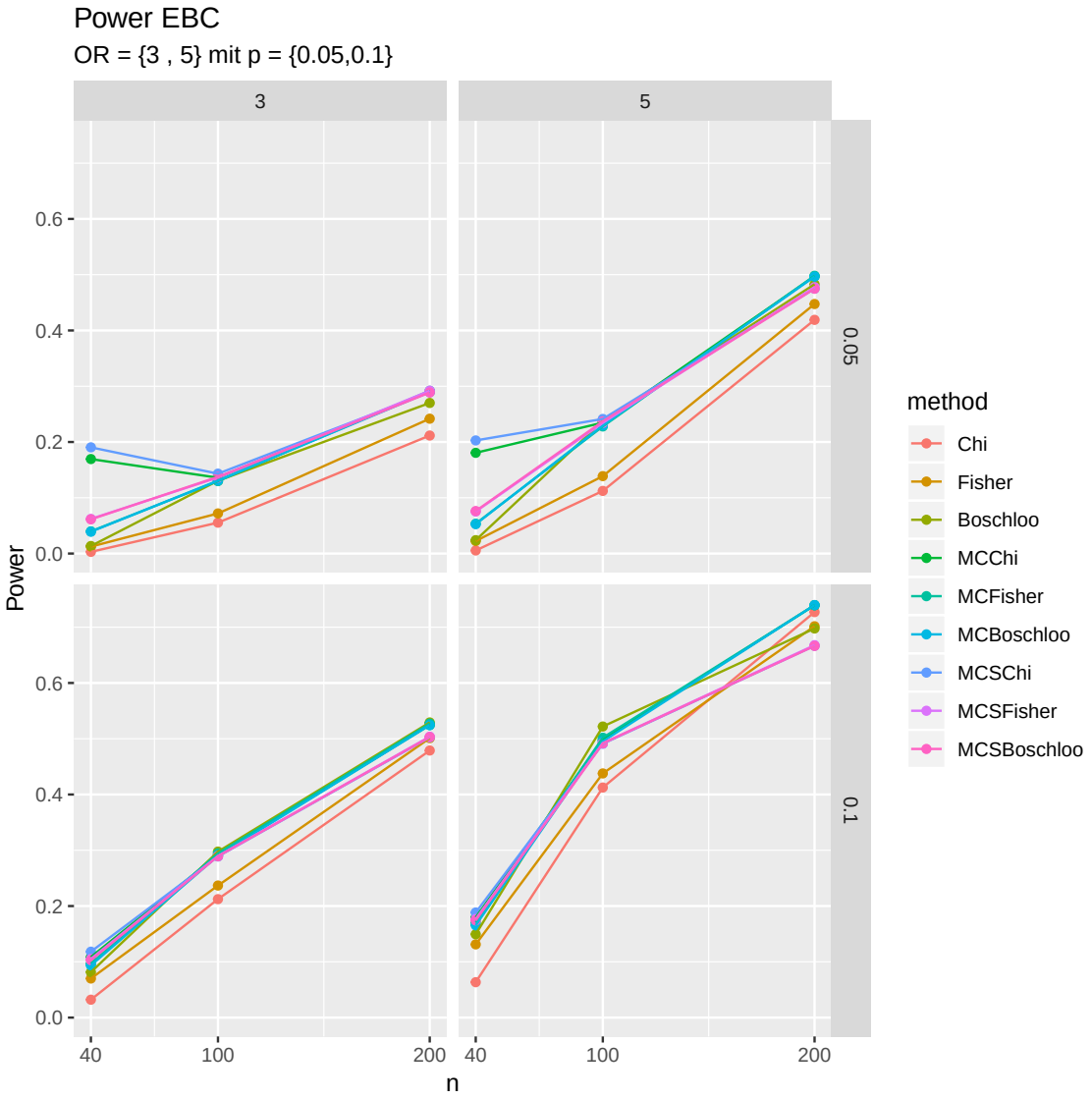

Abbildung 3.2: Power EBC

Die Tabellen 3.4 bis 3.6 Zeigen die Ergebnisse 1-Betaniveaus, also der Power, bei Odds Ratios von 3 und 5, zu den eingestellten Fallzahlen und Erfolgswahrscheinlichkeiten. Die Abbildung 3.2 basiert auf den Daten der Tabellen 3.4 bis 3.6 und dient der Veranschaulichung der Ergebnisse.

Im allgemeinen Vergleich ist deutlich erkennbar das jeder Test bei unterer Erfolgswahrscheinlichkeit, allen Fallzahlen und allen Odds Ratios stark unterpowert. Dabei erreicht selbst bei einer hohen Fallzahl und einem OR von 5 kein Test eine Power über 50%. Mit steigender Erfolgswahrscheinlichkeit steigt auch die Power der Fallzahlen. Die Höchste Power bietet die Methode MCBoschloo mit 0,7397 und der höchsten Fallzahl einem OR von 5. Damit liegt keines der Ergebnisse in einem akzeptablen Bereich (>80%) für klinische Studien. Mit steigender Fallzahl lässt die Streung der Power zwischen den Tests nach. Sie werden somit Präziser und die Unterschiede zwischen den Tests geringer. Die Steigerung des Odds Ratios ist lediglich in mittleren bis hohen Fallzahlen in den Methoden deutlich zu erkennen.

Bei den Standardhypothesentests ist klar zu erkennen, dass diese sehr stark unterpowert sind bei einer Kleinen Fallzahl. Dort übersteigen sie nicht einmal 15%. Ein kleiner Vorteil ist mit der Boschloo Methode erzielbar. Dieser Test schneidet am besten ab in seiner Gruppe. Alle likelihoodbasierten Tests schneiden schlechter oder genauso gut ab wie die randomisierten Tests. Der Chiquadrattest ist deutlich der schlechteste Test.

