Therapievergleichsstudien

Dirk Hasenclever

Überblick: Therapievergleichsstudien

- Themen
  - Systematische Verzerrungen und Techniken zum Herstellen von Vergleichbarkeit
  - Formale Designs mit zwei und mehr als zwei Armen
  - Typen formaler Fragestellungen

- Spätere Vorlesungen
  - Endpunkte
  - Fallzahlplanung

Herstellen von Vergleichbarkeit

Zeitpunkt	Heterogenitätsquellen	Ziel	
bei Eintritt in die Studie	Patientcharakteristika (bekannte und latente Faktoren, die den Therapieerfolg beeinflussen)	Strukturgleichheit	Randomisieren
während der Durchführung der Studie	Behandlung (Strategien bei Unverträglichkeit der Therapie, Zeitschema, supportive Maßnahmen, Pflege) Beobachtung (Zeitschema, Beurteilungskriterien)	Beobachtungs- und Behandlungsgleich-heit	Standardisieren Verblinden Unabhängiger Review
nach Beendigung der Studie	Auswertung (Zählweise nicht protokollgerecht behandelter Patienten)	Auswertungsgleich-heit	Vorab-Analyseplan

Strukturgleichheit

- bedeutet, dass sich die zu vergleichenden Therapiegruppen vor Beginn der Studie in allen wichtigen Störgrößen **nicht** oder **nur zufällig** unterscheiden.
  - Störgrößen sind Größen, die den Therapieerfolg beeinflussen können.
- Wie erreicht man Strukturgleichheit?



Randomisation

Simpson‘s Paradoxon

- Änderung der Assoziation zwischen zwei Merkmalen, wenn bzgl. eines dritten Merkmales stratifiziert wird
- Fiktives Beispiel:

  Vergleich zweier Therapien bei gastro-intestinalen Karzinomen

	mehr als 5 Jahre überlebt		
	ja	**nein**	
Therapie A	30 (30%)	70	100
Therapie B	80 (40%)	120	200
Total	110 (37%)	190	300



Therapie B um 10 % besser als Therapie A

(Quelle: Guggenmoos I., Medizinische Statistik, Blackwell Wissenschaftsverlag Berlin,1996)

Simpson‘s Paradoxon

Stratifikation nach Tumorlokalisation

Magen

Therapie B um 5% schlechter als Therapie A

	mehr als 5 Jahre überlebt		
	ja	**nein**	
Therapie A	20 (25%)	60	80
Therapie B	8 (20%)	32	40
Total	28 (23%)	92	120

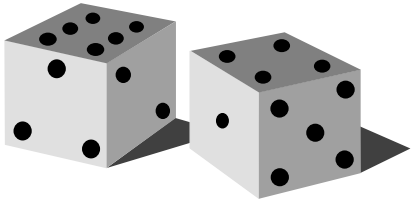
Kolon/Rektum

Therapie B um 5% schlechter als Therapie A

	mehr als 5 Jahre überlebt		
	ja	**nein**	
Therapie A	10 (50%)	10	20
Therapie B	72 (45%)	88	160
Total	82 (46%)	98	180

Randomisation

Randomisation ist die zufällige Zuordnung der Patienten zu den Therapiegruppen



- gleichmäßige Verteilung bekannter und unbekannter Störgrößen (hinreichend große Fallzahl vorausgesetzt) auf die Therapiearme
- gleiche Chance für jeden Patienten in einen der Therapiearme zu kommen

Randomisation

- Randomisationsergebnis darf nicht vorhersagbar sein
  (Verborgene Zuweisung, „Allocation Concealment“)
- keine Beweise für Unterlegenheit einer Therapie
  (Prinzip der vergleichbaren Unsicherheit)
- Therapien ethisch vertretbar und zumutbar
- Aufklärung und schriftliche Zustimmung der Patienten

Techniken zur Randomisation

- **Blockrandomisation**

  gewährleistet annähernd gleiche Fallzahl in den zu vergleichenden Therapiegruppen (Balanciertheit)

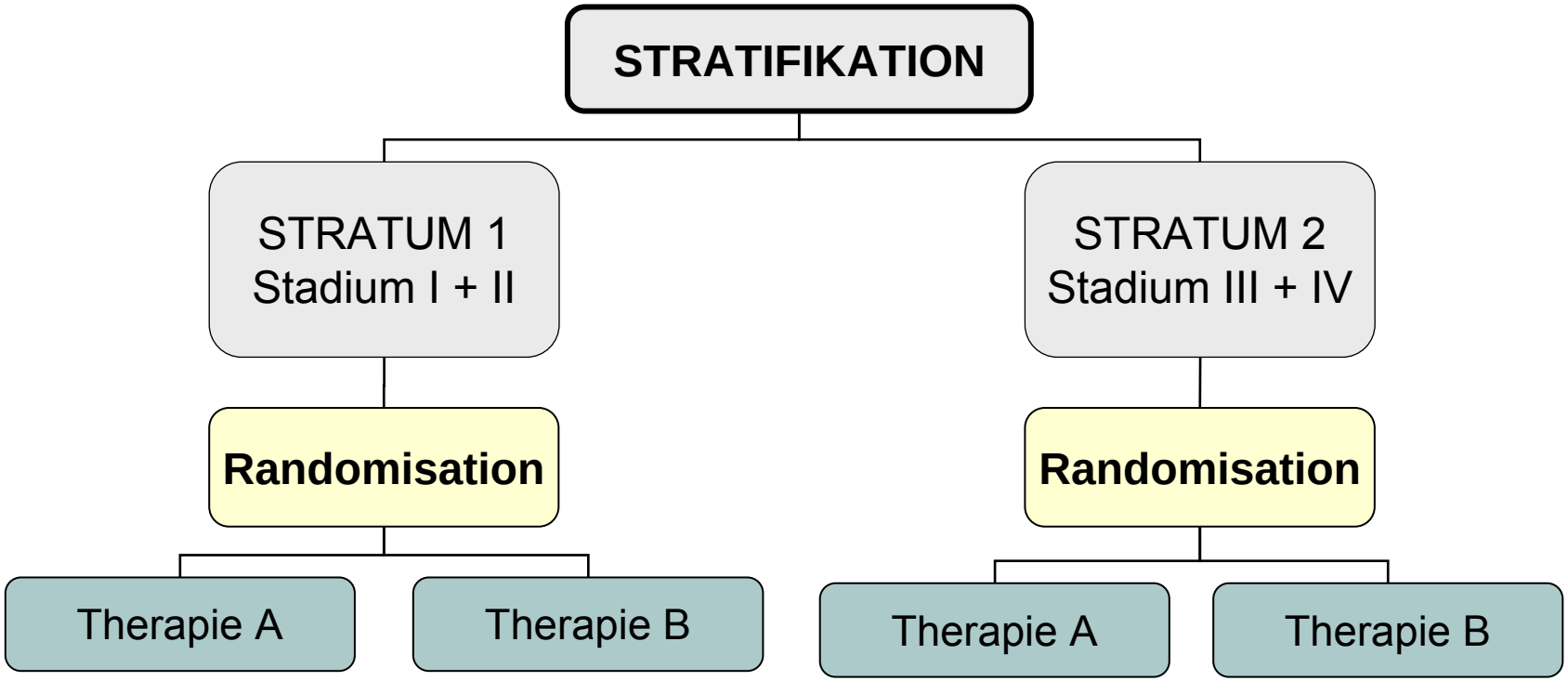
- **Stratifizierte Randomisation**

  gewährleistet annähernd Gleichverteilung von Störgrößen

  Entsprechend der Kategorien bekannter Störgrößen werden **Schichten (Strata)** gebildet. Randomisation erfolgt innerhalb der Schichten.

Stratifizierte Randomisation

Beispiel: Stratifikation nach Stadium der Erkrankung



Behandlungs- und Beobachtungsgleichheit

- bedeutet gleiche Behandlung der Patienten in den zu vergleichenden Therapiegruppen und gleiche Maßstäbe bei der Beurteilung von Therapiedurchführung, Therapieerfolg und Nebenwirkung
- Wie erreicht man **Behandlungs- und Beobachtungsgleichheit ?**



- **Blindversuche**
- **Einsatz von Placebo**
- **Standardisiertes Vorgehen bei Behandlung und Beobachtung**
- **Reviewbeurteilungen**

Blindversuch

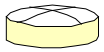
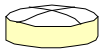
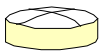
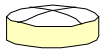
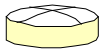
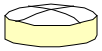
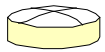
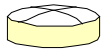
Versuch	Zuordnung der Patienten zu den Therapiearmen unbekannt für		
	Patient	behandelnden Arzt	Beurteiler/ Biometriker
einfach blind	✵		
doppel blind	✵	✵	
dreifach blind	✵	✵	✵

- **Offener Versuch**
  - Die Zuordnung zu den Behandlungsgruppen ist allen bekannt.
- **Blindversuch z.B.**
  - bei leichter Erkennbarkeit der angewandten Therapie
  - bei unterschiedlichen Nebenwirkungen
    - meist nicht möglich bzw. schwierig durchzuführen.

