

# Universitätslehrgänge „Clinical Research“ Epidemiologische Grundlagen



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

---

ao. Univ.-Prof. Mag. Dr. Hanno Ulmer  
*hanno.ulmer@i-med.ac.at*

---

Department für Medizinische Statistik, Informatik und  
Gesundheitsökonomie, Medizinische Universität Innsbruck



# Inhalte des Seminars

---

- Epidemiologie
  - epidemiologische Maßzahlen
  - Altersstandardisierung
  - Studientypen
- Biostatistische Methoden
  - deskriptive Statistik
  - schließende Statistik
  - Regressionsanalyse
- Bias und Confounding
- Literatur lesen

- Übung zur deskriptiven Statistik
- Berechnen Sie das relative Risiko
- Führen Sie einen Chi-Quadrat Test durch
- Epidemiologische Studie lesen und mittels STROBE Checkliste prüfen

# Grundlagen der Epidemiologie



---

MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

---

Hanno Ulmer

---

Department für Medizinische Statistik, Informatik und  
Gesundheitsökonomie, Medizinische Universität Innsbruck





# Epidemiologie, Definition

- Die **Epidemiologie** (von [griech.](#) *epi* „auf, über“, *demos* „[Volk](#)“, *logos* „[Lehre](#)“, ursprünglich: "Seuchenkunde") ist jene wissenschaftliche Disziplin, die sich mit den Ursachen und Folgen sowie der Verbreitung von gesundheitsbezogenen Zuständen und Ereignissen in [Populationen](#) beschäftigt. Die Epidemiologie untersucht somit jene Faktoren, die zu Gesundheit und [Krankheit](#) von Individuen und Populationen beitragen und ist deshalb die Basis aller Maßnahmen, die im Interesse der [Volksgesundheit](#) unternommen werden.
- Im Gegensatz dazu kümmert sich die [Medizin](#) darum, dem einzelnen Menschen in einem konkreten Krankheitsfall zu helfen.

# Bereiche/Teilgebiete der Epidemiologie



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

Infektionsepidemiologie, Epidemiologie allergischer und dermatologischer Erkrankungen, Epidemiologie der Arbeitswelt, Epidemiologische Methoden, Ernährungsepidemiologie, Genetische Epidemiologie, Herz-Kreislauf-Epidemiologie, Krebs epidemiologie, etc.



# Schlüsselfragen der Epidemiologie

---

- Was?
  - Um welches Gesundheitsproblem geht es? Spezifizierung
- Wann?
  - Zu welchem Zeitpunkt oder in welchem Zeitraum?
- Wo?
  - An welchen Orten tritt das Problem auf?
- Wer?
  - Wer ist von dem Problem betroffen? Geschlecht, Alter, Sozialstatus...
- Warum?
  - Welche Ursachen gibt es für das Problem?

# Herz-Kreislauf- Epidemiologie: Beispiel Schlaganfall



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

---

Hanno Ulmer

---

Department für Medizinische Statistik, Informatik und  
Gesundheitsökonomie, Medizinische Universität Innsbruck

# Definition des Schlaganfalls



Allgemein

**Akutes fokales neurologisches Defizit auf Grund  
eines umschriebenen Durchblutungsmangels  
oder einer Blutung des Gehirns**

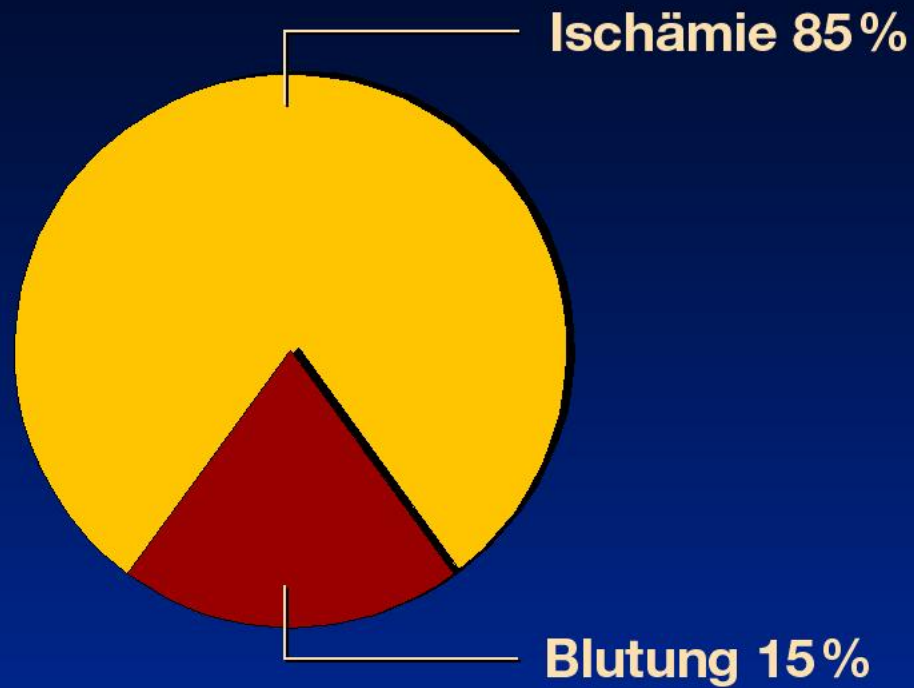
# Definition des Schlaganfalls



## Transitorische Ischämische Attacke

**Die TIA ist ein reversibler ischämischer Schlaganfall, gekennzeichnet durch schlagartig auftretendes fokales neurologisches Defizit mit vollständiger Rückbildung in Minuten bis wenigen Stunden, spätestens innerhalb von 24 Stunden**

# Schlaganfall



# Epidemiologie



## Prävalenz, Inzidenz und Letalität des Schlaganfalls in Deutschland

<b>Prävalenz</b>
<b>400.000 - 600.000</b>
<b>(600 - 700/10.000 Einw.)*</b>

<b>Inzidenz</b>
<b>120.000 - 200.000</b>
<b>(15 - 25/ 10.000/Jahr)</b>

\* > 65 Jahre

<b>Letalität**</b>
<b>15.000 - 30.000/Jahr</b>
<b>(10 - 20%)</b>

\*\* innerh. v. 30 Tagen im Krankenhaus



# Epidemiologie



Risikofaktor	Ischämie	Blutung	
		intrazerebral	subarachnoidal
Hypertonie	++	++	+
Zigarettenrauchen	++	+ -	++
Diabetes mellitus	++	0	0
Alkoholabusus	+ -	++	+
Fettstoff- wechselstörung	+	0	0
Herzkrankheiten	++	0	0

++ enge Beziehung

+ mäßige Beziehung

+ - Beziehung nicht eindeutig

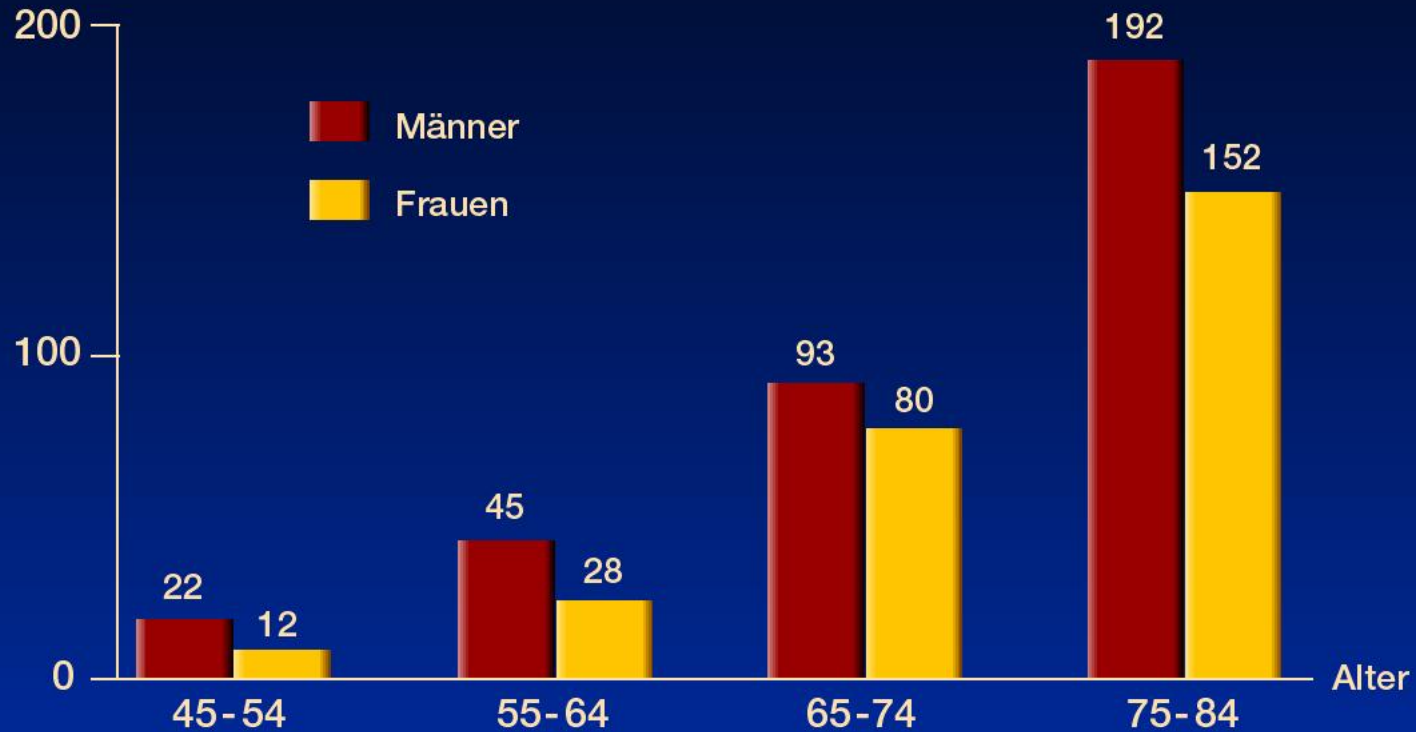
0 keine Beziehung

# Epidemiologie

## Inzidenz des ersten Schlaganfalls\*



pro 10.000 Pers./Jahr



\* einschließlich TIA's

Framingham-Studie

# Epidemiologie



<b>Risikofaktor</b>	<b>Prävalenz</b>
<b>Hypertonie</b>	<b>25 – 40 %</b>
<b>Diabetes mellitus</b>	<b>4 – 8 %</b>
<b>Zigarettenrauchen</b>	<b>20 – 40 %</b>
<b>Alkoholabusus</b>	<b>6 – 30 %</b>
<b>Fettstoff- wechselstörung</b>	<b>5 – 30 %</b>
<b>Herzkrankheiten</b>	<b>10 – 20 %</b>
<b>Vorhofflimmern</b>	<b>1 %</b>

# Epidemiologie



<b>Risikofaktor</b>	<b>Relatives Risiko</b>
<b>Hypertonie</b>	<b>bis 5 fach</b>
<b>Diabetes mellitus</b>	<b>bis 3 fach</b>
<b>Zigarettenrauchen</b>	<b>bis 2 fach</b>
<b>Alkoholabusus</b>	<b>bis 4 fach *</b>
<b>Fettstoff- wechselstörung</b>	<b>bis 3 fach</b>
<b>Herzkrankheiten</b>	<b>bis 4 fach</b>
<b>Vorhofflimmern</b>	<b>bis 17 fach **</b>

\* unter Einschluß der Hirnblutung

\*\* bei Kombination

# Grundbegriffe der Epidemiologie



---

MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

---

Hanno Ulmer

---

Department für Medizinische Statistik, Informatik und  
Gesundheitsökonomie, Medizinische Universität Innsbruck

# Epidemiologische Maßzahlen



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

Inzidenz: Maß für die Anzahl der Neuerkrankungen in einem definierten Zeitraum

Prävalenz: Maß für die Anzahl von Erkrankten zu einem definierten Zeitpunkt

Mortalität: Maß für die Anzahl der Todesfälle

Altersstandardisierung

Inzidenz = Anzahl der Neuerkrankungen im Beobachtungszeitraum / Anzahl der Personen unter Risiko (zu Beginn des Zeitraums)

Beispiel: In einer Stadt leben 100.000 Frauen. Aktuell leiden 800 von ihnen an Brustkrebs. Von den anderen 99.200 erkranken im Laufe eines Jahres 110 an Brustkrebs.

Inzidenz =  $110/99.200 = 0,001109$  oder 110,9 pro 100,000 Frauen

# Epidemiologische Maßzahlen



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

Prävalenz = Anzahl der Erkrankungsfälle in der  
Bevölkerung / Bevölkerungsumfang

Beispiel: Aktuell sind 800 von 100.000 Frauen an  
Brustkrebs erkrankt.

Prävalenz =  $800/100.000 = 0,008$  oder 800 pro  
100,000 Frauen

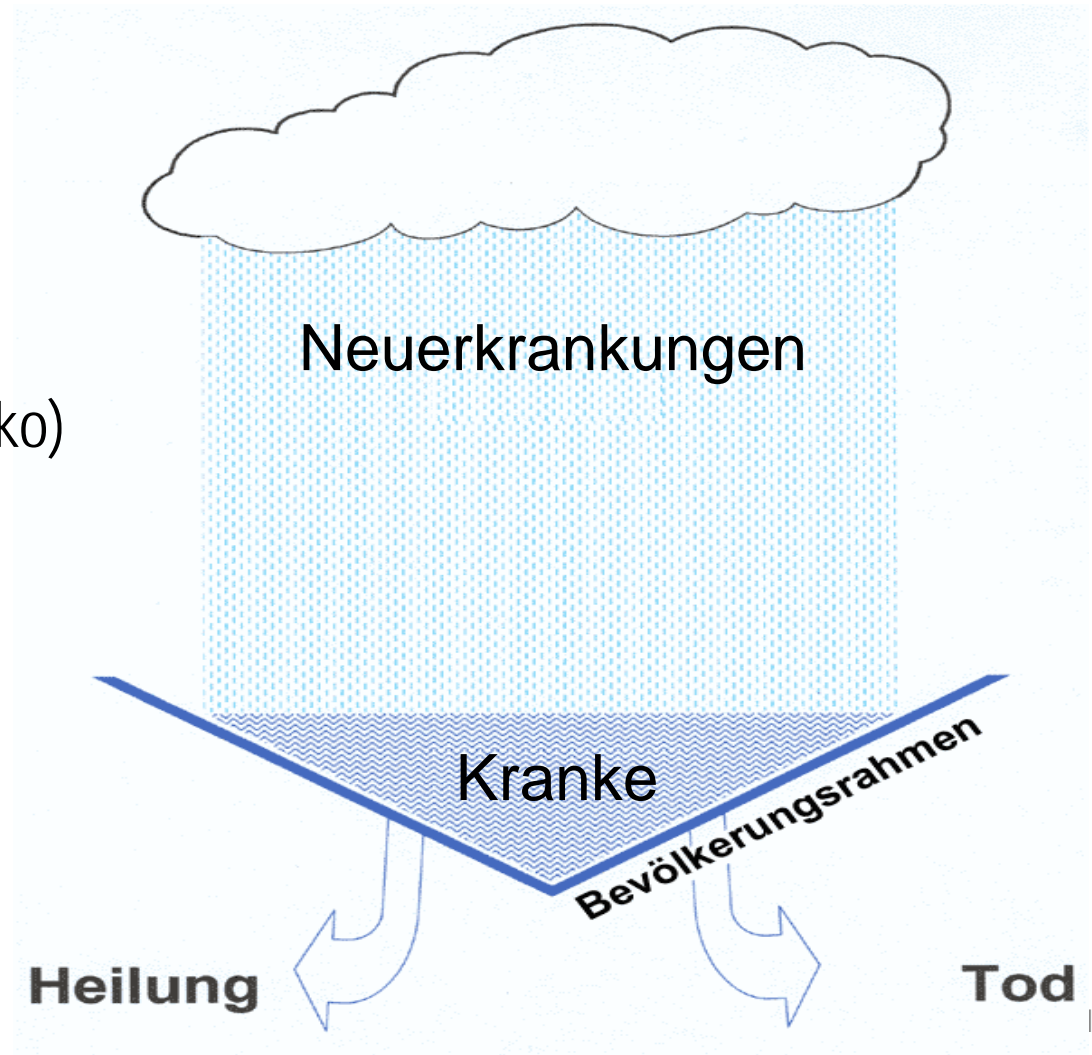
Prävalenz = Inzidenz x Krankheitsdauer



# Prävalenz versus Inzidenz

Inzidenz  
(absolutes Risiko)

Prävalenz



# Epidemiologische Maßzahlen



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

Gesamtmortalität = Anzahl der Todesfälle in einem Zeitraum / Bevölkerungsumfang

Beispiel: Im Laufe eines Jahres versterben in der Beispielstadt 100 Frauen.

Gesamtmortalität =  $100/100.000 = 0,001$  oder 100 pro 100,000 Frauen



# Epidemiologische Maßzahlen

---

Ursachenspezifische Mortalität = Anzahl der Todesfälle nach Ursache in einem Zeitraum / Bevölkerungsumfang

Beispiel: Von den 100 Todesfällen sind 40 auf Brustkrebs zurückzuführen. Somit beträgt die

Brustkrebsmortalität =  $40/100.000 = 0,0004$  oder 40 von 100.000 Frauen.

Altersspezifische Mortalität = Anzahl der Todesfälle in einer bestimmten Altersklasse / Bevölkerungsumfang in dieser Altersklasse

Beispiel: In der betrachteten Stadt sind 16.000 Frauen zwischen 55 und 60 Jahre alt. In dieser Altersklasse versterben im Laufe des Jahres 10 Frauen.

Altersspezifische Mortalität =  $10/16.000 = 0,000625$   
oder 63 von 100.000 Frauen.

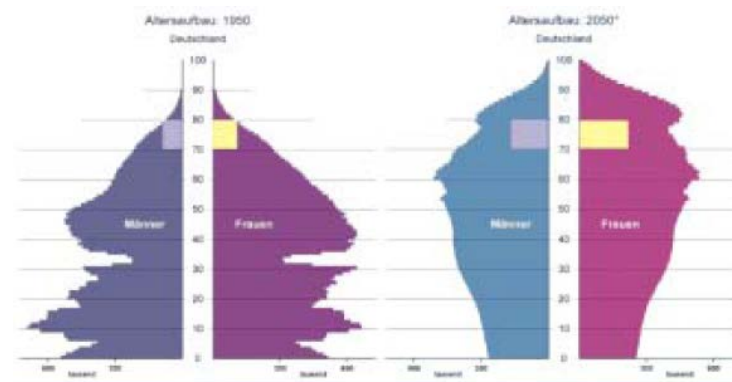
Letalität („Tödlichkeit einer Erkrankung“) = Anzahl der Todesfälle nach Ursache in einem Zeitraum / Anzahl der Neuerkrankungen an dieser Ursache im selben Zeitraum

Beispiel: Von den 110 neu an Brustkrebs erkrankten Frauen sterben im Laufe des Jahres 10 Frauen.

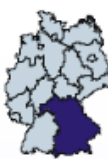
Letalität =  $10/110 = 0,091$  oder 9,1%

# Altersstandardisierung

Ermöglicht den Vergleich von Bevölkerungen mit unterschiedlicher Altersstruktur, indem verzerrende Alterseinflüsse beseitigt werden.



Aufgrund des demographischen Wandels nimmt die Anzahl der über 70-jährigen stark zu. Dies muss bei temporalen Vergleichen berücksichtigt werden



## Altersstandardisierung - Beispiel

Schritt 1: Berechnung der rohen Mortalitätsraten

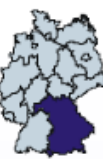
Altersklasse	Kreuzberg				Zehlfendorf			
	Population	Todesfälle			Population	Todesfälle		
0-19	34.000	34			18.000	5		
20-44	75.500	135			32.000	30		
45-64	27.000	299			29.000	165		
65 und älter	15.000	1.167			20.000	1.585		
Insgesamt	<b>151.500</b>	<b>1.635</b>			<b>99.000</b>	<b>1.785</b>		

Rohe Mortalität:

Kreuzberg  $\frac{1.635}{151.500} = 10,8 \text{ pro } 1.000 \text{ Einwohner}$

Zehlfendorf  $\frac{1.785}{99.000} = 18,0 \text{ pro } 1.000 \text{ Einwohner}$

**Beispiel zur Altersstandardisierung,  
entnommen einem Vortrag von A.  
Daugs, Tumorzentrum Erlangen-  
Nürnberg**



## Altersstandardisierung - Beispiel

Schritt 2: Berechnung der altersspezifischen Mortalitätsraten

Altersklasse	Kreuzberg			Zehlendorf		
	Population	Todesfälle	Mortalität*	Population	Todesfälle	Mortalität*
0-19	34.000	34	1,0	18.000	5	0,3
20-44	75.500	135	1,8	32.000	30	0,9
45-64	27.000	299	11,1	29.000	165	5,7
65 und älter	15.000	1.167	77,8	20.000	1.585	79,3
Insgesamt	151.500	1.635		99.000	1.758	

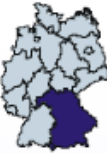
Rohe Mortalität:

$$\text{Kreuzberg } \frac{1.635}{151.500} = 10,8 \text{ pro 1.000 Einwohner}$$

$$\text{Zehlendorf } \frac{1.785}{99.000} = 18,0 \text{ pro 1.000 Einwohner}$$

\* alle Raten pro 1.000 Personen





## Altersstandardisierung - Beispiel

### Schritt 3: Wahl einer Standardpopulation

Altersklasse	Standard	Kreuzberg			Zehlendorf			
		Population	Todesfälle	Mortalität*	Population	Todesfälle	Mortalität*	
0-19	<b>385.000</b>	34.000	34	1,0	18.000	5	0,3	
20-44	<b>850.500</b>	75.500	135	1,8	32.000	30	0,9	
45-64	<b>540.000</b>	27.000	299	11,1	29.000	165	5,7	
65 und älter	<b>356.000</b>	15.000	1.167	77,8	20.000	1.585	79,3	
Insgesamt	<b>2.131.500</b>	151.500	1.635		99.000	1.758		

Rohe Mortalität:

$$\text{Kreuzberg } \frac{1.635}{151.500} = 10,8 \text{ pro 1.000 Einwohner}$$

$$\text{Zehlendorf } \frac{1.785}{99.000} = 18,0 \text{ pro 1.000 Einwohner}$$

\* alle Raten pro 1.000 Personen



## Altersstandardisierung - Beispiel

Schritt 4: Anwendung der altersspezifischen Mortalitätsraten auf die fiktive Standardpopulation

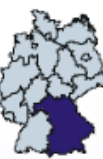
Altersklasse	Standard	Kreuzberg				Zehlendorf			
		Population	Todesfälle	Mortalität	Erw. Fälle	Population	Todesfälle	Mortalität	Erw. Fälle
0-19	385.000	34.000	34	1,0	385	18.000	5	0,3	116
20-44	850.500	75.500	135	1,8	1.531	32.000	30	0,9	765
45-64	540.000	27.000	299	11,1	5.994	29.000	165	5,7	3.078
65 und älter	356.000	15.000	1.167	77,8	27.697	20.000	1.585	79,3	28.231
Insgesamt	2.131.500	151.500	1.635			99.000	1.758		

Rohe Mortalität:

$$\text{Kreuzberg } \frac{1.635}{151.500} = 10,8 \text{ pro 1.000 Einwohner}$$

$$\text{Zehlendorf } \frac{1.785}{99.000} = 18,0 \text{ pro 1.000 Einwohner}$$

\* alle Raten pro 1.000 Personen



## Altersstandardisierung - Beispiel

Schritt 5: Berechnung der altersstandardisierten Mortalitätsraten

Altersklasse	Standard	Kreuzberg				Zehlendorf			
		Population	Todesfälle	Mortalität	Erw. Fälle	Population	Todesfälle	Mortalität	Erw. Fälle
0-19	385.000	34.000	34	1,0	385	18.000	5	0,3	116
20-44	850.500	75.500	135	1,8	1.531	32.000	30	0,9	765
45-64	540.000	27.000	299	11,1	5.994	29.000	165	5,7	3.078
65 und älter	356.000	15.000	1.167	77,8	27.697	20.000	1.585	79,3	28.231
Insgesamt	<b>2.131.500</b>	151.500	1.635	<b>16,7</b>	<b>35.607</b>	99.000	1.758	<b>15,1</b>	<b>32.190</b>

Rohe Mortalität:

$$\text{Kreuzberg } \frac{1.635}{151.500} = 10,8 \text{ pro 1.000 Einwohner}$$

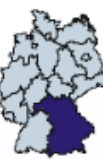
$$\text{Zehlendorf } \frac{1.785}{99.000} = 18,0 \text{ pro 1.000 Einwohner}$$

Altersstandardisierte Mortalität:

$$\text{Kreuzberg } \frac{35.607}{2.131.500} = 16,7 \text{ pro 1.000 Einwohner}$$

$$\text{Zehlendorf } \frac{32.190}{2.131.500} = 15,1 \text{ pro 1.000 Einwohner}$$

\* alle Raten pro 1.000 Personen



## Altersstandardisierung - Beispiel

Der Vergleich der altersstandardisierten mit der rohen Mortalität zeigt die verzerrenden Alterseinflüsse.

Altersklasse	Standard	Kreuzberg				Zehlendorf			
		Population	Todesfälle	Mortalität	Erw. Fälle	Population	Todesfälle	Mortalität	Erw. Fälle
0-19	385.000	34.000	34	1,0	385	18.000	5	0,3	116
20-44	850.500	75.500	135	1,8	1.531	32.000	30	0,9	765
45-64	540.000	27.000	299	11,1	5.994	29.000	165	5,7	3.078
65 und älter	356.000	15.000	1.167	77,8	27.697	20.000	1.585	79,3	28.231
Insgesamt	2.131.500	151.500	1.635	16,7	35.607	99.000	1.758	15,1	32.190

Rohe Mortalität:

$$\text{Kreuzberg } \frac{1.635}{151.500} = 10,8 \text{ pro 1.000 Einwohner}$$

$$\text{Zehlendorf } \frac{1.785}{99.000} = 18,0 \text{ pro 1.000 Einwohner}$$

Altersstandardisierte Mortalität:

$$\text{Kreuzberg } \frac{35.607}{2.131.500} = 16,7 \text{ pro 1.000 Einwohner}$$

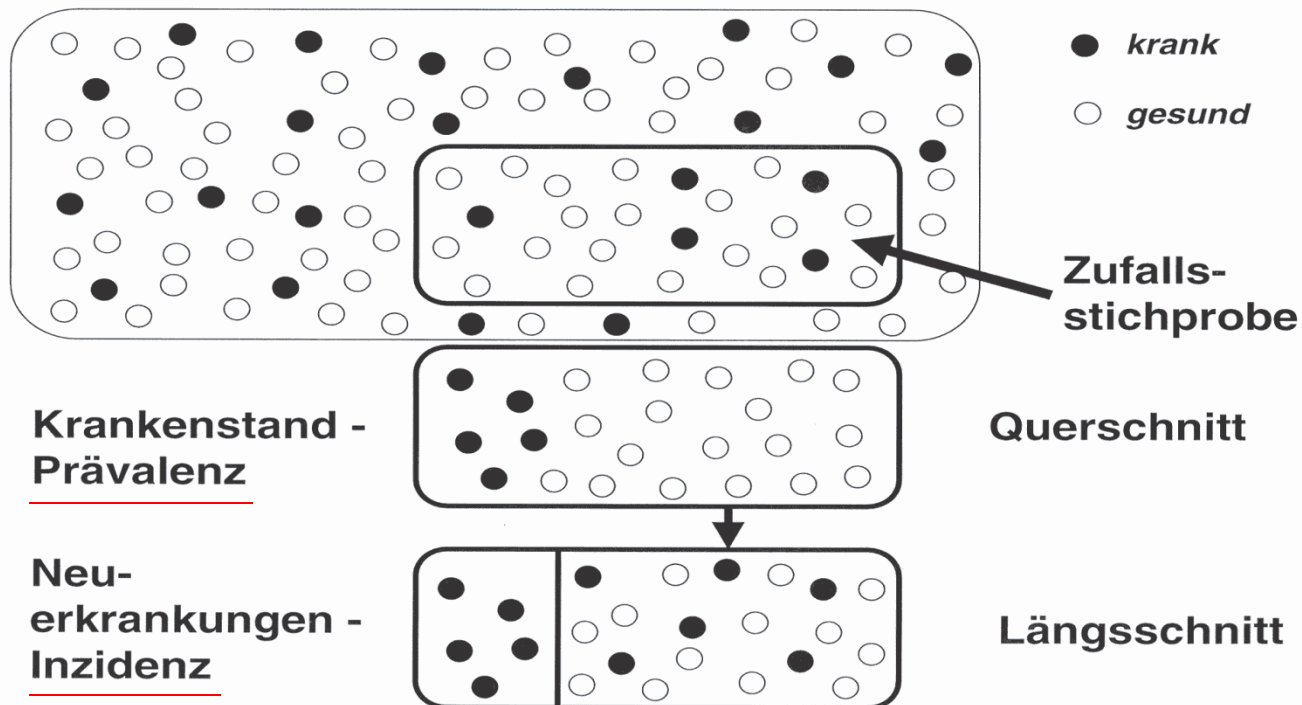
$$\text{Zehlendorf } \frac{32.190}{2.131.500} = 15,1 \text{ pro 1.000 Einwohner}$$

\* alle Raten pro 1.000 Personen

# Prävalenz/Inzidenz schätzen



## HÄUFIGKEIT VON KRANKHEITEN IN DER BEVÖLKERUNG



# Die wichtigsten Studententypen



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

---

Dr. Hanno Ulmer

[hanno.ulmer@i-med.ac.at](mailto:hanno.ulmer@i-med.ac.at)

---

Department für Medizinische Statistik, Informatik und  
Gesundheitsökonomie, Medizinische Universität Innsbruck

# EPI DEMOS – ‚Was auf dem Volke liegt‘



MEDIZINISCHE UNIVERSITÄT

## Gesundheitsproblem beschreiben

- Kasuistik
- Fallserie
- Survey
- Register

↓  
**quantifizieren**



↓  
**analysieren**

- Kontrollierte Studie
- RCT



## Fazit für die Praxis

- Anwendungsbeobachtung
- Meta-Analyse, system. Review
- Leitlinien-, HTA-Bericht





# Welche Studientypen kennen Sie?

---

## **Querschnittstudie**

Klinische Prüfung/RCT

Prävalenzstudie

## **Kohortenstudie**

Fall-Kontroll-Studie

## **Ökologische Studie**

Kasuistik/Fallserie

## **Interventionsstudie**

Meta-Analyse



# Medizinische Statistik



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

---

ao. Univ.-Prof. Mag. Dr. Hanno Ulmer  
[hanno.ulmer@i-med.ac.at](mailto:hanno.ulmer@i-med.ac.at)

---

Department für Medizinische Statistik, Informatik und  
Gesundheitsökonomie, Medizinische Universität Innsbruck

# Grundlagen



---

MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

---

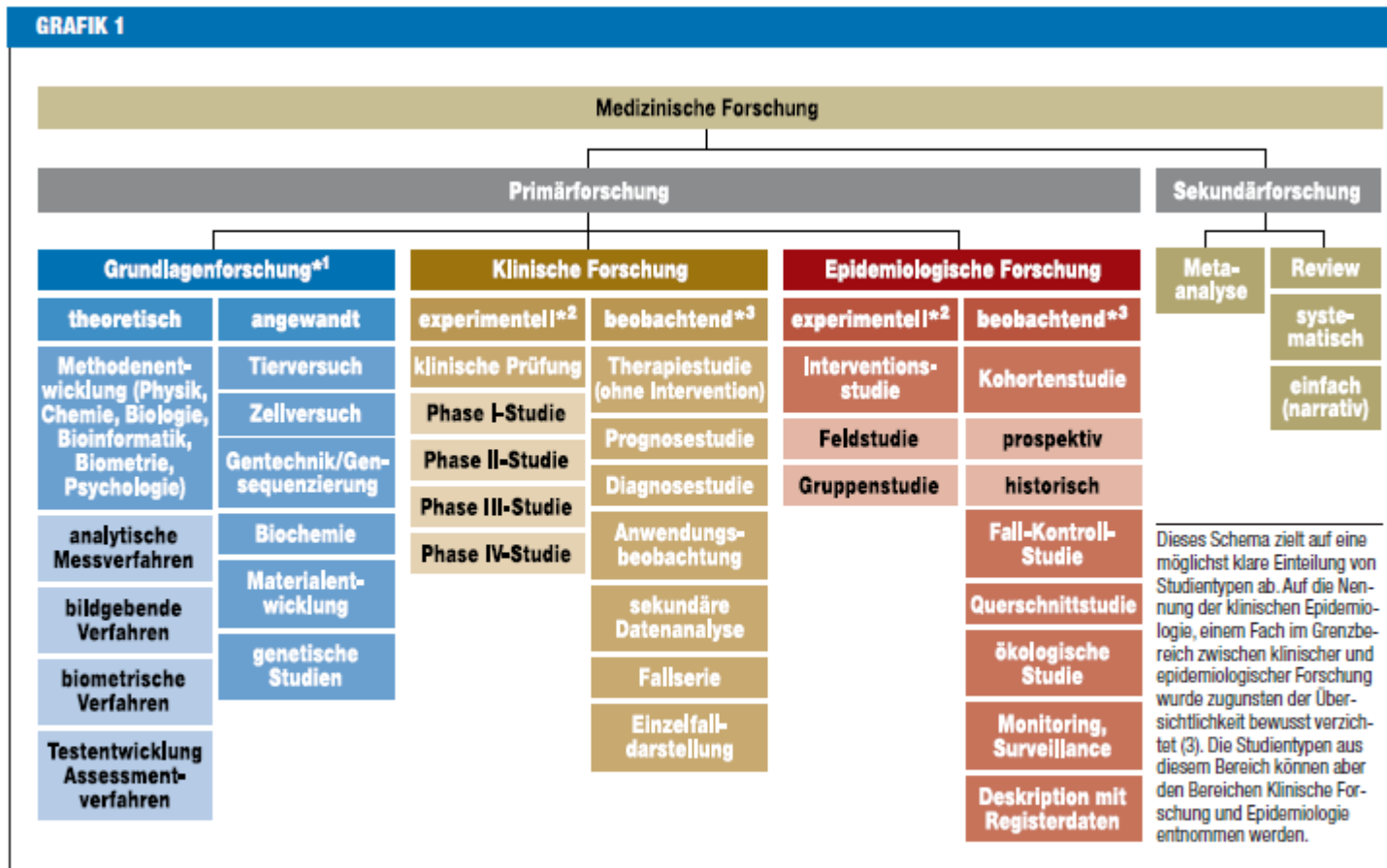
Hanno Ulmer

*hanno.ulmer@i-med.ac.at*

---

Department für Medizinische Statistik, Informatik und  
Gesundheitsökonomie, Medizinische Universität Innsbruck

# Medizinische Forschung



**Einteilung verschiedener Studientypen**

\*<sup>1</sup> häufig synonym verwendet: Experimentelle Forschung; \*<sup>2</sup> analoger Begriff: interventionell; \*<sup>3</sup> analoger Begriff: nicht interventionell/nicht experimentell

# Statistik in medizinischen Top Journals



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK



*The American Statistician, February 2007, Vol. 61, No. 1*

MEDICINE

## **The Use of Statistics in Medical Research: A Comparison of *The New England Journal of Medicine* and *Nature Medicine***

Alexander M. STRASAK, Qamruz ZAMAN, Gerhard MARINELL, Karl P. PFEIFFER, and Hanno ULMER



# Methodik der Fallstudie

- Alle Originalarbeiten publiziert im ersten Halbjahr 2004:
  - Vol. 350 No. 1–26 of NEJM
  - Vol. 10 No. 1–6 of NatMedwurde für die bibliometrische Analyse ausgewählt
- Editorials, Letters, Case Reports wurden nicht analysiert
- Zusätzlich wurden die Wiener klinische und die Wiener medizinische Wochenschrift untersucht:

AUSTRIAN JOURNAL OF STATISTICS  
Volume 36 (2007), Number 2, 141–152

**The Use of Statistics in Medical Research:  
A Comparison of Wiener Klinische Wochenschrift  
and Wiener Medizinische Wochenschrift**

Alexander M. Strasak<sup>1</sup>, Qamruz Zaman<sup>1</sup>, Gerhard Marinell<sup>2</sup>,  
Karl P. Pfeiffer<sup>1</sup>, and Hanno Ulmer<sup>1</sup>

<sup>1</sup>Dept. of Medical Statistics, Informatics and Health Economics,  
Innsbruck Medical University, Austria

<sup>2</sup>Inst. of Statistics, University of Innsbruck, Austria

# Statistische Verfahren



Kategorien nach Emerson/Colditz 1985	New England Journal of Medicine (n = 91)		Nature Medicine (n = 34)	
	n	%	n	%
<b>Types and Frequencies of Statistical Methods†</b>				
No statistical methods	2	2.2	1	2.9
Descriptive statistics only	3	3.3	5	14.7
Inferential methods	86	94.5	28	82.4
t-tests	32	35.2	14	41.2
Contingency table analysis				
Basic ( $\chi^2$ -, Fishers Exact test)	42	46.2	0	0.0
Advanced	6	6.6	0	0.0
Non-parametric tests	24	26.4	7	20.6
Analysis of Variance				
Basic (one-way ANOVA)	6	6.6	9	26.5
Advanced	6	6.6	1	2.9
Correlation coefficients	12	13.2	2	5.9
Regression				
Basic (simple-linear regression)	4	4.4	1	2.9
Advanced	27	29.7	0	0.0
Epidemiologic methods	25	27.5	0	0.0
Survival Analysis	39	42.9	4	11.8
Other methods	15	16.5	5	14.7
Unidentified method/test	1	1.1	10	29.4
Confidence intervals	61	67.0	0	0.0

# Nature Medicine versus New England Journal of Medicine

---



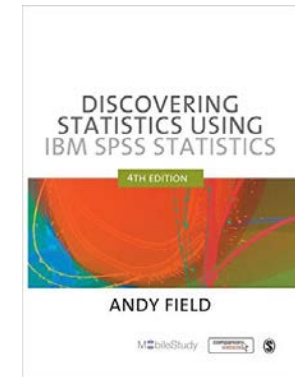
- In über 95% der Originalarbeiten wurden statistische Methoden verwendet
- Statistische Methodik unterscheidet sich zwischen
  - Grundlagenwissenschaft (Nature Medicine)
  - Klinische Forschung (NEJM)
- Methodik komplexer in NEJM
- Keine Fallzahlschätzung bzw. Poweranalyse in NatMed
- Dokumentation der Methoden mangelhaft, fehlt fast völlig in Nature Medicine
- Chi<sup>2</sup> Test, Überlebenszeitanalyse in NEJM, t-test in Nature Medicine

# Literaturhinweise



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

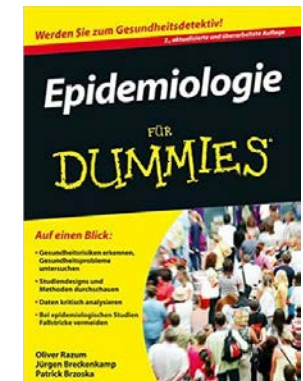
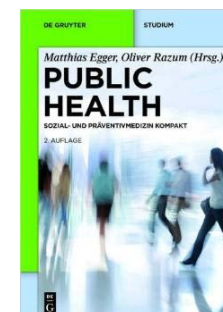
- Grundlagenwissenschaft



- Klinische Forschung



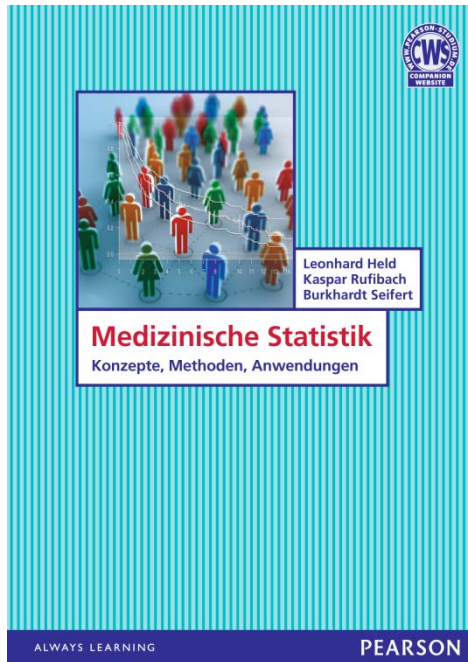
- Epidemiologische Forschung





# Prof. Dr. Leonhard Held / Prof. Dr. Burkhardt Seifert / Dr. Kaspar Rufibach Medizinische Statistik

## Inhalte aus Kapitel 1



Prof. Dr. Leonhard Held / Prof. Dr. Burkhardt Seifert /  
Dr. Kaspar Rufibach  
Medizinische Statistik

ISBN 978-3-8689-4100-5

448 Seiten | 2-farbig

Juli 2013

€ 34,95 [D] | € 36,00 [A] | SFR 46,70

[www.pearson-studium.de](http://www.pearson-studium.de)

[www.pearson.ch](http://www.pearson.ch)



# Quellen, Internet, Software

---

- British Medical Journal Statistics at Square One
- Deutsches Ärzteblatt Bewertung wissenschaftlicher Publikationen
- Deutsche Medizinische Wochenschrift Statistik-Serie
- Jumbo Münster
- Graphpad Quickcalcs
- R, Stata, SPSS, SAS, Statistica, GraphPad, MedCalc
- nQuery Advisor, East, PASS, StudySize

# Häufig verwendete statistische Methoden



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

- **Studienplanung**
  - Fragestellung
  - Hypothesen
  - Studiendesign
  - Fallzahlschätzung
  - Datenerhebung
  
- **Statistische Auswertung**
  - Deskriptive Statistik
  - Inferenzstatistik I: Schätzen von Parametern mittels Konfidenzintervallen
  - Inferenzstatistik II: Unterschiede, Hypothesenprüfung mittels Signifikanztests
  - Inferenzstatistik III: Zusammenhänge, Korrelations- und Regressionsanalysen

## Didgeridoo playing as alternative treatment for obstructive sleep apnoea syndrome: randomised controlled trial

Milo A Puhani, Alex Suarez, Christian Lo Cascio, Alfred Zahn, Markus Heitz, Otto Braendli

### Abstract

**Objective** To assess the effects of didgeridoo playing on daytime sleepiness and other outcomes related to sleep by reducing collapsibility of the upper airways in patients with moderate obstructive sleep apnoea syndrome and snoring.

**Design** Randomised controlled trial.

**Setting** Private practice of a didgeridoo instructor and a single centre for sleep medicine.

**Participants** 25 patients aged > 18 years with an apnoea-hypopnoea index between 15 and 30 and who complained about snoring.

**Interventions** Didgeridoo lessons and daily practice at home with standardised instruments for four months. Participants in the control group remained on the waiting list for lessons.

**Main outcome measure** Daytime sleepiness (Epworth scale from 0 (no daytime sleepiness) to 24), sleep quality (Pittsburgh quality of sleep index from 0 (excellent sleep quality) to 21), partner rating of sleep disturbance (visual analogue scale from 0 (not disturbed) to 10), apnoea-hypopnoea index, and health related quality of life (SF-36).

**Results** Participants in the didgeridoo group practised an average of 5.9 days a week (SD 0.86) for 25.3 minutes (SD 3.4). Compared with the control group in the didgeridoo group daytime sleepiness (difference -3.0, 95% confidence interval -5.7 to -0.3,  $P=0.03$ ) and apnoea-hypopnoea index (difference -6.2, -12.3 to -0.1,  $P=0.05$ ) improved significantly and partners reported less sleep disturbance (difference -2.8, -4.7 to -0.9,  $P<0.01$ ). There was no effect on the quality of sleep (difference -0.7, -2.1 to 0.6,  $P=0.27$ ).

The combined analysis of sleep related outcomes showed a moderate to large effect of didgeridoo playing (difference between summary z scores -0.78 SD units, -1.27 to -0.28,  $P<0.01$ ). Changes in health related quality of life did not differ between groups.

**Conclusion** Regular didgeridoo playing is an effective treatment alternative well accepted by patients with moderate obstructive sleep apnoea syndrome.

**Trial registration** ISRCTN: 31571714.

### Introduction

Snoring and obstructive sleep apnoea syndrome are two highly prevalent sleep disorders caused by collapse of the upper airways.<sup>1,2</sup> The most effective intervention for these disorders is continuous positive airway pressure therapy, which reduces daytime sleepiness<sup>3</sup> and the risk of cardiovascular morbidity and mortality in the most severely affected patients (apnoea-hypopnoea index (measured as episodes per hour) > 30).<sup>2</sup> For

moderately affected patients (apnoea-hypopnoea index 15-30) who complain about snoring and daytime sleepiness, however, continuous positive airway pressure therapy may not be suitable and other effective interventions are needed.<sup>1,6,7</sup>

AS, a didgeridoo instructor, reported that he and some of his students experienced reduced daytime sleepiness and snoring after practising with this instrument for several months. In one person, the apnoea-hypopnoea index decreased from 17 to 2. This might be due to training of the muscles of the upper airways, which control airway dilation and wall stiffening.<sup>8,9</sup> We tested the hypothesis that training of the upper airways by didgeridoo playing reduces daytime sleepiness in moderately affected patients.

### Methods

#### Participants and methods

We included German speaking participants aged > 18 years with self reported snoring and an apnoea-hypopnoea index of 15-30 (determined by a specialist in sleep medicine within the past year). Exclusion criteria were current continuous positive airway pressure therapy, use of drugs that act on the central nervous system (such as benzodiazepines), current or planned intervention for weight reduction, consumption of  $\geq 14$  alcoholic drinks a week or  $\geq 2$  a day, and obesity (body mass index  $\geq 30$  kg/m<sup>2</sup>).

We recruited patients at our study centre (Zuercher Hoehenklinik Wald, Wald, Switzerland) and one private practice in Zurich. Physicians at the study centre assessed all potential participants for eligibility. Those willing to participate provided written informed consent. After study enrolment, all patients completed a baseline assessment.

We randomised enrolled patients into an intervention group with didgeridoo training or a control group. We used STATA software (STATA 8.2, College Station, Tx) to generate the randomisation list (ralloc command) with stratification for disease severity (apnoea-hypopnoea index 15-21 or 22-30 and Epworth score < 12 or  $\geq 12$ ). The randomisation list was concealed from the recruiting physicians and the didgeridoo instructor in an administrative office otherwise not involved in the study. We used a central telephone service, which the didgeridoo instructor used to obtain group allocation.

#### Intervention and control

Participants in the intervention group started their didgeridoo training after the instructor received group allocation. The instructor (AS) gave the first individual lesson immediately after randomisation. In the first lesson, participants learnt the lip technique to produce and hold the keynote for 20-30 seconds. In the second lesson (week 2) the instructor explained the concept and



Fig 1 Man playing didgeridoo

technique of circular breathing. Circular breathing is a technique that enables the wind instrumentalist to maintain a sound for long periods of time by inhaling through the nose while maintaining airflow through the instrument, using the cheeks as bellows. In the third lesson (week 4) the didgeridoo instructor taught the participants his technique to further optimise the complex interaction between the lips, the vocal tract, and circular breathing so that the vibrations in the upper airway are more readily transmitted to the lower airways.<sup>8</sup> In the fourth lesson, eight weeks after randomisation, the instructor and the participants repeated the basics of didgeridoo playing and made corrections when necessary. Participants had to practise at home for at least 20 minutes on at least five days a week and recorded the days with practice and the practice time (answer options for 0, 20, or 30 minutes).

Participants received a standardised acrylic plastic didgeridoo that was developed by the instructor in collaboration with (Greasyl) GmbH (Ebnatnangen, Zurich, Switzerland), and costs 680 (€43; \$94, fig 1). The didgeridoo is 130 cm long with a diameter of 4 cm and an elliptical embouchure with a diameter of 2.8-3.2 mm. Acrylic didgeridoos are easier for beginners to learn on than conventional wooden didgeridoos.

Participants in the control group remained on a waiting list to start their didgeridoo training after four months. They were not allowed to start didgeridoo playing during these four months.

#### Outcome measures

Our primary outcome was daytime sleepiness as measured by the Epworth scale, which has been validated in German speaking patients.<sup>10</sup> Scores range from 0 (no daytime sleepiness) to 24, and scores > 11 represent excessive daytime sleepiness.

Secondary outcomes included three additional sleep related outcomes measures: the apnoea-hypopnoea index, the Pittsburgh quality of sleep index, and a partner's rating for sleep disturbance.

The cardiorespiratory sleep study was performed at the sleep laboratory of the study centre with a computerised system (SleepLab Pro, Jaeger, Hoechberg, Germany), according to the guidelines of the German Society for Sleep Medicine.<sup>11</sup> We

measured airflow using nasal and oral thermistors and a nasal cannula with a differential pressure flow sensor. We defined episodes of apnoea as cessation of airflow of > 10 seconds with decrements of blood oxygen saturation of  $\geq 4\%$ . Hypopnoea was defined as a reduced airflow for at least 10 seconds with decrements of blood oxygen saturation of  $\geq 4\%$  or waking, or both. The person who analysed the sleep recordings was blinded to group allocation throughout the trial.

The Pittsburgh quality of sleep index is a self administered questionnaire with 19 items to determine sleep quality, latency, duration, and disturbance within the past four weeks.<sup>12</sup> The global score ranges from 0 to 21, with higher values representing worse quality of sleep. A score of  $\geq 5$  represents impaired sleep quality. We used a validated German version.<sup>13</sup>

The partners (when present) rated their sleep disturbance by the participants' snoring during the previous seven nights on a visual analogue scale from 0 to 10. The visual analogue scale was similar to a Borg scale and had verbal descriptors for 0 (not disturbed at all) to 5 (severely disturbed), 7 (very severely disturbed), 9 (very, very severely disturbed), and 10 (extremely disturbed). The partners completed the scale independently from the participants and sent it back to the study centre.

Finally, we used the German SF-36 to assess generic health related quality of life.<sup>14</sup>

#### Analysis

We analysed all data on an intention to treat basis. For the primary analysis we compared change scores (differences between baseline and follow-up) between groups using two sample *t* tests. We also performed an analysis of covariance with the primary and secondary continuous end points at four months after randomisation as the dependent variables and their baseline values, markers of severity of disease (apnoea-hypopnoea index and Epworth score), weight change, and group allocation as independent variables.

We selected the Epworth scale as our primary outcome but also considered the three other sleep related outcomes (apnoea-hypopnoea index, Pittsburgh quality of sleep index, and partner rating). To provide an overall estimate of the effects of didgeridoo playing on the four outcome measures we used a summary measure described by Schouten.<sup>15</sup> Briefly, for each patient and outcome we calculated a z score (difference of individual change minus overall mean change score/overall SD of change score) and then a summary score as the average of the four z scores. We compared these summary scores between the groups using a two sample *t* test.

For all analyses, we used 95% confidence intervals and considered  $P \leq 0.05$  as significant. All statistical analyses were performed with SPSS (12.0.1, Chicago, Ill).

### Results

Figure 2 shows the study flow from screening of potential participants to the final assessment. We included 25 patients from August 2004 to April 2005, all of whom completed the trial. Table 1 shows the participants' characteristics and the baseline values of the outcomes measures. Most patients were men, aged about 50, and had an average apnoea-hypopnoea index of 21 and excessive daytime sleepiness (mean Epworth scores 11.8 in the didgeridoo group and 11.1 in the control group). The Pittsburgh quality of sleep index indicated impaired sleep quality (5.2 and 5.8) and the partners of the study participants on average had severely disturbed sleep (5.6 and 5.5). The SF-36 scores were in the range of the normal population with exception of the mental component and vitality scores, which were lower



# Beispielstudie

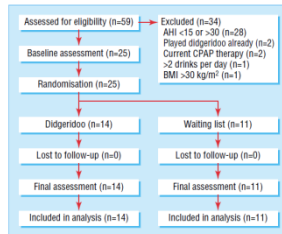


Fig 2 Flow of participants through study

(referred scores of 50 for mental component and 63.3 for vitality).

On average, participants in the didgeridoo group practised on 5.9 days a week (SD 0.86, range 4.6-6.9) for 25.3 minutes (3.4). There were no adverse or unexpected events in either group. Table 2 shows the effects of didgeridoo playing on the four sleep related outcomes. The primary outcome (daytime sleepiness as measured by the Epworth scale) improved significantly in the didgeridoo group compared with the control group (difference -3.0 units, 95% confidence interval -5.7 to -0.3,  $P=0.03$ ). Figure 3 shows the individual responses in daytime sleepiness in the two groups.

The quality of sleep did not differ significantly between groups (difference -0.7 units, -2.1 to 0.6,  $P=0.27$ ), but the

Table 1 Characteristics of participants according to allocation to intervention (didgeridoo) or control. Numbers are means (SD) except for absolute values

	Didgeridoo group (n=14)	Control group (n=11)
Age (years)	49.9 (6.7)	47.0 (8.9)
Men	12	9
Years of snoring	8.7 (8.0)	8.9 (3.5)
Body mass index	25.8 (4.0)	25.9 (2.4)
Systemic blood pressure	133.7 (14.0)	133.3 (14.0)
Diastolic blood pressure	80.9 (7.1)	77.3 (8.4)
Used any medication	3	1
Played wind instrument	0	2
Drinks/week	2.2 (3.1)	2.2 (1.8)
Reason for study participation:		
Snoring	14	10
Intolerance to CPAP therapy	0	1
Apnoea-hypopnoea index	22.3 (5.0)	19.9 (4.7)
Epworth scale	11.8 (3.5)	11.1 (6.4)
Pittsburgh quality of sleep index	5.2 (1.7)	5.8 (2.8)
Partner's rating of sleep disturbance	5.8 (2.4)	5.5 (2.3)
SF-36:		
Physical component score	52.7 (7.4)	52.7 (7.0)
Mental component score	41.1 (12.1)	44.8 (8.6)
Role physical	88.9 (11.3)	92.5 (8.9)
Role physical	76.2 (25.1)	82.5 (20.6)
Bodily pain	79.2 (22.0)	80.9 (29.1)
General health	70.4 (17.1)	69.9 (16.8)
Vitality	48.6 (15.2)	53.0 (11.1)
Social functioning	66.4 (20.6)	69.1 (14.7)
Role emotional	72.2 (27.9)	83.5 (17.4)
Mental health	66.9 (19.5)	68.4 (19.9)

\*One participant in the didgeridoo group did not have a partner.

Table 2 Effects of intervention on sleep related outcomes

Outcome	Didgeridoo group	Control group	Raw difference* (95% CI)	Adjusted difference† (95% CI)
Epworth scale				
At 4 months	7.4 (2.3)	9.6 (6.0)		
Change from baseline	-4.4 (3.7)	-1.4 (2.6)	-3.0 (-5.7 to -0.3), $P=0.03$	-2.8 (-5.4 to -0.2), $P=0.04$
Pittsburgh quality of sleep index				
At 4 months	4.3 (2.1)	5.6 (2.7)		
Change from baseline	-0.9 (1.6)	-0.2 (1.7)	-0.7 (-2.1 to 0.6), $P=0.27$	-0.8 (-2.3 to 0.8), $P=0.30$
Partner rating of sleep disturbance				
At 4 months	2.3 (1.4)	4.8 (2.2)		
Change from baseline	-3.4 (2.4)	-0.6 (1.9)	-2.8 (-4.7 to -0.9), $P=0.01$	-2.7 (-4.2 to -1.2), $P=0.01$
Apnoea-hypopnoea index				
At 4 months	11.6 (8.1)	15.4 (9.8)		
Change from baseline	-10.7 (7.7)	-4.5 (6.9)	-6.2 (-12.3 to -0.1), $P=0.05$	-6.6 (-13.3 to -0.1), $P=0.05$

\*Two sample *t* tests.

†Analysis of covariance with adjustment for severity of disease (apnoea-hypopnoea index and Epworth scale) and weight change during study period.

partners of those in the didgeridoo group reported less sleep disturbance (difference -2.8 units, -4.7 to -0.9,  $P<0.01$ ). We also observed a significant effect of didgeridoo playing on apnoea-hypopnoea (difference for apnoea-hypopnoea index -6.2, -12.3 to -0.1,  $P=0.05$ ). Didgeridoo playing did not have a significant effect on any domain of the SF-36. Adjustment for severity of the condition and weight change during the study did not alter the results substantially for any outcome.

Figure 4 shows the combined analysis of the four sleep related outcomes. The summary *z* scores differed by -0.78 (-1.27 to -0.28,  $P<0.01$ ), favouring the didgeridoo over the control group.

## Discussion

In this randomised controlled trial we found that four months of training of the upper airways by didgeridoo playing reduces daytime sleepiness in people with snoring and obstructive sleep apnoea syndrome. The reduction of the apnoea-hypopnoea index by didgeridoo playing indicated that the collapsibility of the upper airways decreased. In addition, the partners of participants in the didgeridoo group were much less disturbed in their sleep.

Earlier studies about the effects of electrical neurostimulation or training of the respiratory muscles showed no improvement

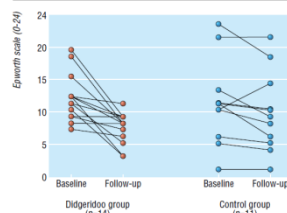


Fig 3 Individual responses in daytime sleepiness, showing direction of change

## Research

in daytime sleepiness<sup>10</sup> or the apnoea-hypopnoea index<sup>16</sup> or were limited by the lack of a control group.<sup>16</sup> Our results are well to show that training the upper airways significantly improves sleep related outcomes. The larger effects we observed may be due to the longer duration of our intervention and the training of the whole vocal tract instead of only single muscles.

### Comparison with continuous positive airway pressure therapy

A meta-analysis of trials evaluating continuous positive airway pressure therapy in patients with moderate to severe obstructive sleep apnoea syndrome showed an average effect of -3.9 units on the Epworth scale.<sup>1</sup> The minimum important difference on this scale for severely affected patients is around 4 units.<sup>19</sup> In our trial, the mean change score in the didgeridoo group was -4.4 units and the difference between the intervention and control group was -3.0 units. Thus the effect of didgeridoo playing seems to be slightly smaller than with CPAP therapy. However, we expected smaller effects because our patients were only moderately affected so that results are likely to be less pronounced.

One of the challenges in the treatment of sleep disorders is 'poor compliance'.<sup>20</sup> Thus new treatments not only need to be effective but also those that people are motivated enough to use. Didgeridoo playing seems to meet these requirements. Participants were highly motivated during the trial and practised, on average, almost six days a week, which was even more than the protocol asked for.

### Strengths and limitations of trial

Strengths of our trial include the long duration of the training so that effects could develop. Also, we blinded outcomes assessors when possible (sleep studies) and controlled for confounding by restricting the study sample to non-obese patients with little alcohol and drug consumption. A limitation is that those in the control group were simply put on a waiting list because a sham intervention for didgeridoo playing would be difficult. A control intervention such as playing a recorder would have been an option, but we would not be able to exclude effects on the upper airways and compliance might be poor. Another limitation is that the sample size was small. We conducted a proof of concept study and larger trials with more diverse study populations are needed to provide more precise estimates of the treatment effect of upper airway training.

In conclusion, didgeridoo playing improved daytime sleepiness in patients with moderate snoring and obstructive sleep apnoea and reduced sleep disturbance in their partners. Larger trials are needed to confirm our preliminary findings, but our results may give hope to the many people with moderate obstructive sleep apnoea syndrome and snoring, as well as to their partners.

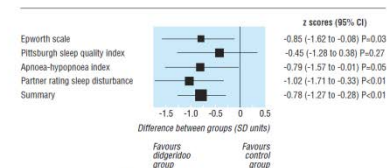


Fig 4 Effects of didgeridoo playing on measure of sleep related outcomes

### What is already known on this topic

Snoring and obstructive sleep apnoea syndrome are highly prevalent sleep disorders associated with substantial morbidity and mortality and rising costs.

Continuous positive airways pressure therapy can reduce daytime sleepiness, but compliance with this treatment is often poor.

Training or electrostimulation of the muscles of the upper airway might reduce collapsibility of the upper airways during sleep.

### What this study adds

Regular playing of a didgeridoo reduces daytime sleepiness and snoring in people with moderate obstructive sleep apnoea syndrome and also improves the sleep quality of partners.

Severity of disease, expressed by the apnoea-hypopnoea index, is also substantially reduced after four months of didgeridoo playing.

Contributors: MAP, AS, and OB designed and organised the study. AS assigned the intervention. CLC, OB, MH, and AZ collected the data. MAP supervised data collection, analysed data, and wrote the first draft. AS, CLC, AZ, MH, and OB critically reviewed the manuscript, and MAP and OB prepared the final version. OB is guarantor.

Funding: Zurich Lung Association, Zuercher Hoehheinklinik Wald.

Competing interests: AS is a professional didgeridoo instructor and teaches 'fat chi' and 'qi gong'.

Ethical approval: Ethics committee of the University Hospital of Zurich.

- Caples SM, Ganali AS, Somers VK. Obstructive sleep apnoea. *Ann Intern Med* 2003;139:167-97.
- Doran J, Lonsdale S, Rafteri R, Innes A. Obstructive sleep apnoea-hypopnoea and related clinical features in a population-based sample of subjects aged 50 to 70 yr. *Am J Respir Crit Care Med* 2001;163:685-9.
- Young T, Palta M, Dempsey J, Skatrud J, Weber S, Badr S. The occurrence of sleep-disordered breathing among middle-aged adults. *N Engl J Med* 1995;328:1230-5.
- White J, Cates C, Wright J. Continuous positive airway pressure for obstructive sleep apnoea. *Cochrane Database Syst Rev* 2002;(2):CD001106.
- Martin BJ, Carrizo SJ, Vicente E, Agusti AG. Long-term cardiovascular outcomes in men with obstructive sleep apnoea-hypopnoea with or without treatment with continuous positive airway pressure: an observational study. *Lancet* 2003;361:1046-53.
- Lewis KE, Soole L, Bartle B, Watkins AJ, Eldon P. Early predictors of CPAP use for the treatment of obstructive sleep apnoea. *Sleep* 2004;27:1344-8.
- Pepas JL, Krieger J, Rosenow JA, Cornejo A, Sierra E, Delgado E, et al. Effective compliance during the first 3 months of continuous positive airway pressure: A European prospective study of 121 patients. *Am J Respir Crit Care Med* 1999;160:1129-9.
- Manni EA, Barnett T, Corwell S, Ludlow CL. The effect of neuromuscular stimulation of the genioglossus on the hypoplasia-aryepiglottic airway. *Laryngoscope* 2002;112:51-6.
- Randall-Walker W, Domanski U, Weckmann R, Rühle KH. Tongue-muscle training by intraoral electrical neurostimulation in patients with obstructive sleep apnoea. *Sleep* 2001;27:254-9.

# „Wie lese ich eine Studie in 10 Minuten“

---



- Grundlegender Aufbau eines Studienberichts (papers):

## IMRaD Schema

- Weitere Strukturierungen
- CONSORT Statement für Interventionsstudien (RCTs)
- STROBE Statement für Beobachtungsstudien (epidemiologische Studien: Kohortenstudie, Fall-Kontroll-Studie u. Querschnittstudie)
- STARD Statement für diagnostische Studien

	Item number	Descriptor	Reported on page number
<b>Title and abstract</b>	1	How participants were allocated to interventions (eg, "random allocation", "randomised", or "randomly assigned").	
<b>Introduction</b>			
Background	2	Scientific background and explanation of rationale.	
<b>Methods</b>			
Participants	3	Eligibility criteria for participants and the settings and locations where the data were collected.	
Interventions	4	Precise details of the interventions intended for each group and how and when they were actually administered.	
Objectives	5	Specific objectives and hypotheses.	
Outcomes	6	Clearly defined primary and secondary outcome measures and, when applicable, any methods used to enhance the quality of measurements (eg, multiple observations, training of assessors, &c).	
Sample size	7	How sample size was determined and, when applicable, explanation of any interim analyses and stopping rules.	
Randomisation			
Sequence generation	8	Method used to generate the random allocation sequence, including details of any restriction (eg, blocking, stratification).	
Allocation concealment	9	Method used to implement the random allocation sequence (eg, numbered containers or central telephone), clarifying whether the sequence was concealed until interventions were assigned.	
Implementation	10	Who generated the allocation sequence, who enrolled participants, and who assigned participants to their groups.	
Blinding (masking)	11	Whether or not participants, those administering the interventions, and those assessing the outcomes were aware of group assignment. If not, how the success of masking was assessed.	
Statistical methods	12	Statistical methods used to compare groups for primary outcome(s); methods for additional analyses, such as subgroup analyses and adjusted analyses.	
<b>Results</b>			
Participant flow	13	Flow of participants through each stage (a diagram is strongly recommended). Specifically, for each group, report the numbers of participants randomly assigned, receiving intended treatment, completing the study protocol, and analysed for the primary outcome. Describe protocol deviations from study as planned, together with reasons.	
Recruitment	14	Dates defining the periods of recruitment and follow-up.	
Baseline data	15	Baseline demographic and clinical characteristics of each group.	
Numbers analysed	16	Number of participants (denominator) in each group included in each analysis and whether the analysis was by "intention to treat". State the results in absolute numbers when feasible (eg, 10/20, not 50%).	
Outcomes and estimation	17	For each primary and secondary outcome, a summary of results for each group, and the estimated effect size and its precision (eg, 95% CI).	
Ancillary analyses	18	Address multiplicity by reporting any other analyses performed, including subgroup analyses and adjusted analyses, indicating those prespecified and those exploratory.	
Adverse events	19	All important adverse events or side-effects in each intervention group.	
<b>Discussion</b>			
Interpretation	20	Interpretation of the results, taking into account study hypotheses, sources of potential bias or imprecision and the dangers associated with multiplicity of analyses and outcomes.	
Generalisability	21	Generalisability (external validity) of the trial findings.	
Overall evidence	22	General interpretation of the results in the context of current evidence.	

### Checklist of items to include when reporting a randomised trial

## First European Multicenter Results With a New Transcutaneous Bone Conduction Hearing Implant System: Short-Term Safety and Efficacy

\*Georg Sprinzl, †Thomas Lenarz, ‡Arneborg Ernst, §Rudolf Hagen,  
\*Astrid Wolf-Magele, †Hamidreza Mojallal, ‡Ingo Todt, §Robert Mlynski,  
and ||Mario D. Wolframm

*\*Department of Otorhinolaryngology, Medical University Innsbruck, Austria; †Department of Otorhinolaryngology, Hannover Medical School, Germany; ‡Clinic for Ears, Nose, and Throat, Unfallkrankenhaus Berlin, Germany; §Department of Otorhinolaryngology, Plastic, Aesthetic and Reconstructive Head and Neck Surgery, University Clinic Würzburg, Germany; and ||Vibrant MED-EL, Innsbruck, Austria*



**TABLE 1.** *Demographic data and medical parameter disease factors of the 12 study participants*

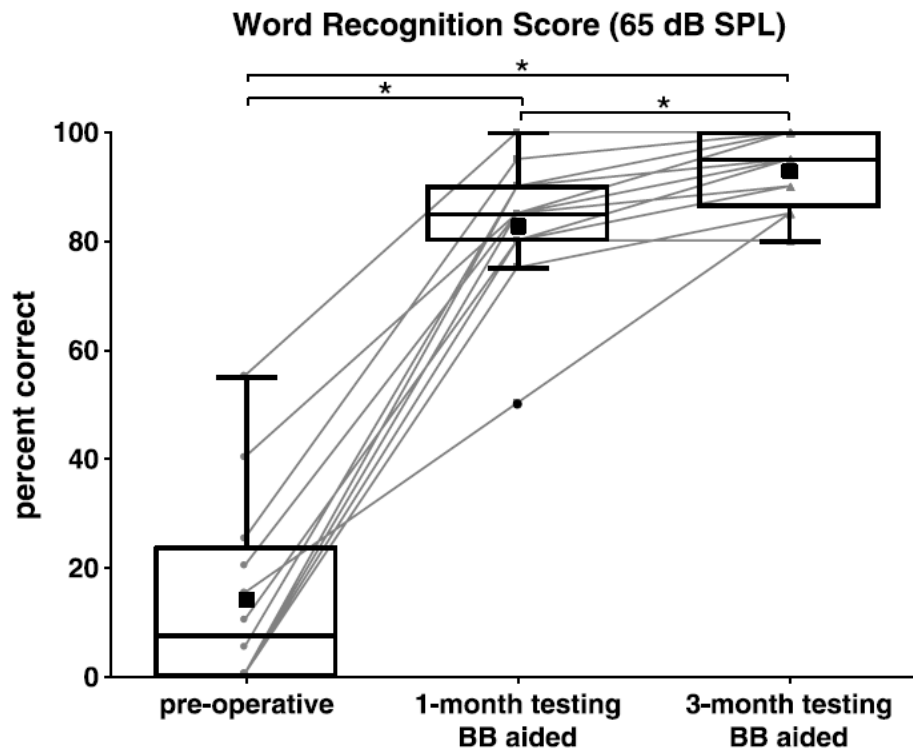
Demographics				Disease factors and medical history						
Subject no.	Age at surgery	Sex	Study site	Implanted ear	No. previous ear surgeries	Duration of HL (yr)	Type of HL	Etiology	PTA <sub>4</sub> BC implanted ear (dB HL)	PTA <sub>4</sub> AC implanted ear (dB HL)
1	69	M	Berlin	R	2	60	CHL	Cholesteatoma	5	45
2	69	F	Berlin	R	4	60	CHL	Cholesteatoma	19	46
3	44	F	Berlin	R	2	9	Mixed	Otosclerosis	35	50
4	28	M	Hannover	R	2	15	CHL	COM	6	30
5	65	F	Hannover	R	1	2	CHL	Glomus tumor	6	66
6	65	F	Hannover	L	1	1	Mixed	Chronic mastoiditis	14	53
7	63	F	Hannover	L	3	22	Mixed	COM	18	67
8	35	M	Würzburg	R	5	35	CHL	Cholesteatoma	8	49
9	20	F	Innsbruck	L	2	20	CHL	Atresia auris	11	73
10	19	F	Innsbruck	R	2	19	Mixed	Cholesteatoma	21	61
11	28	F	Innsbruck	R	0	28	Mixed	Atresia auris	25	93
12	27	F	Innsbruck	R	1	27	CHL	Atresia auris	15	73

## Study Design

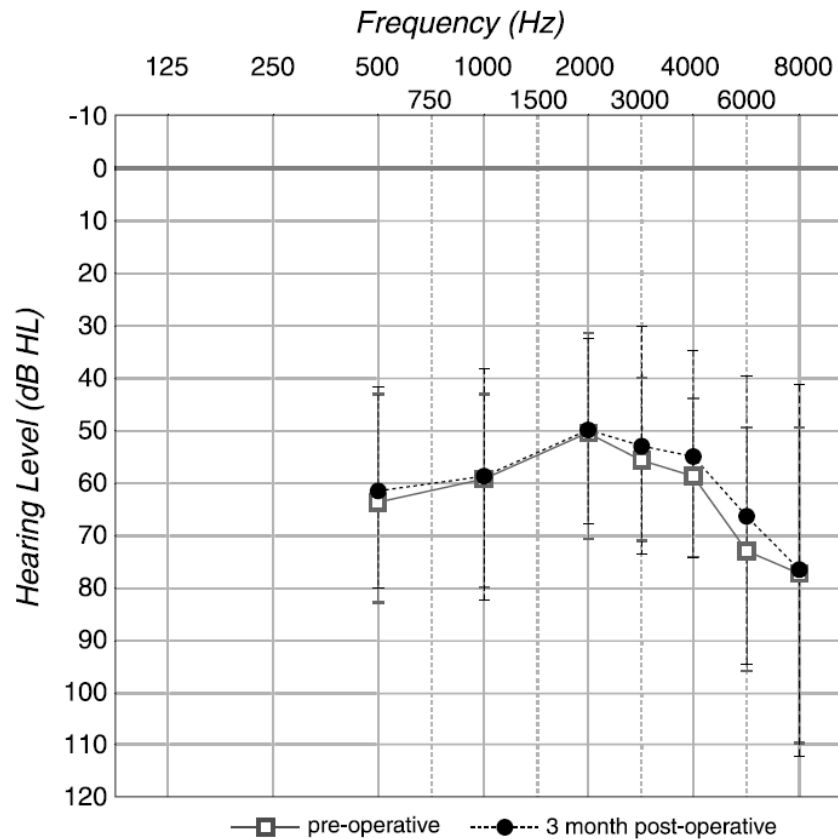
The study was a prospective, single-subject repeated-measures design, in which each subject served as his/her own control. Performance on audiometric tests preoperatively was compared with the aided 3 month postoperative condition using the Bonebridge. This type of design has been applied frequently to the evaluation of implantable hearing devices in multicenter clinical trials (35–37). It minimizes the effect of variability inherent to the population to the evaluation of treatment outcomes. Standardized evaluation methods were used to assure the reliability of the data across different investigational centers.