Die Randomisierungstests holen eine leicht höhere Power in kleinen Fallzahlen und der unteren Erfolgswahrscheinlichkeit raus. Hier ist zu erkennen, dass die Methoden MCChi und MCSChi von den anderen Randomisierungsmethoden abheben. Bei steigender Fallzahl und Erfolgswahrscheinlichkeit relativiert sich dieser Unterschied. Die Methoden MCFisher und MCBoschloo bzw. auch MCSFischer und MCSBoschloo weisen keinen Unterschied zwischen einander auf. Aus diesem Grund sind die Fisher Methoden in der Grafik nicht erkennbar. Die beiden Grafen der Boschloomethoden können als Grafen für die Fisher Methoden angenommen werden.

3.3 Permuted Block Randomisierung

3.3.1 Fehler 1. Art

Tabelle 3.7: Alphafehler PBR bei n=40

		Likelihood			Monte-Carlo			Monte-Carlo-Summe		
OR	p	Chi	Fisher	Boschloo	MCChi	MCFisher	MCBoschloo	MCSChi	MCSFisher	MCSBoschloo
1	0.05	0.0006	0.0030	0.0034	0.1499	0.021	0.0203	0.1646	0.0432	0.0354
1	0.10	0.0067	0.0157	0.0196	0.0411	0.0302	0.0290	0.0486	0.0411	0.0366

Tabelle 3.8: Alphafehler PBR bei n=100

		Likelihood			Monte-Carlo			Monte-Carlo-Summe		
OR	p	Chi	Fisher	Boschloo	MCChi	MCFisher	MCBoschloo	MCSChi	MCSFisher	MCSBoschloo
1	0.05	0.0076	0.0076	0.0204	0.016	0.0095	0.0095	0.016	0.0095	0.0095
1	0.10	0.0172	0.0171	0.0301	0.0201	0.0201	0.0201	0.0201	0.0201	0.0201

Tabelle 3.9: Alphafehler PBR bei n=200

		Likelihood			Monte-Carlo			Monte-Carlo-Summe		
OR	p	Chi	Fisher	Boschloo	MCChi	MCFisher	MCBoschloo	MCSChi	MCSFisher	MCSBoschloo
1	0.05	0.0201	0.0222	0.0301	0.0218	0.0218	0.0218	0.0218	0.0218	0.0218
1	0.10	0.0265	0.0297	0.0384	0.0296	0.0296	0.0296	0.0296	0.0296	0.0296

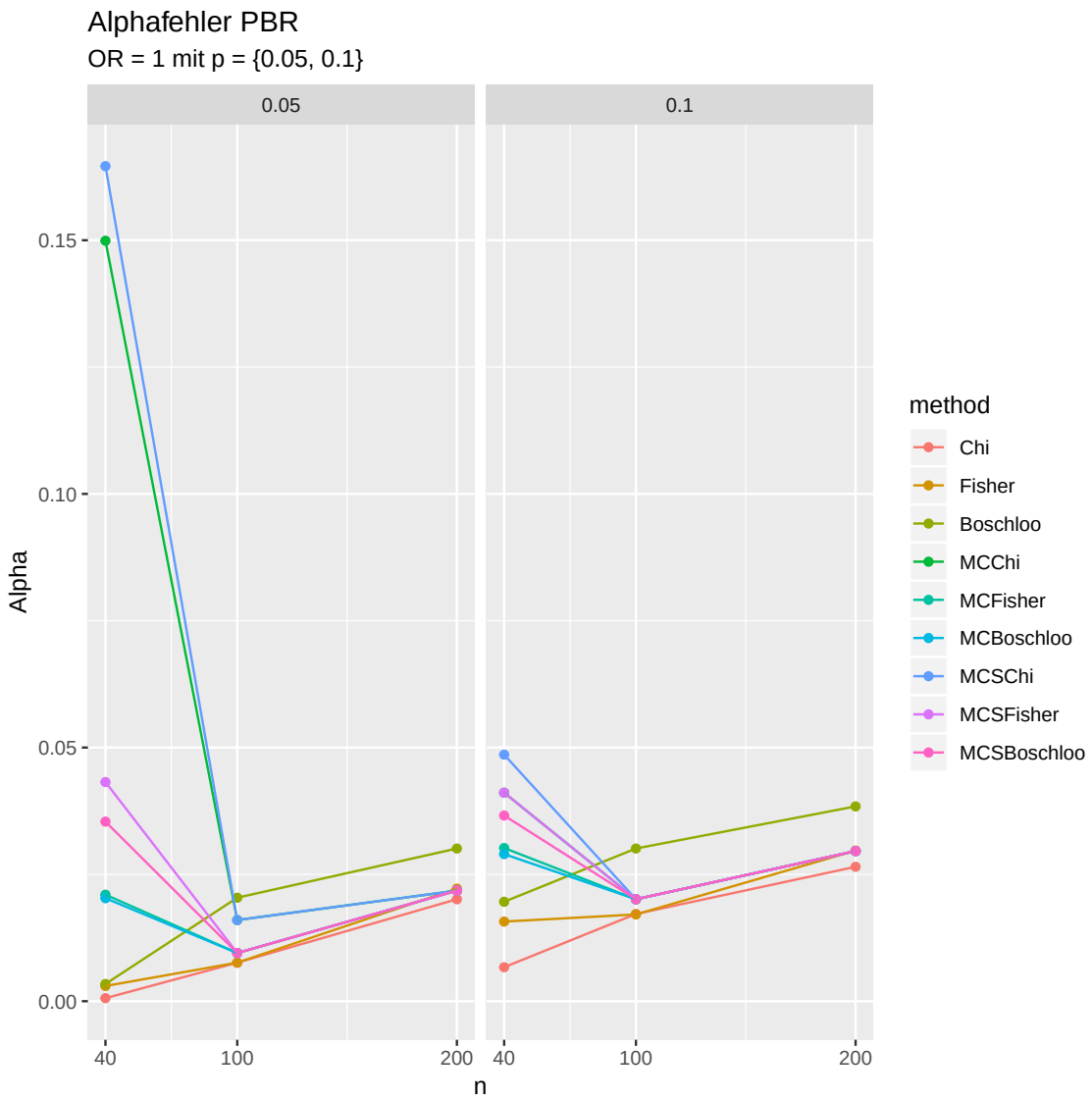

Abbildung 3.3: Alphafehler PBR

Die Tabellen 3.7 bis 3.9 zeigen die Ergebnisse des Alphaniveaus nach Fallzahl für die Permuted Block Randomisierung mit festen Blocklängen zu den zwei verschiedenen Erfolgswahrscheinlichkeiten. Die Abbildung 3.3 beruht auf den Daten der Tabellen 3.7 bis 3.9.