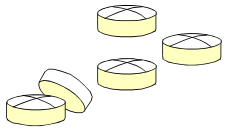
Blindversuch

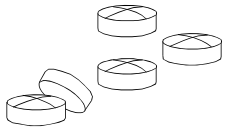
„double dummy“-Technik

Zum Beispiel: Bei unterschiedlichem Zeitplan der Applikation

	Tag 1	Tag 2	Tag 3	Tag 4	Tag 5
Therapie A					
Therapie B					

Zum Beispiel: Bei unterschiedlichen Applikationsformen

Therapie A:  + 

Therapie B:  + 

Placebo

- placebo (lat.): „Ich werde gefallen.„
- Placebobehandlung (Scheinbehandlung)
  - mit Präparat, das keinen Wirkstoff enthält (Placebo)
  - durch Placebo – Operation („Placebo-Chirurgie ist ethisch und klinisch notwendig“, Seiler et al, Dt. Ärzteblatt, 2004)
- Eine Placebobehandlung muss in **allen** Details der Therapie-modalität gleichen
  - Applikationsform, zeitliche Abfolge der Behandlung, Betreuung durch Pflegepersonal, etc.
- Placeboeffekt
  - durch Fremd- und Autosuggestion

Beispiel: Placebo-Chirurgie

- Randomisierter Vergleich zweier Arthroskopie – Verfahren und eines Placeboeingriffs bei Patienten mit chronischen Kniegelenkschmerzen (N Engl J Med, 347,81-88,2002)
  - primärer Endpunkt: postoperativer Schmerz nach 24 Monaten

Placebogruppe

- Standardarthroskopie - Verfahren wurde simuliert:
  Hautschnitt, verwendete OP - Instrumente, Dauer im OP etc.
- Postoperative Betreuung analog Verumgruppen
  Wundverband, Gehhilfen, Krankenhausaufenthalt, etc.

Beispiel: Placebo-Chirurgie

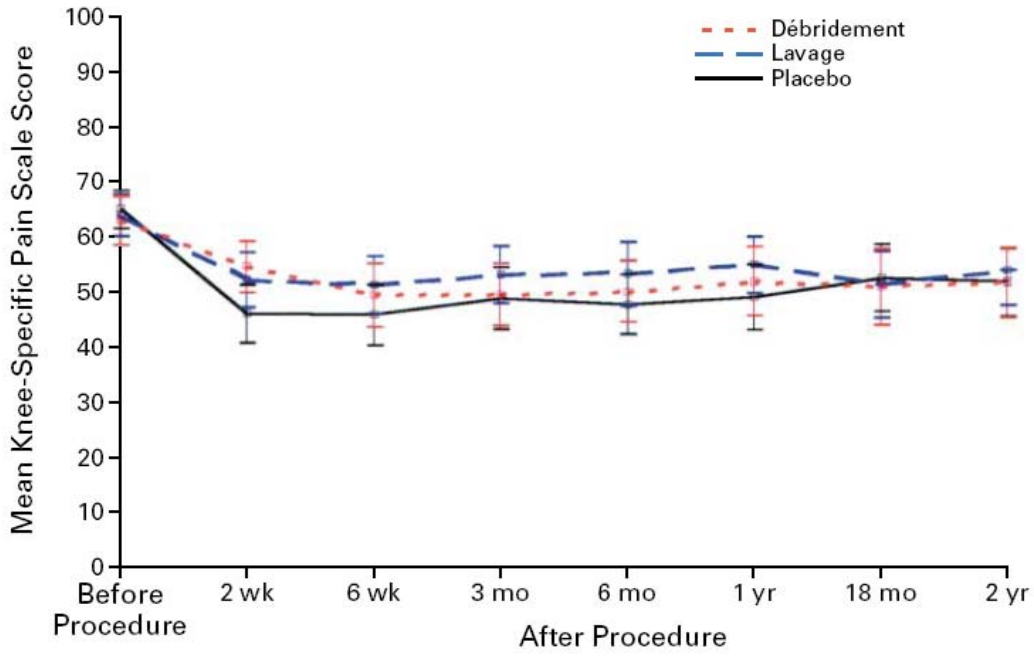
- Pflegepersonal kannte Gruppenzugehörigkeit der Patienten nicht

➧ **BEHANDLUNGSGLEICHHEIT GESICHERT**

- Bewertung der Endpunkte erfolgte durch Beurteiler, die Gruppenzugehörigkeit nicht kannten

➧ **BEOBACHTUNGSGLEICHHEIT GESICHERT**

Beispiel: Placebo-Chirurgie



No. at Risk								
Placebo	60	59	57	56	57	53	52	55
Lavage	61	59	57	59	59	57	56	55
Débridement	58	59	59	58	56	50	51	53

Figure 1. Mean Values (and 95 Percent Confidence Intervals) on the Knee-Specific Pain Scale.
Assessments were made before the procedure and 2 weeks, 6 weeks, 3 months, 6 months, 12 months, 18 months, and 24 months after the procedure. Higher scores indicate more severe pain.

(Moseley N Engl J Med, 347,81-88,2002)

Reviewbeurteilungen

- Reviewbeurteilungen sind Beurteilungen von nicht oder nur schwer objektivierbaren diagnostischen oder therapeutischen Merkmalen durch **qualifizierte Gutachter**, die hinsichtlich **aller anderer Merkmale in Unkenntnis** gelassen werden.

Auswertungsgleichheit

- bedeutet Gleichbehandlung der Daten in den Therapiegruppen bei allen Auswertungen
- Wie sichert man **Auswertungsgleichheit** ?



Nicht protokollgerecht verlaufende Fälle müssen in den Therapiegruppen mit gleichen Maßstäben bewertet werden

Auswertungsgleichheit

- **Protokollabweichungen**

  z.B.: aufgrund von

  - Nebenwirkungen
  - anderen medizinischen Erwägungen
  - Versehen

  erschweren Schätzung des „wahren“ Therapieeffektes

- **Studienabrecher (Drop-outs)**

  z.B.:aufgrund von

  - Widerruf der Teilnahme durch Patienten
  - Verlust des Patienten aus dem Monitoringprozess
  - Tod des Patienten aus unabhängiger Ursache



Informative Drop-outs führen zu einer Verzerrung des Therapievergleichs

Protokollabweichungen / Studienabbrecher

It is always better to have no dropouts and certainly the optimal solution to the problem of missing values is not to have any.

(Everitt 1996)

- **So gering wie möglich halten**

  durch Berücksichtigung bei Studienplanung und Monitoringmaßnahmen

- **Einheitliche Bewertung/Zählung**

  durch Festlegung vor Beginn der Auswertung (möglichst im Studienprotokoll)

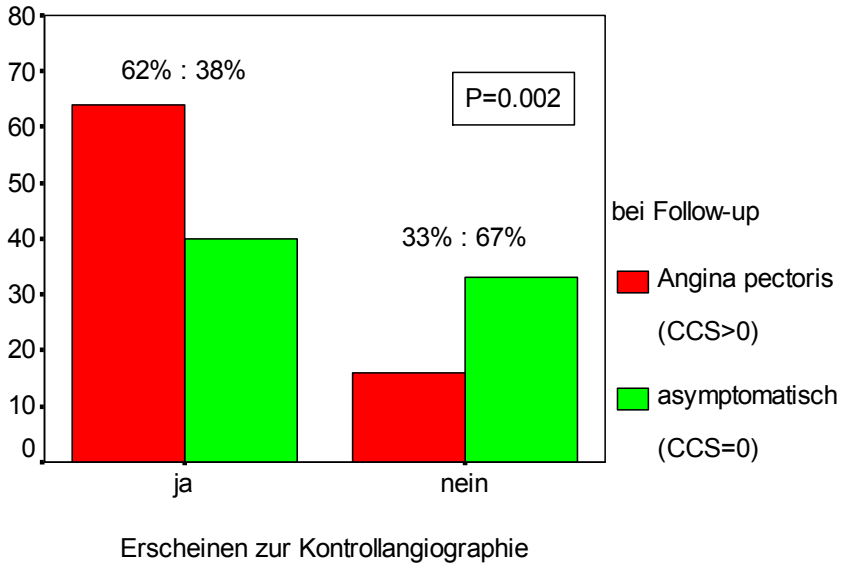
- **Nicht aus Auswertung ausschließen**

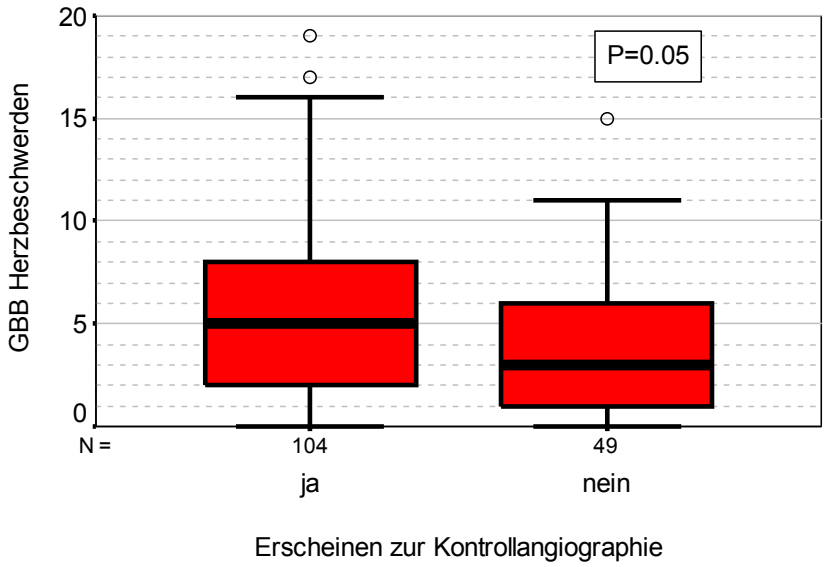
  Patient kann Studie verlassen, aber nicht die Auswertung.