# Medizinproduktstudie

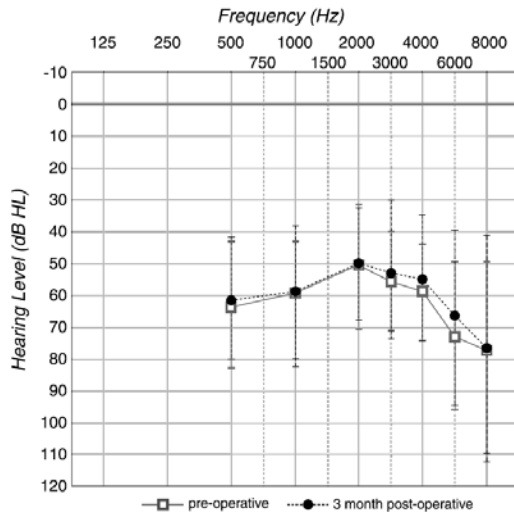
Statistical analyses were performed using IBM SPSS Statistics 19 (IBM, Armonk, NY, USA). One-way repeated-measure ANOVAs with time as factor were performed (significance was accepted at  $p \leq 0.05$ ) and followed by post hoc pairwise comparisons to examine significant differences between the single test intervals. For each ANOVA, Mauchly's test of sphericity was applied. If sphericity could not be assumed, a Greenhouse-Geisser correction was used as part to the ANOVA. P-values of the pairwise comparisons were adjusted with the Holm-Sidak method. Box-Whisker Plots represent the whole data set. Whiskers extend to the maximum value within 1.5 times the interquartile range (IQR) above the third quartile or the minimum value within 1.5 times the IQR below the first quartile. Values outside this range are considered to be outliers, depicted as individual dots. Tukey box-whisker plots were generated using GraphPad Prism 5 (<http://www.graphpad.com>).



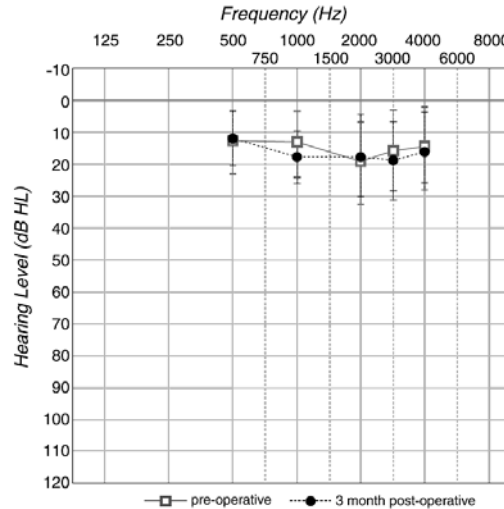
**FIG. 3.** Word recognition scores in quiet (Freiburger mono-syllables) for the implanted ear: preoperative, 1-month postoperative and 3-months postoperative. Both postoperative scores are significantly improved from preoperative scores ( $p < 0.001$ ) and from each other ( $p = 0.010$ ),  $n = 12$ , BB = Bonebridge.



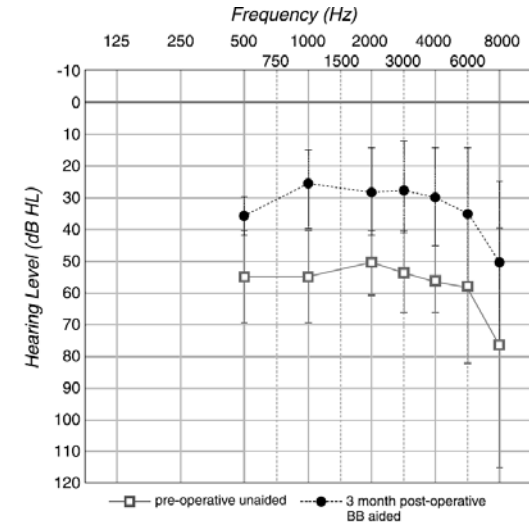
**FIG. 5.** Mean air conduction thresholds for the implanted ear: preoperative unaided testing compared with 3-month postoperative tests. Error bars represent T 1 SD (n = 12).



**FIG. 5.** Mean air conduction thresholds for the implanted ear: preoperative unaided testing compared with 3-month postoperative tests. Error bars represent  $\pm 1$  SD ( $n = 12$ ).



**FIG. 6.** Mean bone conduction thresholds for the implanted ear: preoperative unaided testing compared with 3-month postoperative tests. Error bars represent  $\pm 1$  SD ( $n = 12$ ).



**FIG. 7.** Mean soundfield thresholds (warble tones) for the implanted ear: preoperative unaided testing compared with 3-month postoperative aided tests. Error bars represent  $\pm 1$  SD ( $n = 12$ ), BB = Bonebridge.

**TABLE 2.** *F* statistics and *p* values from analysis of variance of audiometric tests (preoperative, 1-month postoperative and 3-month postoperative;  $n = 12$ )

	500 Hz		1 kHz		2 kHz		3 kHz		4 kHz		6 kHz		8 kHz	
	$F_{(2,22)}$	<i>p</i>	$F_{(2,22)}$	<i>p</i>	$F_{(2,22)}$	<i>p</i>	$F_{(2,22)}$	<i>p</i>	$F_{(2,22)}$	<i>p</i>	$F_{(2,22)}$	<i>p</i>	$F_{(2,22)}$	<i>p</i>
Air conduction (Fig. 5)	0.394	0.68	0.555	0.58	0.681	0.52	0.723	0.50	1.00	0.38	0.726	0.50	0.032	0.97
Bone conduction (Fig. 6)	0.919	0.41	1.00	0.38	1.16	0.33	1.61	0.22	1.46	0.25	—	—	—	—
Soundfield (Fig. 7)	14.7	<0.001	48.1	<0.001	24.3	<0.001	64.8	<0.001	61.8	<0.001	28.7	<0.001	5.00	0.036



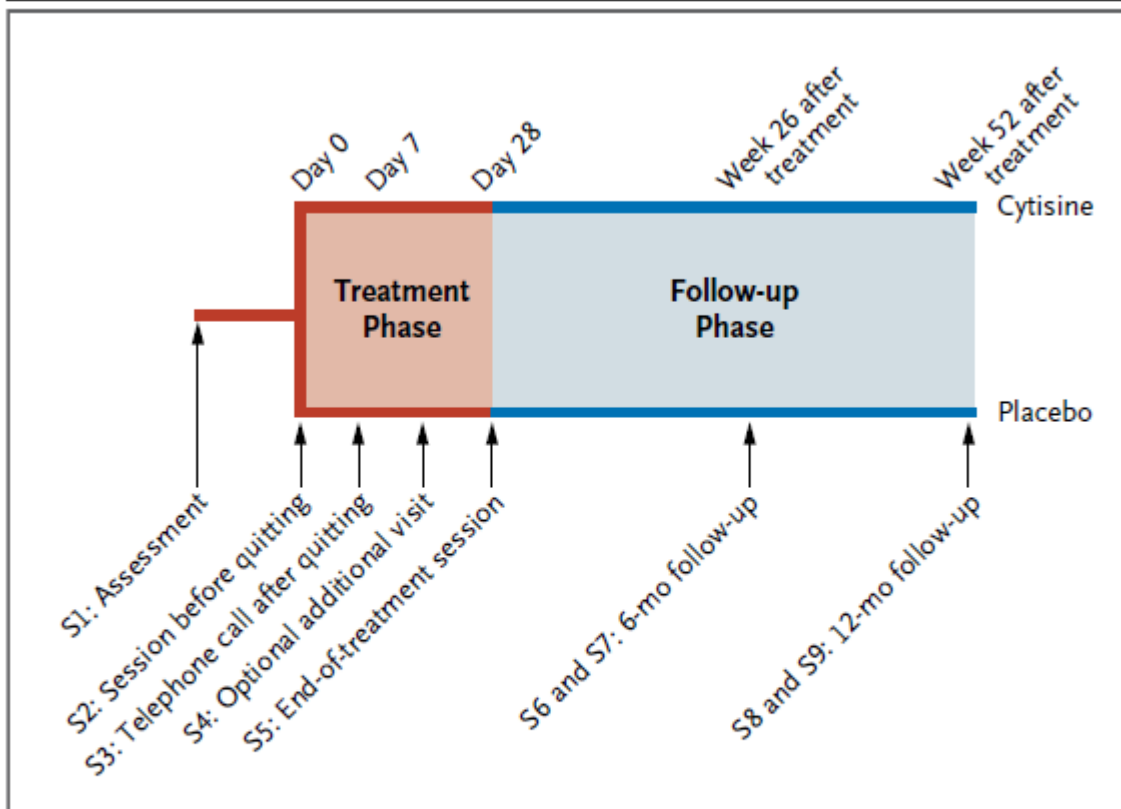
*The* NEW ENGLAND JOURNAL *of* MEDICINE

ORIGINAL ARTICLE

## Placebo-Controlled Trial of Cytisine for Smoking Cessation

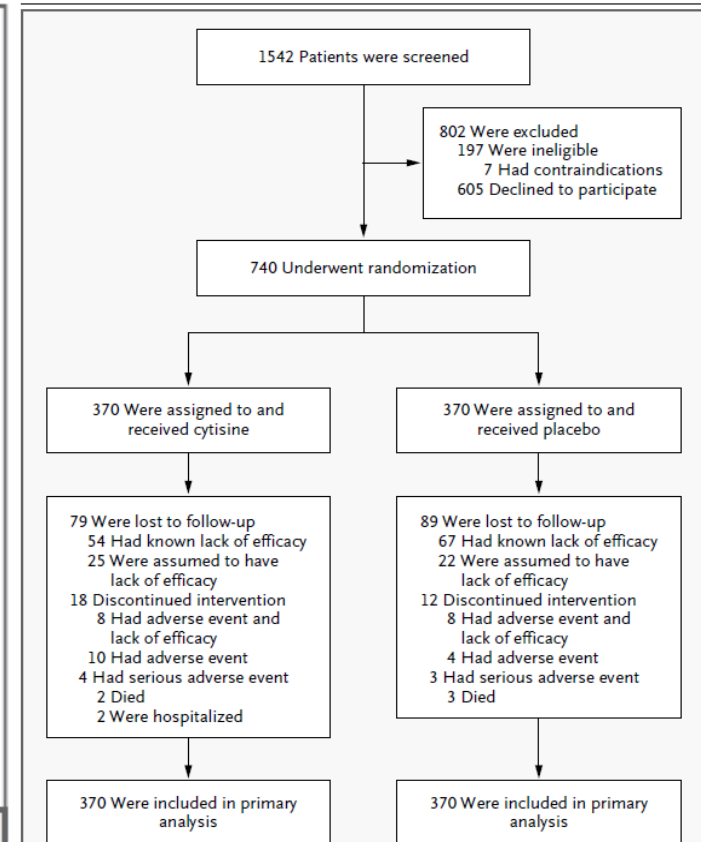
Robert West, Ph.D., Witold Zatonski, M.D., Magdalena Cedzynska, M.A.,  
Dorota Lewandowska, Ph.D., M.D., Joanna Pazik, Ph.D., M.D.,  
Paul Aveyard, Ph.D., M.D., and John Stapleton, M.Sc.

# Arzneimittelstudie



**Figure 1. Timing of Study Procedures.**

Session numbers are indicated by S1 through S9. Sessions 3, 6, and 8 were telephone sessions. The others were clinic visits.



**Figure 2. Numbers of Patients Who Were Enrolled in the Study and Included in the Primary Analysis.**

The patients lost to follow-up included some patients who discontinued the intervention or who had a serious adverse event.



**Table 1. Characteristics of the Study Participants.\***

Characteristic	Cytisine (N = 370)	Placebo (N = 370)
Male sex — no. (%)	183 (49.5)	161 (43.5)
Age — yr	47.8±12.6	48.5±12.6
Married — no. (%)†	190 (51.4)	207 (56.1)
Employment involving manual labor — no. (%)‡	196 (54.3)	178 (50.0)
Tried to stop smoking previously — no. (%)	307 (83.0)	301 (81.4)
No. of cigarettes smoked daily	23.0±8.7	22.5±9.6
Carbon monoxide in exhaled breath — ppm	19.2±8.7	18.2±9.0
Duration of smoking — yr	28.1±11.6	28.6±11.7
FTND score§	6.3±2.1	6.1±2.2
Beck Depression Inventory score¶	10.5±7.5	10.7±7.9

**Table 2. Effect of Cytisine on Smoking Cessation.\***

Outcome	Cytisine (N= 370)	Placebo (N= 370)	Percentage-Point Difference (95% CI)	Relative Rate (95% CI)†
	<i>percent (number)</i>			
Primary outcome: abstinence for 12 mo	8.4 (31)	2.4 (9)	6.0 (2.7–9.2)‡	3.4 (1.7–7.1)
Abstinence for 6 mo	10.0 (37)	3.5 (13)	6.5 (2.9–10.1)‡	2.9 (1.5–5.3)
Point prevalence at 12 mo	13.2 (49)	7.3 (27)	5.9 (1.6–10.3)§	1.8 (1.2–2.8)

**Table 3. Adverse Events Reported by 10 or More Study Participants.\***

Event	Cytisine (N = 370)	Placebo (N = 370)	Percentage-Point Difference (95% CI) <sup>†</sup>	Relative Rate (95% CI) <sup>‡</sup>
	<i>percent (number)</i>			
Any gastrointestinal event	13.8 (51)	8.1 (30)	5.7 (1.2 to 10.2) <sup>§</sup>	1.7 (1.1 to 2.6)
Upper abdominal pain	3.8 (14)	3.0 (11)	0.8 (-1.8 to 3.4)	1.3 (0.6 to 2.8)
Nausea	3.8 (14)	2.7 (10)	1.1 (-1.5 to 3.6)	1.4 (0.6 to 3.1)
Dyspepsia	2.4 (9)	1.1 (4)	1.4 (-0.5 to 3.2)	2.2 (0.7 to 7.2)
Dry mouth	2.2 (8)	0.5 (2)	1.6 (0 to 3.3)	4.0 (0.9 to 18.7)
Any psychiatric event	4.6 (17)	3.2 (12)	1.4 (-1.4 to 4.2)	1.4 (0.7 to 2.9)
Dizziness	2.2 (8)	1.1 (4)	1.1 (-0.7 to 2.9)	2.0 (0.6 to 6.6)
Somnolence	1.6 (6)	1.1 (4)	0.5 (-1.1 to 2.2)	1.5 (0.4 to 5.3)
Any nervous system event	2.7 (10)	2.4 (9)	0.3 (-2.0 to 2.6)	1.1 (0.5 to 2.7)
Headache	1.9 (7)	2.2 (8)	-0.3 (-2.3 to 1.8)	0.9 (0.3 to 2.4)
Skin and subcutaneous tissue	1.6 (6)	1.4 (5)	0.3 (-1.5 to 2.0)	1.2 (0.4 to 3.9)

## STATISTICAL ANALYSIS

With the use of previous trial data as a guide, we estimated that we would need to enroll 740 participants (370 in each group) to detect a between-group difference of 6 percentage points (6% vs. 12%) for the primary outcome, with 80% power and at an alpha level of 0.05.

The analyses of outcomes were based on the intention-to-treat principle, with treatment considered to have failed in participants who were lost to follow-up.<sup>21</sup> The absolute percentage-point difference between participants who met the criteria for abstinence in the two groups was tested with the use of Fisher's exact test. The relative rate of abstinence (the percentage of patients in the cytosine group who met the abstinence criteria divided by the percentage in the placebo group) was also calculated. The 95% confidence interval was calculated for all measures. The relative rates and percentage-point differences were calculated for adverse events reported by 10 or more participants. Logistic regression was used to examine efficacy, with adjustment for baseline characteristics.

# Deskriptive Statistik



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

---

Hanno Ulmer

*hanno.ulmer@i-med.ac.at*

---

Department für Medizinische Statistik, Informatik und  
Gesundheitsökonomie, Medizinische Universität Innsbruck

# Deskriptive Statistik, Merkmalstypen



## DEFINITION 2.1

### Qualitative Daten

Bei **qualitativen** Variablen werden die Individuen bestimmten Kategorien zugeordnet, z. B. {rot, grün, blau}, {0, A, B, AB} oder {männlich, weiblich}. Synonyme für „qualitativ“ sind **kategoriell**, **kategorial** oder **diskret**. Auch die Bezeichnung **Faktor** oder **Faktorvariable** ist gebräuchlich für eine qualitative Variable.

## DEFINITION 2.2

### Quantitative Daten

**Quantitative** Variablen können nur numerische Werte annehmen. Sie heißen deshalb auch **numerische** oder **stetige** Daten.

## DEFINITION 2.3

### Zähldaten

**Zähldaten** sind spezielle diskrete Daten, die zählen, wie oft ein Ereignis aufgetreten ist. Zähldaten können folglich die Werte 0, 1, 2, ... annehmen.

# Beispieldaten, BMJ 2005

Patient	Geschlecht (in Jahren)	Alter	Körpergröße (in m)	Behandlungsgruppe	Epworth-Index
1	Mann	45	1,74	Kontrollgruppe	11
2	Frau	41	1,70	Didgeridoo	11
3	Mann	41	1,80	Kontrollgruppe	5
4	Mann	39	1,79	Didgeridoo	8
5	Mann	33	1,80	Kontrollgruppe	1
6	Frau	55	1,67	Didgeridoo	9
7	Frau	50	1,65	Kontrollgruppe	10
8	Frau	65	1,64	Kontrollgruppe	11
9	Mann	43	1,76	Didgeridoo	7
10	Mann	59	1,85	Didgeridoo	9

**Tabelle 2.1:** Daten der ersten 10 Patienten der Didgeridoo-Studie.

Variable	Datentyp	Kategorien
Geschlecht	binär	Mann / Frau
Alter (in Jahren)	stetig	
Körpergröße (in m)	stetig	
Behandlungsgruppe	binär	Didgeridoo / Kontrolle
Epworth-Index	approximativ stetig	

**Tabelle 2.2:** Datentypen einiger Variablen der Didgeridoo-Studie.

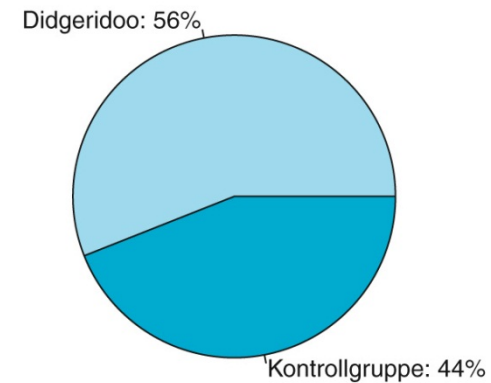
Didgeridoo playing as alternative treatment for obstructive sleep apnoea syndrome: randomised controlled trial  
Milo A Puhan, Alex Suarez, Christian Lo Cascio, Alfred Zahn, Markus Heitz, Otto Braendli, BMJ 2005



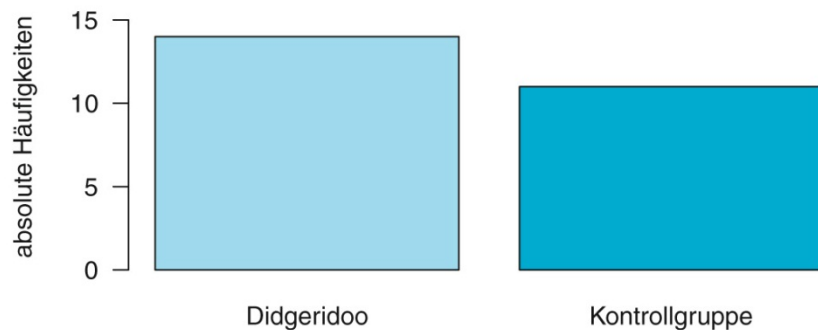
# Deskription qualitativer Daten

	abs. Häufigkeit	rel. Häufigkeit (%)
Didgeridoo	14	56
Kontrollgruppe	11	44
Total	25	100

**Tabelle 2.4:** Anzahl Patienten in den Behandlungsgruppen der Didgeridoo-Studie.



**Abbildung 2.2:** Kuchendiagramm der Behandlungsrückhäufigkeiten.



**Abbildung 2.1:** Balkendiagramm der Behandlungsrückhäufigkeiten.

	Didgeridoo	Kontrollgruppe	Total
Frauen	12	9	21
Männer	2	2	4
Total	14	11	25

**Tabelle 2.5:** Anzahl Patienten in den Behandlungsgruppen der Didgeridoo-Studie nach Geschlecht



# Deskription quantitativer Daten

Intervall	(30, 35]	(35, 40]	(40, 45]	(45, 50]	(50, 55]	(55, 60]	(60, 65]
Anzahl Werte im Intervall	1	2	7	4	6	4	1
Relativer Anteil	4 %	8 %	28 %	16 %	24 %	16 %	4 %

**Tabella 2.3:** Verteilung des Alters in der Didgeridoo-Studie.

## DEFINITION 2.14

### Histogramm

Mit einem Histogramm wird die Verteilung von stetigen Daten visualisiert. Dazu wird der Bereich der Daten in gleiche, anliegende aber sich nicht überlappende Intervalle (Zellen, Klassen) zerlegt. Dann zählt man die Anzahl der Beobachtungen in jedem Intervall und erstellt ein Balkendiagramm.

# Deskription quantitativer Daten

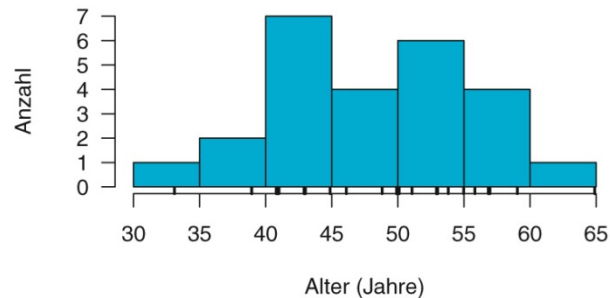


Abbildung 2.5: Histogramm des Alters der Patienten der Didgeridoo-Studie.

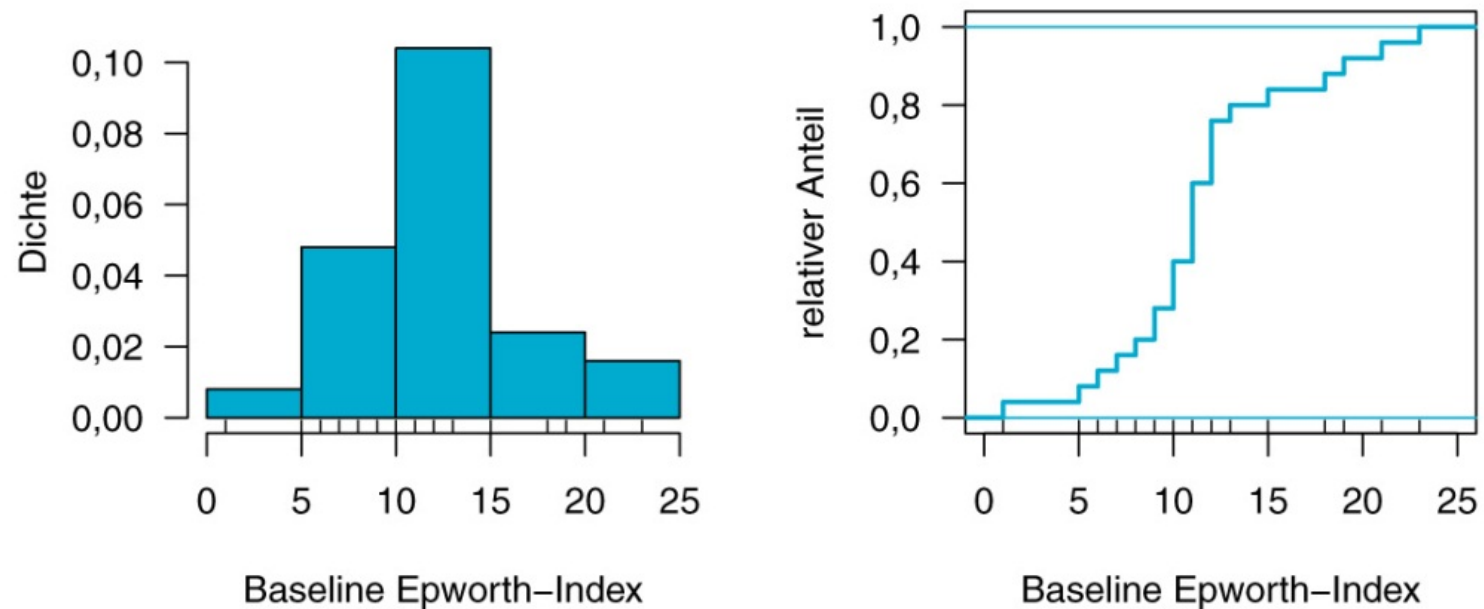
## DEFINITION 2.5

### Normalverteilung

Die drei Haupteigenschaften der Normalverteilung sind:

- Symmetrie, d. h., Abweichungen nach links und nach rechts vom Zentrum sind gleich häufig und gleich groß.
- „Große“ Abweichungen vom Zentrum sind selten.
- Die Normalverteilung lässt sich durch lediglich zwei Parameter eindeutig beschreiben: den Mittelwert und die Standardabweichung.

# Deskription quantitativer Daten



**Abbildung 2.10:** Histogramm und empirische Verteilungsfunktion der Baseline Epworth-Index Daten.

# Statistische Kennwerte

## DEFINITION 2.6 Mittelwert

Der Mittelwert (engl. mean) von  $n$  Beobachtungen  $x_1, x_2, \dots, x_n$  ist

$$\bar{x} = (x_1 + x_2 + \dots + x_n) / n = \frac{1}{n} \sum_{i=1}^n x_i.$$

## DEFINITION 2.7 Perzentil

Eine Zahl  $p_k$  heißt  $k$ -tes Perzentil ( $k$  bezeichnet hierbei eine ganze Zahl zwischen 1 und 99) einer Variablen, wenn mindestens  $k$  % der Beobachtungen der Variable kleiner oder gleich  $p_k$  und mindestens  $(100 - k)$  % größer oder gleich  $p_k$  sind.

## DEFINITION 2.12 Interquartilsabstand

Der **Interquartilsabstand** (engl. interquartile range, IQR) ist die Differenz zwischen dem 75. und dem 25. Perzentil.

# Statistische Kennwerte

## DEFINITION 2.13 Spannweite

Die Spannweite (engl. range) ist die Differenz zwischen Maximum und Minimum und gibt den Bereich an, in dem die Daten liegen.

## DEFINITION 2.8 Varianz

Die Varianz (engl. variance) ist die mittlere quadratische Abweichung der Beobachtungen vom Mittelwert:

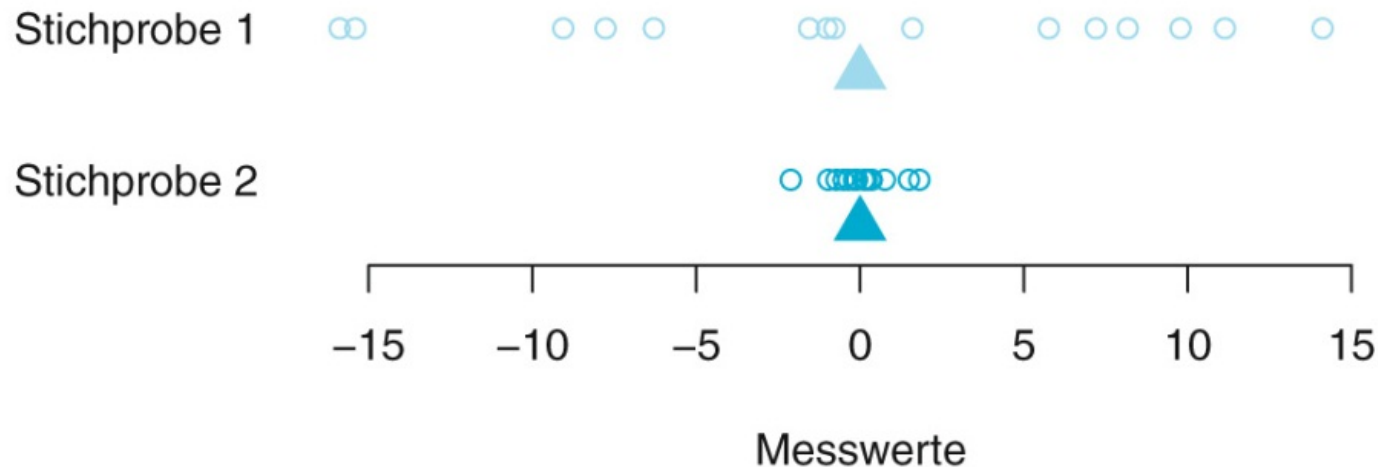
$$\begin{aligned} s^2 &= \{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2\} / (n - 1) \\ &= \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})^2. \end{aligned}$$

## DEFINITION 2.9 Standardabweichung

Die Standardabweichung (SD) ist die Wurzel aus der Varianz:

$$s = \sqrt{s^2}.$$

# Streuung



**Abbildung 2.4:** Zwei Stichproben mit demselben Mittelwert, aber verschiedener Streuung.

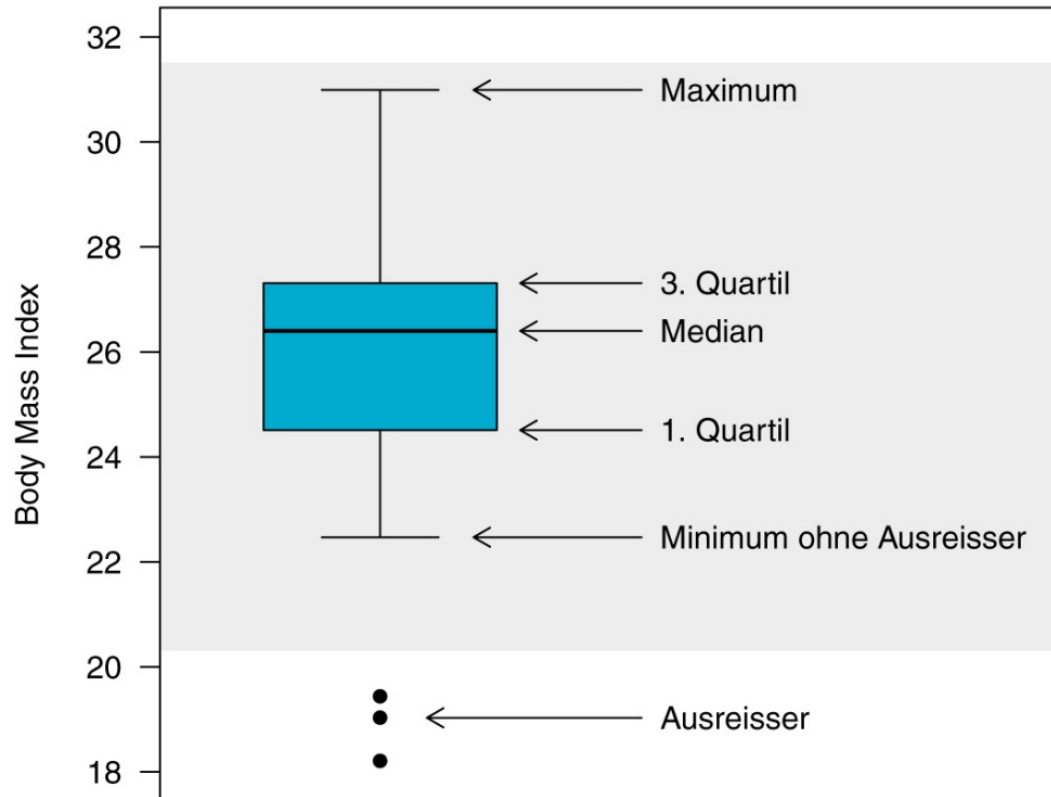
## DEFINITION 2.11

### Variationskoeffizient

Der **Variationskoeffizient** (CV) ist der Quotient von Standardabweichung und Mittelwert.

$$CV = s/\bar{x}$$

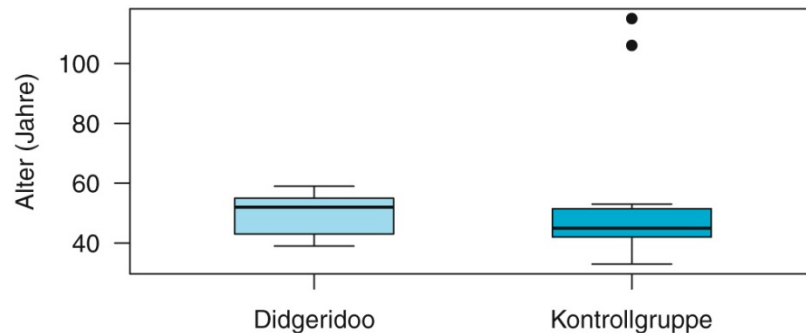
# Boxplot



**Abbildung 2.7:** Boxplot des Body-Mass-Index in der Didgeridoo-Studie. Das graue Band beschreibt den Bereich 1. Quartil  $-1,5 \cdot \text{IQR}$  bis zum 3. Quartil  $+1,5 \cdot \text{IQR}$ .



# Vergleiche mit Boxplot



**Abbildung 2.9:** Boxplots des Alters aller Patienten nach Behandlungsgruppe mit den fehlerhaften Daten aus Abbildung 2.3 rechts.