Beim ganzheitlichen Vergleich aller Methoden ist zu erkennen, dass die Methoden MCChi und MCSChi das Alphanivea bei einer Fallzahl von 40 und einer Erfolgswahrscheinlichkeit von 0,05 nicht einhalten. Ab einer Erfolgswahrscheinlichkeit von 0,1 oder einer Fallzahl von 100 halten alle Tests das Alphaniveau ein. Bei der Erhöhung der Erfolgswahrscheinlichkeit vergrößert sich das Alphaniveau pro Fallzahl

minimal.

In den likelihoodbasierten Tests ist erkennbar, dass das Alpha-Niveau mit steigender Fallzahl stetig zunimmt. Außerdem hebt sich hier der Boschloo-Test deutlich von den anderen ab und ist im gesamten der beste Test ab einer mittleren Fallzahl.

Die Randomisierungstests schneiden auch hier für eine kleine Fallzahl, zu beiden Erfolgswahrscheinlichkeiten, besser ab, als die likelihoodbasierten Tests. Alerrdings sinkt ihr Alpha-Niveau bei einer mittleren Fallzahl fast um die Hälfte und nimmt beim weiteren Anstieg von n nur mäßig wieder zu. Die Methode "Boschloo"übersteigen sie nicht mehr. Zwischen den Randomethoden sind ab der mittleren Fallzahl kaum bis gar keine Unterschiede mehr zu verzeichnen. Des Weiteren ist kein Unterschied zwischen den Randomisierungsmethoden, Fisher und Chiquadrattest mehr zu finden ab einer mittleren Fallzahl.

3.3.2 Fehler 2. Art

Tabelle 3.10: Power PBR bei n=40

		Likelihood			Monte-Carlo			Monte-Carlo-Summe		
OR	p	Chi	Fisher	Boschloo	MCChi	MCFisher	MCBoschloo	MCSChi	MCSFisher	MCSBoschloo
3	0.05	0.0030	0.0124	0.0135	0.1684	0.0391	0.0382	0.1898	0.0617	0.0607
3	0.10	0.0321	0.0701	0.0815	0.1082	0.0962	0.0929	0.1187	0.1055	0.1034
5	0.05	0.0055	0.0224	0.024	0.1801	0.0528	0.0523	0.2036	0.0763	0.0770
5	0.10	0.0635	0.1313	0.1494	0.1805	0.1707	0.1666	0.188	0.1757	0.1736

Tabelle 3.11: Power PBR bei n=100

		Likelihood			Monte-Carlo			Monte-Carlo-Summe		
OR	p	Chi	Fisher	Boschloo	MCChi	MCFisher	MCBoschloo	MCSChi	MCSFisher	MCSBoschloo
3	0.05	0.0556	0.0556	0.1141	0.0715	0.0663	0.0663	0.0715	0.0663	0.0663
3	0.10	0.2039	0.2020	0.2771	0.2194	0.2193	0.2193	0.2194	0.2193	0.2193
5	0.05	0.1153	0.1153	0.2116	0.1391	0.1341	0.1341	0.1391	0.1341	0.1341
5	0.10	0.4105	0.4033	0.4991	0.4201	0.4201	0.4201	0.4201	0.4201	0.4201

Tabelle 3.12: Power PBR bei n=200

		Likelihood			Monte-Carlo			Monte-Carlo-Summe		
OR	p	Chi	Fisher	Boschloo	MCChi	MCFisher	MCBoschloo	MCSChi	MCSFisher	MCSBoschloo
3	0.05	0.2109	0.2218	0.2615	0.2176	0.2176	0.2176	0.2176	0.2176	0.2176
3	0.10	0.4671	0.4784	0.5167	0.4692	0.4692	0.4692	0.4692	0.4692	0.4692
5	0.05	0.418	0.4256	0.4678	0.4199	0.4197	0.4197	0.4199	0.4197	0.4197
5	0.10	0.7248	0.6973	0.6872	0.7171	0.7171	0.7171	0.7171	0.7171	0.7171