Problem informative Drop-outs

Informative Drop-outs verzerren Therapieeffekte.
Dies lässt sich nicht durch höhere Fallzahlen korrigieren.

Beispiel: Erscheinen zur Kontrollangiographie nach PTCA





Validität von Studien

... ist der Grad, in welchem Schlussfolgerungen aus Studien gezogen werden können.

- **Interne Validität**
  - Methodisch korrekte Planung, Durchführung und Analyse, so dass beobachtete Effekte auch der zu untersuchenden Intervention zugeschrieben werden können

    ➞ **Glaubwürdigkeit der Ergebnisse**
  - Struktur-, Behandlungs-, Beobachtungs- und Auswertungsgleichheit sind Kriterien der internen Validität
- **Externe Validität** (Generalisierbarkeit)
  - Die Studienergebnisse lassen sich unverzerrt unter Praxisbedingungen reproduzieren

    ➞ **Übertragbarkeit der Ergebnisse**

Selektion von Studienpatienten

Coronary Artery Study

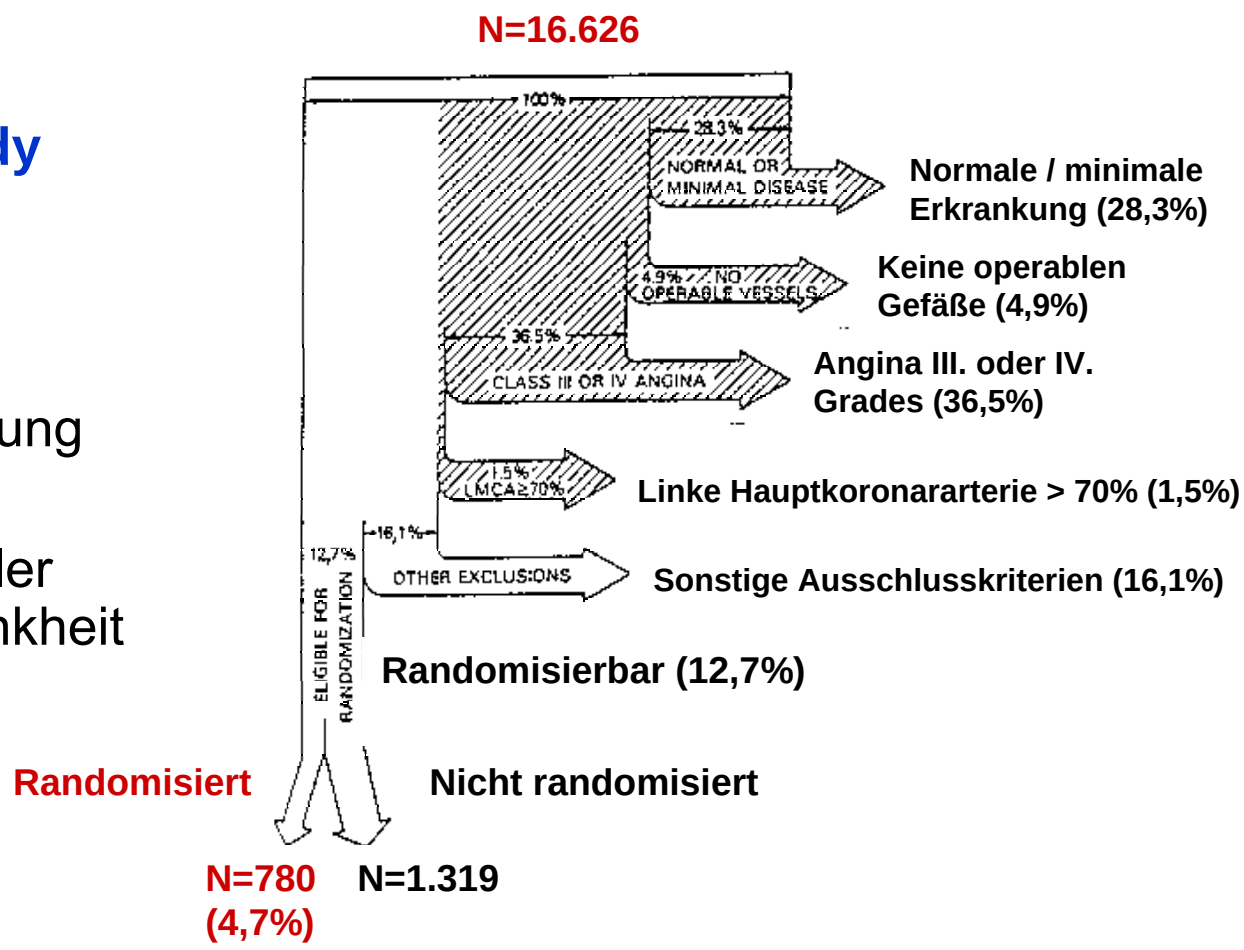
Bypass Operation

vs

Nichtinvasive Behandlung

bei Patienten mit stabiler ischämischer Herzkrankheit

Circulation 68: 939-950 (1983)

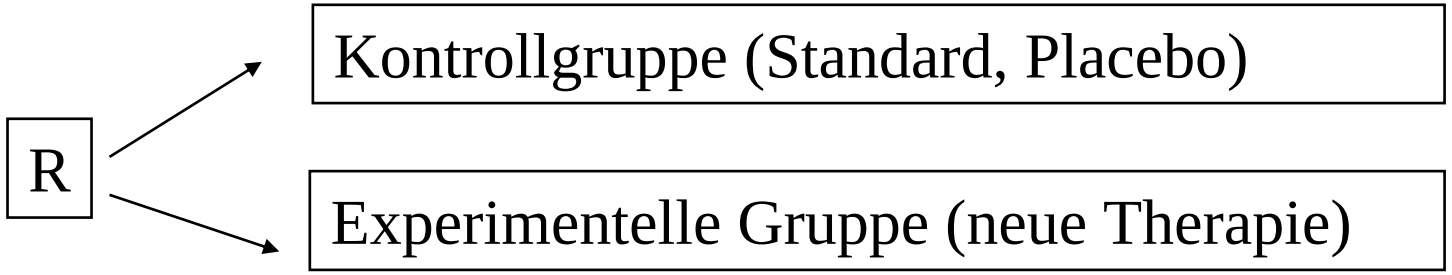


Typen von Vergleichstudien

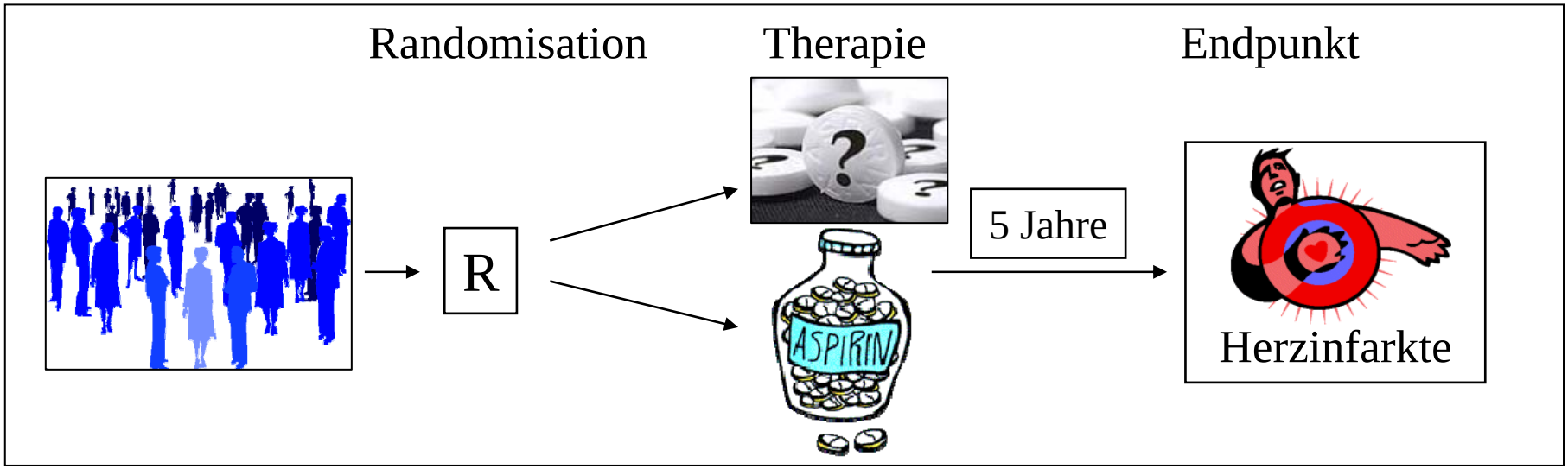
- Zwei Arme
  - Parallelgruppendesign
  - Cross-Over-Design