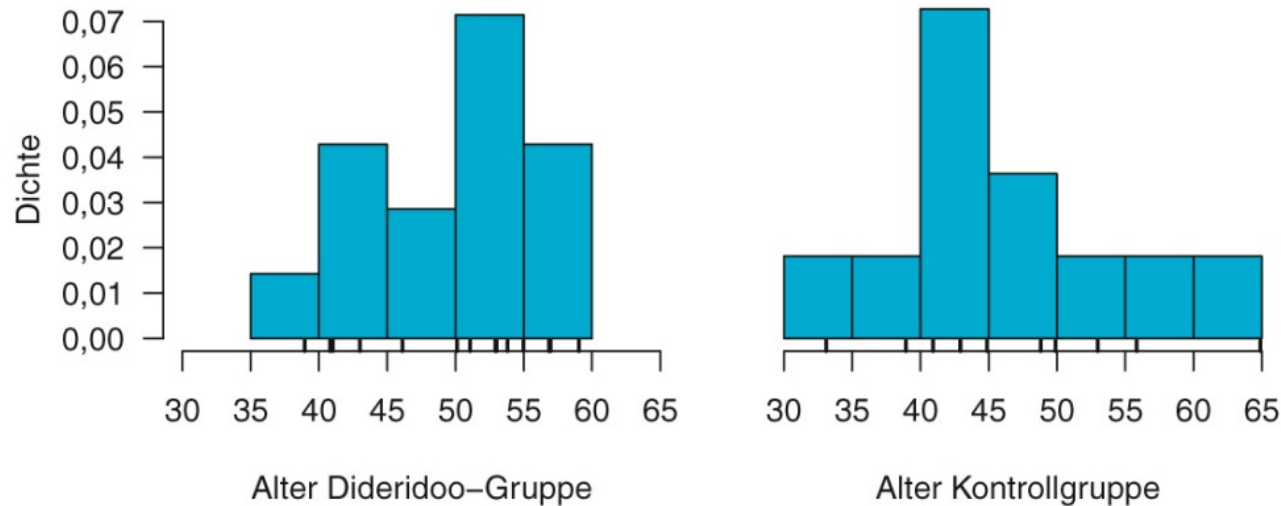
## DEFINITION 2.15

### Boxplot

Die „Box“ im **Boxplot** (engl. box and whiskers plot) gibt den Bereich vom 25. zum 75. Perzentil an, der horizontale Strich in der Box den Median. Die Stäbe (engl. whiskers), die aus der Box herausragen, sind nicht einheitlich definiert. Bei einfachen Boxplots reichen sie zum Minimum und zum Maximum. Eine verbreitete Definition, die auf John W. Tukey zurückgeht, besteht darin, die Länge der Whiskers auf maximal das 1,5-fache der Boxlänge zu beschränken. Beobachtungen außerhalb dieses Bereichs werden als „Ausreißer“ gekennzeichnet.

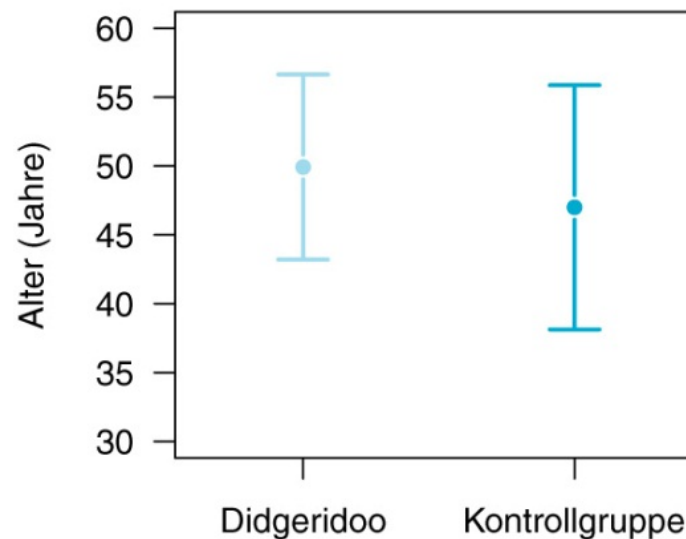


# Deskriptive Vergleiche Didgeridoo Studie



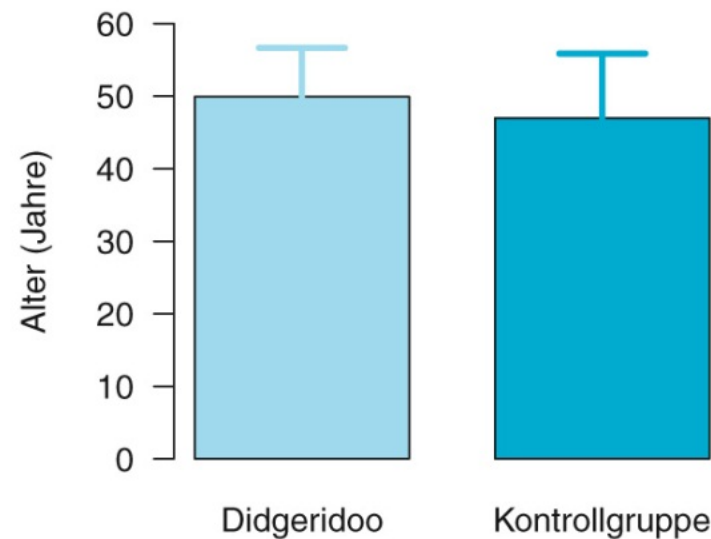
**Abbildung 2.6:** Histogramme des Alters aller Patienten in der Didgeridoo-Studie, nach Behandlungsgruppe.

# Deskriptive Vergleiche Didgeridoo Studie



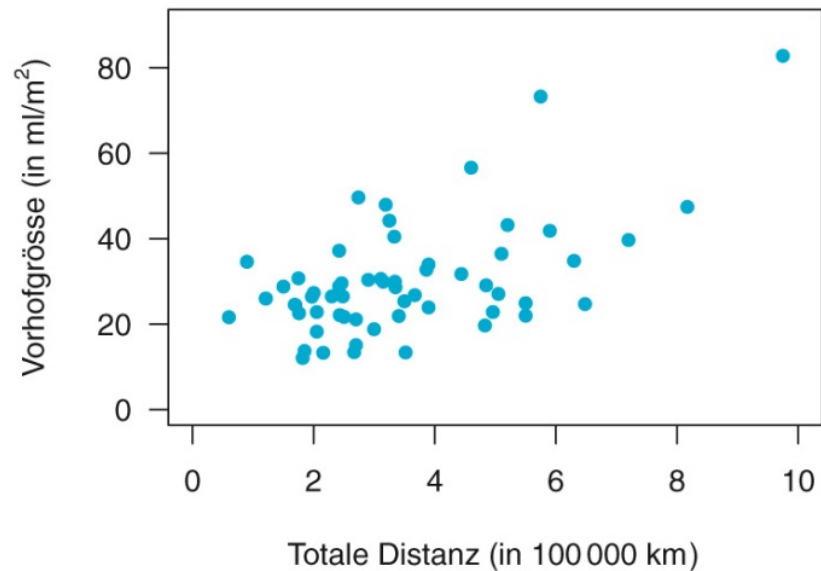
**Abbildung 2.12:** Fehlerbalken-Diagramm des Alters nach Behandlungsgruppe (Mittelwert  $\pm$  Standardabweichung).

# Deskriptive Vergleiche Didgeridoo Studie



**Abbildung 2.11:** Balkendiagramme des Alters nach Behandlungsgruppe (Mittelwert  $\pm$  Standardabweichung).

# Streudiagramm und Korrelation



**Abbildung 6.1:** Streudiagramm der Größe des linken Vorhofs gegen die totale Distanz.

Variable	<i>n</i>	Min.	Median	$\bar{x}$	Max.	<i>s</i>
Totale Distanz (in 100 000 km)	61	0,6	3,1	3,4	9,8	1,8
Vorhofgröße (in $ml/m^2$ )	61	12,1	27,1	30,2	82,8	13,0

**Tabelle 6.1:** Deskriptive Statistiken in der Rennfahrerstudie.

# Korrelationskoeffizient

## DEFINITION 6.1 Pearson-Korrelation

Die Korrelation einer Stichprobe von  $n$  Paaren von Beobachtungen  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  ist

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x \cdot s_y}, \quad (6.1)$$

wobei

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{und} \quad s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} \quad (6.2)$$

die geschätzten Standardabweichungen der Variablen  $x$  und  $y$  bezeichnen, und

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

die geschätzte **Kovarianz** zwischen beiden Variablen.

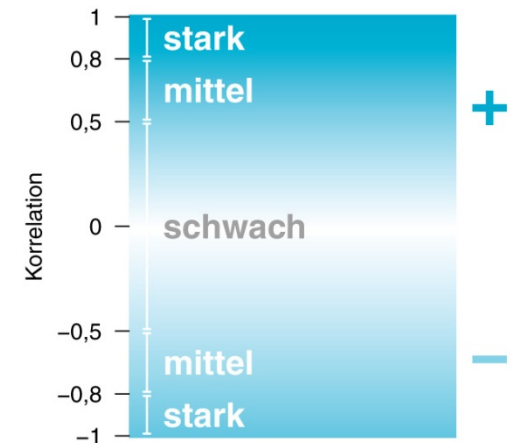


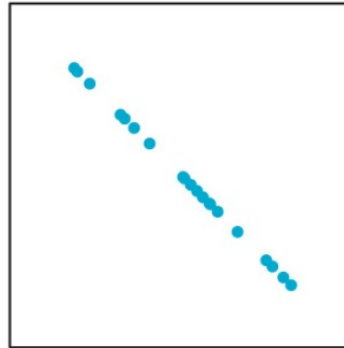
Abbildung 6.2: Die Bewertung von Korrelationskoeffizienten.

# Korrelationskoeffizient

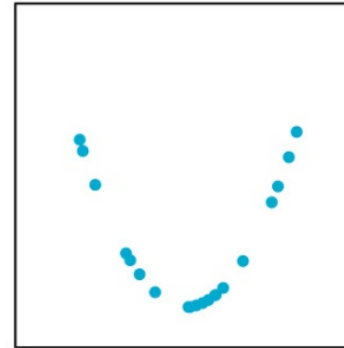
Grafik 1:  $r = 1$



Grafik 2:  $r = -1$



Grafik 3:  $r = 0$



Grafik 4:  $r = 0,7$



Grafik 5:  $r = 0,7$

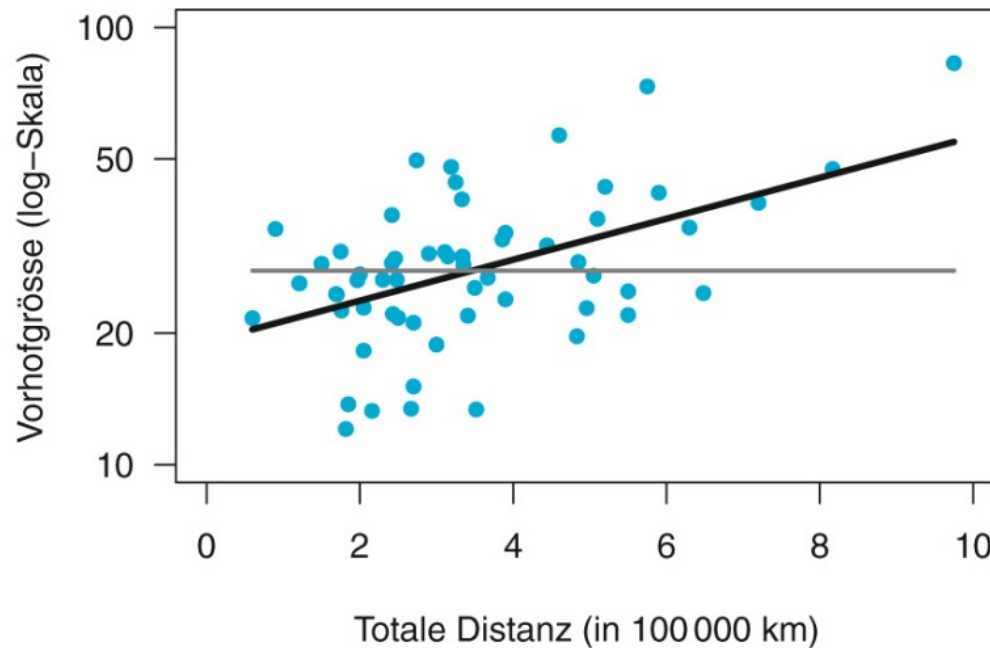


Grafik 6:  $r = 0,7$



**Abbildung 6.3:** Illustration von speziellen Korrelationen.

# Regression



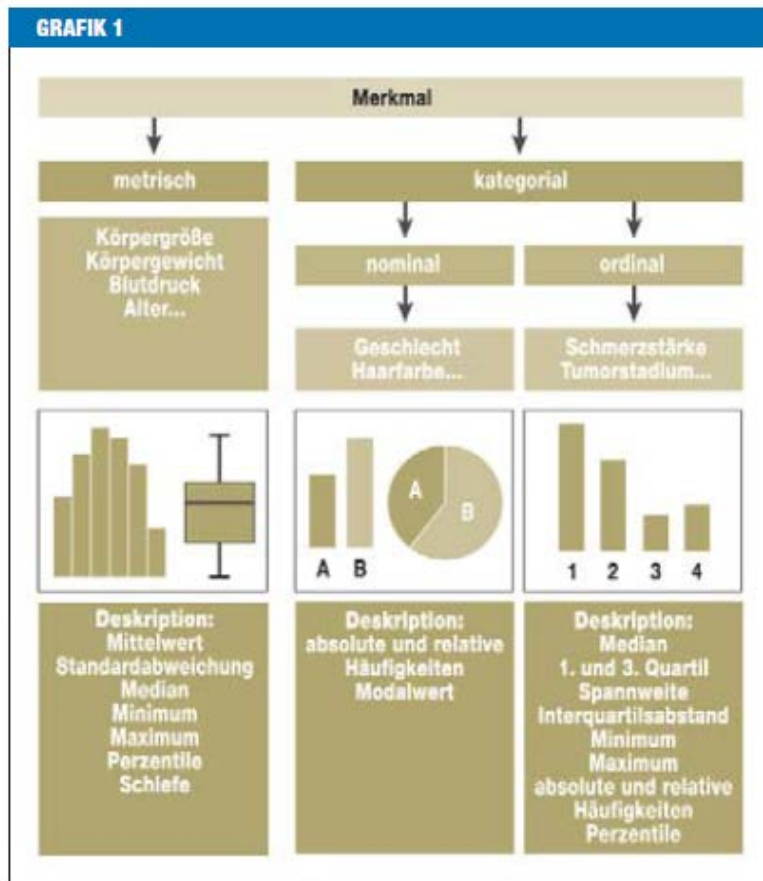
**Abbildung 6.9:** Streudiagramm der logarithmierten Größe des linken Vorhofs gegen die totale Distanz mit Regressionsgerade.

# Deskriptive Statistik

Qualitative Merkmale	Quantitative Merkmale normalverteilt	beliebig verteilt
n, %	Arithmetischer Mittelwert	Median
	Standardabweichung	Perzentile
Kreis-, Balkendiagramm	Histogramm	Boxplot
Beispiele:		
Geschlecht, Ansprechen auf Therapie, etc.	Blutdruck	CRP, GGT Schweregrade



# Merkmaltypen und statistische Maßzahlen, Grafiken

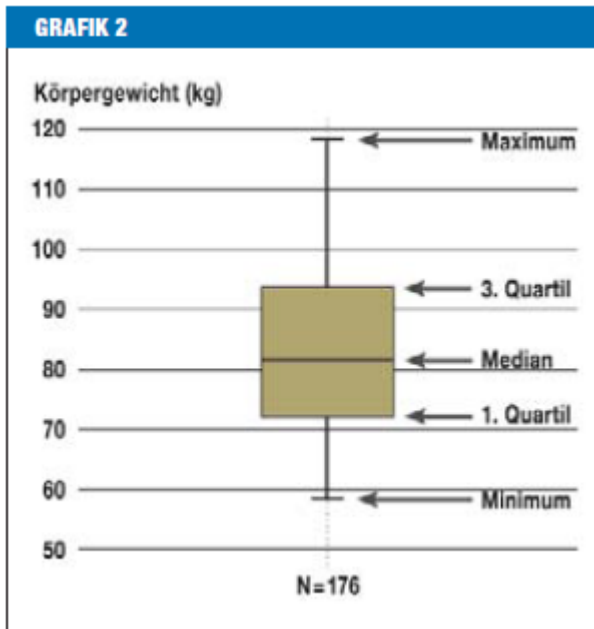


Schema der Merkmaltypen und geeignete statistische Maßzahlen zur deskriptiven Darstellung

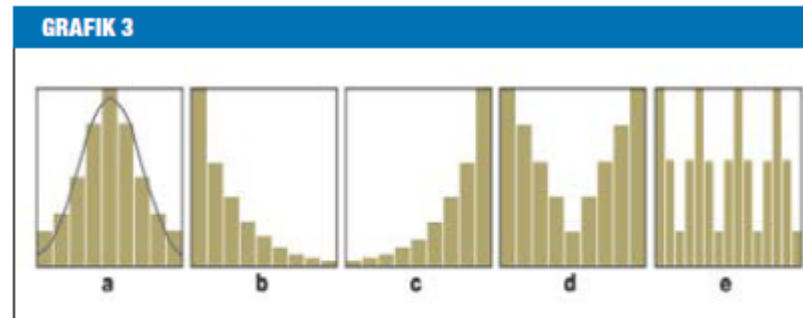
**Table 1. Baseline Characteristics of Patients in the Intention-to-Treat Population.\***

Characteristic	Capecitabine (N=1004)	Fluorouracil plus Levocovorin (N=983)
<b>Sex (%)</b>		
Male	54	54
Female	46	46
<b>Age (yr)</b>		
Median	62	63
Range	25–80	22–82
<b>Age group (%)</b>		
<70 yr	81	79
≥70 yr	19	21
<b>ECOG performance score (%)</b>		
0	85	85
1	15	15
<b>Nodal status — (%)</b>		
N1	69	71
N2	31	29
<b>Tumor stage (%) †</b>		
T1 or 2	10	10
T3	76	76
T4	14	14
<b>Carcinoembryonic antigen level (%)</b>		
≤ULN	83	85
>ULN	9	7
Missing data	8	8

# Statistische Grafiken

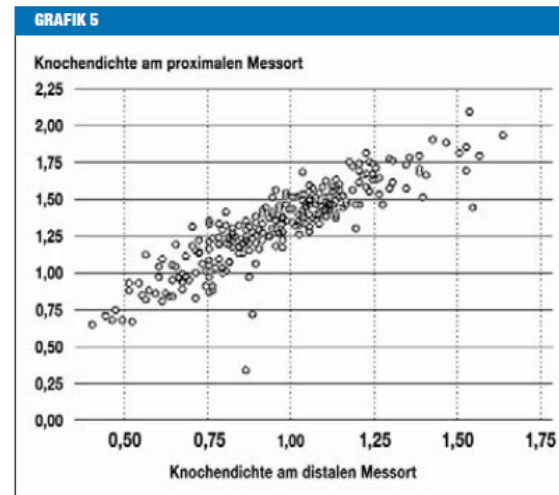


Beispiel für einen Boxplot

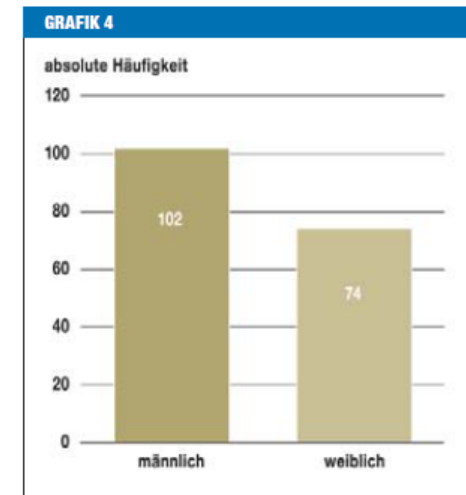


Beispiele für Verteilungsformen in Histogrammen

a) Normalverteilung (symmetrisch), b) linksgipflig (= rechtsschief); c) rechtsgipflig (= links-schief); d) zweigipflig (symmetrisch); e) mehrgipflig



Beispiel für ein Streudiagramm



# Deskriptive Statistik

## KASTEN

Mittelwert	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Varianz	$\text{Var} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
Standardabweichung	$s = \sqrt{\text{Var}}$
Schiefe	$g = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$
Median	$\tilde{x} = x_{(n+1)/2}$ falls $n$ ungerade $\tilde{x} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$ falls $n$ gerade
Spannweite	$R = x_{\max} - x_{\min}$

mit

$n$  = Stichprobenumfang beziehungsweise Fallzahl

$x_i$  = Messwert für  $i$ -tes Stichprobenelement  
beziehungsweise  $i$ -ten Fall, wobei  $i = 1, \dots, n$

$x_{(i)}$  = bezeichnet den  $i$ -ten Wert in der aufsteigend  
geordneten Reihe der Messwerte, wobei  
 $i = 1, \dots, n$

## Median oder Mittelwert?

St. Lange<sup>1</sup>, R. Bender<sup>2</sup>

<sup>1</sup> Abteilung für Medizinische Informatik, Biometrie und Epidemiologie der Ruhr-Universität Bochum

<sup>2</sup> Fakultät für Gesundheitswissenschaften, AG Epidemiologie und medizinische Statistik, Universität Bielefeld

Tab.2 Übersetzungen (deutsch – englisch)

(arithmetischer) Mittelwert	(arithmetic) mean
Median	Median
Ausreißer	Outlier
Stichprobenumfang	Sample size
schiefe Verteilung	Skewed distribution
zensierte Daten	Censored data

## Variabilitätsmaße

St. Lange<sup>1</sup>, R. Bender<sup>2</sup>

<sup>1</sup> Abteilung für Medizinische I

<sup>2</sup> Fakultät für Gesundheitswis

Tab.1 Übersetzungen (deutsch – englisch)

Spannweite	range
Standardabweichung	standard deviation
Varianz	variance
Standardfehler des Mittelwertes	standard error of the mean
Variabilitätsmaß	measure of variability
Spannweite	range
Interquartilsabstand	interquartile range
Summe der Abweichungsquadrate	sum of squares

# Mittelwert vs. Median - Beispiel



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

Wann verwendet man den Mittelwert, wann den Median?

---

**Daten:** 12  
14  
16  
18  
20

**Mittelwert:**  $\frac{12+14+16+18+20}{5} = 16$

**Median:** mittlere Wert der Rangliste: 16

→ Bei Vorliegen einer Normalverteilung sind Mittelwert und Median gleich.

**Daten mit einem Ausreißer:** 12  
14  
16  
18  
20  
40

**Mittelwert:**  $\frac{12+14+16+18+20+40}{6} = 20$

**Median:** mittlerer Wert der Rangliste – bei einer geraden Anzahl an Werten wird der Mittelwert der beiden mittleren

Werte berechnet →  $\frac{16+18}{2} = 17$

→ Bei Vorliegen von nicht normalverteilten Daten sind Mittelwert und Median nicht gleich.

→ Median ist robust gegen Ausreißer, Mittelwert nicht.

# Standardabweichung oder Standardfehler?

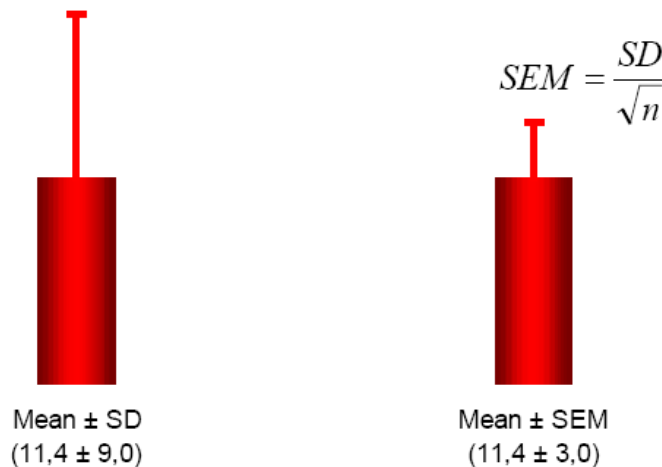
- Standardfehler beschreibt nicht die Daten, sondern gibt die Genauigkeit des Mittelwerts als Schätzwerts an.

$1-p=0.$

**SD > SEM**

$$SD = \sqrt{\frac{(\bar{x} - x_1)^2 + (\bar{x} - x_2)^2 + (\bar{x} - x_3)^2 + \dots + (\bar{x} - x_n)^2}{n-1}}$$

Alter von 9 Kindern





# Übung: Statistische Maßzahlen

- Berechnen Sie bitte den arithmetischen:
  - Mittelwert
  - Median
  - Varianz
  - Standardabweichung
  - Spannweite
  - Interquartilsabstand
  - und den Variationskoeffizient

Erstellen Sie Histogramme und Boxplots für:



- Punktezahlen von 20 Studenten  
6,3,7,5,6,4,4,6,7,3,5,9,6,4,2,7,5,5,8,6
- Anzahl der Angestellten in 20 Apotheken  
2,3,3,3,4,4,4,4,5,5,5,5,5,5,6,6,6,8,10,15
- Krankheitstage von 20 Personen  
0,0,0,0,0,0,0,0,1,1,1,1,1,2,2,2,3,3,15,76

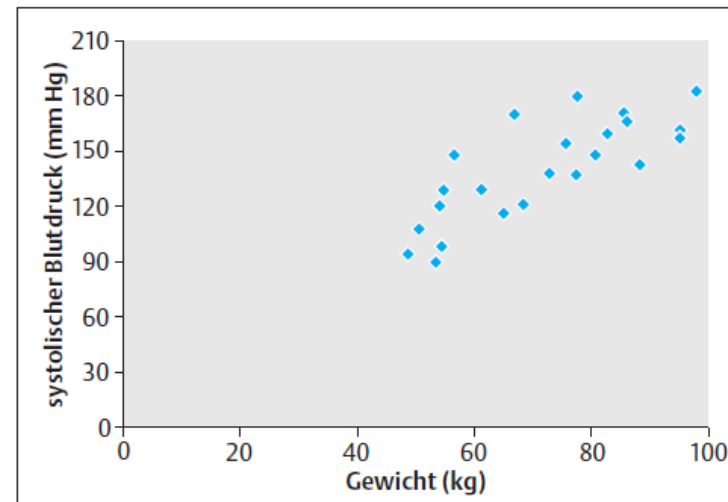
## (Lineare) Regression/Korrelation

St. Lange<sup>1</sup>, R. Bender<sup>2</sup>

<sup>1</sup> Abteilung für Medizinische Informatik, Biometrie und Epidemiologie der Ruhr-Universität Bochum

<sup>2</sup> Fakultät für Gesundheitswissenschaften, AG Epidemiologie und medizinische Statistik, Universität Bielefeld

Neben der univariaten, das heißt auf ein einzelnes Merkmal bezogenen Analyse von Daten aus einer klinischen Studie, ist man häufig daran interessiert, den Zusammenhang zwischen zwei (bivariat) oder mehreren (multivariat) Variablen zu betrachten. Bei Betrachtung von zwei quantitativen Merkmalen bietet sich als anschauliche, graphische Darstellungsweise die Punktwolke an, bei der die Wertepaare durch einen Punkt in einem Koordinatensystem abgebildet werden (**Abb. 1**). Damit wird sofort visuell erfassbar, ob überhaupt ein Zusammenhang besteht, und wenn ja, wie stark er ist. **Tab. 1** enthält die Werte für den systolischen Blutdruck und das Körpergewicht von 24 zufällig ausgewählten Patienten einer dermatologischen Ambulanz. **Abb. 1** zeigt die dazugehörige Punktwolke, die einen recht deutlichen Zusammenhang zwischen den beiden Merkmalen erkennen lässt.





# Inferenzstatistik III: Zusammenhänge, Korrelations- und Regressionsanalysen



## Erraten von Korrelationen

Ende Neustart Hilfe!

Plot A: 0.57	Plot B: -0.86	Plot C: 0.22	Plot D: -0.67	
$r = -0.86$ <input type="radio"/> A	<input checked="" type="radio"/> B	<input type="radio"/> C	<input type="radio"/> D	Neue Plots
$r = -0.67$ <input type="radio"/> A	<input type="radio"/> B	<input type="radio"/> C	<input checked="" type="radio"/> D	Neustart
$r = 0.22$ <input type="radio"/> A	<input type="radio"/> B	<input checked="" type="radio"/> C	<input type="radio"/> D	Serie: 4 Richtige in Folge
$r = 0.57$ <input checked="" type="radio"/> A	<input type="radio"/> B	<input type="radio"/> C	<input type="radio"/> D	
Sie haben 4 richtig.		4 richtig von 16 Plots – 25%.		

# Correlation

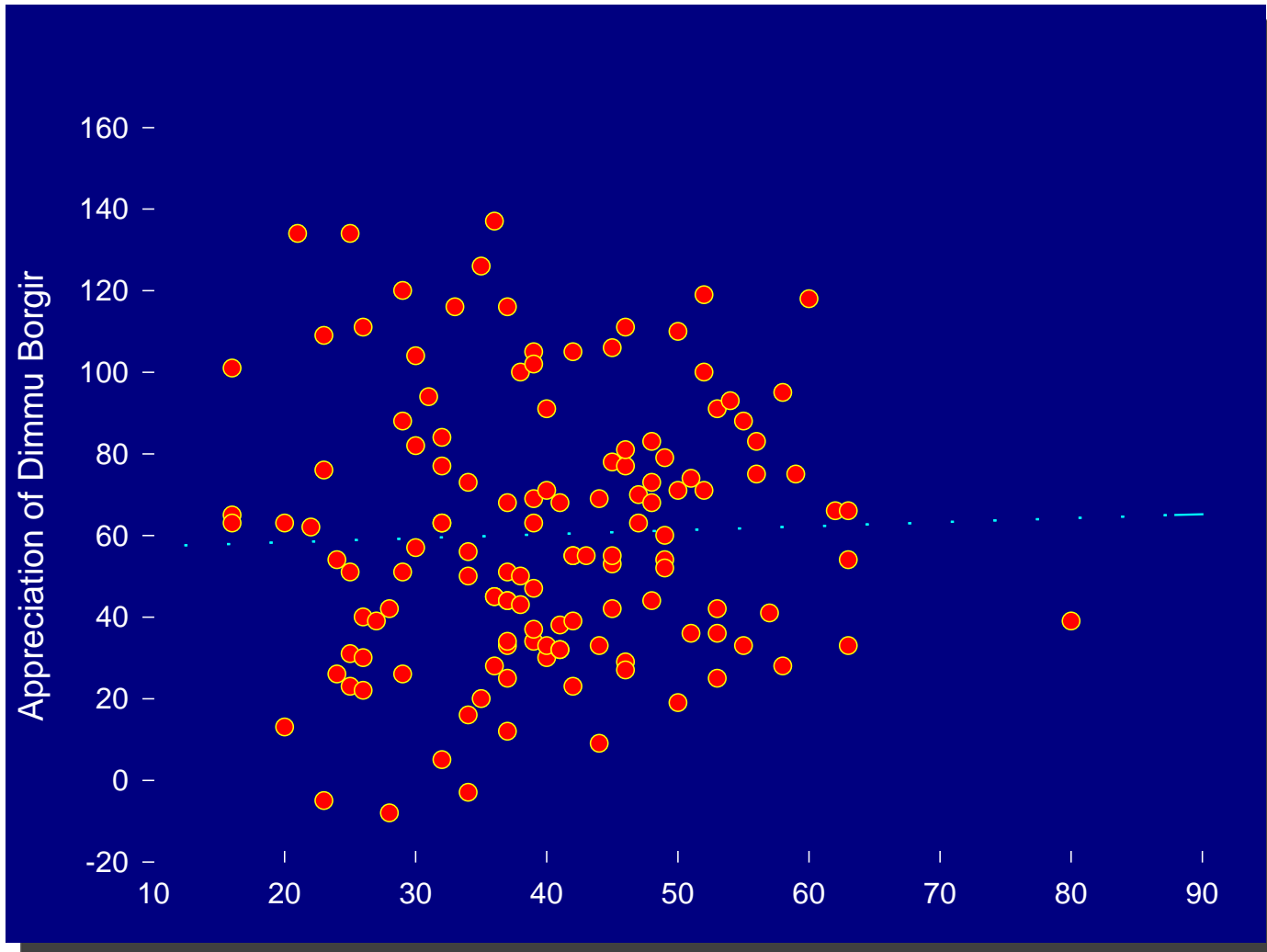


- Measuring Relationships
  - Scatterplots
  - Covariance
  - Pearson's Correlation Coefficient
- Nonparametric measures
  - Spearman's Rho
  - Kendall's Tau
- Interpreting Correlations
  - Causality

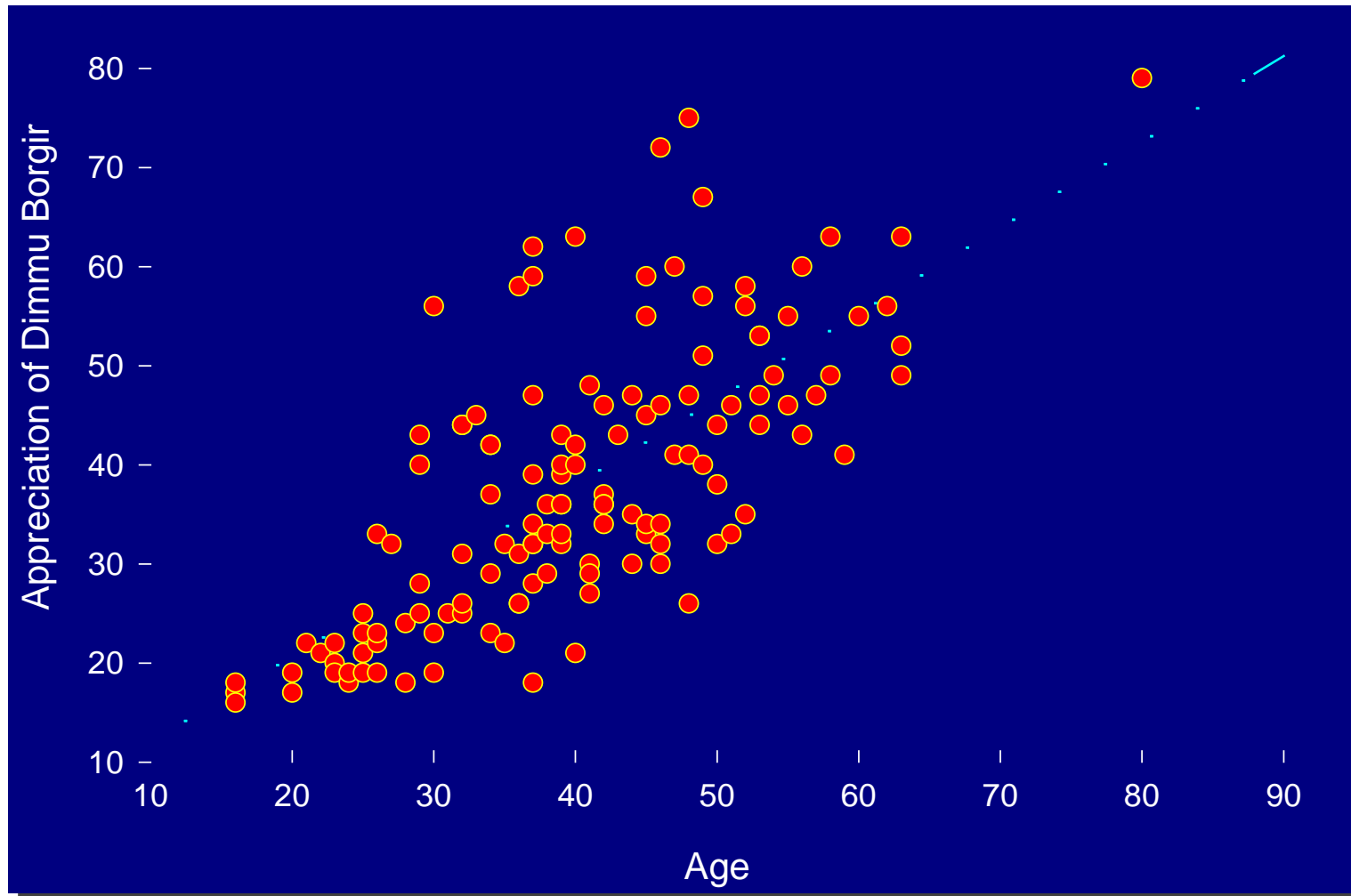
# What is a Correlation?