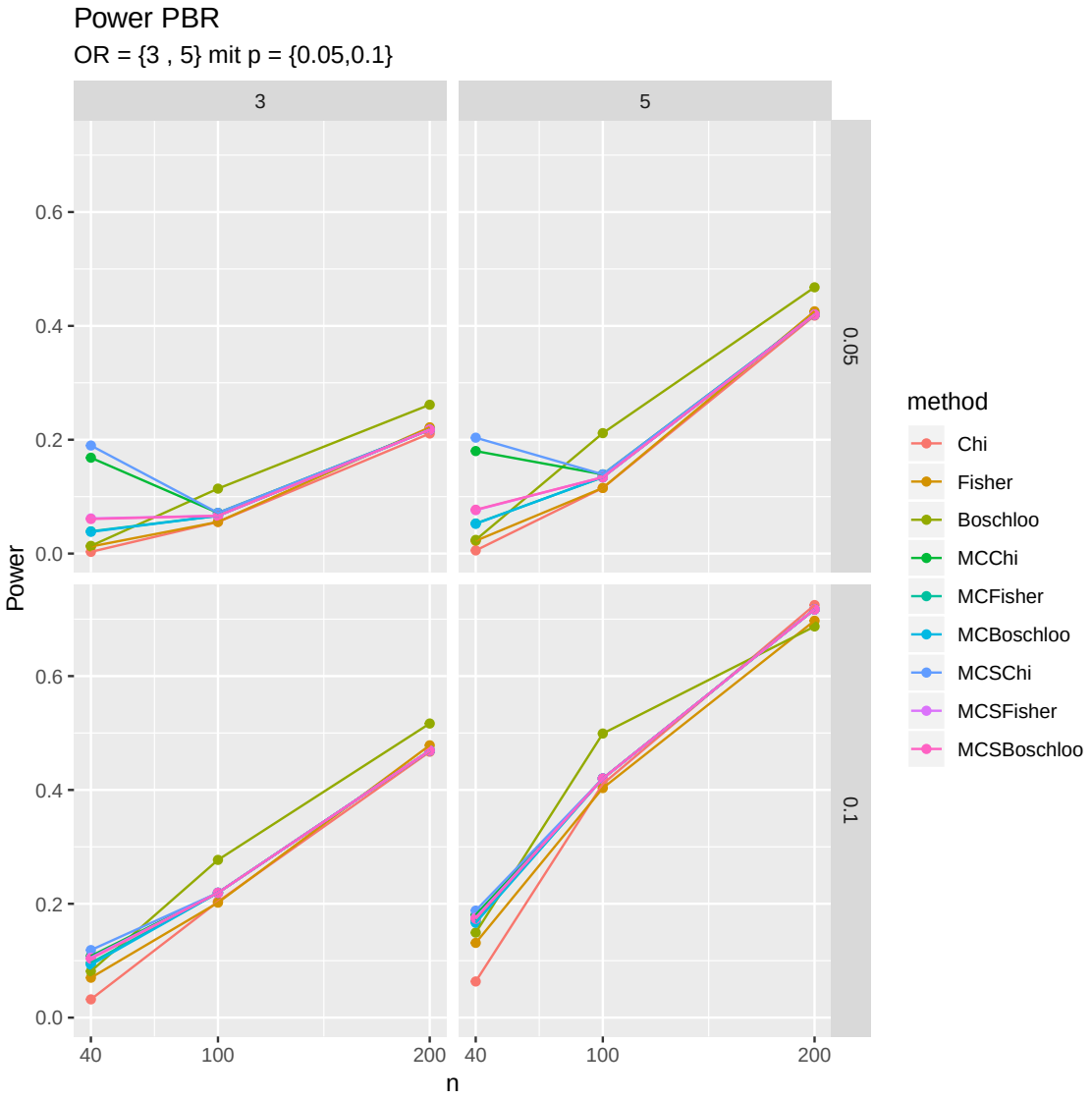

Abbildung 3.4: Power PBR

Die Tabellen 3.10 bis 3.13 zeigen die Ergebnisse der Power nach der Fallzahl für die Permuted Block Randomisierung mit festen Blocklängen zu den zwei verschiedenen Erfolgswahrscheinlichkeiten. Die Abbildung 3.4 beruht auf den Daten der Tabellen 3.10 bis 3.12.

Ganzheitlich liefern alle Tests mit zunehmendem n die gleiche Power, das heißt für hohe Fallzahlen unterscheiden sich die Tests nur minimal. Einen kleinen Vorsprung liefert der Boschlootest ab mittleren Fallzahlen. Für kleine Fallzahlen lässt sich der MSCChi Test als der Beste Test annehmen. Im Allgemeinen sind aber auch hier alle Tests unterpowert. Die maximale Power wird erreich mit dem Chiquadrat-Test bei einer Fallzahl von 200, einem OR von 5 und einer Erfolgswahrscheinlichkeit von 0,1. Für fast jeden Test erhöht sich die Power bei der Steigerung des Odds Ratios von 3 auf 5.

Bei den likelihoodbasierten Tests schneidet auch hier der Boschlootest am besten ab. Mit steigender Fallzahl gleichen sich die Tests an und werden somit präziser. Es lässt sich daher kein sehr deutlicher Unterschied zwischen den likelihoodbasierten Tests finden.

Die Randomisierungstests liefern mit steigendem n fast komplett gleiche Ergebnisse. Eine Unterscheidung zwischen den Randomisierungstestmethoden ist ab einer mittleren Fallzahl oder einer Erfolgswahrscheinlichkeit von 0,1 nicht erkennbar. Lediglich bei einer kleinen Fallzahl und der unteren Erfolgswahrscheinlichkeit schneiden MCChi und MCSChi besser ab.

3.4 Efrons Biased Coin vs. Permuted Block Randomisierung

Die Nachfolgenden Tabellen sollen den Unterschied der beiden Randomiseirungsmethoden näher darstellen in Bezug auf die Tests. Dazu zeigen die Tabellen 3.13 bis 3.15 die Differenz der Ergebnisse. Bei einem negativen Ergebnis ist die Permuted Block Methode besser. Bei einem positiven Ergebnis hat die Efrons biased Coin Methode ein besseres α- oder 1-β-Niveau.

Tabelle 3.13: Differenz von EBC und PBR bei n=40

			Likelihood			Monte-Carlo			Monte-Carlo-Summe		
OR	p	Fehler	Chi	Fisher	Boschloo	MCChi	MCFisher	MCBoschloo	MCSChi	MCSFisher	MCSBoschloo
1	0.05	Alpha	0	0	0	-3e-04	-3e-04	-0.0003	9e-04	0.0003	9e-04
1	0.10	Alpha	0	0	0	5e-04	-2e-04	0.0004	-4e-04	-0.0001	-5e-04
3	0.05	Power	0	0	0	1e-03	7e-04	0.0009	6e-04	0.0004	6e-04
3	0.10	Power	0	0	0	4e-04	7e-04	0.0007	-5e-04	-0.0003	-5e-04
5	0.05	Power	0	0	0	5e-04	7e-04	0.0004	-8e-04	-0.0010	-8e-04
5	0.10	Power	0	0	0	-8e-04	-5e-04	-0.0012	4e-04	0.0012	6e-04