- Mehrere Arme (Vorsicht Multiplizitätsproblem)
  - Mehrere Parallelgruppen(3-armig, 4-armig)
  - Faktorielles Design
  - Orthogonales Design

Parallelgruppendesign

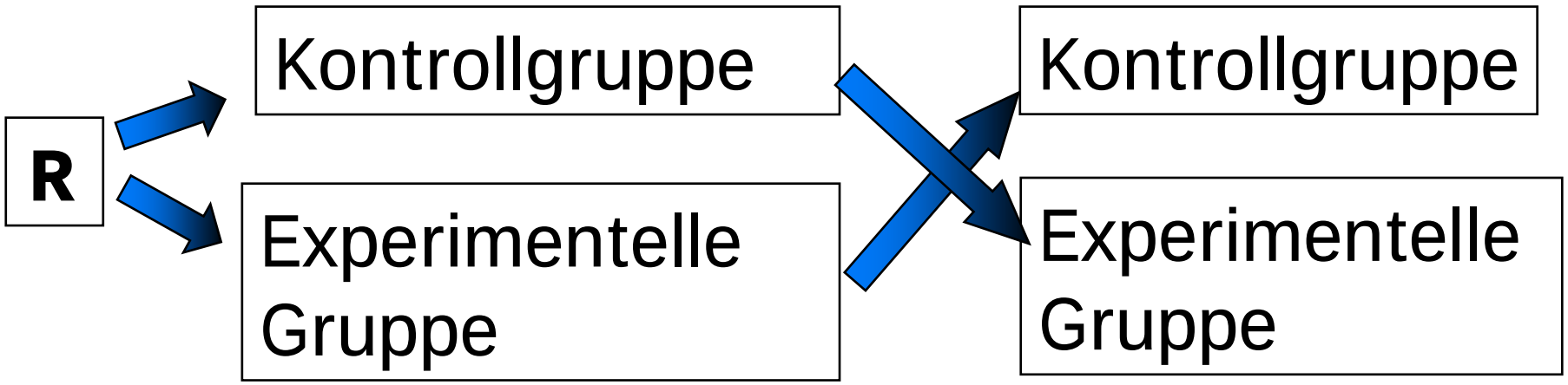


Beispiel: Physician Health Study (1989)



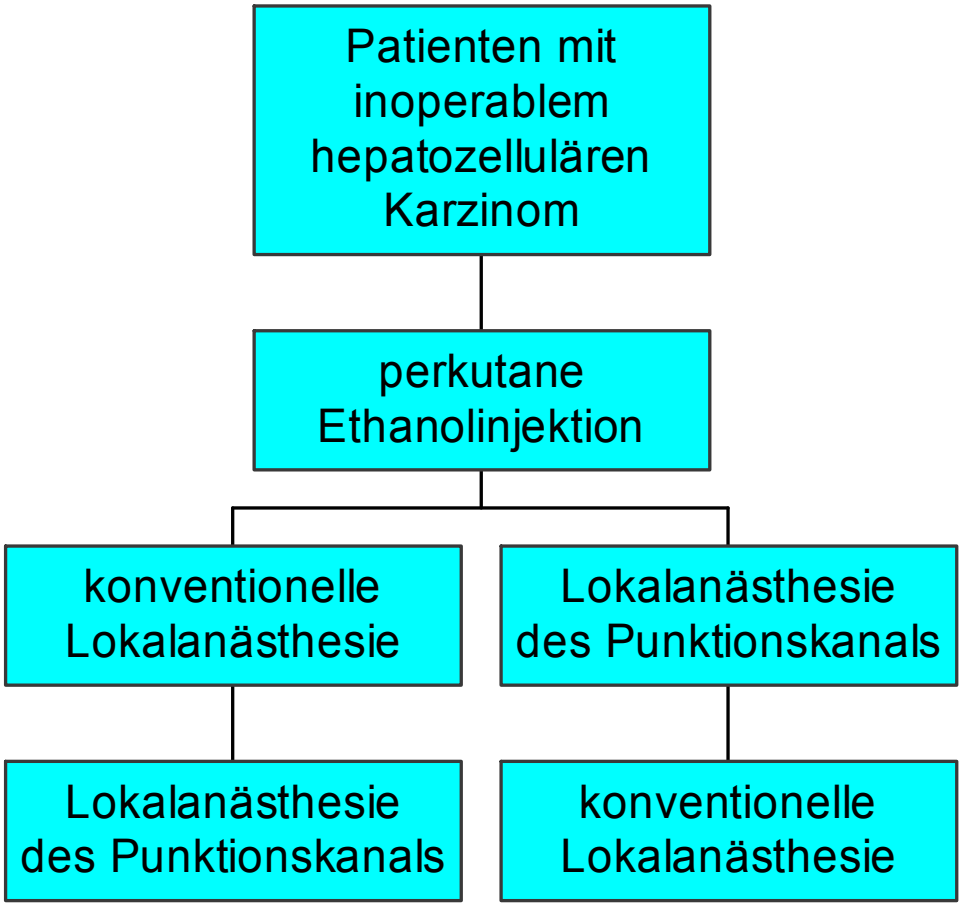
Cross – Over - Design

Prospektive kontrollierte klinische Studie mit sequentiellen Vergleichsgruppen:



Cross-Over-Design

Beispiel:



Studie des KKSL

Cross-Over-Design

Vorteile:

- keine Imbalanzen prognostischer Faktoren zwischen den Therapiegruppen
- Intraindividueller Vergleich möglich
- geringere Patientenzahl erforderlich
- höhere Bereitschaft zur Randomisation

Nachteil:

- Therapieabbrüche wegen Nebenwirkungen oder mangelnder Compliance nach der ersten Therapie führen zu Informationsverlust

Cross-Over-Design

Zu beachten

- Therapieerfolg muss unmittelbar nach Therapieende feststellbar sein
- wird typischerweise bei chronischen Erkrankungen angewendet
- Benutzung von wash-out Perioden, um ggf. den carry-over Effekt (Überhangeffekt) zu minimieren
- Kurze Behandlungsperioden, um Therapieabbrüche zu verhindern

Fallzahlreduktion: Innersubjekt-Design

- Messwert = Populations-mittelwert + reproduzierbarer individueller Wert + nicht reproduzierbarer Fehler

  z.B.: Therapieeffekt allg. Compliance „Tagesform“

  ←——— „Rauschen“ ———→

- Kann man jeden Patienten einmal mit Therapie A und einmal mit Therapie B versorgen, so verschwindet bei der Differenzbildung für die Therapieergebnisse (Messwert A minus Messwert B) die reproduzierbare Fehlerkomponente.

- **„Rauschen“ kleiner → benötigte Fallzahl kleiner**
- **wenn möglich, Innersubjekt-Design (z.B. Cross-over)**

Multiplizitätsprobleme

- Das **Multiplizitätsproblem** tritt immer dann auf, wenn man
  - innerhalb einer Studie oder einem Experiment mehrere Tests durchführt,
  - man aber den Fehler erster Art über alle geplanten Tests,
    - d.h. die Wahrscheinlichkeit in der Studie insgesamt eine oder mehrere falsch positive Testentscheidungen zu treffen,

  zum Signifikanzniveau α kontrollieren möchte.

- Beispiele:
  - Hochdimensionale Daten:
    - Selektionsverfahren (vgl. Vorlesung Hochdimensionale Daten)
  - Testen auf akkumulierenden Daten
    - Sequenzialpläne
  - Studien mit mehreren Vergleichsgruppen
    - Abschlusstestprinzip, Hierarchisierung von Kontrasten

Multiplizitätproblem bei mehrarmige Studien I

- Bei 3-armigen Studien mit den Armen A,B,C kann man drei Paarvergleiche anstellen: A-B, A-C, B-C
- Bei 4-armigen Studien sind es schon 6 Möglichkeiten:
  A-B, A-C, A-D, B-C, B-D, C-D

- Diese Tests sind nicht unabhängig, da die Daten der Arme mehrfach verwendet werden.