- In statistics, **dependence** or **association** is any statistical relationship, whether causal or not, between two random variables or bivariate data.
  - This means that the marginal distribution of a random variable A is different from the distribution of A knowing B
- **Correlation** is a way of measuring the relationship between two metric or ordinal variables
- Correlation is a measure of association
- In **common usage**, it most often refers to the extent to which two variables have a **linear relationship** with each other
- (Linear) correlation  $\Rightarrow$  Dependence
- But: Dependence  $\not\Rightarrow$  (Linear) correlation

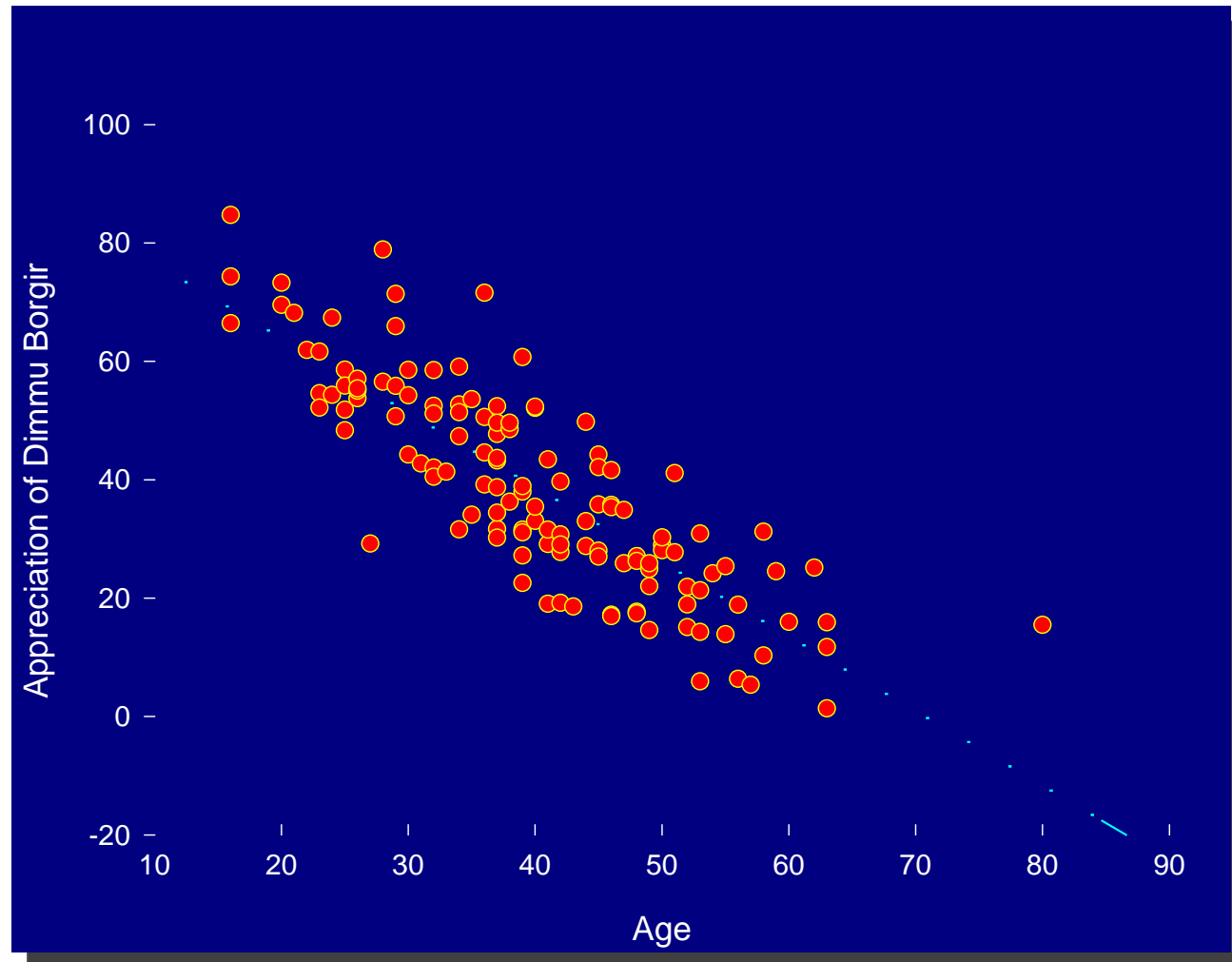
# Very small relationship



# Positive relationship



# Negative relationship

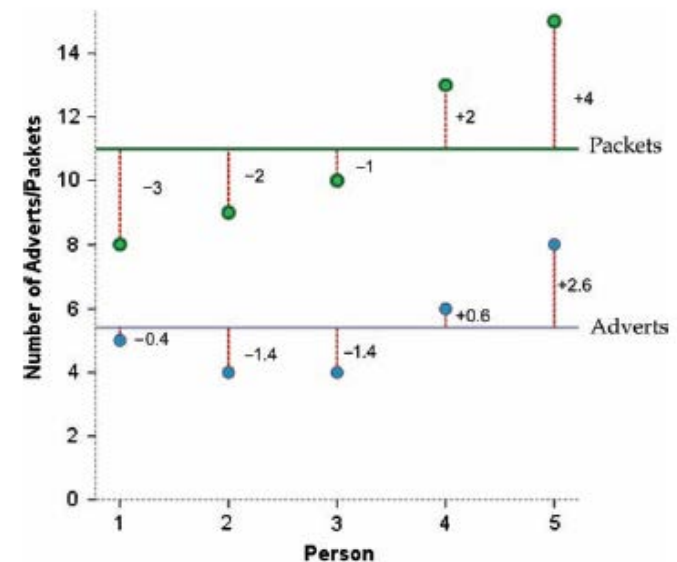


# Measuring (linear) relationships

- We need to see whether as one variable increases, the other increases, decreases or stays the same.
- This can be done by calculating the **covariance**.
  - We look at how much each score deviates from the mean.
  - If both variables deviate from the mean by the same amount, they are likely to be related.

- Example:

Participant:	1	2	3	4	5	Mean	s
Adverts Watched	5	4	4	6	8	5.4	1.67
Packets Bought	8	9	10	13	15	11.0	2.92



# Covariance

- The **variance** tells us by how much scores deviate from the mean for a single variable.
- It is closely linked to the sum of squares.

$$\begin{aligned}\text{Variance} &= \frac{\sum(x_i - \bar{x})^2}{N-1} \\ &= \frac{\sum(x_i - \bar{x})(x_i - \bar{x})}{N-1}\end{aligned}$$

- **Covariance** is similar – it tells us by how much scores on two variables differ from their respective means.

$$\text{Cov}(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$



# Example



<i>Participant:</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>Mean</i>	<i>s</i>
Adverts Watched	5	4	4	6	8	5.4	1.67
Packets Bought	8	9	10	13	15	11.0	2.92

$$\begin{aligned}\text{cov}(x, y) &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1} \\ &= \frac{(-0.4)(-3) + (-1.4)(-2) + (-1.4)(-1) + (0.6)(2) + (2.6)(4)}{4} \\ &= \frac{1.2 + 2.8 + 1.4 + 1.2 + 10.4}{4} \\ &= \frac{17}{4} \\ &= 4.25\end{aligned}$$

# Pearson correlation coefficient (1)

- It depends upon the units of measurement.
  - E.g. The Covariance of two variables measured in Miles might be 4.25, but if the same scores are converted to Km, the Covariance is 11.
- One solution: standardise it!
  - Divide by the standard deviations of both variables.
- The standardised version of Covariance is known as the **Pearson correlation coefficient**.

$$r = \frac{Cov_{xy}}{s_x s_y}$$
$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

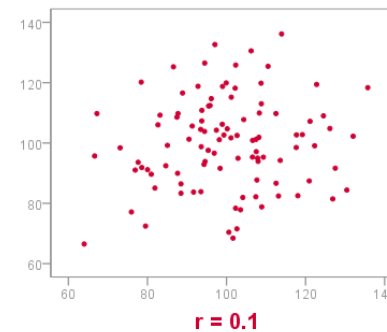
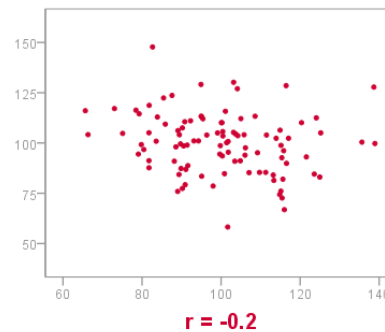
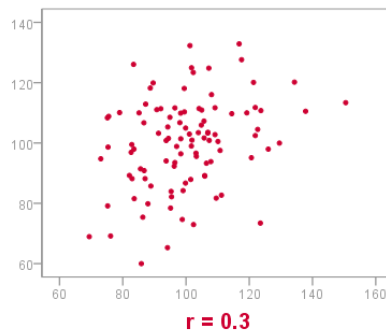
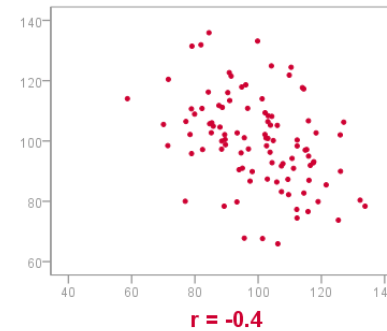
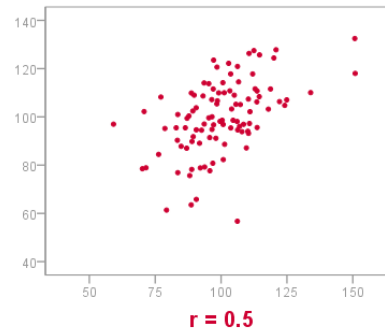
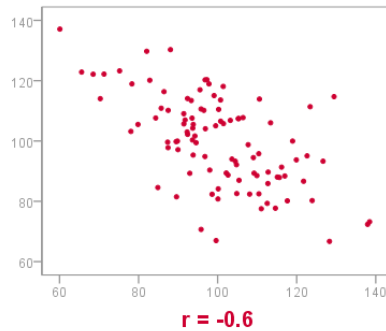
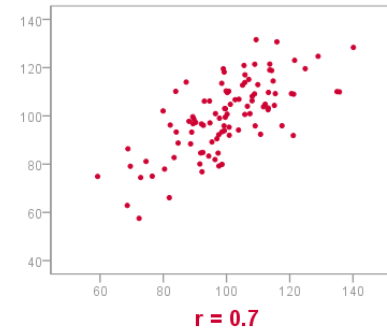
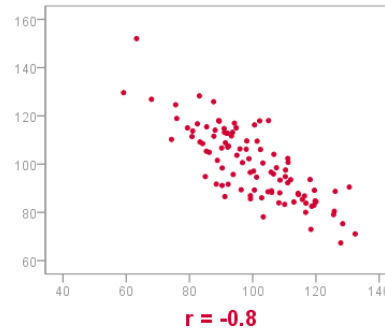
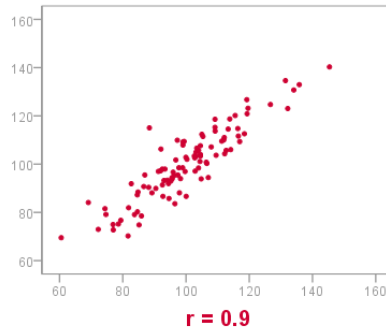
In the example:

$$r = \frac{Cov_{xy}}{s_x s_y}$$
$$= \frac{4.25}{1.67 \times 2.92}$$
$$= .87$$

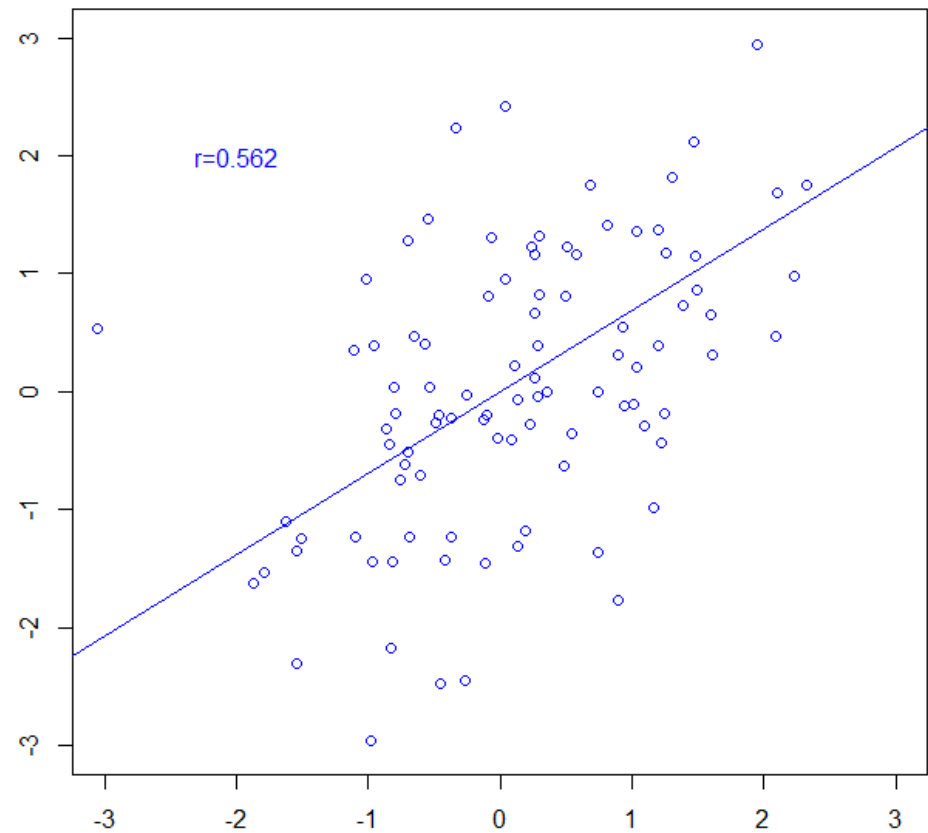
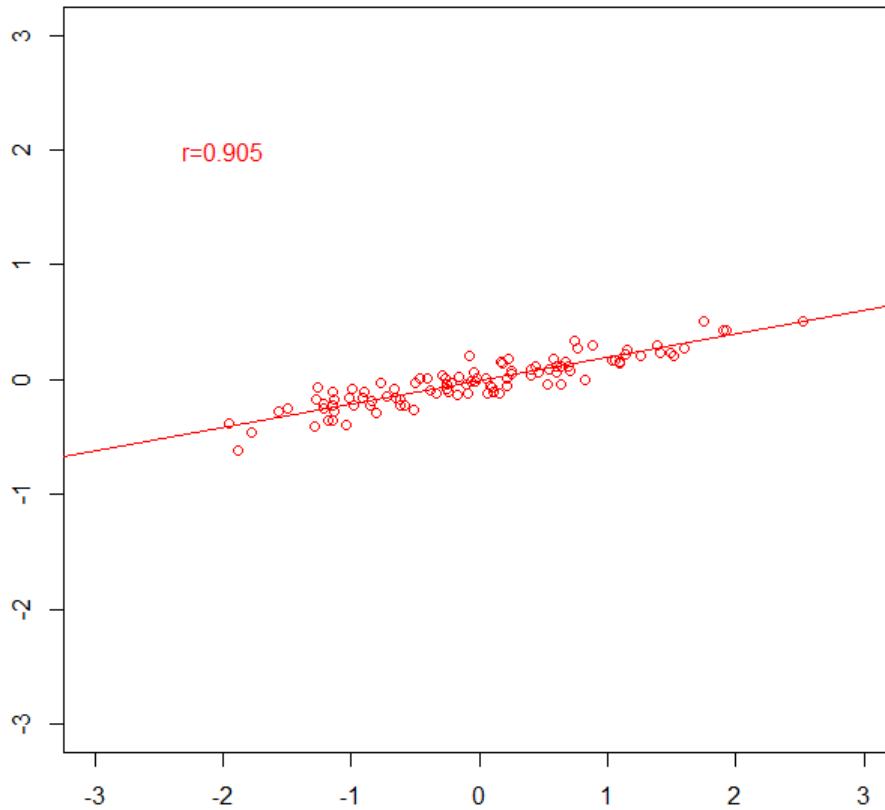
# Pearson Correlation Coefficient (2)

- Measure of linear dependence
- It varies between -1 (perfect negative) and +1 (perfect positive)
  - 0 = no relationship
- It is an effect size
  - $\pm 0.1$  = small effect
  - $\pm 0.3$  = medium effect
  - $\pm 0.5$  = large effect
- **Coefficient of determination,  $r^2$** 
  - By squaring the value of  $r$  you get the proportion of variance in one variable shared by the other.
- In our example:  $0.87^2$ , i.e. 75.7% of the variability in “Packets bought” can be explained by “Adverts watched”

# Examples

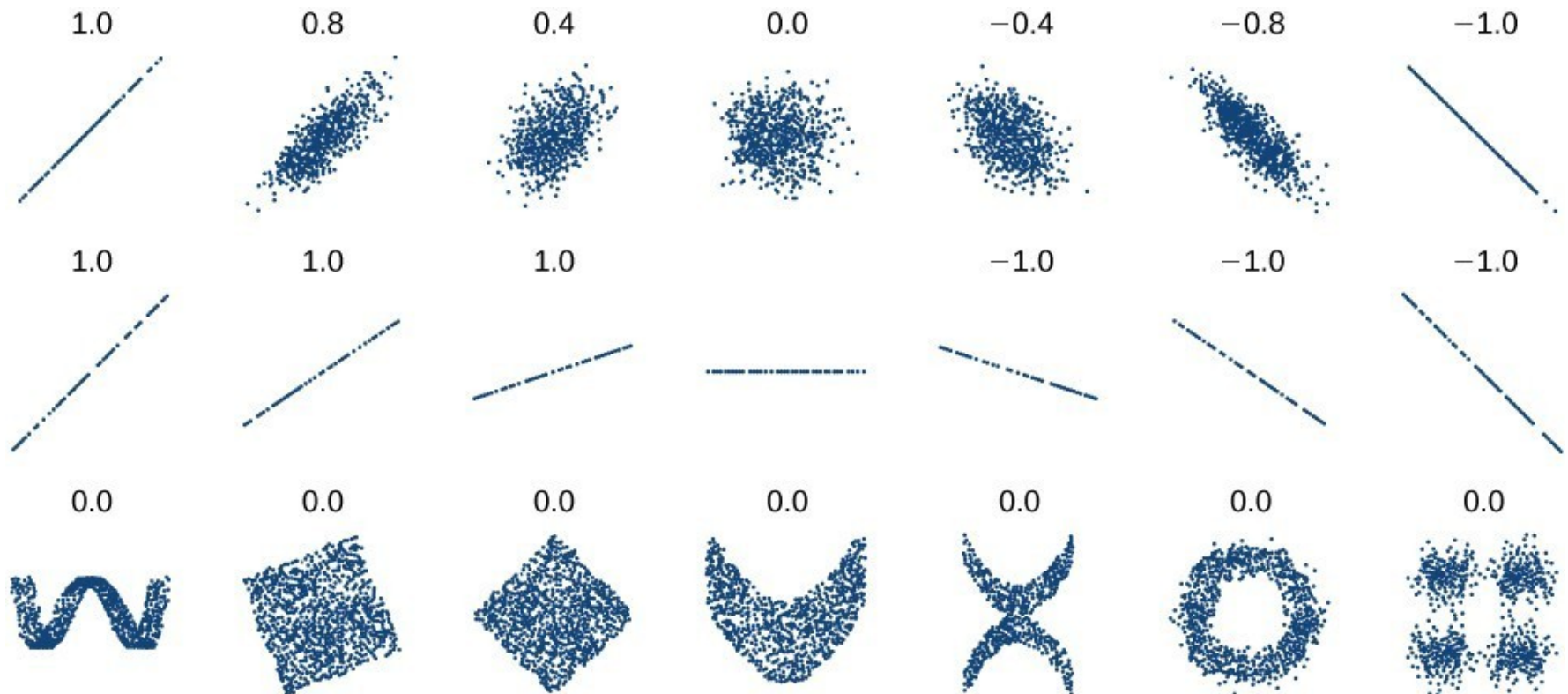


# Steepness of slope $\neq$ Strength of correlation!



# Non-linear dependencies

- Pearson correlation coefficient not meaningful for non-linear dependencies!

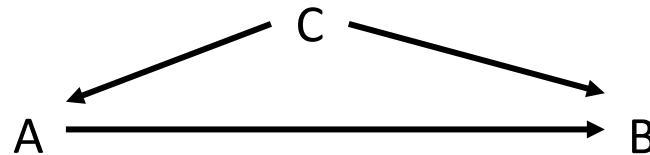


# Correlation/Association and causality (1)



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

- Direction of causality:
  - Correlation coefficients say nothing about which variable causes the other to change
- The third-variable problem:
  - in any correlation, causality between two variables cannot be automatically assumed because there may be other measured or unmeasured variables (“**confounders**”) affecting the results.



# Correlation/Association and causality (2)

- Due to confounders, the estimated of the effect can be distorted so severely that even the estimate shows even in the opposite direction than the true effect
- **Simpson's paradox**
- Appleton et al. (The American Statistician 1996;50(4))
- **UC Berkeley gender bias**

	Applicants	Admitted
Men	8442	44%
Women	4321	35%

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	<b>82%</b>
B	560	63%	25	<b>68%</b>
C	325	<b>37%</b>	593	34%
D	417	33%	375	<b>35%</b>
E	191	<b>28%</b>	393	24%
F	373	6%	341	<b>7%</b>





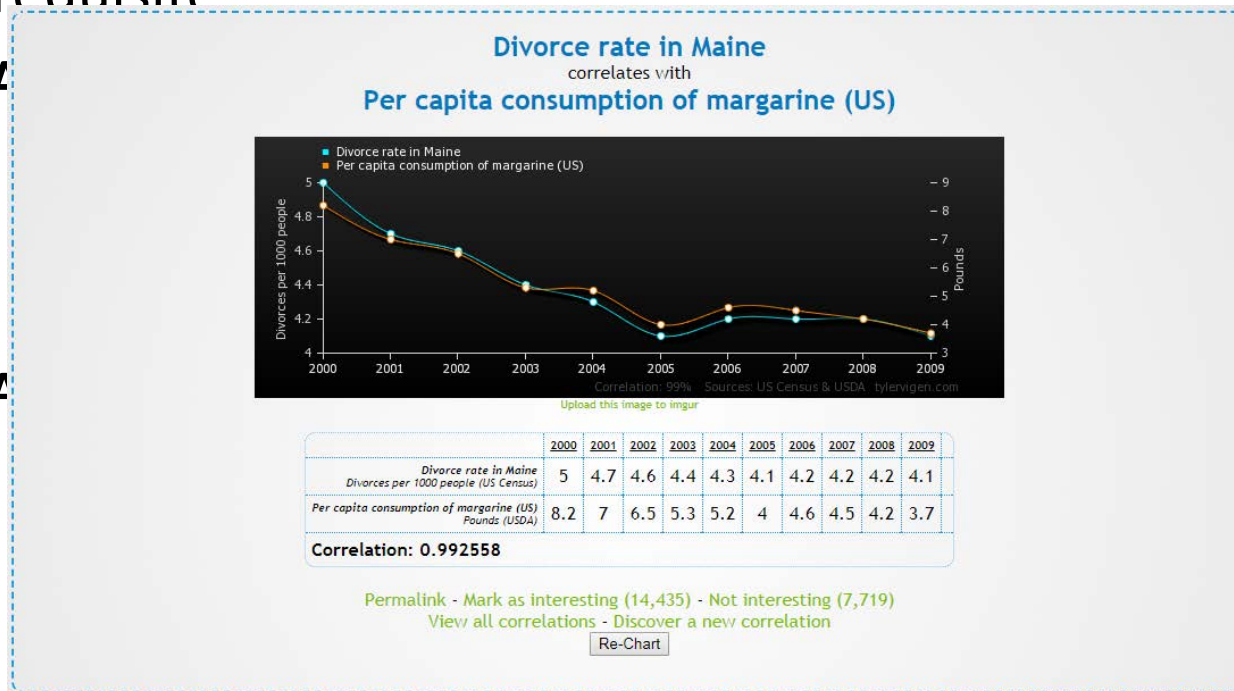
# Correlation/Association and causality (3)

- In many experiments/studies, the main focus is on establishing **causal relationships** rather than mere associations
- Prerequisite:

[http://tylervigen.com/view\\_correlation?id=1703](http://tylervigen.com/view_correlation?id=1703)

– A

• A



ables

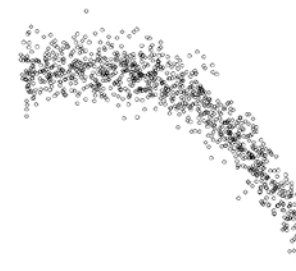
# Nonparametric Correlation

- **Spearman's Rho**
  - Pearson's correlation on the ranked data
- **Kendall's Tau**
  - Better than Spearman's for small samples
- Rho and Tau also appropriate for ordinal data if the assumed dependence is monotonic

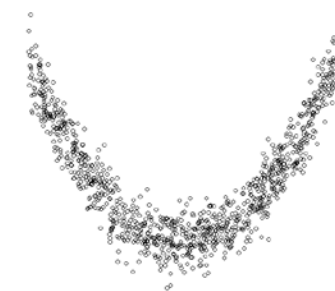
<i>Participant:</i>	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>Mean</i>	<i>s</i>
Adverts Watched	5	4	4	6	8	5.4	1.67
Packets Bought	8	9	10	13	15	11.0	2.92

$$R = 0.871$$

$$\text{Rho} = 0.667$$



Monotonic



Non-monotonic

# Conducting Correlation Analysis

---



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

Check  
assumptions/bias

# Reporting the results of a correlation analysis



**TABLE 7.2** An example of reporting a table of correlations

	<i>Exam Performance</i>	<i>Exam Anxiety</i>	<i>Revision Time</i>
Exam Performance	1	-.44*** [-.564, -.301]	.40*** [.245, .524]
Exam Anxiety	103	1	-.71*** [-.863, -.492]
Revision Time	103	103	1

ns = not significant ( $p > .05$ ), \*  $p < .05$ , \*\*  $p < .01$ , \*\*\*  $p < .001$ . BCa bootstrap 95% CIs reported in brackets.

# Exercise

---



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

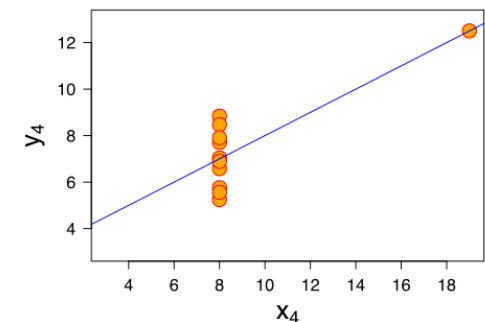
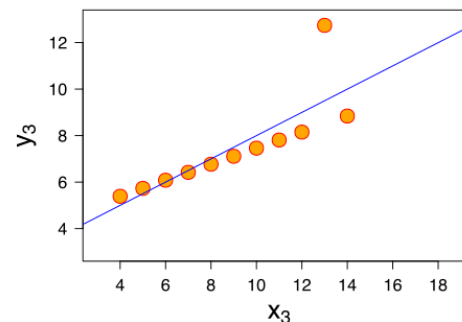
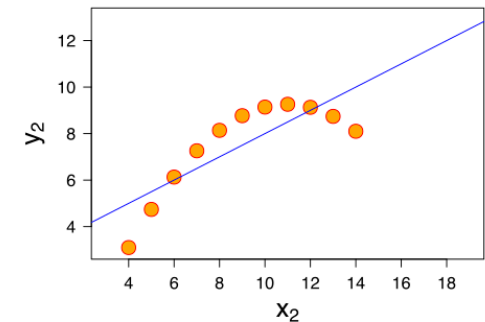
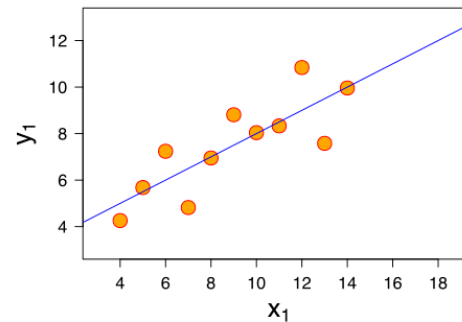
**Analyse the data in Data\_correlation.csv!**

# Importance of looking at a set of data graphically



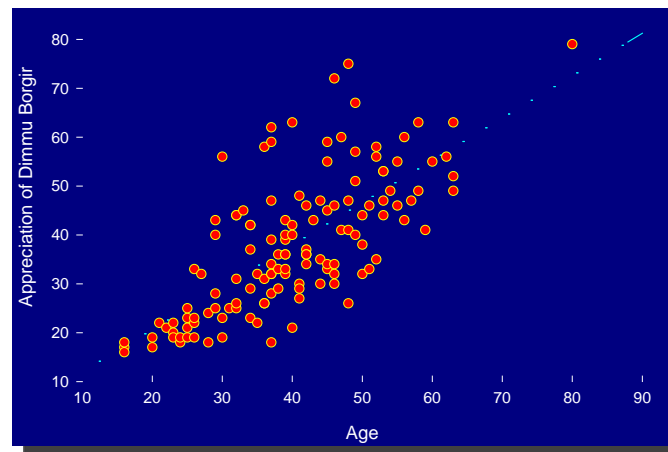
## Anscombe's quartet

- All four sets ( $x_1$  and  $y_1$ ,  $x_2$  and  $y_2$ ,  $x_3$  and  $y_3$ ,  $x_4$  and  $y_4$ ) are identical when examined using only summary statistics
- Mean of  $x$ : 9, Variance of  $x$ : 11
- Mean of  $y$ : 7.5, Variance of  $y$ : 4.125
- Correlation between  $x$  and  $y$ : 0.816
- Coefficient of determination: 0.67
- Linear regression line:  $y=3+0.5*x$
- **However, they vary considerably when graphed**



# (Simple) linear regression (1)

- **A procedure to predict the value of one variable Y from another variable X**
  - X – independent variable, predictor
  - Y – dependent variable, outcome
  - Because there is only one predictor -> simple linear regression
  - More than one predictor -> multiple linear regression
- It is a hypothetical model of the relationship between two variables.
- **The model used is a linear one**

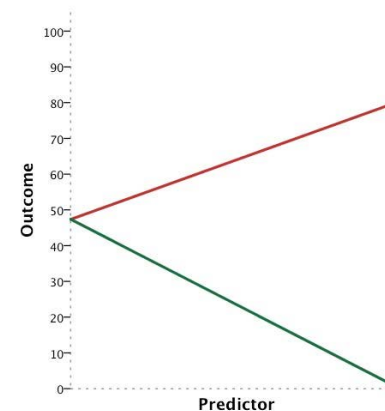


## (Simple) linear regression (2)

- Therefore, we describe the relationship using the equation of a straight line

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

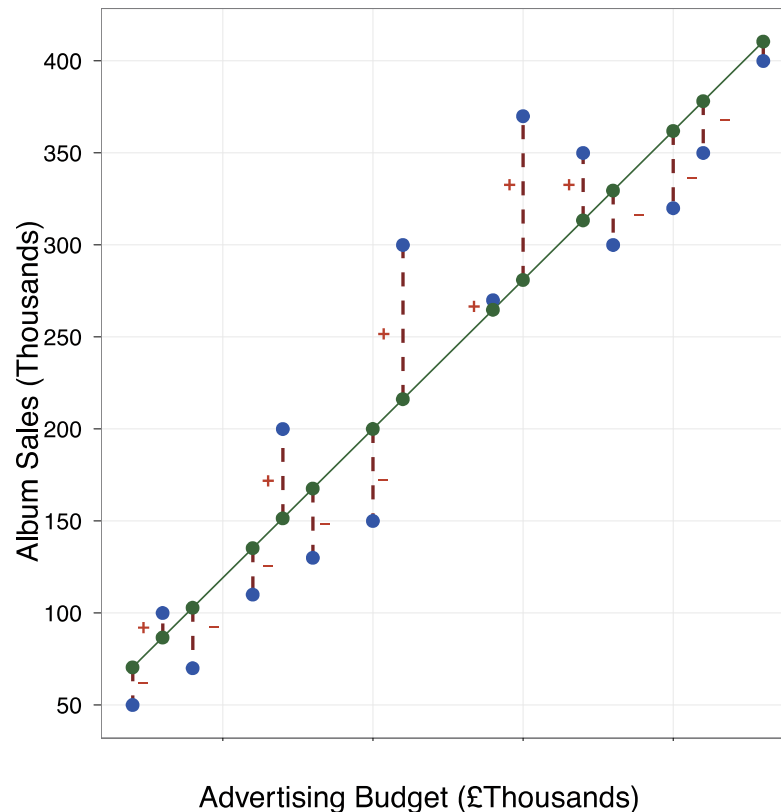
- $b_1$ 
  - **Regression coefficient** for the predictor
  - **Gradient** (slope) of the regression line
  - Direction/Strength of Relationship
- $b_0$ 
  - **Intercept** (value of Y when X = 0)
  - Point at which the regression line crosses the Y-axis (ordinate)





# The Method of Least Squares (1)

- From all possible lines choose the one which **minimizes the sum of the squared residuals** (differences between the prediction from the line and the true y-value)



$\sum$   
(or residuals)  
between the  
line and the  
actual data

# The Method of Least Squares (2)

$$\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \rightarrow \text{Min!}$$

Coefficients  $b_0$  and  $b_1$  of this minimization task are given by:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{x,y}}{s_{x,x}}$$
$$b_0 = \bar{y} - b_1 \bar{x}$$

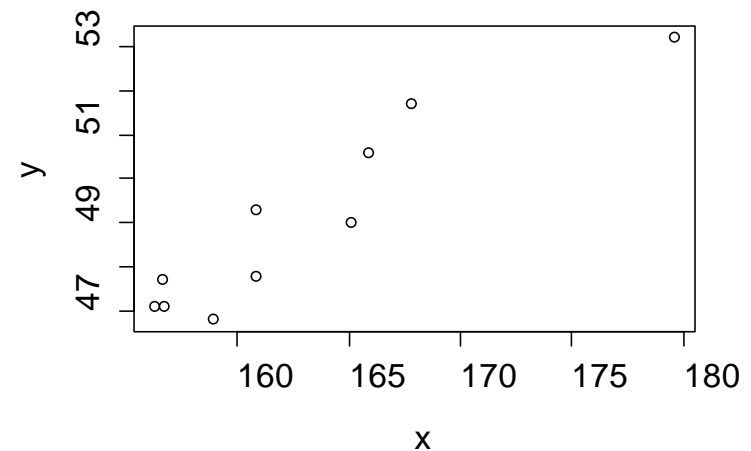
$b_1$  can also be formulated with the Pearson correlation coefficient:

$$b_1 = r_{xy} * \frac{s_y}{s_x}$$

# Example

PERSON $i$	1	2	3	4	5	6	7	8	9	10
KÖRPERGRÖSSE $x$	156.3	158.9	160.8	179.6	156.6	165.1	165.9	156.7	167.8	160.8
RINGGRÖSSE $y$	47.1	46.8	49.3	53.2	47.7	49.0	50.6	47.1	51.7	47.8

- Calculate variances of  $x$  and  $y$ , covariance and the Pearson correlation coefficient!
- Calculate the regression line according to the least squares approach!
- Calculate the variance of the residuals ( $\hat{y}-y$ )!