Tabelle 3.14: Differenz von EBC und PBR bei n=100

			Likelihood			Monte-Carlo			Monte-Carlo-Summe		
OR	p	Fehler	Chi	Fisher	Boschloo	MCChi	MCFisher	MCBoschloo	MCSChi	MCSFisher	MCSBoschloo
1	0.05	Alpha	0.0000	0.0036	0.0069	0.0226	0.0270	0.0224	0.0321	0.0329	0.0322
1	0.10	Alpha	0.0006	0.0059	0.0073	0.0178	0.0198	0.0169	0.0208	0.0214	0.0212
3	0.05	Power	-0.0002	0.0163	0.0164	0.0645	0.0636	0.0645	0.0717	0.0700	0.0712
3	0.10	Power	0.0085	0.0348	0.0205	0.0725	0.0751	0.0708	0.0695	0.0704	0.0700
5	0.05	Power	-0.0030	0.0235	0.0260	0.0956	0.0938	0.0956	0.1022	0.1002	0.1029
5	0.10	Power	0.0020	0.0345	0.0228	0.0781	0.0818	0.0752	0.0709	0.0719	0.0716

Tabelle 3.15: Differenz von EBC und PBR bei n=200

			Likelihood			Monte-Carlo			Monte-Carlo-Summe		
OR	p	Fehler	Chi	Fisher	Boschloo	MCChi	MCFisher	MCBoschloo	MCSChi	MCSFisher	MCSBoschloo
1	0.05	Alpha	-0.0026	0.0017	-0.0004	0.0163	0.0168	0.0164	0.0210	0.0227	0.0209
1	0.10	Alpha	0.0006	0.0036	0.0027	0.0130	0.0141	0.0126	0.0174	0.0186	0.0175
3	0.05	Power	0.0007	0.0200	0.0085	0.0729	0.0742	0.0724	0.0717	0.0744	0.0711
3	0.10	Power	0.0118	0.0227	0.0125	0.0554	0.0579	0.0551	0.0334	0.0346	0.0339
5	0.05	Power	0.0010	0.0218	0.0150	0.0773	0.0783	0.0768	0.0551	0.0579	0.0551
5	0.10	Power	0.0023	0.0044	0.0105	0.0224	0.0222	0.0226	-0.0504	-0.0493	-0.0507

In kleinen Fallzahlen sind im gesamten keine Unterschiede zwischen den Hypothesentests basierend auf EBC oder PBR zu erkennen. In den likelihoodbasierten Tests liegt der Unterschied sogar exakt bei 0. In den Randomisierungstests bezieht sich die Differenz auf die vierte Nachkommastelle oder ist sogar noch geringer. Bei steigender Fallzahl erhöht sich der Unterschied, aber übersteigt für Alpha keinen

Wert von 3,3% und für die Power keinen Wert von 10%. Es ist dennoch erkennbar, dass ab einer mittleren Fallzahl die Randomisierungstests, bei Probanden welche mit EBC randomisiert werden, die Niveaus besser ausschöpfen.
In Standard Hypothesentests ist kein Unterschied bei der Verwendung der Randomisierungsmethode erkennbar.

3.5 Berechnungszeiten

Für die Berechnung aller Datensätze war ein Zeitraum von ca. 2 Wochen notwendig. Mit steigender Fallzahl, dauerte die Berechnung natürlich länger. Die Datensätze für $n = 40$ und $n = 200$, nahmen dabei einen Zeitraum von einer Woche ein. Zusätzlich war die Berechnung von 6 Datensäten mit einer Fallzahl von $n=400$ geplant. Diese war im zeitlichen Rahmen dieser Arbeit nicht mehr möglich, da sie alleine ca. 3-4 Wochen in Anspruch genommen hätte.

4 Diskussion

4.1 Vorteil von Randomisierungstests

Randomisierungstests bieten lediglich einen Vorteil für sehr kleine Studien ($n<100$). Dort kann das Alphaniveau sehr gut ausgeschöpft werden und auch die kann gehoben werden im Vergleich zu den likelihoodbasierten Tests. In der Realität kommen solche kleinen Studien aber sehr selten vor, da sie bei der geringen Power von max. 50% bei einem OR von 5 und 100 Patienten sehr unökonomisch sind. Vertretbar wäre dies bei klinischen Studien, welche nur sehr wenige Patienten, auf Grund von seltenen Krankheiten, haben. Dabei können Randomisierungstests ein Vorteil in Bezug auf den Alphafehler bieten und die Power je nach Unterschied in den Gruppen deutlich verbessern. Dennoch sollte bei diesen Studien ein großer Unterschiet bzw. OR zwischen den Gruppen vorhanden sein, um überhaupt eine geringe Chance zu haben, einen Unterschied nachzuweisen.

Bei großen Studien (ab $n>200$) ist abzusehen, dass sich Randomisierungstests nicht mehr lohnen, da sie sich in Bezug auf den Alpha-Fehler und die Power zu Standard Hypothesentests nicht mehr deutlich abheben. Betrachtet man den Rechenaufwand und die notwendigen Technischen Mittel, die alleine für 200 Patienten eingeplant werden müssen, ist diese Methode für den geringen Vorteil nicht effizient. Bei klinischen Studien mit einem Design, wie in der hier beschriebenen Simulationsstudie, ist eine Fallzahl von 200 Patient sehr unüblich. Spekulativ lässt sich vermuten das bei noch höheren Fallzahlen, gar kein Unterschied in den Hypothesentests erkennen lässt.

4.2 Randomisierungsmethoden

Die Unterschiede der Randomisierungsmethoden in Bezug auf die Hypothesentests ist in Fallzahlen größer 100 zu erkennen. Dort funktionieren Randomisierungstests deutlich besser bei der Efrons biased Coin Methode. Das bedeutet bei einer Studie, welche eine Permutationsmethode wie Permuted Block verwendet, lassen sich

Randomisierungstests gar nicht mehr vorteilhaft Anwenden, da sie schlechter abschneiden im Vergleich zu likelihoodbasierten Tests. Es zeigt sich deutlich das die likelihoodbasierten Tests ab *n=100* sowohl im Alphafehler als auch in der Power besser abschneiden. Die Aussage des Papers, dass Randomisierungstests und Permutationstests nicht dasselbe sind, lässt sich demnach unterstützen.