- Will man für das Gesamtexperiment maximal 5% Verfahrens-wahrscheinlichkeit für eine (oder mehrere) falsch-positive Entscheidung haben, darf man nicht einfach 3 bzw. 6 Zweigruppentests durchführen.

Multiplizitätproblem bei mehrarmige Studien II

- Mögliches Ziel: Aussagen zu alle Paarvergleichen
  mit Kontrolle des Fehlers erster Art
  für das Gesamtexperiment.

- Abschluss-Testprinzip für dreiarmige Studien:
  - Teste Globale Nullhypothese zum Niveau α : A=B=C
    - Falls nicht signifikant: STOP
    - Falls signifikant: Teste: A=B, A=C, B=C
      zum nominalen Niveau α
  - Satz: Mit diesem Verfahren gilt Pr(≥ 1 falsch positive Entscheidung) ≤ α

Multiplizitätproblem bei mehrarmige Studien III

- Abschluss-Testprinzip für vierarmige Studien:
  - Teste Globale Nullhypothese zum Niveau α : A=B=C=D
    - Falls nicht signifikant: STOP
    - Falls signifikant: Teste alle vier Dreier-Kombinationen:
      - A=B=C, A=B=D, A=C=D, B=C=D zum nominalen Niveau α
      - Falls alle nicht signifikant: STOP
      - Falls signifikant:
        - Teste zum nominalen Niveau α nur solche Zweierkombinationen, für die alle Dreierkombinationen, in denen die beiden Arme enthalten sind, signifikant waren.
  - Satz: Mit diesem Verfahren gilt Pr(≥ 1 falsch positive Entscheidung) ≤ α

Mehrarmigkeit erfordert Fallzahlaufschlag

- Welche Fallzahlen N braucht man für das Abschlusstestprinzip um in allen Paarvergleichen eine vorgegebene Power zu haben?

- Im Allgemeinen schwierig, hängt von spezieller Konstellation der spezifischen Alternativhypothesen ab. Simulation!

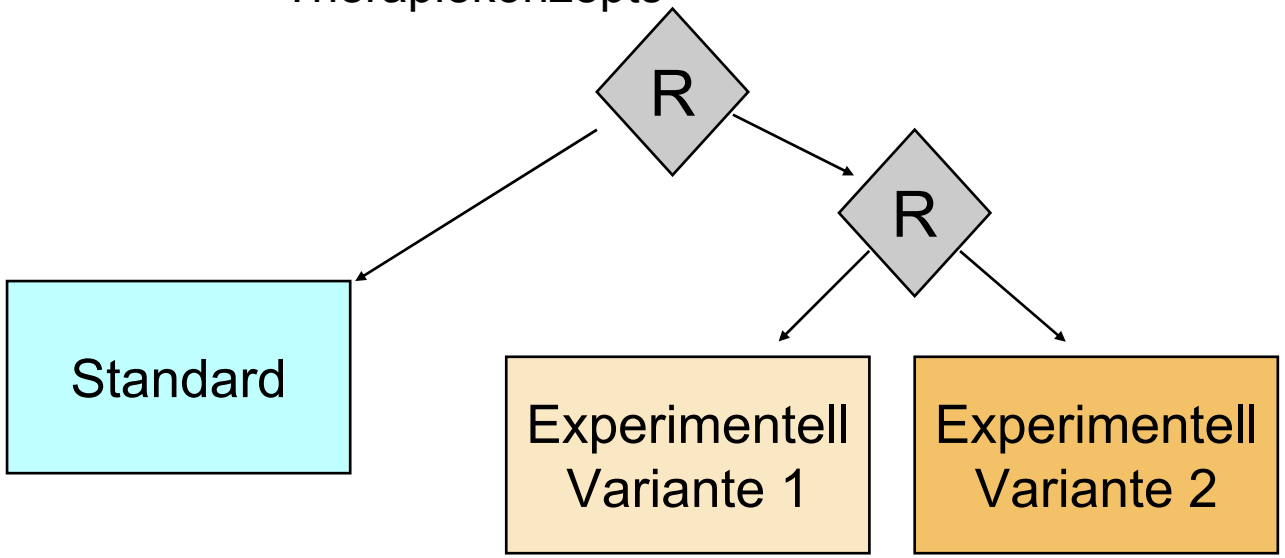
- **Grobe Faustregel:**
  - **Plane Zweigruppenvergleiche und bestimme nötige Fallzahl n pro Arm:**
  - **K-armige Studie:**

$$N = k \cdot n \cdot \sqrt{k-1}$$

Aufschlag für Multiplizität

Mehrarmige Studien, bei denen nicht alle Paarvergleiche interessieren

- Interessieren nicht alle Paarvergleiche, kann man die Anzahl der Tests reduzieren:
- Orthogonales Design:
  - Frage 1: Prinzipielle Wirksamkeit eines neuen Therapiekonzepts
    - Standard versus Experimentell (gepoolt)
  - Frage 2: Zwei Versionen der Implementation des neuen Therapiekonzepts



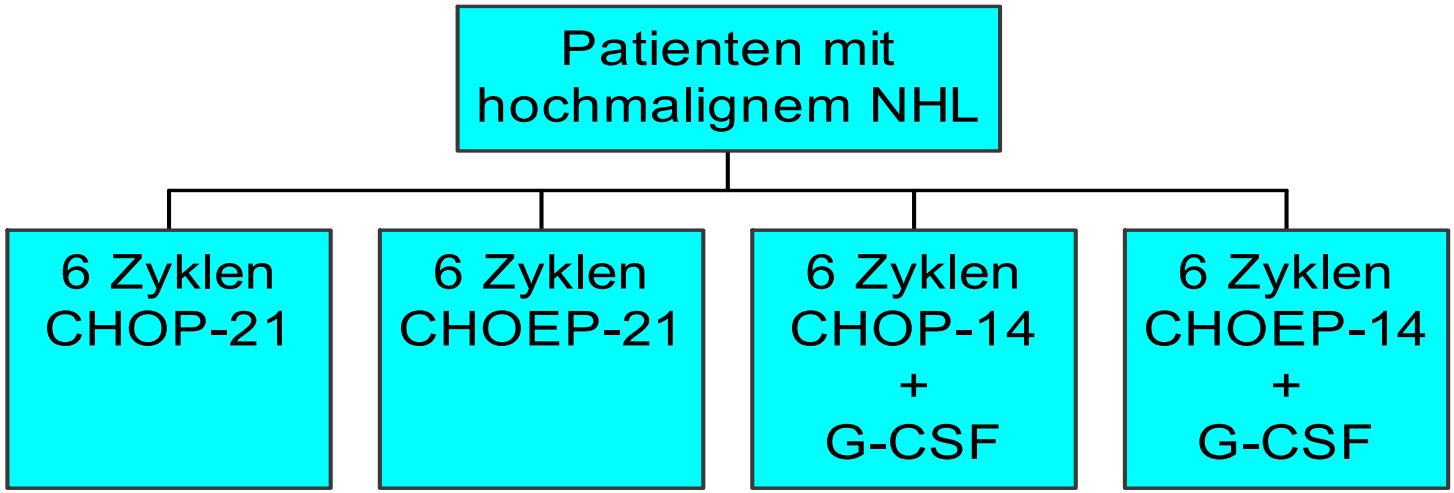
Randomisation 1:1:1

1:2 hat fast soviel Power wie 3/2 : 3/2

Zusatzfrage 2 wird nur getestet, wenn Vergleich 1 signifikant.
So keine Multiplzitätskorrektur nötig

Faktorielles Design

Beispiel: (2x2 faktorielles Design)



Studie der Deutschen Studiengruppe Hochmaligne Non-Hodgkin Lymphome

Faktorielles Design

Gestattet die gleichzeitige Beantwortung von 2 Fragen

1. Frage:

Führt die Hinzunahme von Etoposid zu
einer Verbesserung der Therapieergebnisse ?
CHOP-21 / CHOP-14 vs. CHOEP-21 / CHOEP-14

2. Frage:

Führt die Verkürzung der Therapieintervalle zu einer Verbesserung der Therapieergebnisse?