# Example

$$\bar{x} = 162.85$$

$$\bar{y} = 49.03$$

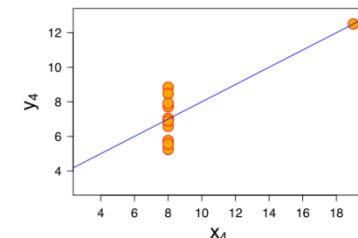
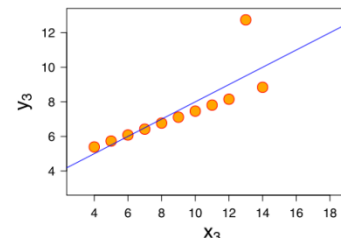
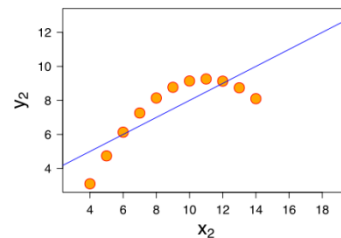
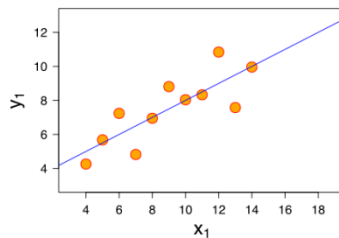
PERSON $i$	1	2	3	4	5	6	7	8	9	10
$(x_i - \bar{x})$	-6.55	-3.95	-2.05	16.75	-6.25	2.25	3.05	-6.15	4.95	-2.05
$(y_i - \bar{y})$	-1.93	-2.23	0.27	4.17	-1.33	-0.03	1.57	-1.93	2.67	-1.23
$(x_i - \bar{x}) \cdot (y_i - \bar{y})$	12.64	8.81	-0.55	69.85	8.31	-0.07	4.79	11.87	13.22	2.52
$(x_i - \bar{x})^2$	42.90	15.60	4.20	280.56	39.06	5.06	9.30	37.82	24.50	4.20

$$b_1: \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{131.39}{463.2} = 0.2836$$

$$b_0: 49.03 - 0.2836 \cdot 162.85 = 2.84$$

# How Good is the model?

- The regression line is only a model based on the data.
- This model might not reflect reality.
- **We need some way of testing how well the model fits the observed data.**
  - Graphically: Deviation from linearity?
  - Diagnostics of residuals
  - **Coefficient of determination** via decomposition of variance



# Exercise: Album\_Sales.sav

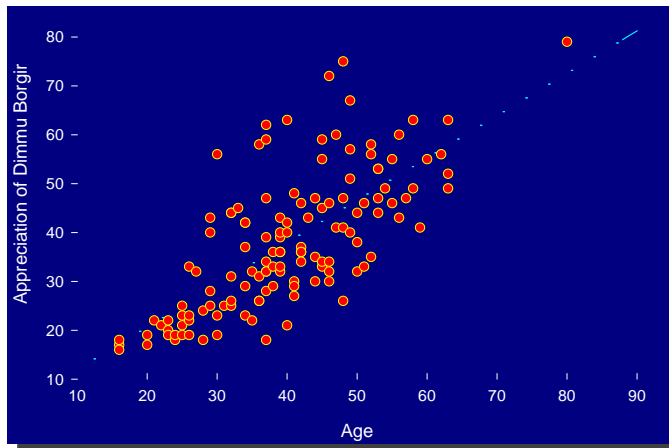
---

- A record company boss was interested in predicting album sales from advertising.
- Data
  - 200 different album releases
- Outcome variable:
  - Sales (CDs and Downloads) in the week after release
- Predictor variable:
  - The amount (in £s) spent promoting the album before release.

**Analyse the data and perform a regression analysis (in SPSS)!**

# What is Multiple Regression?

- (Simple) Linear Regression is a model to predict the value of one variable from another
- **Multiple Regression** is a natural extension of this model:
  - We use it to predict values of an outcome from *several* predictors
  - It is a hypothetical model of the relationship between several variables



Simple regression - line

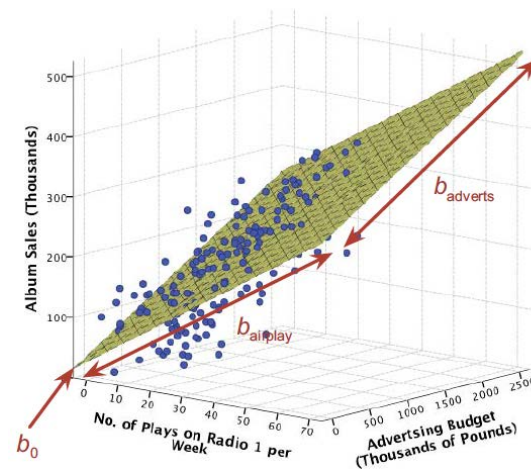


FIGURE 8.3  
Scatterplot of  
the relationship  
between  
album sales,  
advertising  
budget and  
radio play

Multiple regression with 2 predictors - plane

# Multiple regression: an example

---

- A record company boss was interested in predicting album sales from advertising.
- Data
  - 200 different album releases
- Outcome variable:
  - Sales (CDs and Downloads) in the week after release
- Predictor variables
  - The amount (in £s) spent promoting the album before release (see last lecture)
  - Number of plays on the radio (new variable)



# Multiple Regression as an Equation

- With multiple regression the relationship is described using a variation of the equation of a straight line

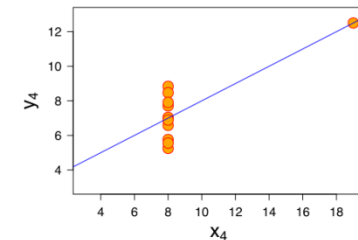
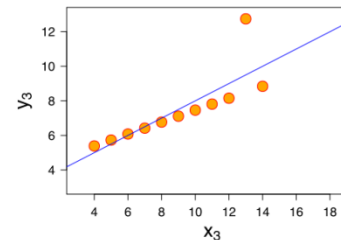
$$y = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n + \varepsilon_i$$

- $b_0$  is the intercept
  - the intercept is the value of the Y variable when all  $X_s = 0$
- $b_1$  is the regression coefficient for variable 1
- $b_2$  is the regression coefficient for variable 2
- $b_n$  is the regression coefficient for  $n^{\text{th}}$  variable.
- $\varepsilon_i$  is the residual (error term)

# Model fit (1)

- As in the simple linear regression model
- $SS_{tot} = SS_{reg} + SS_{res}$
- Coefficient of determination,  $R^2$ 
  - The proportion of variance accounted for by the model.
- $R^2$  equals the square of the Pearson correlation coefficient between  $y$  and  $\hat{y}$
- $0 \leq R^2 \leq 1$
- **Diagnostics of residuals:**
  - Standardized residuals
- **Influential cases:**
  - Cook's distance

$$R^2 = \frac{SS_{reg}}{SS_{tot}}$$



# Model fit (2)

- **Diagnostics of residuals:**

- **Standardized residuals**

- In an average sample, 95% of standardized residuals should lie between  $\pm 2$ .

- 99% of standardized residuals should lie between  $\pm 2.5$ .

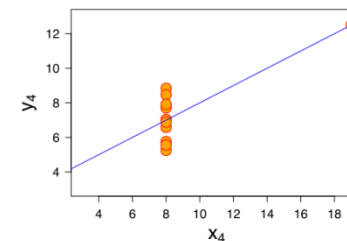
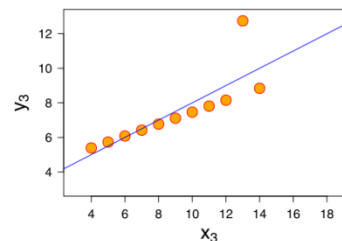
- Outliers: any case for which the absolute value of the standardized residual is 3 or more, is likely to be an outlier.

- **Influential cases:**

- **Cook's distance**

- Measures the influence of a single case on the model as a whole.

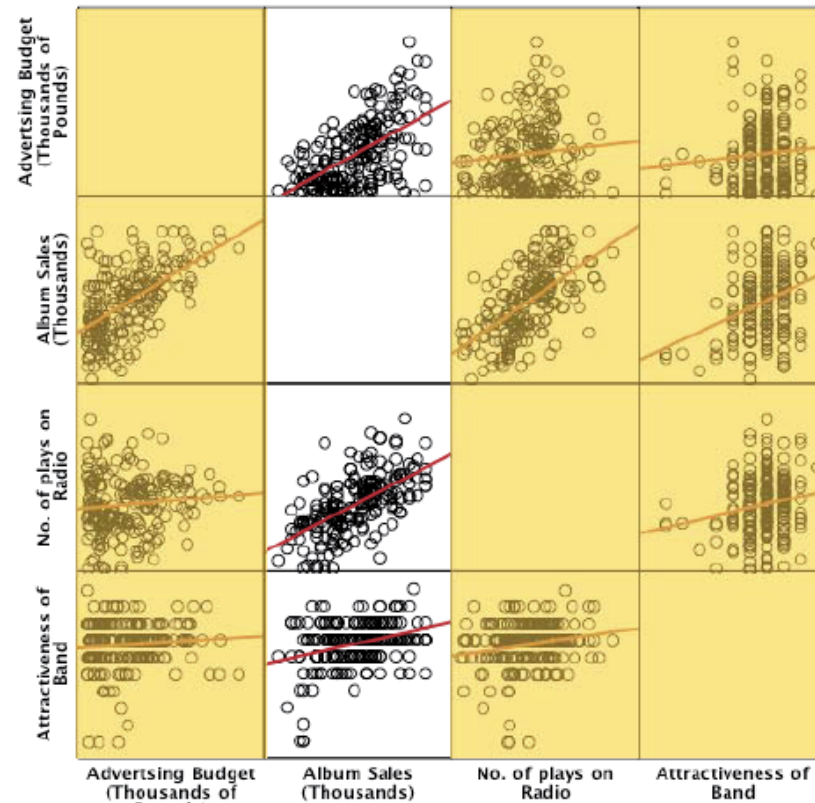
- Weisberg (1982): absolute values greater than 1 may be cause for concern



# Which variables to include?

- First step: scatterplot

**FIGURE 8.14**  
Matrix  
scatterplot  
of the  
relationships  
between  
advertising  
budget,  
airplay, and  
attractiveness  
of the band  
and  
album sales





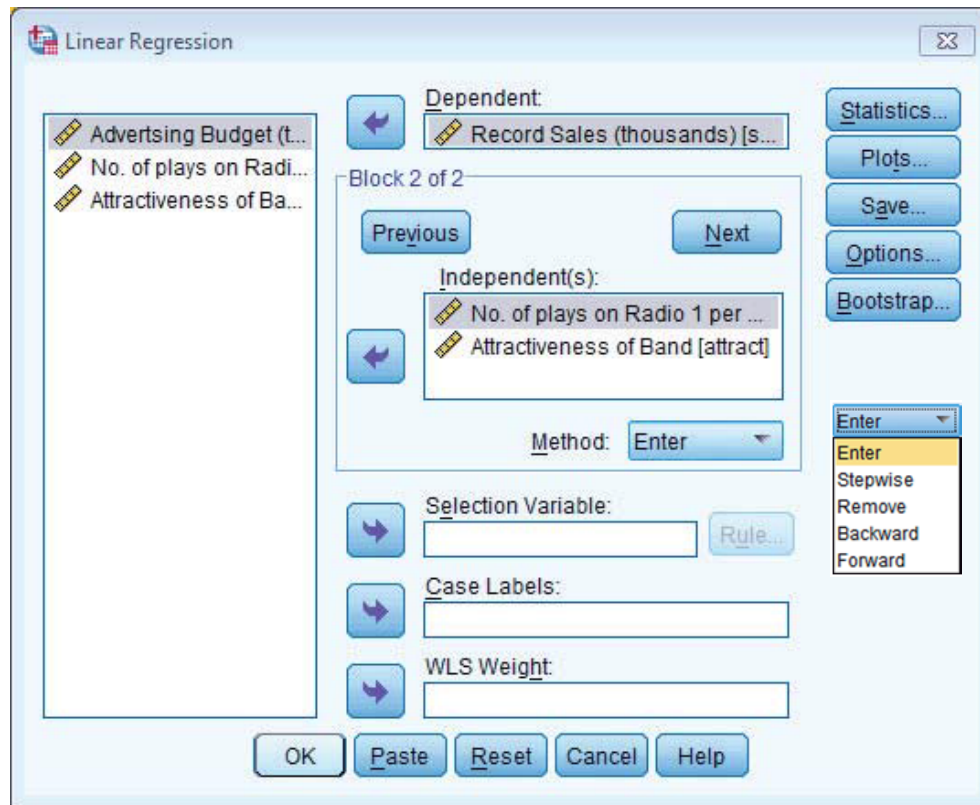
# Method 1: Forced entry

- All variables are entered into the model simultaneously
- The results obtained depend on the variables entered into the model
  - It is important, therefore, to have good theoretical reasons for including a particular variable

# Method 2: Stepwise entry

- Variables are entered into the model based on mathematical criteria
- Computer selects variables in steps.
- Step 1: Choose the predictor that can explain the most variance in the outcome variable.
- Step 2: Having selected the 1<sup>st</sup> predictor, a second one is chosen from the remaining predictors.
- The semi-partial correlation is used as a criterion for selection.
  - The semi-partial correlation measures the relationship between two variables controlling for the effect that a third variable has on only one of the others
  - It measures the *unique contribution* of a predictor to explaining the variance of the outcome
- Drawbacks of stepwise entry:
  - Rely on a mathematical criterion.
    - Variable selection may depend upon only slight differences in the semi-partial correlation
    - These slight numerical differences can lead to major theoretical differences
  - Should be used only for exploration

# Doing Multiple Regression in SPSS (1)



**Linear Regression**

**Dependent:** Record Sales (thousands) [s...]

**Block 2 of 2**

**Independent(s):**

- No. of plays on Radio 1 per ...
- Attractiveness of Band [attract]

**Method:** Enter

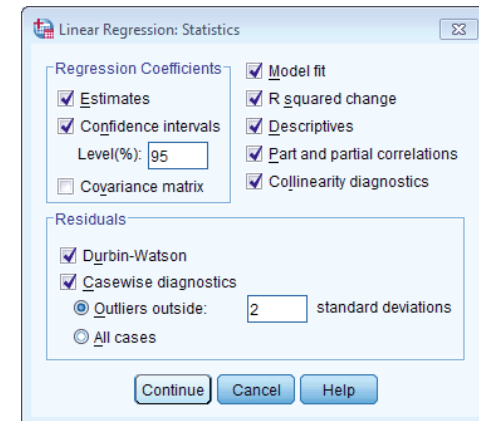
**Selection Variable:**

**Case Labels:**

**WLS Weight:**

Buttons: Previous, Next, Statistics..., Plots..., Save..., Options..., Bootstrap...

Buttons: OK, Paste, Reset, Cancel, Help



**Linear Regression: Statistics**

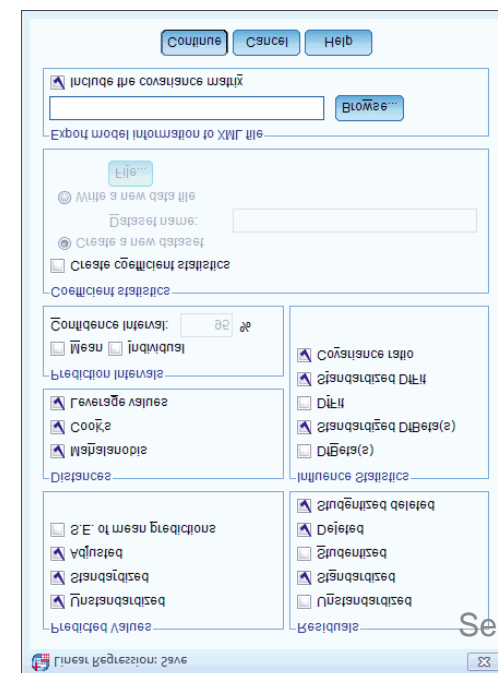
**Regression Coefficients**

- Estimates
- Confidence intervals
- Level(%): 95
- Covariance matrix
- Model fit
- R squared change
- Descriptives
- Part and partial correlations
- Collinearity diagnostics

**Residuals**

- Durbin-Watson
- Casewise diagnostics
- Outliers outside: 2 standard deviations
- All cases

Buttons: Continue, Cancel, Help



Buttons: Continue, Cancel, Help

Include the covariance matrix

Export to non-SPSS file

File:

- Write a new file
- Create a new file name
- Create coefficient statistics

**Coefficient statistics**

Confidence interval: 95

- Mean
- Individual
- Confidence ratio
- Standardized DIF-II
- DIF-II
- Standardized DIF-III(e)
- DIF-III(e)

**Distances**

- Z.E. of mean predictions
- Adjusted
- Standardized
- Influence
- Unadjusted

Buttons: Continue, Cancel, Help

# Doing Multiple Regression in SPSS (2)

Model Summary<sup>c</sup>

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.578 <sup>a</sup>	.335	.331	65.991	.335	99.587	1	198	.000	
2	.815 <sup>b</sup>	.665	.660	47.087	.330	96.447	2	196	.000	1.950

a. Predictors: (Constant), Advertsing Budget (Thousands of Pounds)

b. Predictors: (Constant), Advertsing Budget (Thousands of Pounds), Attractiveness of Band, No. of plays on Radio

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	134.140	7.537		17.799	.000	119.278	149.002
	Advertsing Budget (Thousands of Pounds)	.096	.010	.578	9.979	.000	.077	.115
2	(Constant)	-26.613	17.350		-1.534	.127	-60.830	7.604
	Advertsing Budget (Thousands of Pounds)	.085	.007	.511	12.261	.000	.071	.099
	No. of plays on Radio	3.367	.278	.512	12.123	.000	2.820	3.915
	Attractiveness of Band	11.086	2.438	.192	4.548	.000	6.279	15.894

a. Dependent Variable: Album Sales (Thousands)

$$y = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

$$\text{Sales} = -26.6 + 0.085\text{adverts} + 3367\text{plays} + 11.1\text{attractiveness}$$



# How to interpret beta values?

- **Beta values:**

- the change in the outcome associated with a unit change in the predictor
- $b_1 = 0.085$ : as advertising increases by £1, album sales increase by 0.085 units.
- $b_2 = 3367$ : each time (per week) a song is played on the radio its sales increase by 3367 units.

- **Standardised beta values:**

- tell us the same but expressed as standard deviations
- $\beta_1 = 0.511$ : As advertising increases by 1 standard deviation, album sales increase by 0.511 of a standard deviation.
- $\beta_2 = 0.512$ : When the number of plays on the radio increases by 1 SD its sales increase by 0.512 standard deviations.

# Reporting the model

**TABLE 8.2** Linear model of predictors of album sales, with 95% bias corrected and accelerated confidence intervals reported in parentheses. Confidence intervals and standard errors based on 1000 bootstrap samples

	<i>b</i>	<i>SE B</i>	$\beta$	<i>p</i>
Step 1				
Constant	134.14 (120.11, 148.79)	7.95		$p = .001$
Advertising Budget	0.10 (0.08, 0.11)	0.01	.58	$p = .001$
Step 2				
Constant	-26.61 (-55.40, 8.60)	16.30		$p = .097$
Advertising Budget	0.09 (0.07, 0.10)	0.01	.51	$p = .001$
Plays on BBC Radio 1	3.37 (2.82, 3.91)	0.32	.51	$p = .001$
Attractiveness	11.09 (6.28, 15.89)	2.22	.19	$p = .001$

Note.  $R^2 = .34$  for Step 1;  $\Delta R^2 = .33$  for Step 2 ( $ps < .001$ ).

# Generalization (1)

---

- When we run regression, we hope to be able to generalize the sample model to the entire population
- To do this, several assumptions must be met
- Violating these assumptions stops us generalizing conclusions to our target population
- Variable Type:
  - Outcome must be continuous
  - Predictors can be continuous or dichotomous
- Non-Zero Variance:
  - Predictors must not have zero variance
- Linearity:
  - The relationship we model is, in reality, linear
- Independence:
  - All values of the outcome should come from a different person

# Generalization (2)

---

- **No Multicollinearity:**
  - Predictors must not be highly correlated.
  - Collinearity diagnostics
- **Homoscedasticity:**
  - For each value of the predictors the variance of the error term should be constant.
  - plot ZRESID against ZPRED
- **Independent Errors:**
  - For any pair of observations, the error terms should be uncorrelated
- **Normally-distributed Errors**
  - Normal probability plot

# Regression plots

ZResid vs. ZPred

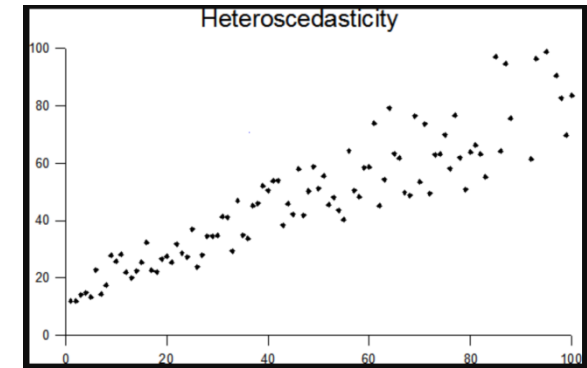
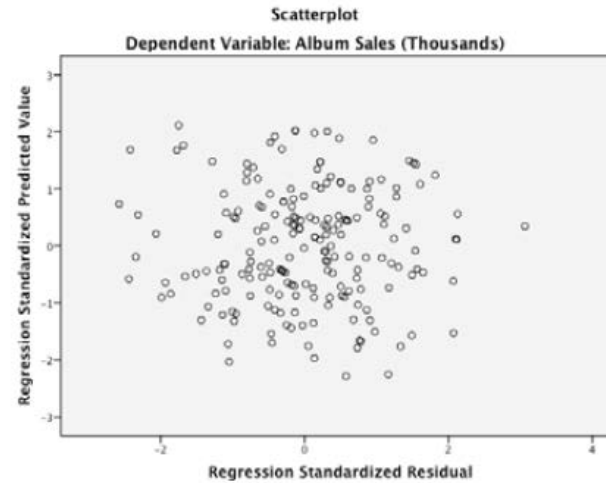
Linear Regression: Plots

DEPENDENT  
\*ZPRED  
\*ZRESID  
\*DRESID  
\*ADJPRED  
\*SRESID  
\*SDRESID

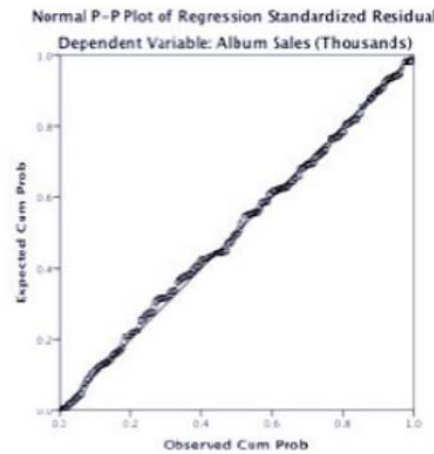
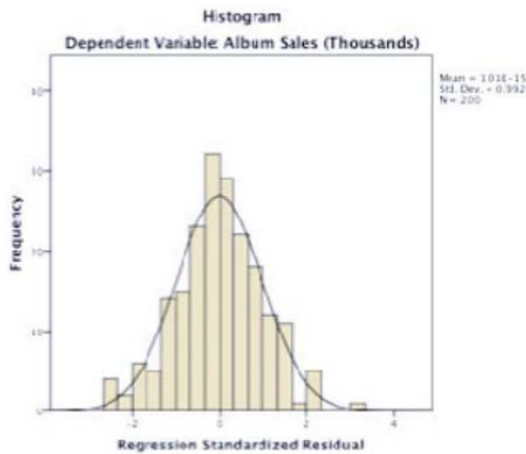
Scatter 1 of 1  
Previous Next  
Y: \*ZRESID  
X: \*ZPRED

Standardized Residual Plots  
 Histogram  
 Normal probability plot  
 Produce all partial plots

Continue Cancel Help



Homoscedasticity: ZRESID vs. ZPRED



Normality of Errors:  
Histograms and P-P plots

# Collinearity diagnostics

**Coefficients<sup>a</sup>**

Model		Correlations			Collinearity Statistics	
		Zero-order	Partial	Part	Tolerance	VIF
1	Advertsing Budget (Thousands of Pounds)	.578	.578	.578	1.000	1.000
2	Advertsing Budget (Thousands of Pounds)	.578	.659	.507	.986	1.015
	No. of plays on Radio	.599	.655	.501	.959	1.043
	Attractiveness of Band	.326	.309	.188	.963	1.038

a. Dependent Variable: Album Sales (Thousands)

- Tolerance should be more than 0.2 (Menard, 1995)
- VIF should be less than 10 (Myers, 1990)

# Signifikanztests und Konfidenzintervalle



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

---

Hanno Ulmer

*hanno.ulmer@i-med.ac.at*

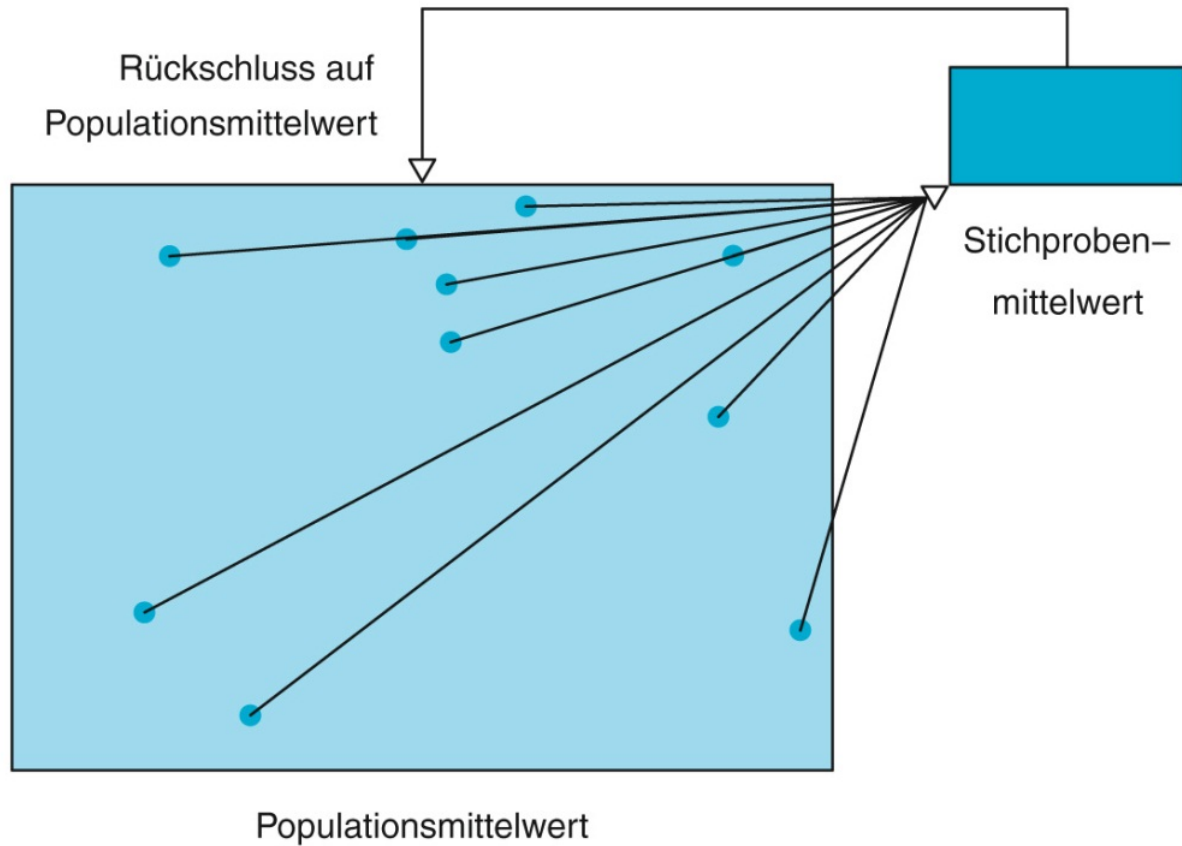
---

Department für Medizinische Statistik, Informatik und  
Gesundheitsökonomie, Medizinische Universität Innsbruck

- Deskriptive Statistik
- Inferenzstatistik I: Schätzen von Parametern mittels Konfidenzintervallen
- Inferenzstatistik II: Unterschiede, Hypothesenprüfung mittels Signifikanztests
- Inferenzstatistik III: Zusammenhänge, Korrelations- und Regressionsanalysen



# Schließende Statistik



**Abbildung 4.1:** Das Prinzip des statistischen Schließens.

# Ziel der schließenden Statistik

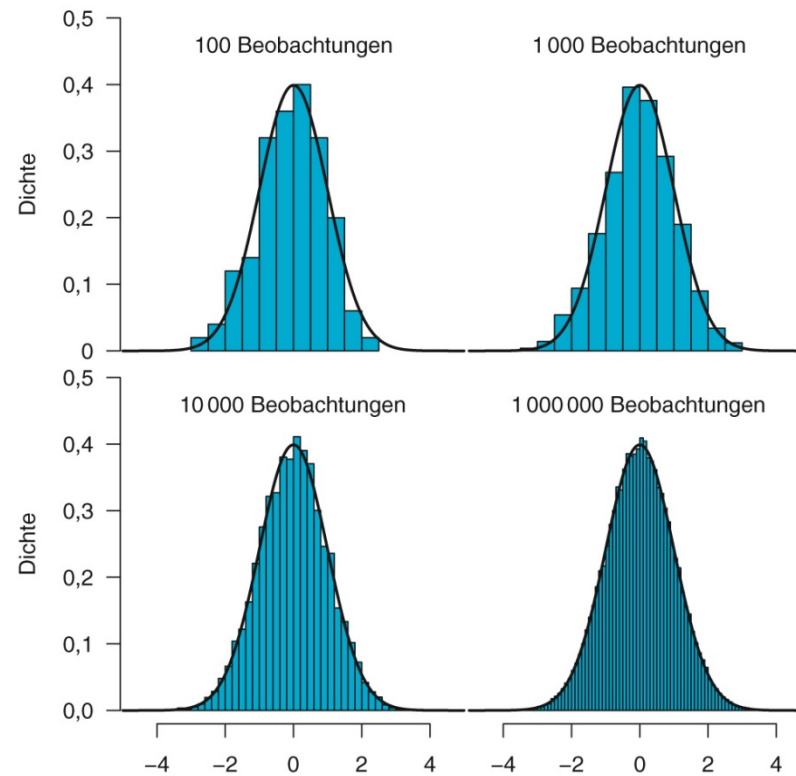
- Aufgrund der Stichprobe Aussagen über die Grundgesamtheit machen
- Von der Stichprobe auf die Grundgesamtheit schließen
- **Grundgesamtheit**
  - ist eine entsprechend dem jeweiligen Untersuchungsziel abgegrenzte Menge von Personen oder Objekten, über die man eine Aussage machen will.
- **Stichprobe**
  - ist eine (kleinere) Menge von Einheiten, die aus der Grundgesamtheit ausgewählt werden und die man misst oder beobachtet, um aus ihnen Schlüsse mit Gültigkeit auch für den nicht ausgewählten Teil der Grundgesamtheit zu ziehen.
- **Repräsentativ**
  - hinsichtlich der für die Untersuchung relevanten Merkmale die Struktur der Grundgesamtheit widerspiegeln

# Grundlagen der schließenden Statistik: Verteilungsannahmen



- Gleichverteilung
  - Bsp.: Würfel, Münze
- Binomialverteilung
  - Bsp.: Ziehen mit Zurücklegen
- Poissonverteilung
  - Bsp.: Seltene Ereignisse mit grossen Stichprobenumfängen und konstanter Wahrscheinlichkeit
- Exponentialverteilung
  - Bsp.: Überlebenszeiten
- Gauß- oder Normalverteilung
  - Wichtigste Verteilung siehe die folgenden Folien
- Testverteilungen
  - Chi-Quadrat-Verteilung
  - T-Verteilung
  - F-Verteilung
  - Anwendung bei Signifikanztests und Konfidenzintervallen

# Normalverteilung



**Abbildung 3.4:** Histogramme für 100, 1000, 10 000 und 1 000 000 Beobachtungen aus einer Normalverteilung.

# Dichtefunktion

## DEFINITION 3.6

## Dichtefunktion

Liegen von einer stetigen Variablen sehr viele, mit beliebiger Genauigkeit gemessene Beobachtungen vor und beschreiben wir diese Daten mit einem Histogramm, dann könnten wir eine extrem kleine Intervalllänge verwenden. Durch die sehr große Zahl der Beobachtungen ließe sich das Histogramm nicht mehr von einer glatten Kurve unterscheiden. Falls wir auf der y-Achse die **Dichte**-Skala abtragen, nennt man ein derartiges „Histogramm“ eine **Wahrscheinlichkeitsdichte**.

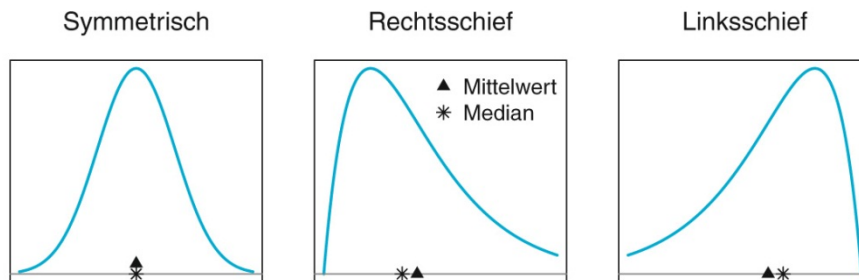


Abbildung 3.5: Dichtefunktion einer symmetrischen, einer rechts- und einer linksschiefen Verteilung.

# Gauss- oder Normalverteilung

## DEFINITION 3.7

## Normalverteilung

Die Dichtefunktion der Normalverteilung lautet:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right\}$$

Dabei ist  $\pi = 3,14159\dots$  die Kreiszahl  $\pi$  und  $\exp$  die Exponentialfunktion. Der Erwartungswert  $\mu$  und die Standardabweichung  $\sigma$  sind die Parameter der Normalverteilung.

# Gauss- oder Normalverteilung

- $-\infty < X < +\infty$
- Parameter Erwartungswert  $\mu$  und Varianz  $\sigma^2$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), x \in \mathbb{R}$$

$$E(X) = \mu$$

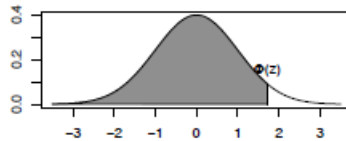
$$\text{Var}(X) = \sigma^2$$

$$N(\mu, \sigma^2)$$

- tabelliert ist die Standardnormalverteilung  $N(0,1)$ 
  - $N(0,1)$ ...Mittelwert 0, Streuung 1
    - Da eine Transformation möglich ist

$$Z = \frac{X - \mu}{\sigma}$$

# Quantile der Standardnormalverteilung



$\Phi(z) = P(Z \leq z)$ ,  $Z \sim \mathcal{N}(0, 1)$ , für  $z < 0$  gilt  $\Phi(z) = 1 - \Phi(-z)$ , Ablesbeispiel  $\Phi(1.72) = 0.95728$ .

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51596	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57928	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61028	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99908	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976

## Wichtige Quantile der Standardnormalverteilung

0.5	0.75	0.9	0.95	0.975	0.99	0.995
0	0.67449	1.2816	1.6449	1.9600	2.3263	2.5758



# Die 68-95-99.7% Regel:

---

Für alle Normalverteilungen gilt

- $\approx 68\%$  aller Beobachtungen liegen innerhalb einer Standardabweichung um den Mittelwert (zwischen  $\mu - \sigma$  und  $\mu + \sigma$ )
- $\approx 95\%$  aller Beobachtungen liegen innerhalb zweier Standardabweichung (zwischen  $\mu - 2\sigma$  und  $\mu + 2\sigma$ )
- $99.7\%$  aller Beobachtungen liegen innerhalb dreier Standardabweichung (zwischen  $\mu - 3\sigma$  und  $\mu + 3\sigma$ )



# Rolle der Normalverteilung

---

- Viele Messwerte sind annähernd normalverteilt
  - Z.B.: Gewicht einer bestimmten Gruppe
- Manche Messwerte können durch Transformation an eine Normalverteilung angenähert werden
  - Auch diskrete Verteilungen, wie z.B. die Binomialverteilung können durch die Normalverteilung approximiert werden
- Wichtige (und günstige) statistisch-theoretische Eigenschaften
- Die Normalverteilung ist Grundlage wichtiger Testverteilungen

# Exercise 1



Entry to a certain University is determined by a national test. The scores on this test are normally distributed with a mean of 500 and a standard deviation of 100. Tom wants to be admitted to this university and he knows that he must score better than at least 70% of the students who took the test. Tom takes the test and scores 585. Will he be admitted to this university?

$$Z = (585 - 500) / 100 = 0.85$$

$$P = [\text{area to the left of } z = 0.85] = 0.8023 = 80.23\%$$

Tom scored better than 80.23% of the students who took the test and he will be admitted to this University.