4.3 Monte Carlo Prozeduren

Beim Vergleich der Monte Carlo Prozeduren „Standard“ und „Summe“ zeigt sich, beim Efrons Biased Coin design (für alle n) und beim Permuted Block für *n=40*, ein kleiner Vorsprung von 1-2% beim Aplpha-fehler und der Power beim Verwenden des Summenverfahrens. Falls die Randomisierungstetsts verwendet werden, sollte dementsprechend dieses Verfahren vorgezogen werden.

4.4 Komplette Randomisierungstests

Im Rahmen dieser Arbeit wurden zunächst die kompletten Randomisierungstests versucht zu implementieren und zu berechnen. Dabei wurde schnell klar, das auf Monte-Carlo Prozeduren umgestiegen werden muss. Ein gewöhnlicher Computer scheitert bei den vollständigen Tests ab einer Patientenanazahl von *n= 100*. Auch bei einem Cluster ist die Berechnung von *n=40* nicht in einer annehmbaren Zeit möglich. Geht man von der Behauptung des Papers aus, es existiert heute genug Rechenleistung, ist diese Aussage nur eingeschränkt zu unterstützen. Die Rechenleistung ist vorhanden, aber steht Ressourcentechnisch bei kaum einer Studie zur Verfügung und sie schafft die Berechnung in keiner annehmbaren Zeit von weniger als 2 Wochen. Die kompletten Tests werden damit vorerst weiterhin ungenutzt bleiben und sich nicht integrieren. In der der heutigen Zeit der Digitalisierung ist es aber nicht auszuschließen, dass die kompletten Tests in Zukunft akkurat möglich sein werden. In diesem Fall, sollten die Tests erneut und für höhere Fallzahlen gegen die Standard Hypothesentests geprüft werden.

4.5 Benötigte Rechenleistung

Für die Rechenleistung ist zu diskutieren, das für die Simulationsstudie 10.000 Experimente mit 200 Patienten durchgeführt wurden. In der Realität wird eine Studie nur einmal durchgeführt, wodurch die Berechnung von Monte Carlo Randomisie-

rungstests natürlich deutlich schneller funktionieren würde. Dennoch wird der rechenaufwand spekulativ sehr hoch, da in normalen Studien die Patientenanzahlen auch deutlich höher sind. Ein likelihoodbasierter Test ist dabei sehr viel schneller und mit einem normalen Computer berechenbar. Für ein Randomisierungstest wird ein Cluster und viel Programmierungsaufwand notwendig. Für einen einer derart kleinen Unterschied in höheren Fallzahlen ist dies nicht sinnvoll.

4.6 Fazit

Die Frage, welchen Vorteil Randomisierungstests bieten, die am Anfang dieser Arbeit gestellt wurde, soll neben den fünf Aussagen des zu Grunde liegenden Papers von „Rosenberger et al. (2018) Leitfaden des Fazits sein.

Der Vorteil von Randomisierungstests ist kritisch zu betrachten, zumal für die Simulationsstudie kein kompletter Test durchgeführt werden konnte und daher die Monte Carlo Prozeduren verwendet wurden. Für einen kompletten Test lassen sich daher keine Aussagen über den Vorteil treffen. Für Monte Carlo Randomisierungstests lässt sich zusammengefasst sagen, dass sie für kleine Fallzahlen einen gut erkennbaren Vorteil liefern im Vergleich zu Likelihoodbasierten Tests und noch akkurat berechenbar sind. Hierbei bietet das Monte-Carlo-Summenverfahren einen größeren Vorteil. Ab einer Fallzahl die größer als 200 ist, relativiert sich der Unterschied in den Ergebnissen und ein likelihoodbasierter Test kann genauso präzise angewendet werden wie ein Monte Carlo-Test. Dabei sollte der beste Likelihoodtest für das Setting verwendet werden. Im Fall der Simulationsstudie der Fisher-Boschloo-Test.

Die fünf Aussagen des Papers werden an dieser Stelle differenziert betrachtet. Die erste Aussage lässt sich anhand der Simulationsstudie nur für kleine Fallzahlen unterstützen. Für größere Fallzahlen, liefern alle Tests ähnliche Ergebnisse. Die zweite These kann durch die Recherchen unterstützt werden, da mit Monte Carlo-Verfahren und dem besten Test zur Stichprobe Randomisierungsverfahren für jeden Outcome angewendet werden können. Die dritte Aussage wird definitiv verneint. Mit Monte Carlo Funktionen ist die Berechnung zwar einfacher und schneller möglich, aber nicht in Sekunden. Bei kompletten Tests funktioniert es gar nicht in einer annehmbaren Zeit. Die vierte Aussage lässt sich für kleine Stichproben, oder den Randomisierungstests basierend auf EBC Randomisierung bestätigen. Die fünfte Aussage konnte Anhand der Recherchen bestätigt werden und ist in der Simulationsstudie erkennbar.

Eidesstaatliche Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer ent-nommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel sind angegeben. Die Arbeit hat mit gleichem bzw. in wesentlichen Teilen gleichem Inhalt noch keiner Prüfungsbehörde vorgelegen.