CHOP-14 / CHOEP-14 vs. CHOP-21 / CHOEP-21

Faktorielles Design

Vorteil:

- Halbierung der Fallzahl im Vergleich zum Parallelgruppendesign, wenn es keine Therapieinteraktionen gibt

- ermöglicht die Analyse von Therapieinteraktionen (z. B. Synergieeffekte) zwischen den Therapien

Faktorielles Design

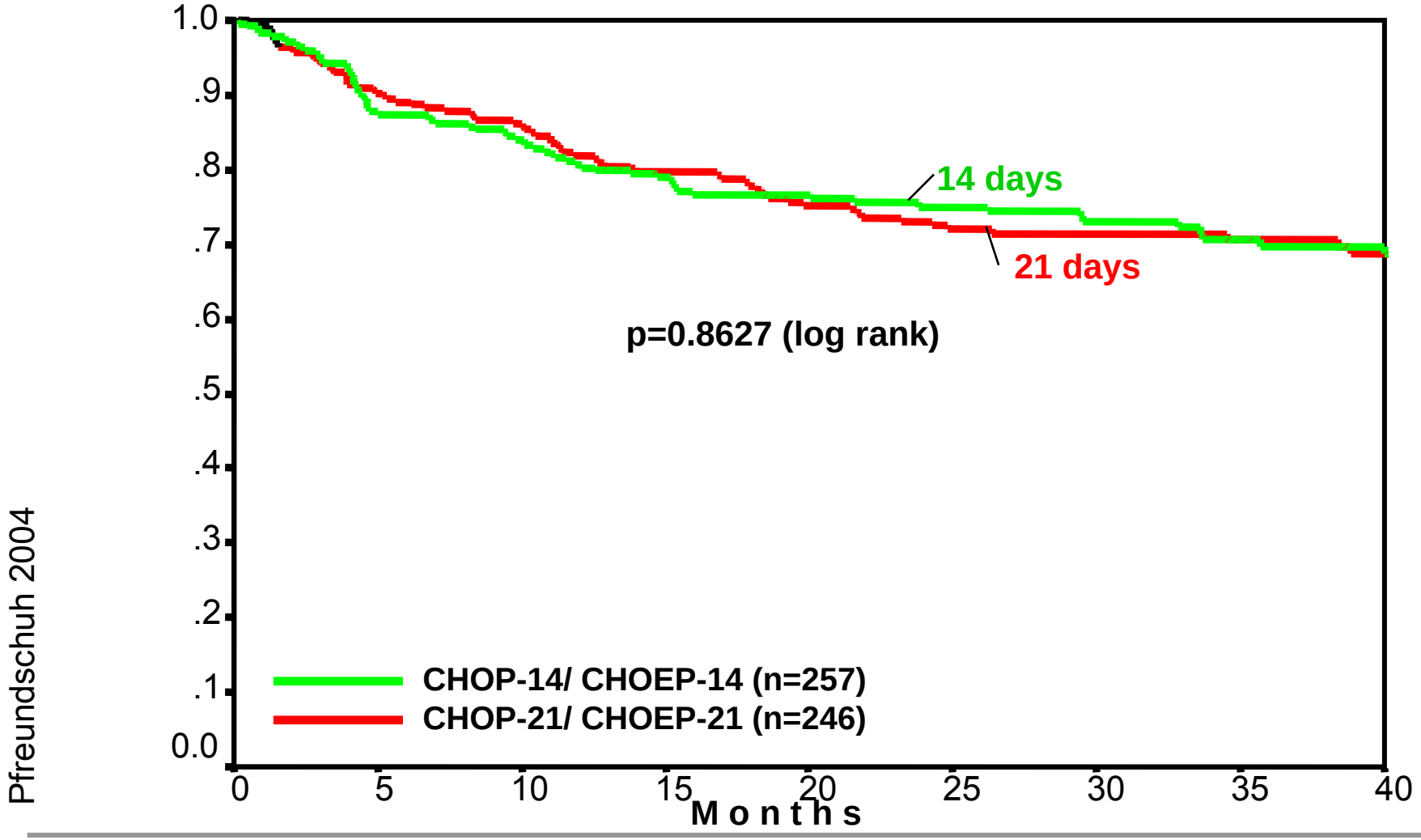
zu beachten:

- es muss ethisch vertretbar sein weder A noch B zu geben
- die Kombination von A und B muss mit vertretbaren Nebenwirkungen gegeben werden können
- eine erwartete Interaktion zwischen den beiden Therapiestrategien sollte gering sein

  (Fallzahl-relevant)