## Exercise 2

The length of life of an instrument produced by a machine has a normal distribution with a mean of 12 months and standard deviation of 2 months. Find the probability that an instrument produced by this machine will last a) less than 7 months. )  
between 7 and 12 months.

$$P(x < 7) = P(z < -2.5) = 0.0062$$

$$P(7 < x < 12) = P(-2.5 < z < 0) = 0.4938$$

# Inferenzstatistik I: Schätzen von Parametern (Konfidenzintervalle)



- Ein Konfidenzintervall ist ein Vertrauensbereich oder eigentlich besser ein Unsicherheitsbereich für die Schätzung eines bestimmten, nicht bekannten Parameters der Grundgesamtheit.
- Konfidenzintervall – ein Bereich, in welchem der zu schätzende, unbekannte Parameter der Grundgesamtheit mit Wahrscheinlichkeit  $(1 - \alpha)$  liegt
- Der interessierende Parameter kann ein Anteil, ein Mittelwert, ein relatives Risiko etc. sein.
- Ein 95%Konfidenzintervall beispielsweise enthält den gesuchten Parameter mit einer Wahrscheinlichkeit von 95%.

# Der Weg zum Konfidenzintervall

- Punktschätzer aus einer Stichprobe liefert nur einen einzigen Wert
  - Verschiedene Stichproben aus ein und derselben Grundgesamtheit liefern unterschiedliche numerische Werte für den zu schätzenden Parameter
- Berücksichtigung der Standardabweichung der Schätzstatistik (= **Standardfehler**)

## DEFINITION 4.1

### Standardfehler

Der Standardfehler eines Schätzers einer Kennzahl der Population ist dessen geschätzte Standardabweichung. Der Standardfehler des Mittelwertes ist  $s/\sqrt{n}$ , wobei  $s$  die Stichprobenstandardabweichung und  $n$  der Stichprobenumfang ist. Bei anderen Kennzahlen ergeben sich analoge Formeln für die zugehörigen Standardfehler.

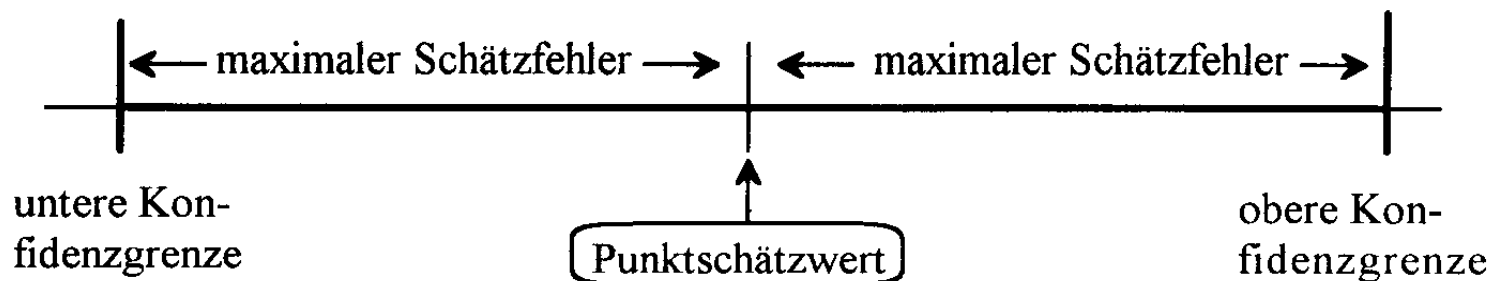
# Inferenzstatistik I: Schätzen von Parametern (Konfidenzintervalle)



- Beispiel: Rule of three (adverse Events)
  - Keine AE in  $n=10$  Patienten  
95% KI (0-30%)
  - Keine AE in  $n=100$  Patienten  
95% KI (0-3%)
  - Keine AE in  $n=1000$  Patienten  
95% KI (0-0,3%)

“A confidence interval is used when estimating an unknown parameter from sample data. The interval gives a range for the parameter – and a confidence level that the range covers the true value.”

Freedman et al. (1991), p. 385.

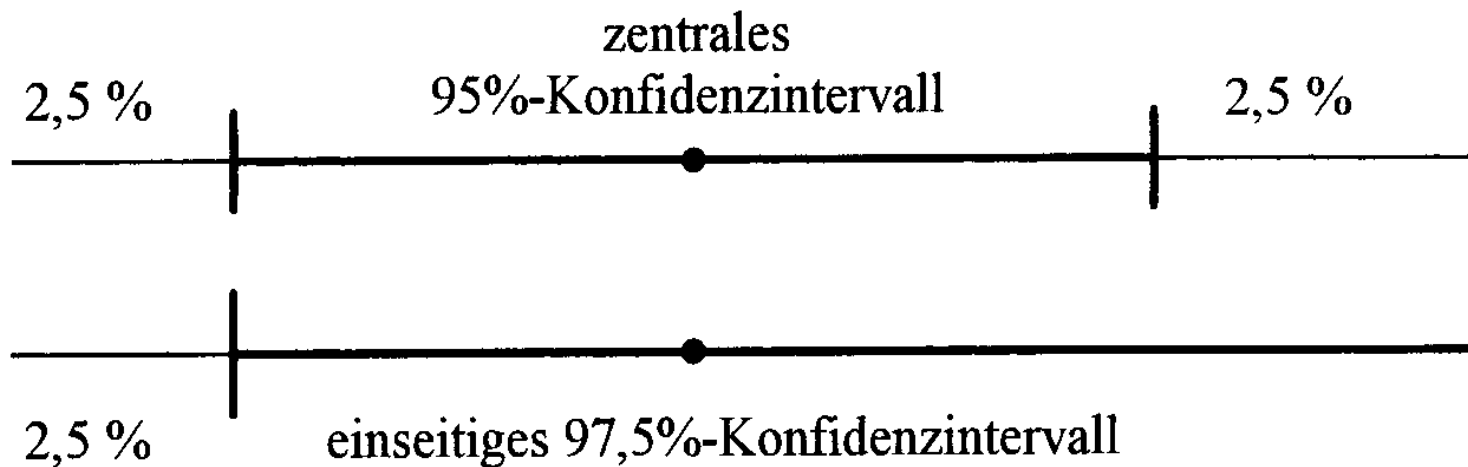




# $(1-\alpha)$ -Konfidenzintervall

- Punktschätzer aus einer Stichprobe liefert nur einen einzigen Wert
  - Verschiedene Stichproben aus ein und derselben Grundgesamtheit liefern unterschiedliche numerische Werte für den zu schätzenden Parameter
- Berücksichtigung der Standardabweichung der Schätzstatistik
- Konfidenzintervall – ein Bereich, in welchem der zu schätzende, unbekannte Parameter der Grundgesamtheit mit Wahrscheinlichkeit  $(1-\alpha)$  liegt

# Zwei- und Einseitige Konfidenzintervalle



# $(1-\alpha)$ -Konfidenzintervall

- Einseitig
  - Bestimmung einer Ober- bzw. Untergrenze
- Zweiseitig
  - symmetrisch
- Üblich Werte für  $\alpha$ 
  - $\alpha=0.05$ ,  $\alpha=0.01$ ,  $\alpha=0.001$
- Grundsätzlich kann für alle statistisch geschätzten Parameter ein Konfidenzintervall berechnet werden
  - Mittelwert, Standardabweichung, Häufigkeit, Korrelationskoeffizient, ...

# Konfidenzintervall für den Mittelwert $\mu$

- Bei **unbekannter** Varianz  $\sigma^2$ , geschätzt durch  $s^2$
- Gegeben:  $X_1, \dots, X_N$  unabhängige normalverteilte Zufallsvariable  $N(\mu, \sigma^2)$
- Obere bzw. untere Grenze des Konfidenzintervalls

$$G_{oben, unten} = \bar{X} \pm t_{(N-1), 1-\alpha/2} s / \sqrt{N}$$

- $t_{(N-1), 1-\alpha/2}$ ...t-Verteilung mit N-1 Freiheitsgraden
- Allgemein gilt:  $t_{(N-1), 1-\alpha} > z_{1-\alpha}$

$$N = (2t_{1-\alpha/2, N-1} s / L)^2$$

t...Quantil der t - Verteilung  
für N - 1 Freiheitsgrade und  $1 - \alpha$

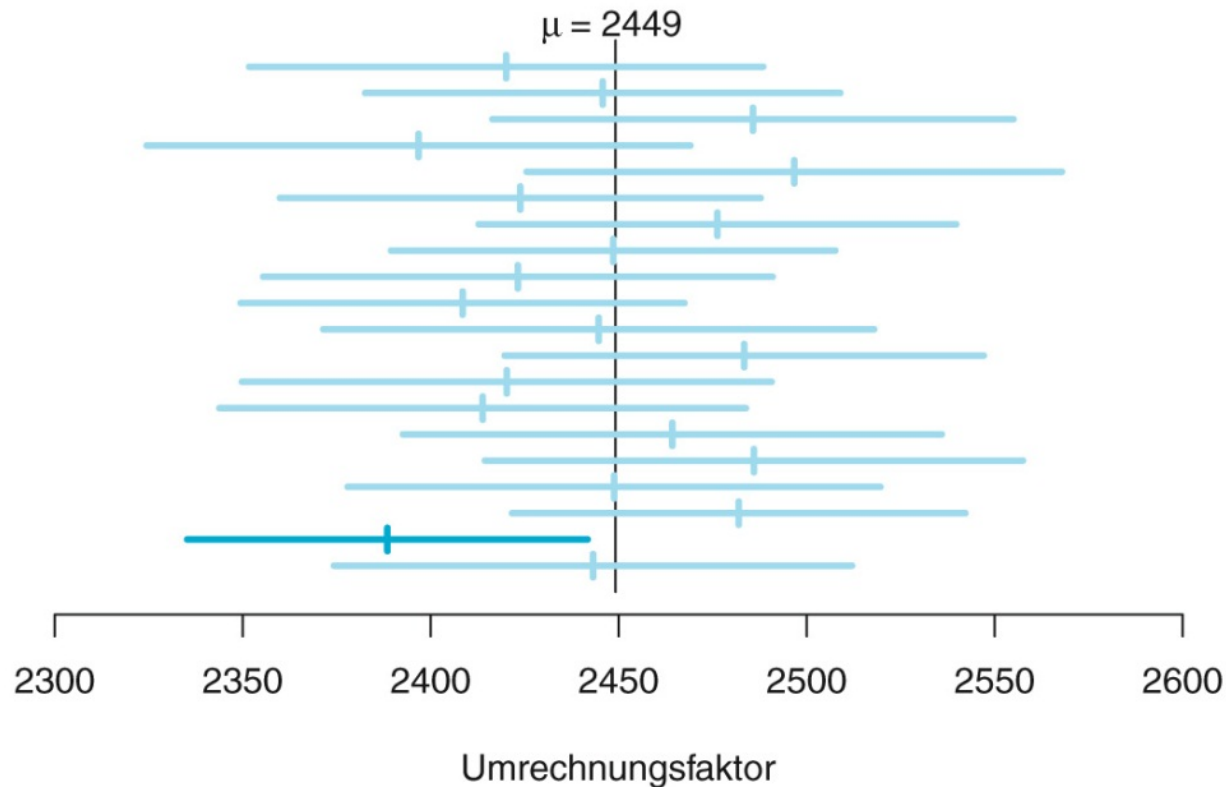
# Konfidenzintervall für einen Anteil $\pi$

- Wahrscheinlichkeit  $\pi$  wird durch  $p$  aus einer Stichprobe vom Umfang  $N$  geschätzt
  - Das Konfidenzintervall ist asymmetrisch
- Approximation durch Normalverteilung
  - Wenn  $N > 30$

$$\left[ p - z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{N}}, p + z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{N}} \right]$$

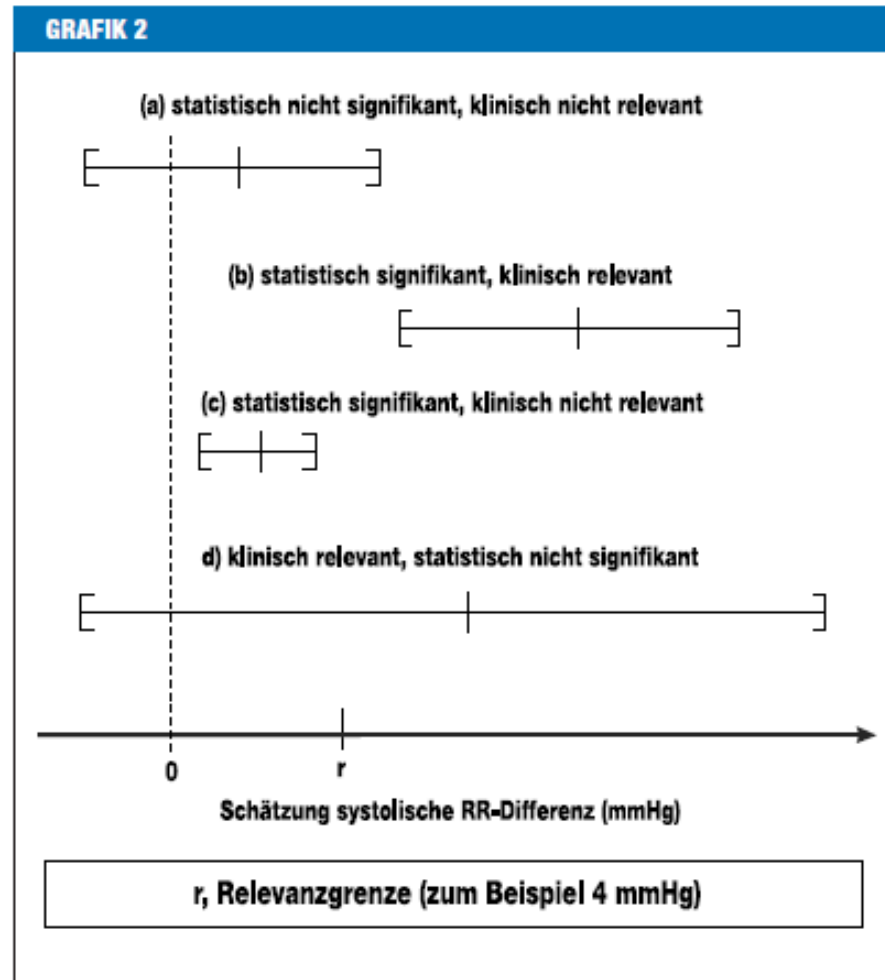
- Approximation nach Wald
- Approximation nach Pearson-Clopper

# 95% Konfidenzintervalle



**Abbildung 4.4:** Illustration des Konfidenzniveaus. Von 20 95 %-Konfidenzintervallen überdeckt eines (das dunkelblaue) den wahren Wert nicht.

# Klinische Relevanz versus statistische Signifikanz



# Beispiel: Schätzen von Parametern (Konfidenzintervalle)



**Table 2. Effect of Cytisine on Smoking Cessation.\***

Outcome	Cytisine (N= 370)	Placebo (N= 370)	Percentage-Point Difference (95% CI)	Relative Rate (95% CI)†
	<i>percent (number)</i>			
Primary outcome: abstinence for 12 mo	8.4 (31)	2.4 (9)	6.0 (2.7–9.2)‡	3.4 (1.7–7.1)
Abstinence for 6 mo	10.0 (37)	3.5 (13)	6.5 (2.9–10.1)‡	2.9 (1.5–5.3)
Point prevalence at 12 mo	13.2 (49)	7.3 (27)	5.9 (1.6–10.3)§	1.8 (1.2–2.8)



- Eine vorgegebene Annahme (**Nullhypothese  $H_0$** ) wird anhand von Daten überprüft. Wenn die Daten "stark" von dem abweichen, was man unter der Nullhypothese erwartet, lässt man die Nullhypothese fallen.
- Im statistischen Test wird dieses Vorgehen formalisiert.
- Nachdem die **Nullhypothese  $H_0$**  und die **Alternativhypothese  $H_1$**  so formuliert sind, dass sie sich gegenseitig ausschließen und keine dritte Möglichkeit zulassen, ergibt sich ein einfaches Entscheidungsschema mit 4 Möglichkeiten
- Der **Fehler 1. Art** ist der Fehler, die Nullhypothese zu **verwerfen**, obwohl sie **richtig** ist.
- Der **Fehler 2. Art** ist der Fehler, die Nullhypothese zu **behalten**, obwohl sie **falsch** ist.

- Ergebnis eines Signifikanztests ist die **Teststatistik**.
- Der Wertebereich der Teststatistik wird in zwei Teilmengen zerlegt, den **Verwerfungsbereich** und den **Annahmebereich**. Wenn die Prüfgröße in den Verwerfungsbereich fällt, wird die Nullhypothese verworfen, ansonsten wird sie behalten.
- Man wählt den Verwerfungsbereich so, dass unter  $H_0$  seine Wahrscheinlichkeit unter einen vorgegebenen Wert  $\alpha$  fällt.  $\alpha$ , das sogenannte **Signifikanzniveau** des Tests, ist damit die **Obergrenze** für die Wahrscheinlichkeit, den **Fehler 1. Art** zu begehen.  $\alpha$  wird vom Versuchsleiter vorgegeben. **Übliche** Werte für  $\alpha$  sind **0.05, 0.01 und 0.001**.
- Welches  $\alpha$  man wählt, hängt von den Konsequenzen ab, die der Fehler 1. Art hat. Der naheliegende Wunsch,  $\alpha = 0$  zu wählen, scheitert daran, dass dann  $\beta$ , die Wahrscheinlichkeit für den **Fehler 2. Art**, groß wird.

# Definitionen

## DEFINITION 5.1

### Nullhypothese

In der **Nullhypothese**  $H_0$  formulieren wir das Gegenteil der **wissenschaftlichen Hypothese** und nehmen an, dass Unterschiede, die wir in einer Stichprobe beobachten, lediglich durch Stichprobenvariation zustande gekommen sind.

## DEFINITION 5.2

### $p$ -Wert

Der  $p$ -Wert ist die Wahrscheinlichkeit, unter Annahme von  $H_0$  ein Resultat **so groß wie das beobachtete oder noch extremer** zu erhalten. „Extremer“ bedeutet dabei „weiter weg vom Nullwert“.

## DEFINITION 5.3

### Statistische Power

Die **Power** eines Hypothesentests ist definiert als  $1 - \beta$ , also die Wahrscheinlichkeit, tatsächlich auf die Alternativhypothese zu entscheiden, wenn diese gilt. Das deutsche Wort für Power ist **Trennschärfe**, welches aber selten verwendet wird. Eine hohe Power eines Hypothesentests entspricht folglich einer hohen Wahrscheinlichkeit, bei Vorliegen der Alternative auch tatsächlich auf diese zu entscheiden, das heißt einen tatsächlich vorhandenen Effekt auch zu entdecken.

### Evidenz gegen Nullhypothese

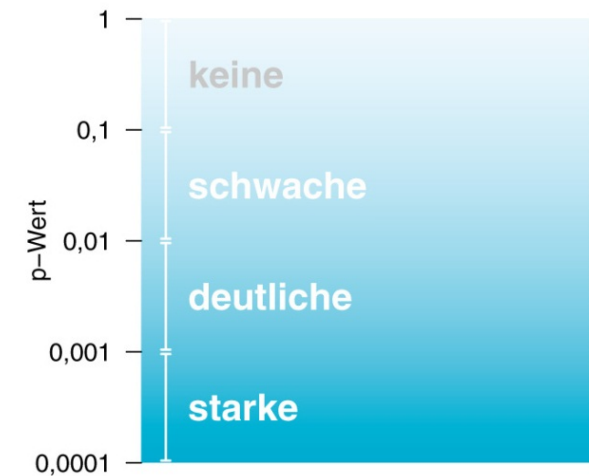


Abbildung 5.4: Interpretation von  $p$ -Werten.

# Mögliche Entscheidungen

In der Stichprobe erhalten wir ein...	In der Population ist...	
	... $H_0$ wahr	... $H_1$ wahr
... nicht signifikantes Testresultat ( $p > \alpha$ )	korrekter Entscheid $1 - \alpha$	falscher Entscheid $\beta$
... signifikantes Testresultat ( $p \leq \alpha$ )	falscher Entscheid $\alpha$	korrekter Entscheid $1 - \beta$

**Tabelle 5.2:** Mögliche Entscheidungen in einem Hypothesentest. In der zweiten Zeile jedes Eintrags steht jeweils die auf die Tabellenspalten bedingte Wahrscheinlichkeit für die entsprechende Entscheidung.



# Type I versus Type II Error

		Decide for	
		$H_0$	$H_1$
Reality	$H_0$	Correct decision	Wrong decision: Type I error ( $\alpha$ )
	$H_1$	Wrong decision: Type II error ( $\beta$ )	Correct decision: Power ( $1-\beta$ )

# Parallelgruppenstudie

- z.B. Untersuchung der Wirksamkeit eines neuen blutdrucksenkenden Medikaments im Vergleich zu einer Standardtherapie oder Placebo (=Kontrolle)
  - **Nullhypothese H0:** die beiden Therapien sind im Mittel gleich wirksam
    - z.B. die Änderung des systolischen Blutdrucks ist im Mittel in beiden Gruppen gleich
    - **Die Ungültigkeit der Nullhypothese ist zu beweisen**
- **Alternativhypothese H1:** die beiden Therapien sind im Mittel unterschiedlich stark wirksam
- **Voraussetzung:** Die beiden Gruppen stimmen in den wesentlichen Merkmalen überein – siehe Randomisierung
  - Nur die Therapien sind unterschiedlich
  - Strukturgleichheit der Gruppen

# Die statistischen Hypothesen

- **Alternativhypothese  $H_A$**

Forschungshypothese

ungerichtet  $H_1: \mu_A \neq \mu_B$

gerichtet  $H_1: \mu_A > \mu_B$

- **Nullhypothese  $H_0$**

geht davon aus, dass das, was mit der Alternativhypothese behauptet wird, nicht zutrifft.

$$H_0: \mu_A = \mu_B$$

Es gibt keinen Behandlungseffekt

# Vorgehen bei der Hypothesenüberprüfung



- Formulierung der Hypothese, Auswahl des Tests
- Datenerhebung
- Analyse der Daten mit Methoden der beschreibenden Statistik (Häufigkeiten, Mittelwert, ....)
- Überprüfung der Daten auf Ihre Verteilung
  - Grafische Darstellung mittels Boxplot und Histogramm
  - Statistische Überprüfung mit dem Kolmogorov-Smirnov Test
- Adaptierung der Testauswahl
- Durchführung des Tests
- Interpretation des Ergebnisses





# Formulating hypothesis & statistical tests

## Steps in conducting a statistical test:

- Quantify the scientific problem from a clinical / biological perspective
- Formulate the model assumptions (distribution of the variable of interest)
- Formulate the problem as a statistical testing problem:  
Nullhypothesis versus alternative hypothesis
- Define the „error“ you are willing to tolerate
- Calculate the appropriate test statistic
- Decide for the null hypothesis or against it

Formulating Hypothesis  
&  
Test statistics  
&  
p-values



# Principles of statistical testing

## Different approaches:

- Test for superiority, standard hypothesis testing
- Test for non-inferiority

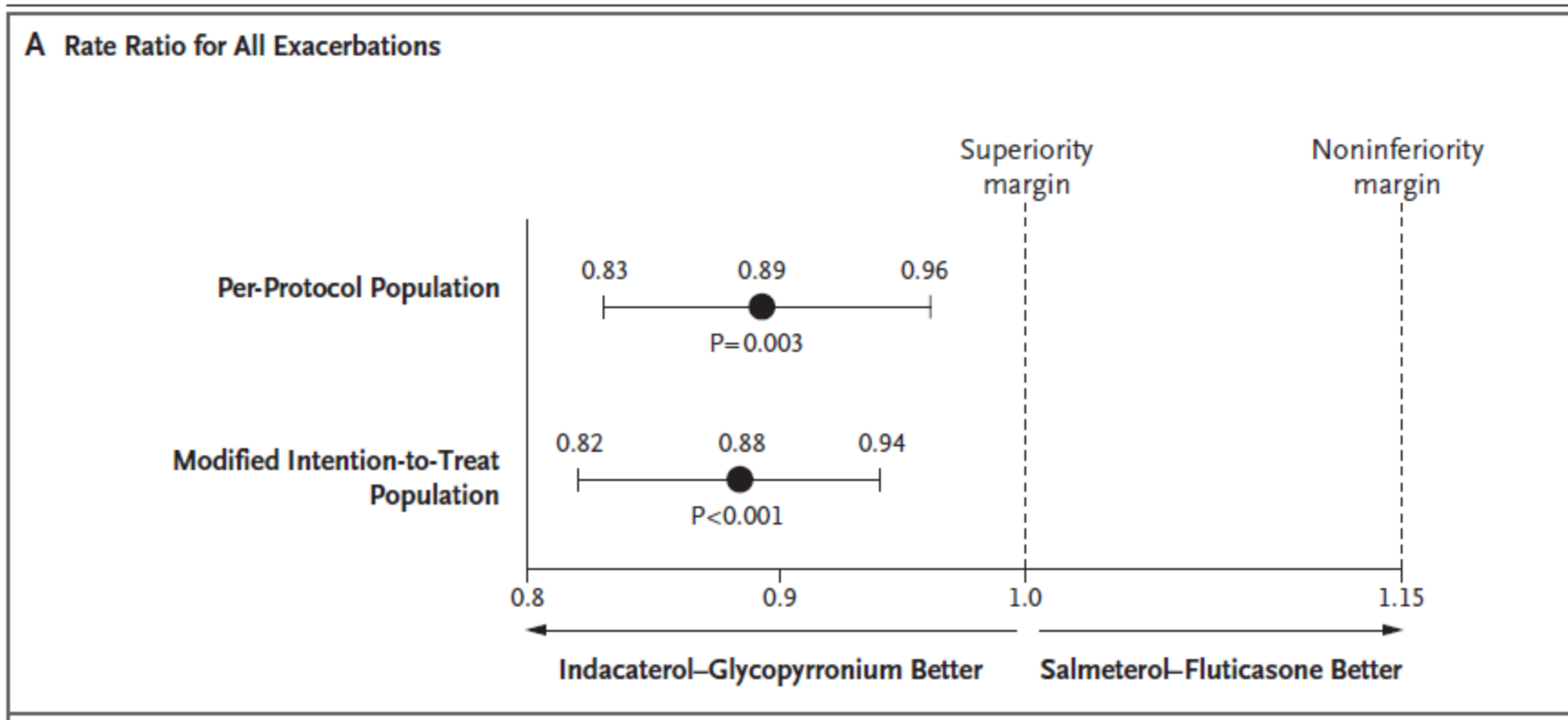
the difference between the new treatment and the standard is less than the smallest clinically meaningful difference, define delta, use confidence intervals

- Test for equivalence

To demonstrate the difference between the new treatment and standard treatment has no clinical importance, define delta

Three primary measures of interest:  
a point estimate,  
a confidence interval, and  
a p-value

# Example



# 2008: 100 Jahre Student's t-Test: William Sealy Gosset



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

VOLUME VI MARCH, 1908 No. 1

## BIOMETRIKA.

### THE PROBABLE ERROR OF A MEAN.

By STUDENT.

#### *Introduction.*

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is better to work with a curve whose area and ordinates are tabled, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error. We are concerned here solely with the first of these two sources of uncertainty.

The usual method of determining the probability that the mean of the population lies within a given distance of the mean of the sample, is to assume a normal distribution about the mean of the sample with a standard deviation equal to  $s/\sqrt{n}$ , where  $s$  is the standard deviation of the sample, and to use the tables of the probability integral.

Biometrika vi

1



<http://www.guinness.com/>

**1936**  
The first GUINNESS®  
brewery overseas is built  
at Park Royal, London.  
William Sealy Gossett,  
the father of modern  
statistics, is appointed  
Head Brewer.

# t-Test

## quantitatives Merkmal

	Therapie	Syst. Blutdruck
1	Therapie A	169
2	Therapie A	139
3	Therapie A	137
4	Therapie A	152
5	Therapie A	142
6	Therapie A	163
7	Therapie A	183
8	Therapie A	143
	$n_1 = 8$	$\bar{x}_1 = 153,5$
1	Therapie B	155
2	Therapie B	154
3	Therapie B	176
4	Therapie B	158
5	Therapie B	156
6	Therapie B	170
7	Therapie B	168
8	Therapie B	179
9	Therapie B	167
10	Therapie B	142
11	Therapie B	163
12	Therapie B	144
	$n_2 = 12$	$\bar{x}_2 = 161$

**H<sub>0</sub>: Es besteht kein Therapieunterschied bzgl. mittleren Blutdrucks ( $\mu_A = \mu_B$ )**

**H<sub>1</sub>: Es besteht ein Unterschied ( $\mu_A \neq \mu_B$ )**

**Teststatistik:**

$$\hat{t} = 1,193$$

**Syst. Blutdruck bei Therapie A größer bzw. kleiner als bei Therapie B => 2-seitiger t-Test**

**Testentscheidung:**  $\hat{t} = 1,193 < 2,101 = t_{18;0,975}$   $p = 0,248$

$t_{n;1-\alpha}$  mit  $n =$  Anzahl der Freiheitsgrade;  $\alpha =$  Signifikanzniveau

**=> Nullhypothese kann nicht abgelehnt werden**

# Student's t-Test für unabhängige Stichproben (1)

- Parametrischer Test zum Vergleich von Mittelwerten

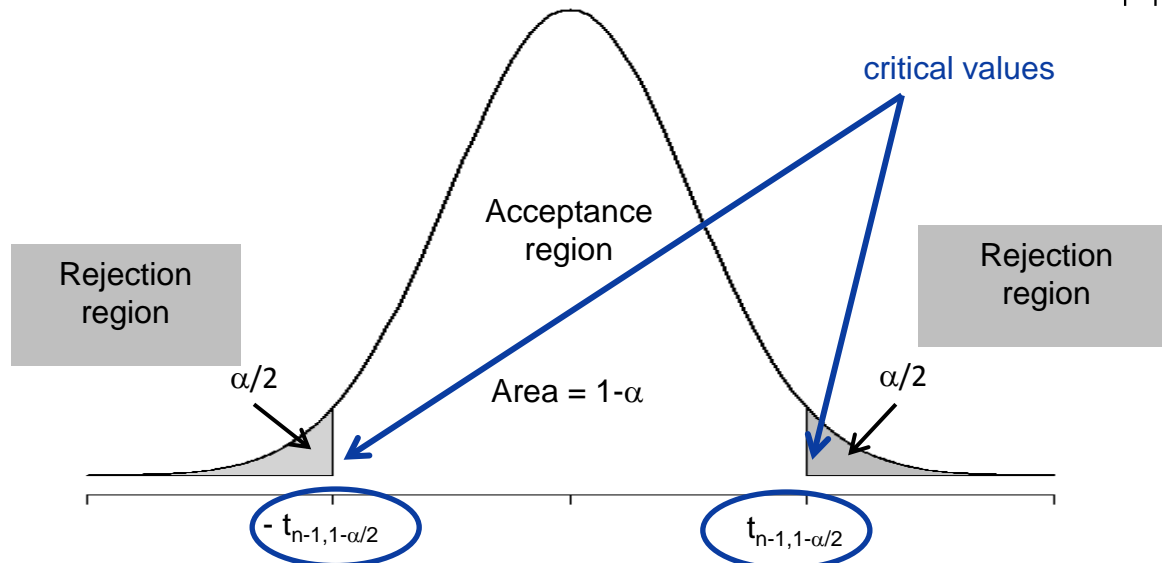
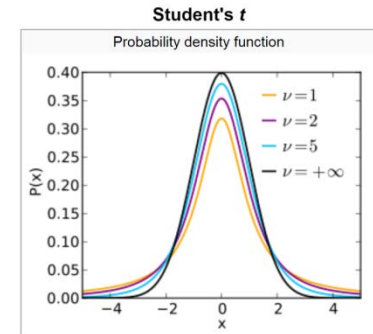
$$H_0 : \mu_A = \mu_B$$

$$H_1 : \mu_A \neq \mu_B$$

- Teststatistik: 
$$\hat{t} = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\left(\frac{1}{N_x} + \frac{1}{N_y}\right) \frac{(N_x - 1)s_x^2 + (N_y - 1)s_y^2}{N_x + N_y - 2}}}$$

- Falls die Nullhypothese gilt, ist die Teststatistik  $t(N_x + N_y - 2)$ -verteilt
- Entscheidung:

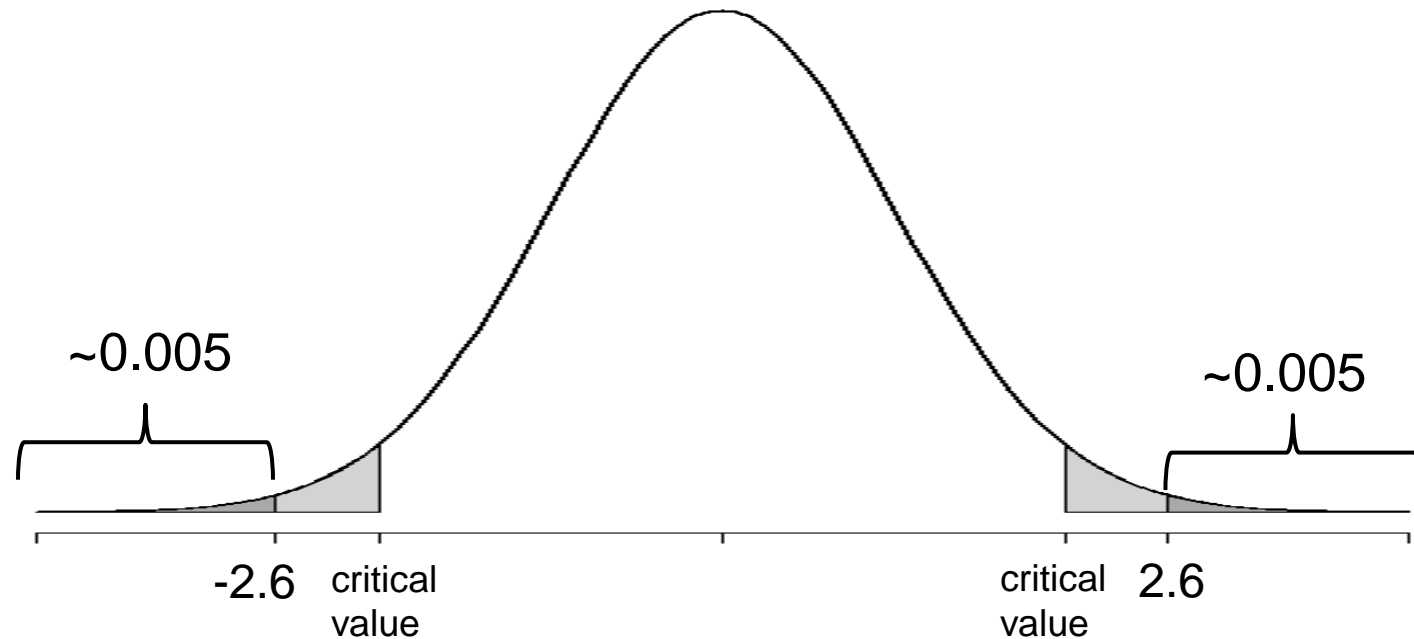
- $H_0$  wird mit Irrtumswahrscheinlichkeit  $\alpha$  verworfen, wenn:  $|\hat{t}| > t_{(N_x + N_y - 2), 1 - \alpha/2}$