Ort, Datum, Unterschrift

Literatur

[1] Rosenberger, WF, Uschner, D, Wang, Y. Randomization- The forgotten component of the randomized clinical trial. Statistics in Medicine. 2019; 38: 1– 12.
Abgerufen von: https://doi.org/10.1002/sim.7901

[2] Randomisierung. [Wikipedia-Artikel]
Abgerufen von: https://de.wikipedia.org/wiki/Randomisierung#Arten_der_Randomisierung [23.01.20, 15:00 Uhr]

[3] F. Krummenauer, C. Baulig, J. Hirsch (2014). Randomisierung in klinischen Studien – Mit Zufall zum Erfolg. Deutscher Ärzte-Verlag, zzi, Z Zahnärztl Impl, 2014; 30 (1)
Abgerufen von: https://www.uni-wh.de/fileadmin/user_upload/03_G/07_Humanmedizin/05_Institute/IMBE/23_-_Randomisation_in_Klinischen_Studien.pdf [23.01.20, 13:20 Uhr]

[4] Rosenberger WF (Juli 2015). Randomization in small clinical Trials [PowerPoint Folien].
Abgerufen von: https://www.newton.ac.uk/files/seminar/20150710110011451-409338.pdf

[5] Testtheorie. [Wikipediaartikel]
Abgerufen von:https://de.wikipedia.org/wiki/Testtheorie_(Statistik)#Testtheorie_als_Entscheidungsproblem [28.01.20, 12:03 Uhr]

[6] Hedderich, Sachs (15.Auflage). Angewandte Statistik. Berlin, Springer Spektrum. Erschienen: 2016. [Fachbuch]

[7] Signifikanztests. [Internetartikel]
Abgerufen von: https://www.statistik-nachhilfe.de/ratgeber/statistik/induktive-statistik/signifikanztests-hypothesentests[28.01.20, 12:32 Uhr]

[8] Zweiseitige Hypothesentests. [Internetseite]
Abgerufen von: https://www.abiturma.de/mathe-lernen/stochastik/hypothesentest/zweiseitige-hypothesentests
[10.02.20, 9:50 Uhr]

[9] Statistischer Test. [Wikipediaartikel]
Abgerufen von: https://de.wikipedia.org/wiki/Statistischer_Test#Asymptotisches_Verhalten_des_Tests [28.01.20, 12:05 Uhr]

[10] Bernoulli-Verteilung. [Wikipediaartikel]
Abgerufen von: https://de.wikipedia.org/wiki/Bernoulli-Verteilung [28.01.20, 12:15 Uhr]

[11] Likelihoodfunktion [Wikipediaartikel]
Abgerufen von: https://de.wikipedia.org/wiki/Likelihood-Funktion [16.02.20, 12:03]

[12] Schätzfunkzion. [Wikipediartikel]
Abgerufen von:https://de.wikipedia.org/wiki/Sch%C3%A4tzfunktion [16.02.20, 12:20]

[13] Chi-Quadrat-Unabhängigkeitstest. [Wikipediartikel]
Abgerufen von:https://de.wikipedia.org/wiki/Chi-Quadrat-Test#Unabh%C3%A4ngigkeitstest [16.02.20, 12:44]

[14] Chi-Quadrat-Unabhängigkeitstest. [Internetartikel]
Abgerufen von:https://www.statistik-nachhilfe.de/ratgeber/statistik/induktive-statistik/signifikanztests-hypothesentests/pruefung-auf-unabhaengigkeit/chi-quadrat-unabhaengigkeitstest [16.02.20, 12:49]

[15] Samuel Kilian (2019). Fallzahlberechnung für den Fisher-Boschloo-Test [PowerPoint Folien]. Kopie Digital auf der CD

[16] Exakter Fisher Test [Internetartikel]
Abgerufen von:https://www.statistik-nachhilfe.de/ratgeber/statistik/induktive-statistik/signifikanztests-hypothesentests/pruefung-auf-unabhaengigkeit/exakter-fisher-test [17.02.20, 12:49]

[17] Klinische Studie [Wikipediaartikel]
Aufgerufen von : https://de.wikipedia.org/wiki/Klinische_Studie [19.02.20, 09:08]

Anhang

R Skript Cluster EBC

```r
#Cluster_EBC.R

Data <- function(data, prob, OR,...){
  #browser()

  #ErfW A und B aus Prob
  a <- function(OR, prob){
    if(OR==1){
      if(prob == 0.05){
        return(2.5)
      }else{
        if(prob==0.1){
          return(5)} else{return(10)}
      }
    }
    else{#OR==3
      if(prob ==0.05){
        p <- (-55*OR -45)/(OR-1)
        q <- (250*OR)/(OR-1)
      }else{
        if(prob==0.1){
          p <- (-60*OR -40)/(OR-1)
          q <- (500*OR)/(OR-1)
        }else{
          if(prob==0.2){
            p <- (-70*OR -30)/(OR-1)
            q <- (1000*OR)/(OR-1)
          }else{print("prob von 0.05 oder 0.1 eingeben!")}}

      }
      a_1 <- -(p/2)  + sqrt((p/2)^2 - q)
      a_2 <- -(p/2)  - sqrt((p/2)^2 - q)

      return(abs(c(a_1, a_2)))
    }
  }
    A_prob <- (min(a(OR, prob)))/50
    B_prob <- ((prob*100) - (min(a(OR, prob))))/50

    params <- ebcPar(data*2, 2/3)
    AB <- as.vector(getRandList(genSeq(params)))

    A_stellen <- grep("A", AB)
    B_stellen <- grep("B", AB)

    Seq <- c(1:data)
    suppressWarnings({
      Seq [A_stellen] <- rbinom(length(A_stellen), 1, A_prob)
      Seq [B_stellen] <- rbinom(length(B_stellen), 1, B_prob)
    })
    Seq <- factor(Seq, levels = c(0,1))

    instance  <-data.frame(AB, Seq)
    return(instance)
}



#m Wiederholungen für Monte Carlo
Algorithmus <- function(data, instance, m, ...){
  #browser()

  #Hypotesentests
  t0 <- table(instance)

  suppressWarnings({
    chi <- chisq.test(t0) #simulate p value nicht auf true gesetzt, da es dann eine form von fishers exact test ist.
    fisher<- fisher.test(t0)
    boschloo<-Exact::exact.test(t0, alternative = "two.sided", npNumbers = 100,
                                np.interval = FALSE, beta = 0.001,
```