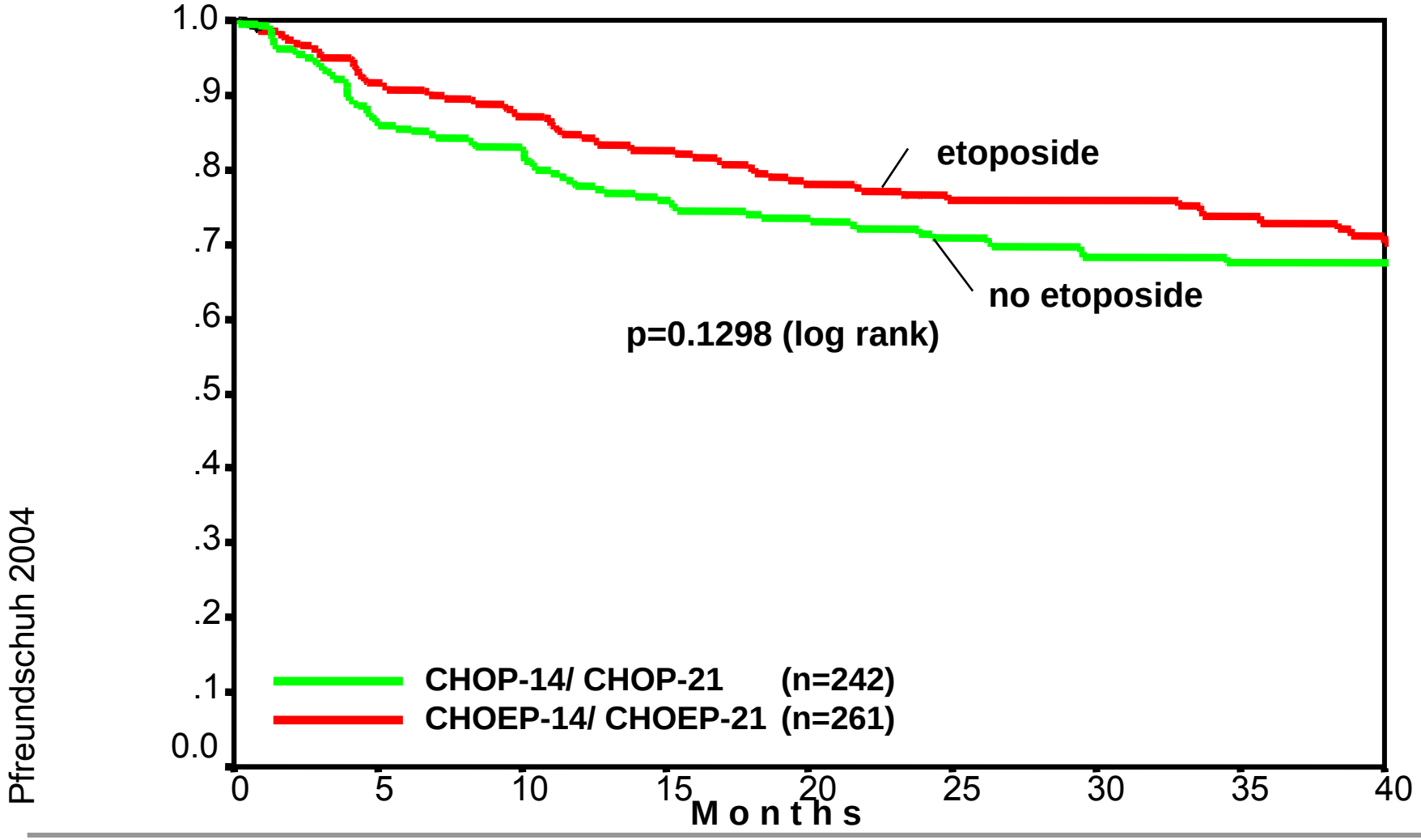
NHL-B: ≤ 60 y + LDH ≤ N

EFFECT OF TIME INTERVALS ON TTTF



NHL-B: ≤ 60 y + LDH ≤ N

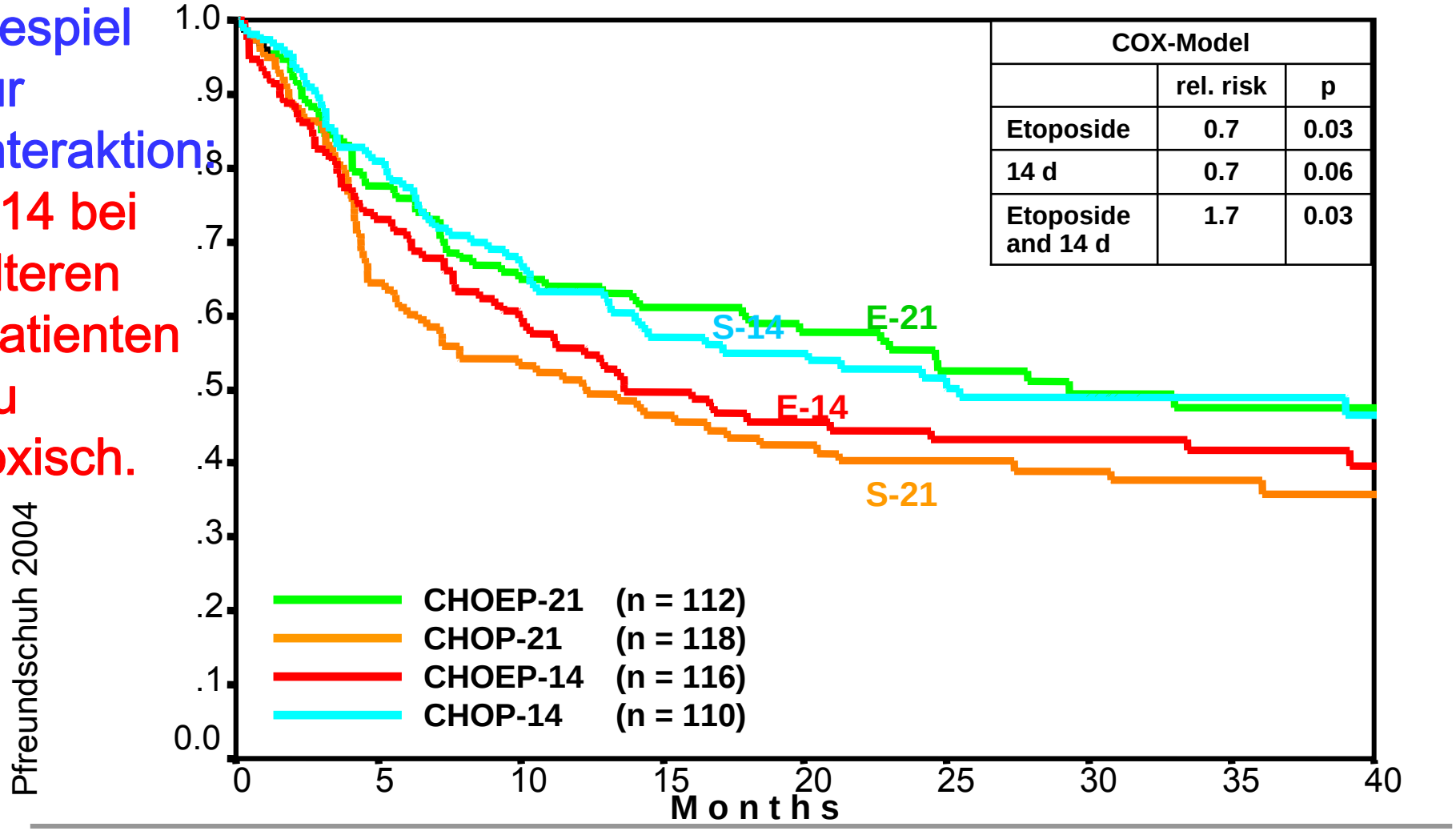
EFFECT OF ETOPOSIDE ON TTTF



NHL-B: > 60 y

TIME TO TREATMENT FAILURE - Interaction

Bespiel für Interaktion: E14 bei älteren Patienten zu toxisch.



COX-Model		
	rel. risk	p
Etoposide	0.7	0.03
14 d	0.7	0.06
Etoposide and 14 d	1.7	0.03

Typen formaler Fragestellungen

- **Differenzstudie**
  Ziel: **Aufdecken** eines **relevanten Unterschieds** durch einen Test

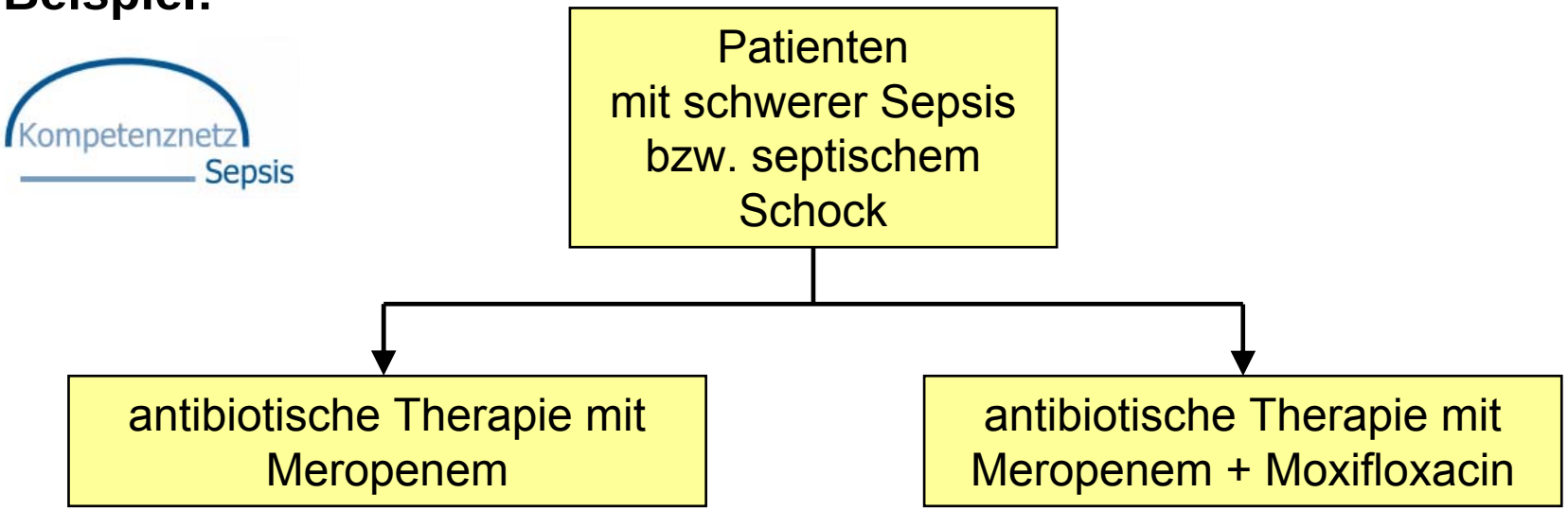
- **Äquivalenzstudie**
  Ziel: **Ausschließen** des Vorhandenseins **eines relevanten Unterschieds vorgegebener Größe**

- **Nicht - Unterlegenheitsstudie (Non - Inferiority)**
  Ziel: **Ausschließen** des Vorhandenseins **einer relevanten Verschlechterung vorgegebener Größe**

- **Studie zum Schätzen eines Effekts**
  Ziel: hinreichend genaue **quantitative Schätzung eines Effekts** mit Angabe eines Konfidenzintervalls

Differenzstudie

Beispiel:

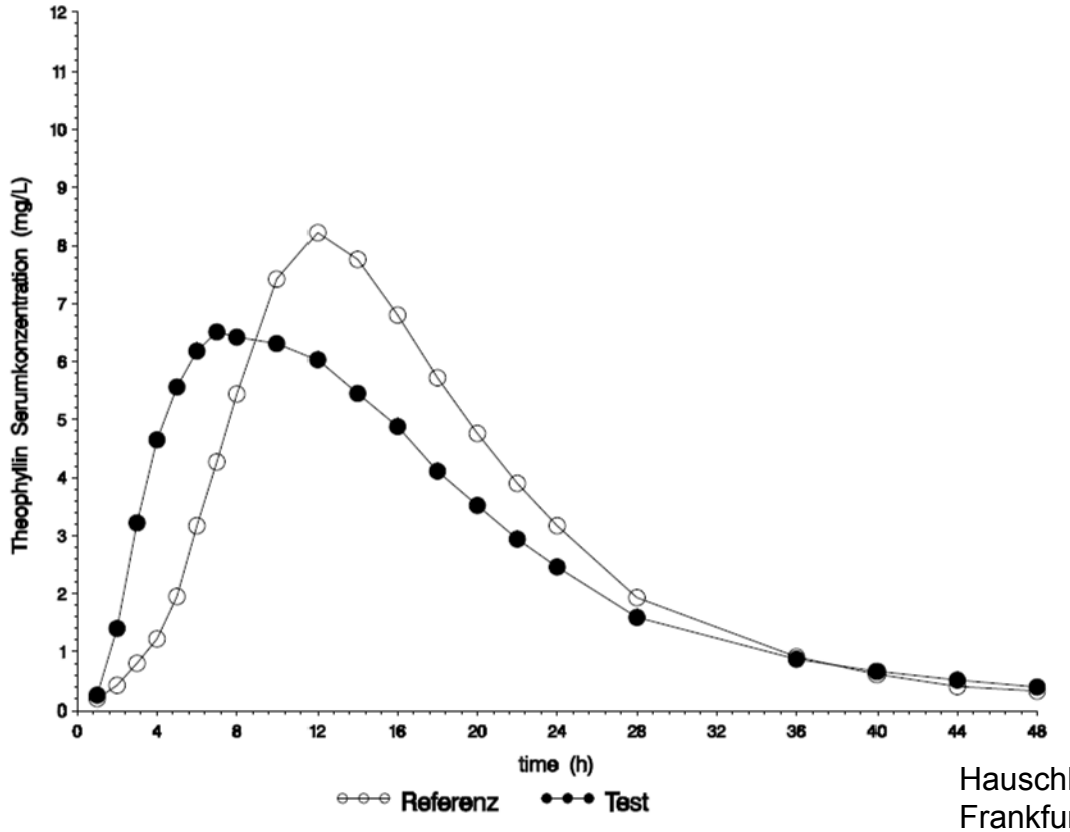


Primäre Fragestellung:

Kann durch eine Kombinationstherapie mit Meropenem plus Moxifloxacin gegenüber einer Monotherapie mit Meropenem der Mittelwert des mittleren totalen SOFA-Scores (über die Studiendauer, maximal jedoch 21 Tage) gesenkt werden

Äquivalenzstudie

Beispiel:



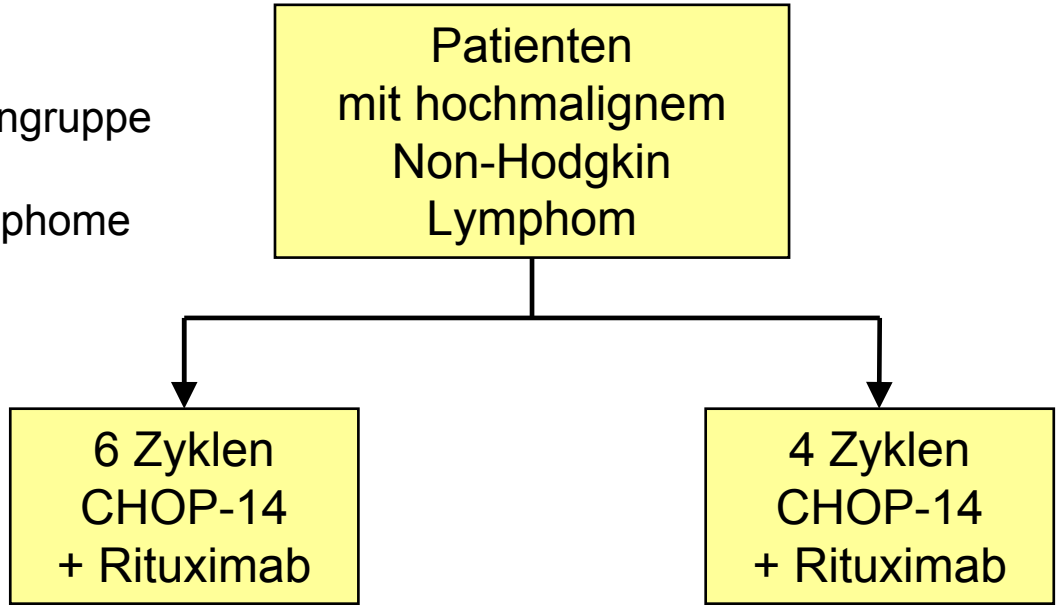
Hauschke, D; Vortrag
Frankfurt 5.9.2007

Ziel: Prüfung auf Bioäquivalenz

(z.B. Vergleich der area under the curve (AUC)

Nicht-Unterlegenheitsstudie

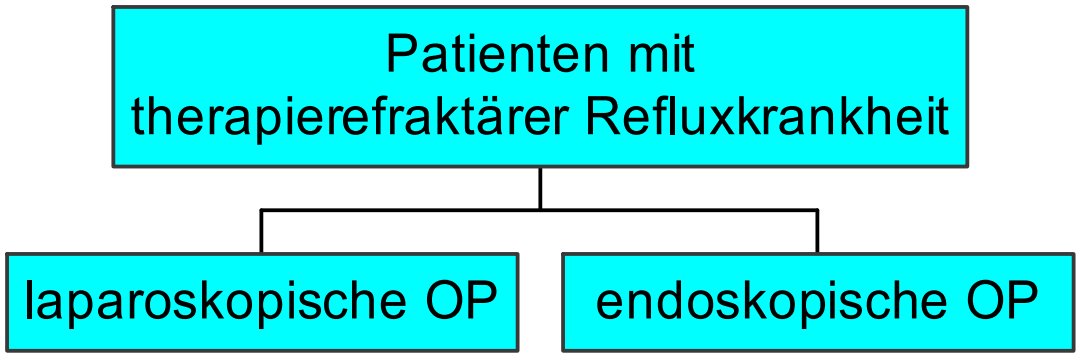
Beispiel:
Studie der
Deutschen Studiengruppe
Hochmaligne
Non-Hodgkin Lymphome



Ziel: Ausschließen des Vorhandenseins einer klinisch relevanten Verschlechterung im 3-Jahres EFS (Event-free Survival) in der Größenordnung von 5 %.

Studie zum Schätzen

Beispiel:

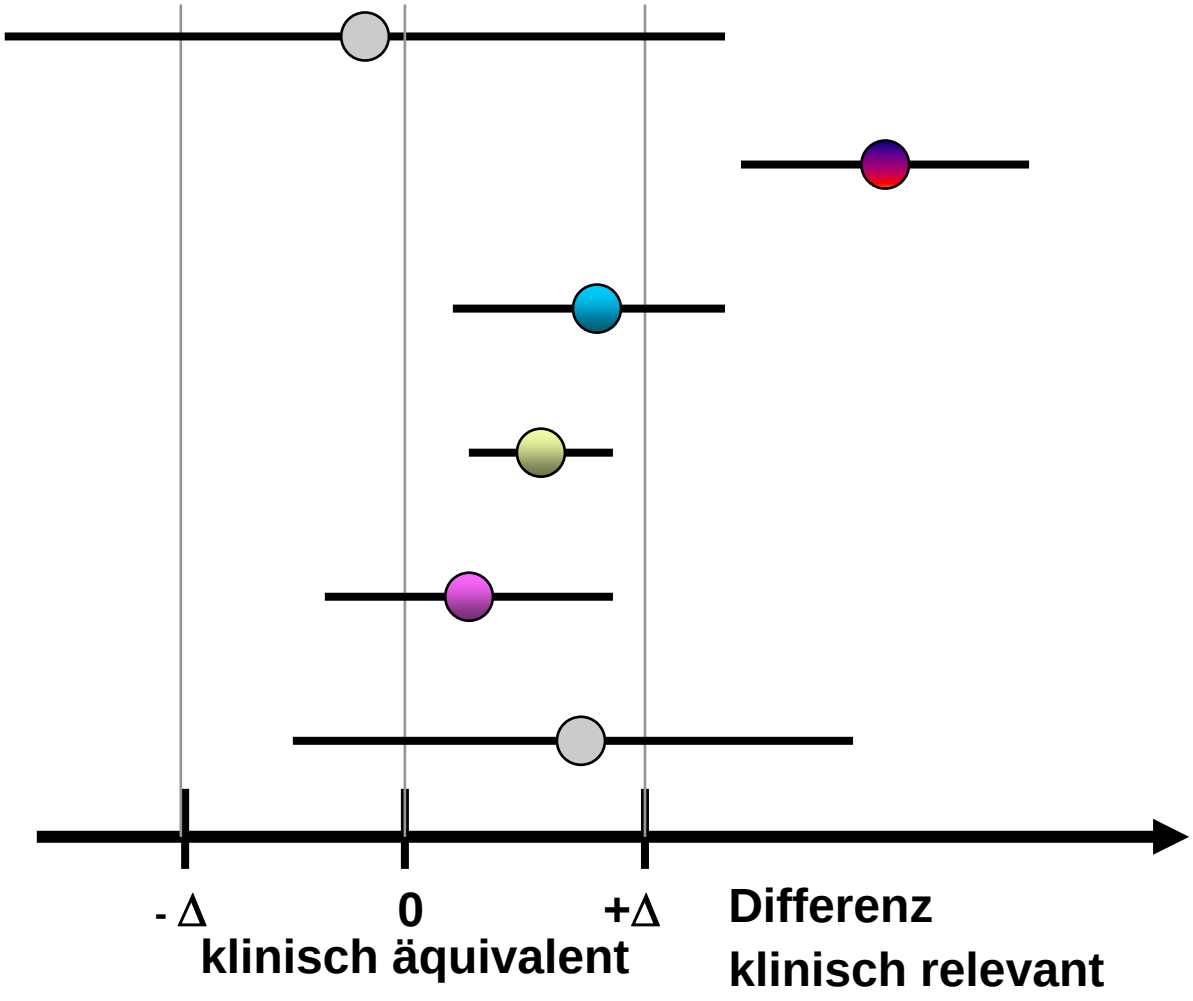


Studie des KKSL

Fragen:
A) Welche Erfolgsrate hat die laparoskopische OP ?
B) Um wieviel ist die laparoskopische OP schlechter?

Ziel: Schätzung der Erfolgswahrscheinlichkeit der endoskopischen OP mit einer Genauigkeit von 10% und der Differenz der Erfolgswahrscheinlichkeit beider Therapien mit ca. 13%.

Studienergebnis – Vergleich zweier Therapien



inkonklusiv

relevant besser

besser
unklar, ob relevant besser

besser,
aber nicht relevant besser

weder relevant besser
noch relevant schlechter

nicht relevant schlechter