R Skript Cluster PBR

```
#Cluster_PBR.R

Data <- function(data, prob, OR,...){
  #browser()

  #ErfW A und B aus Prob
  a <- function(OR, prob){
    if(OR==1){
      if(prob == 0.05){
        return(2.5)
      }else{
        if(prob==0.1){
          return(5)} else{return(10)}
      }
    }
    else{#OR==3
      if(prob ==0.05){
        p <- (-55*OR -45)/(OR-1)
        q <- (250*OR)/(OR-1)
      }else{
        if(prob==0.1){
          p <- (-60*OR -40)/(OR-1)
          q <- (500*OR)/(OR-1)
        }else{
          if(prob==0.2){
            p <- (-70*OR -30)/(OR-1)
            q <- (1000*OR)/(OR-1)
          }else{print("prob von 0.05 oder 0.1 eingeben!")}}

      }
      a_1 <- -(p/2)  + sqrt((p/2)^2 - q)
      a_2 <- -(p/2)  - sqrt((p/2)^2 - q)

      return(abs(c(a_1, a_2)))
    }
  }
  A_prob <- (min(a(OR, prob)))/50
  B_prob <- ((prob*100) - (min(a(OR, prob))))/50

  pbr <- function(k,b, K = 2, ratio = rep(1,K)) {
    #browser()
    ratio <- ratio/sum(ratio)
    AB <- c(NULL)
    for(i in 1:b){
      AB0 <- sample(rep(0:(K-1), times = ratio*k))%>%
        gsub("0", "A", .)%>%
        gsub("1","B",.)
      AB<- c(AB, AB0)
    }
    return(AB)
  }
  AB <- pbr(4, (data*2)/4)

  A_stellen <- grep("A", AB)
  B_stellen <- grep("B", AB)

  Seq <- c(1:data)
  suppressWarnings({
    Seq [A_stellen] <- rbinom(length(A_stellen), 1, A_prob)
    Seq [B_stellen] <- rbinom(length(B_stellen), 1, B_prob)
  })
  Seq <- factor(Seq, levels = c(0,1))

  instance  <-data.frame(AB, Seq)
  return(instance)
}
```

R Skript Cluster Auswertung

```
#Auswertung der Clusterergebnisse
#Cluster Ergebnisdatein auf der CD zur Arbeit
#Auswertungsfunktion:
# wie viel Prozent sind signifikant?
Proportion <- function(x){
  #browser()
  x0 <- gsub(NaN, 1, x)
  t1 <- length(grep(TRUE,x0<0.05))/10000
  t0 <- length(grep(TRUE,x0>=0.05))/10000
  return(cbind(t1,t0))
}

#Clusterdateien befinden sich auf der CD im Buch

#Efrons biased Coin:
E_N20_OR1_5  <- results_EBS20[1:10000,]
E_N20_OR3_5  <- results_EBS20[10001:20000,]
E_N20_OR5_5  <- results_EBS20[20001:30000,]
E_N20_OR1_10 <- results_EBS20[30001:40000,]
E_N20_OR3_10 <- results_EBS20[40001:50000,]
E_N20_OR5_10 <- results_EBS20[50001:60000,]

E_N50_OR1_5  <- results_EBS50[1:10000,]
E_N50_OR3_5  <- results_EBS50[10001:20000,]
E_N50_OR5_5  <- results_EBS50[20001:30000,]
E_N50_OR1_10 <- results_EBS50[30001:40000,]
E_N50_OR3_10 <- results_EBS50[40001:50000,]
E_N50_OR5_10 <- results_EBS50[50001:60000,]

E_N100_OR1_5  <- results_EBS100[1:10000,]
E_N100_OR3_5  <- results_EBS100[10001:20000,]
E_N100_OR5_5  <- results_EBS100[20001:30000,]
E_N100_OR1_10 <- results_EBS100[30001:40000,]
E_N100_OR3_10 <- results_EBS100[40001:50000,]
E_N100_OR5_10 <- results_EBS100[50001:60000,]

#Permuted Block:
P_N20_OR1_5  <- results_PBR20[1:10000,]
P_N20_OR3_5  <- results_PBR20[10001:20000,]
P_N20_OR5_5  <- results_PBR20[20001:30000,]
P_N20_OR1_10 <- results_PBR20[30001:40000,]
P_N20_OR3_10 <- results_PBR20[40001:50000,]
P_N20_OR5_10 <- results_PBR20[50001:60000,]

P_N50_OR1_5  <- results_PBR50[1:10000,]
P_N50_OR3_5  <- results_PBR50[10001:20000,]
P_N50_OR5_5  <- results_PBR50[20001:30000,]
P_N50_OR1_10 <- results_PBR50[30001:40000,]
P_N50_OR3_10 <- results_PBR50[40001:50000,]
P_N50_OR5_10 <- results_PBR50[50001:60000,]

P_N100_OR1_5  <- results_PBR100[1:10000,]
P_N100_OR3_5  <- results_PBR100[10001:20000,]
P_N100_OR5_5  <- results_PBR100[20001:30000,]
P_N100_OR1_10 <- results_PBR100[30001:40000,]
P_N100_OR3_10 <- results_PBR100[40001:50000,]
P_N100_OR5_10 <- results_PBR100[50001:60000,]


spaltennamen <- c("Alpha", "Richtig", "Alpha", "Richtig", "Power", "Beta", "Power", "Beta","Power", "Beta", "Power", "Beta")
zeilennamen <- c("OR1_5", "OR1_5", "OR1_10", "OR1_10", "OR3_5", "OR3_5", "OR3_10","OR3_10","OR5_5", "OR5_5", "OR5_10", "OR5_10")
Spalten <- c("Fehler","Chi", "fisher", "boschloo", "MC_Chi", "MC_fisher", "MC_boschloo", "MCS_Chi", "MCS_fisher", "MCS_boschloo")


#N20 EBC
N20_EBC <-
  rbind(
    OR1_5  <- apply(E_N20_OR1_5,  2,Proportion),
    OR1_10 <- apply(E_N20_OR1_10, 2,Proportion),
    OR3_5  <- apply(E_N20_OR3_5,  2,Proportion),
    OR3_10 <- apply(E_N20_OR3_10, 2,Proportion),
    OR5_5  <- apply(E_N20_OR5_5,  2,Proportion),
```