Rejection region: „ $H_0$  wird abgelehnt, da Daten zu stark von dem abweichen, was man sich unter  $H_0$  erwarten würde“

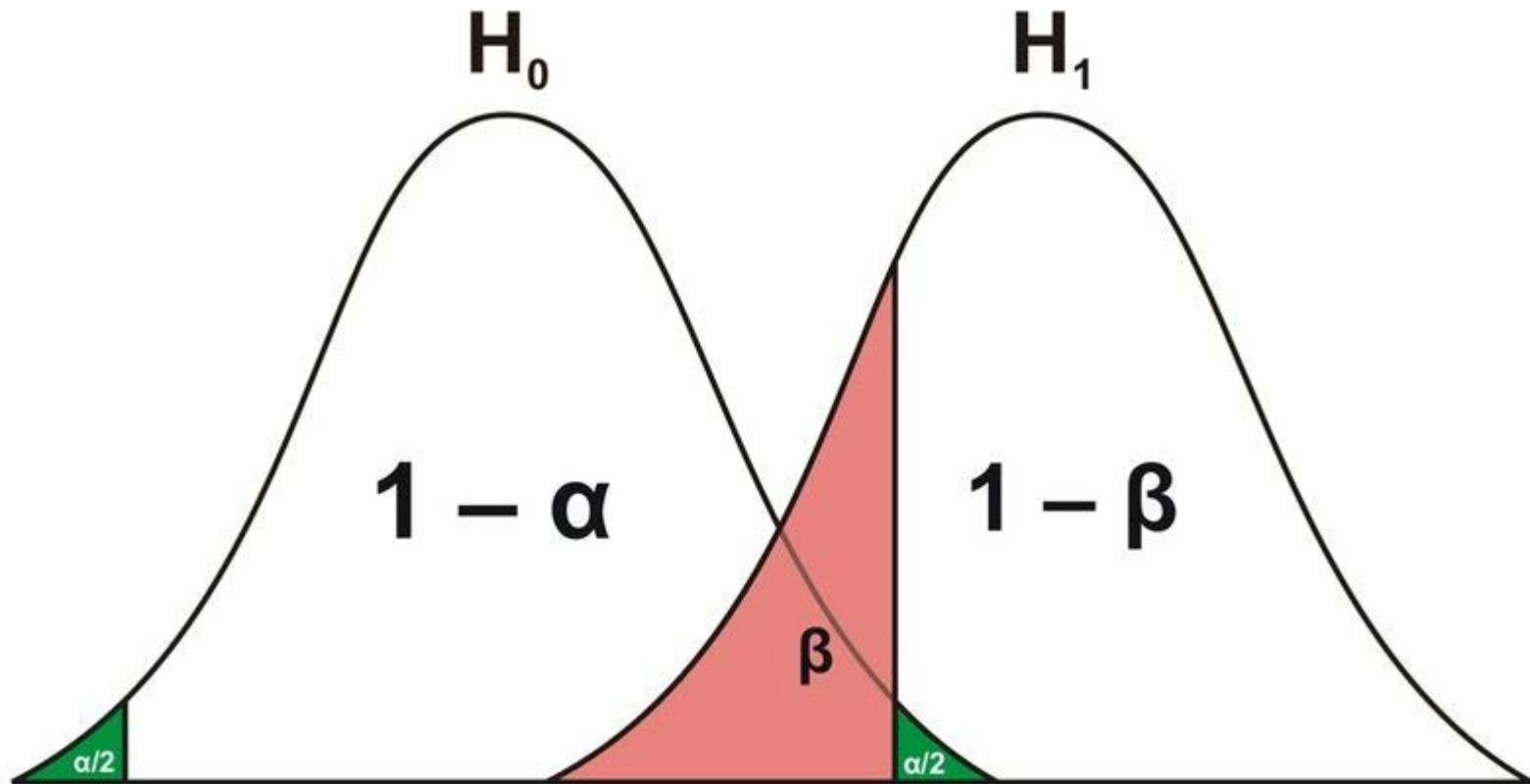
# Student's t-Test für unabhängige Stichproben (2)

- Wie gelangt man von der Teststatistik  $\hat{t}$  zum genauen p-Wert?
- Z.B.  $\hat{t} = 2,6$



- P-value (two-sided) = 0.005 + 0.005 = 0.01 (= Area under the curve)

# Fehler 1. und 2. Art - Zusammenhang





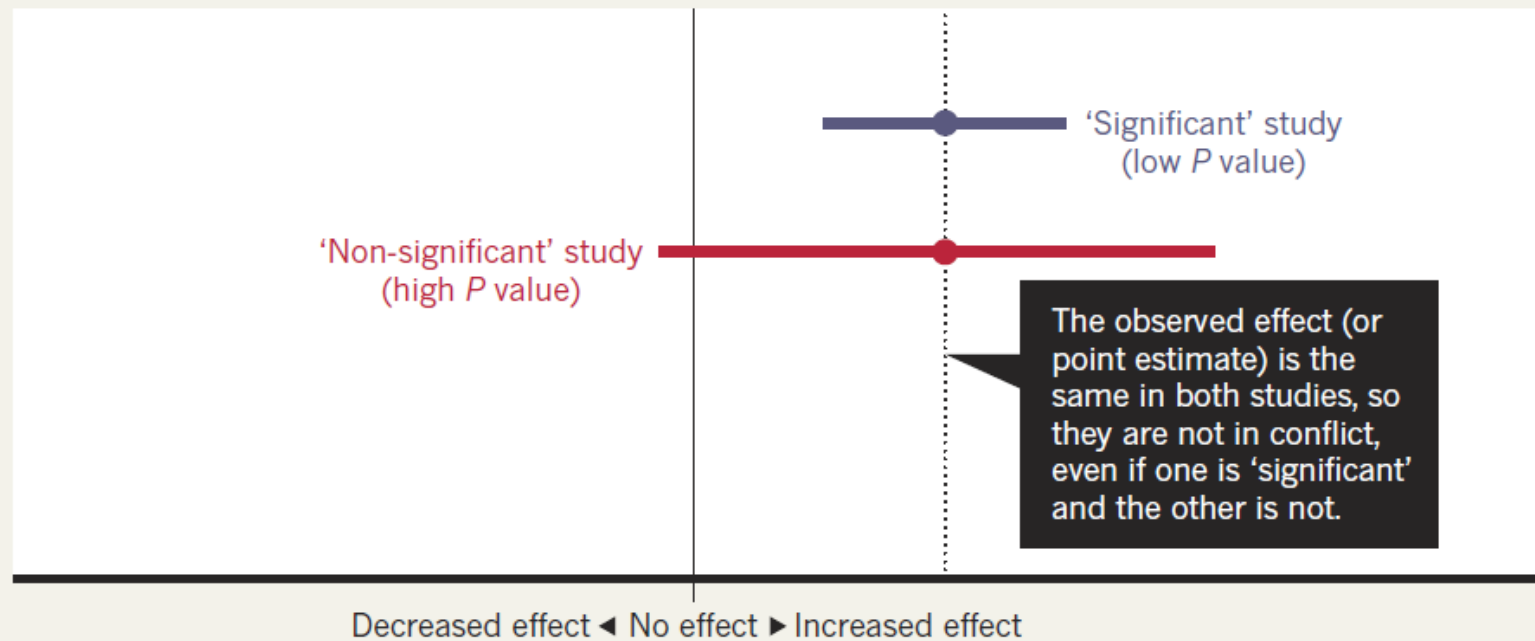
# Was ist der p-Wert

- Wahrscheinlichkeit, daher  $0 < p\text{-Wert} \leq 1$
- Er deutet an, wie wahrscheinlich es ist, ein Stichprobenergebnis wie beobachtet oder extremer zu erhalten, wenn die [Nullhypothese](#) wahr ist.
- Mit dem p-Wert wird also angedeutet, wie extrem das Ergebnis ist: je kleiner der p-Wert, desto mehr spricht das Ergebnis gegen die Nullhypothese
- $p\text{-Wert} = P(\text{Daten so extrem wie beobachtet} | H_0)$
- Häufige Fehlinterpretation: ~~p-Wert gibt an, wie wahrscheinlich die Nullhypothese bei Erhalt eines Stichprobenergebnisses wie beobachtet ist, d.h.  $P(H_0 | \text{Daten so extrem wie beobachtet} | H_0)$~~
- Die Nullhypothese wird verworfen, wenn der p-Wert kleiner als das vom Anwender festgelegte [Signifikanzniveau](#)  $\alpha$  ist (oft 0.01, 0.001, oder 0.05)
- Wenn die Nullhypothese zugunsten der Alternativhypothese verworfen wird, wird das Resultat als statistisch signifikant bezeichnet.
- Die Größe des p-Werts gibt **keine** Aussage über die Größe des wahren Effekts

# Falsche Schlußfolgerungen

## BEWARE FALSE CONCLUSIONS

Studies currently dubbed 'statistically significant' and 'statistically non-significant' need not be contradictory, and such designations might cause genuine effects to be dismissed.



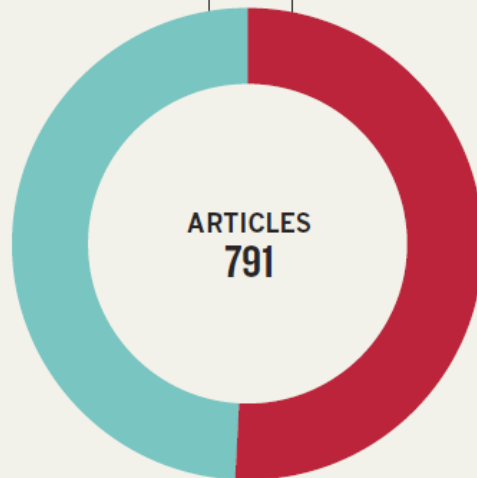
# Falsche Schlußfolgerungen

## WRONG INTERPRETATIONS

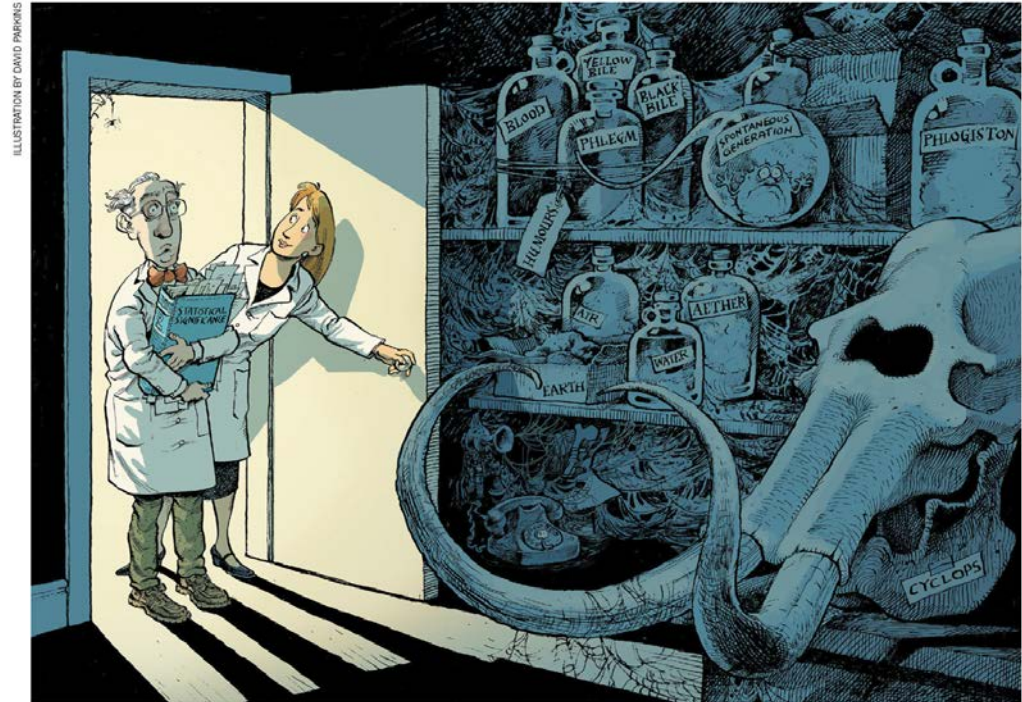
An analysis of 791 articles across 5 journals\* found that around half mistakenly assume non-significance means no effect.

Appropriately interpreted  
49%

Wrongly interpreted  
51%



\*Data taken from: P. Schatz *et al. Arch. Clin. Neuropsychol.* 20, 1053–1059 (2005); F. Fidler *et al. Conserv. Biol.* 20, 1539–1544 (2006); R. Hoekstra *et al. Psychon. Bull. Rev.* 13, 1033–1037 (2006); F. Bernardi *et al. Eur. Sociol. Rev.* 33, 1–15 (2017).



## Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

# Student`s t-Test für unabhängige Stichproben (3)



- Ein zwei-Stichproben t-Test zum Vergleich der Mittelwerte in zwei Gruppen liefert einen Wert der Teststatistik von 2,6, was einen p-Wert von 0.01 liefert.
- Richtige Interpretation:
  - Falls das Experiment 100 Mal wiederholt wird, d.h. 100 Mal zufällig eine Stichprobe gezogen wird, dann Teststatistik berechnet wird etc., und falls es in Wirklichkeit keinen Unterschied zwischen den beiden Gruppen gibt (Mittelwerte sind gleich), dann kann man erwarten, dass nur eine dieser 100 Teststatistiken einen Wert  $\geq |2.6|$  aufweist
- Falsche Interpretation:
  - ~~Die Nullhypothese ist zu 99% falsch~~

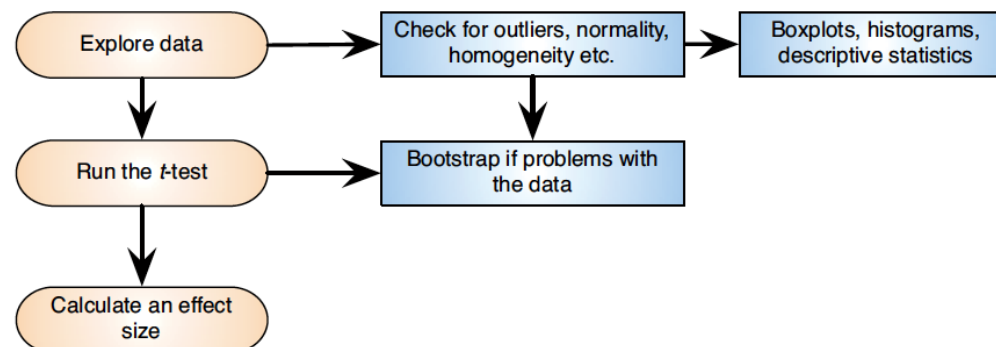
<https://www.graphpad.com/quickcalcs/ttest1.cfm>

# Student's t-Test für unabhängige Stichproben (4)



- Damit die Teststatistik unter  $H_0$  auch tatsächlich t-verteilt ist und der Test valide ist, müssen folgende Voraussetzungen erfüllt sein:
  - Jede der beiden Grundgesamtheiten sollten normalverteilt sein, Abweichungen bei größeren Stichprobenumfängen ( $N \geq 30$ ) zulässig (zentraler Grenzwertsatz)
  - -> Graphische Überprüfung mittels Histogramm
    - -> Kolmogoroff-Smirnov-Test
      - $H_0$ : Daten sind normalverteilt vs.  $H_1$ : Daten sind nicht normalverteilt
  - Gleiche Varianzen in den beiden Gruppen
    - -> Levene-Test

**FIGURE 9.3**  
The general process for performing a t-test



# Gepaarter Zwei-Stichproben t-Test

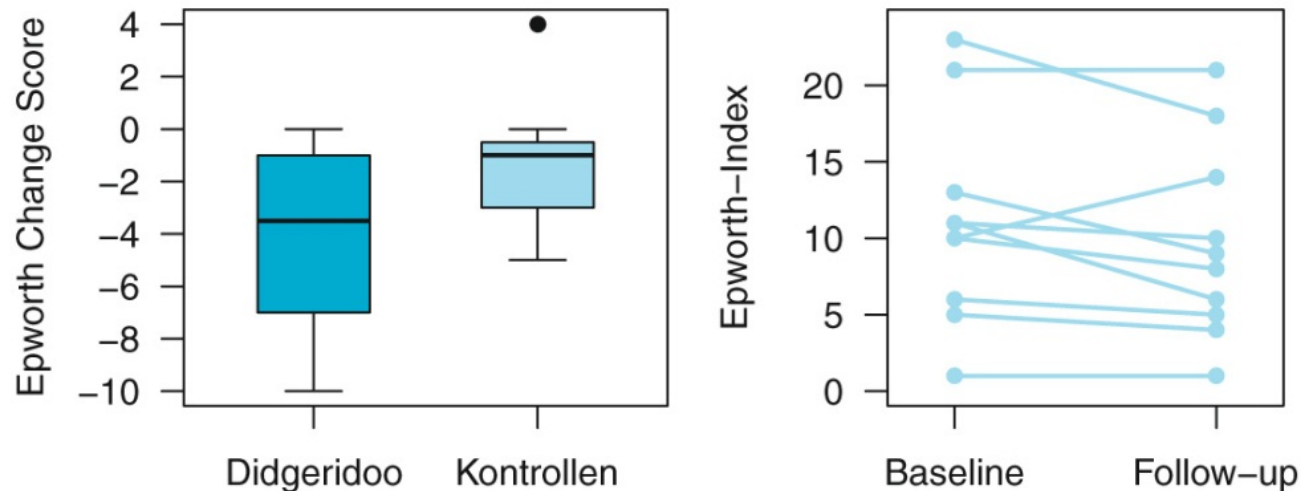
- Vorher-Nachher-Messungen
- Gruppe 1 (vorher) nicht unabhängig von Gruppe 2 (nachher)
- Modifikation der t-Tests notwendig
- -> Abhängiger/Gepaarter t-Test
- Berechnung der Differenz D für jedes Paar

$$H_0 : \mu_D = 0$$

$$H_1 : \mu_D \neq 0$$

$$t = \frac{\bar{D} - \mu_D}{s_D / \sqrt{N}} \sim t_{n-1}$$

# Verbundene versus unverbundene Stichproben



**Abbildung 4.5:** Illustration von unverbundenen (links) und verbundenen (rechts, nur Kontrollgruppe) Stichproben in der Didgeridoo-Studie.



# Beispiel

- Um eine neue Therapie zur Senkung des Cholesterinspiegels zu testen, werden bei zehn Probanden vor und nach der Behandlung die Cholesterinwerte bestimmt. Es ergeben sich die folgenden Messergebnisse:

Vor der Behandlung:	223	259	248	220	287	191	229	270	245	201
Nach der Behandlung:	220	244	243	211	299	170	210	276	252	189
Differenz:	3	15	5	9	-12	21	19	-6	-7	12

$$\bar{d} = 5,9 \quad s_d = 11,3866$$

<https://www.graphpad.com/quickcalcs/ttest1.cfm>

$$t = \sqrt{10} \frac{5,9}{11,3866} = 1,6385$$

Teststatistik

$$t(0,975; 9) = 2,2622$$

Kritischer Wert

$|t| \leq t(0,975; 9)$  Die Nullhypothese, dass die Erwartungswerte der Cholesterinwerte vor und nach der Behandlung gleich sind, kann nicht abgelehnt werden. Wenn die Behandlung überhaupt einen Effekt hat, so ist dieser nicht groß genug, um ihn mit einem so kleinen Stichprobenumfang zu entdecken.



- Analysis of Variance - Streuungszerlegung
- Parametrischer Test zum Vergleich von Mittelwerten
- T-test = Zweigruppenvergleich, ANOVA mehrere Gruppen möglich
- Mittelwertsvergleich erfolgt durch die Zerlegung der Varianz in:
  - Streuung zwischen den Gruppen
  - Streuung innerhalb der Gruppen
- Voraussetzungen
  - Normalverteilung
  - Varianzhomogenität

- Einfache ANOVA,  
ein Faktor mit  $\geq 2$  Stufen
- Mehrweg-ANOVA,  
mehrere Faktoren mit  $\geq 2$  Stufen
- Kovarianzanalyse, auch stetige Einflussfaktoren
- Multivariate ANOVA, mehrere abhängige Variablen
- ANOVA für Messwiederholungen,  
aufeinanderfolgende Beobachtungen sind abhängig

# ANOVA, post hoc Vergleiche



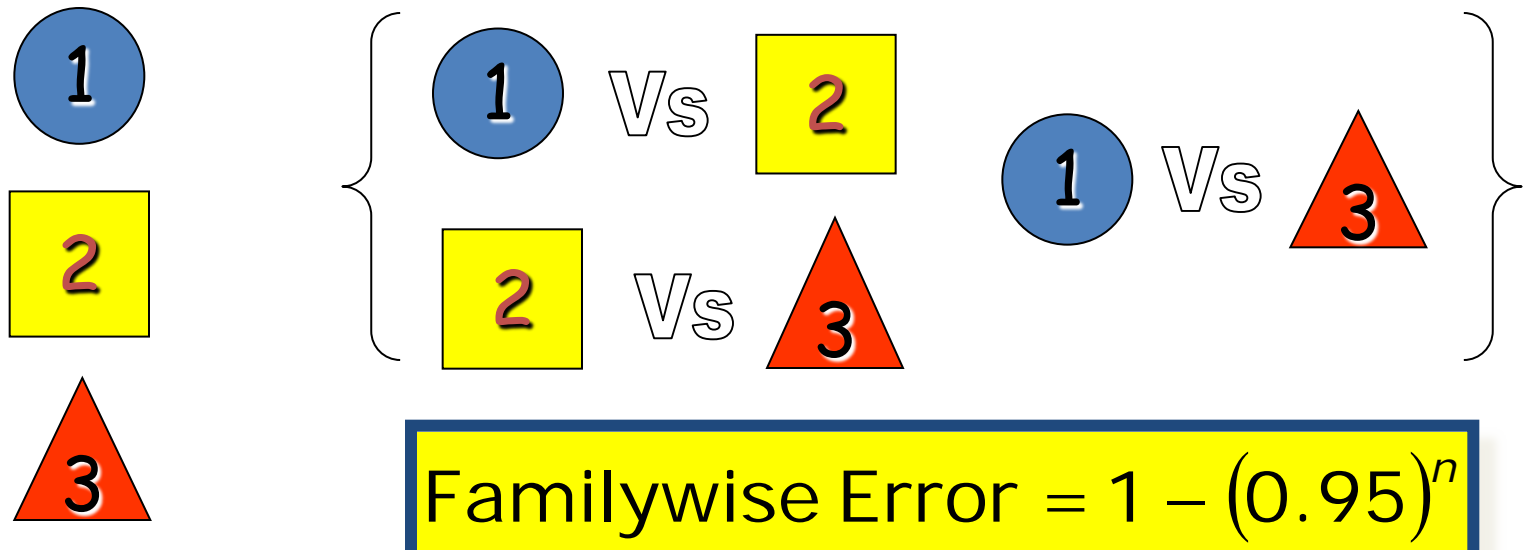
MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

- Bonferroni
- Bonferroni-Holm
- Holm-Sidak
- Tukey
- Scheffe
- LSD
- Dunnett

# Why Not Use Lots of $t$ -Tests?

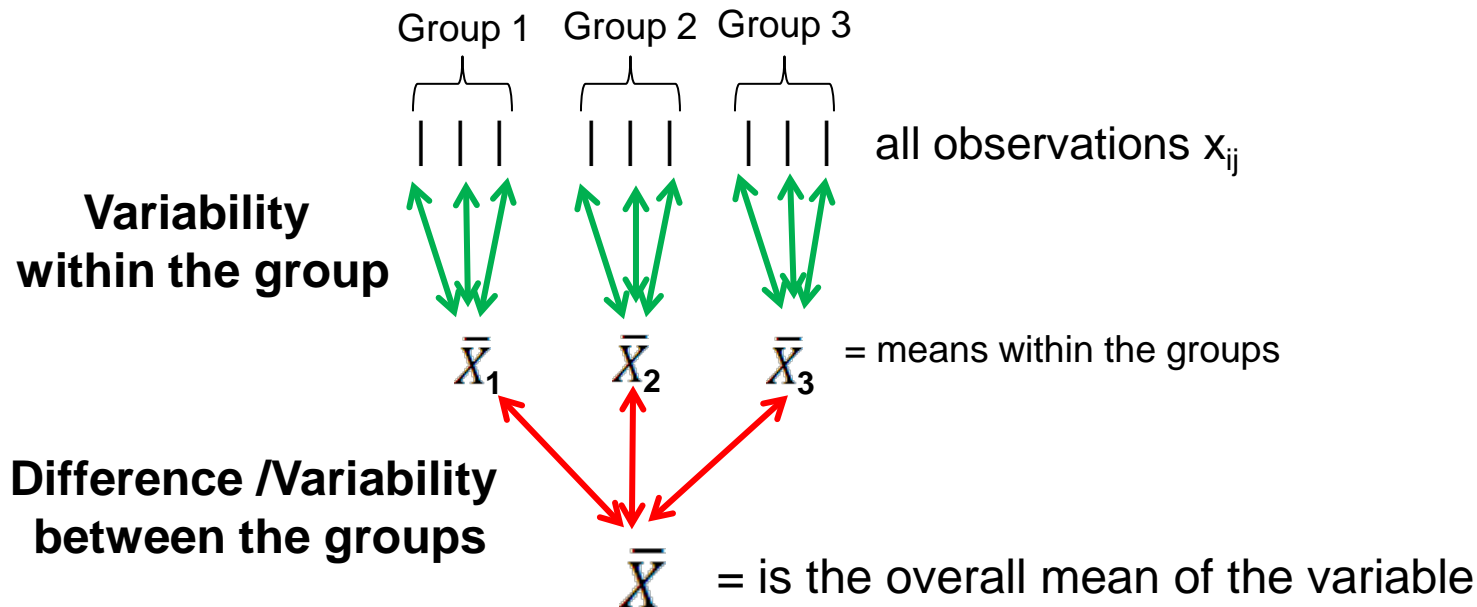
If we want to compare several means why don't we compare pairs of means with  $t$ -tests?

- Can't look at several independent variables.
- Inflates the Type I error rate.



# Analysis of Variance (ANOVA)

- Situation: Compare the means of  $k$  samples ( $k > 2$ )
- Assumption: normal distribution of the population,  $\sigma = \sigma_1 = \sigma_2 = \dots = \sigma_k$
- Hypothesis:  $H_0: \mu_1 = \mu_2 = \dots = \mu_k$  versus  $H_1: \mu_i \neq \mu_j$  ( $i \neq j$ ): At least two of the means differ
- Nowadays, linear mixed effects models are preferred instead of ANOVA



## Pearson's Chi-Quadrat Test

1900

$$X^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

Korrelation und Regression



Karl Pearson (1857-1936)

## Fisher's Exakter Test

Varianzanalyse



Sir Ronald Aylmer Fisher (1890-1962)

# Chi-Quadrat-Test qualitatives Merkmal

Blutdruck	Therapie A	Therapie B	Gesamt
Hypertonie	$n_{1.1}=300$	$n_{1.2}=150$	$n_{1.}=450$
Normotonie	$n_{2.1}=300$	$n_{2.2}=250$	$n_{2.}=550$
Gesamt	$n_{.1}=600$	$n_{.2}=400$	$n=1000$

**H0: Es besteht kein Zusammenhang zwischen Therapie und Blutdruck (  $p_A = p_B$  )**

**H1: Es besteht ein Zusammenhang (  $p_A \neq p_B$  )**

**Teststatistik:**

$$\hat{\chi}^2 = n \cdot \frac{(n_{22} \cdot n_{11} - n_{21} \cdot n_{12})^2}{(n_{2.} \cdot n_{1.} \cdot n_{.2} \cdot n_{.1})} = 1000 \cdot \frac{(300 \cdot 250 - 300 \cdot 150)^2}{450 \cdot 550 \cdot 600 \cdot 400} = 15,15$$

**Testentscheidung:**

$$\hat{\chi}^2 = 15,15 > 3,841 = \chi_{1;0,95}^2 \quad \Rightarrow \quad \text{Ablehnen der Nullhypothese}$$

$\chi_{n;1-\alpha}^2$  mit  $n =$  Anzahl der Freiheitsgrade;  $\alpha =$  Signifikanzniveau  
 hannu.ulmer@i-med.ac.at

# Beispiel



<b>Smoking status</b>	<b>Current Smoker</b>	<b>Ex-Smoker</b>	<b>Never Smoker</b>	<b>Row Total</b>
<b>Gender</b>				
Men	144	310	268	722
Women	117	143	475	735
Column Total	261	453	743	1457

Teststatistik:  $121.9218 \sim \chi^2 ((2-1)*(3-1)) = \chi^2 (2)$

Kritischer Wert von  $\chi^2 (2) = 5.99$

→ Nullhypothese (Rauchverhalten unterscheidet sich nicht zwischen Männern und Frauen) kann abgelehnt werden ( $p = 3.3e-27$ )



# Übung: Führen Sie einen Chi-Quadrat Test durch



**Table 2. Effect of Cytisine on Smoking Cessation.\***

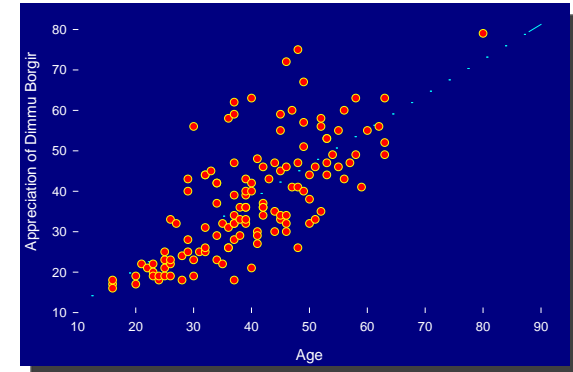
Outcome	Cytisine (N= 370)	Placebo (N= 370)	Percentage-Point Difference (95% CI)	Relative Rate (95% CI)†
	<i>percent (number)</i>			
Primary outcome: abstinence for 12 mo	8.4 (31)	2.4 (9)	6.0 (2.7–9.2)‡	3.4 (1.7–7.1)
Abstinence for 6 mo	10.0 (37)	3.5 (13)	6.5 (2.9–10.1)‡	2.9 (1.5–5.3)
Point prevalence at 12 mo	13.2 (49)	7.3 (27)	5.9 (1.6–10.3)§	1.8 (1.2–2.8)

# Regression – significance testing (1)

$$y = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n + \varepsilon_i$$

$$SS_{tot} = SS_{reg} + SS_{res}$$

$$R^2 = \frac{SS_{reg}}{SS_{tot}} \dots \text{Coefficient of determination}$$



Using the regression model, can we significantly better predict values of the outcome than using the mean?

H0:  $R^2=0$  (alternativ:  $\beta_1 = \beta_2 = \dots = \beta_n = 0$ )

H1:  $R^2 \neq 0$

Test statistic:

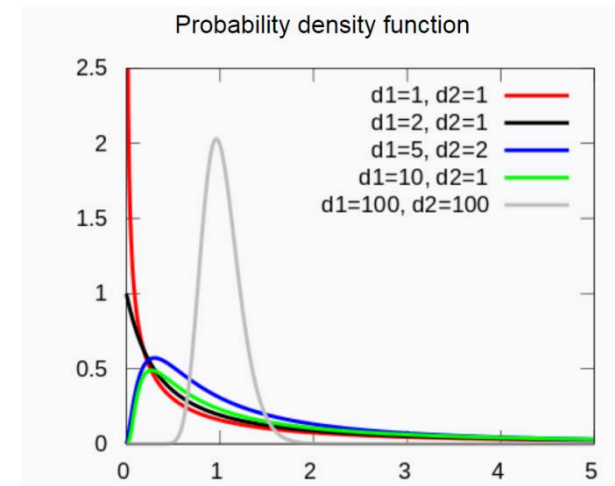
$$F = \frac{MS_M}{MS_R}$$

F-distribution:

MS ... Mean Squares (averages of total values)

$F \sim F(n, N - (n + 1))$ -distributed

ANOVA test – **AN**alysis **Of** **VA**riance



# Regression – significance testing (2)

---



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

Critical values for the F-distribution:

[http://www.socr.ucla.edu/applets.dir/f\\_table.html](http://www.socr.ucla.edu/applets.dir/f_table.html)

Critical value for  $F(1,198)$  for  $\alpha=0.05$ :  $\sim 3.9$

# Regression – significance testing (3)

- To test the significance of individual regression coefficients

– t-test

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

– Test statistic:  $T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t(N-2)$ -distributed

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	134.140	7.537		17.799	.000	119.278	149.002
	Advertising Budget (Thousands of Pounds)	.096	.010	.578	9.979	.000	.077	.115
2	(Constant)	-26.613	17.350		-1.534	.127	-60.830	7.604
	Advertising Budget (Thousands of Pounds)	.085	.007	.511	12.261	.000	.071	.099
	No. of plays on Radio	3.367	.278	.512	12.123	.000	2.820	3.915
	Attractiveness of Band	11.086	2.438	.192	4.548	.000	6.279	15.894

a. Dependent Variable: Album Sales (Thousands)

# Regression – significance testing (4)

---

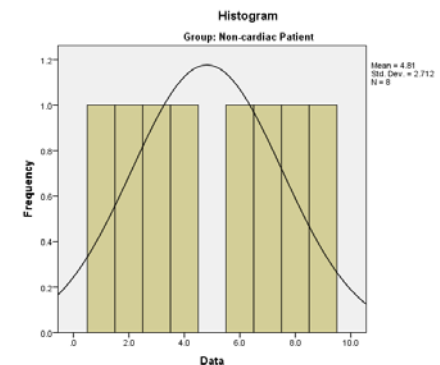
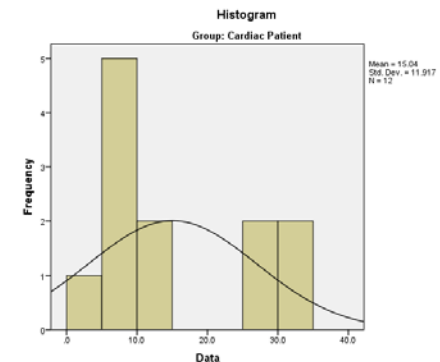
- Prerequisites that F-test (ANOVA) and t-test for regression are valid
  - Homoscedasticity:
    - For each value of the predictors the variance of the error term should be constant.
  - Independent Errors:
    - For any pair of observations, the error terms should be uncorrelated
  - Normally-distributed Errors
    - Normal probability plot

# Nicht-parametrische Tests

## Mann-Whitney U Test (1)

- 20 Patienten einer Klinik werden untersucht. 12 davon sind in kardiologischer Behandlung, während 8 dies nicht sind. Sie alle beantworten einen Fragebogen zum allgemeinen Wohlbefinden (Werte von 0 bis 35, 0 steht für ein sehr hohes, 35 für ein sehr geringes Wohlbefinden). Es soll geprüft werden, ob es Unterschiede hinsichtlich der zentralen Tendenz des Wohlbefindens zwischen den Herzpatienten und den übrigen Patienten gibt.
- *Mann-Whitney.sav*
- T-Test: Nur zulässig, falls Daten normalverteilt
- Falls Voraussetzungen für T-Test nicht erfüllt, dann:
- **Mann-Whitney U Test:** zulässig für alle Verteilungen, solange  $F_Y(x) = F_X(x - a)$  gilt (Verteilung in den beiden Gruppen bis auf Verschiebung gleich)

$$H_0 : a = 0 \text{ vs. } H_1 : a \neq 0$$



# Nicht-parametrische Tests

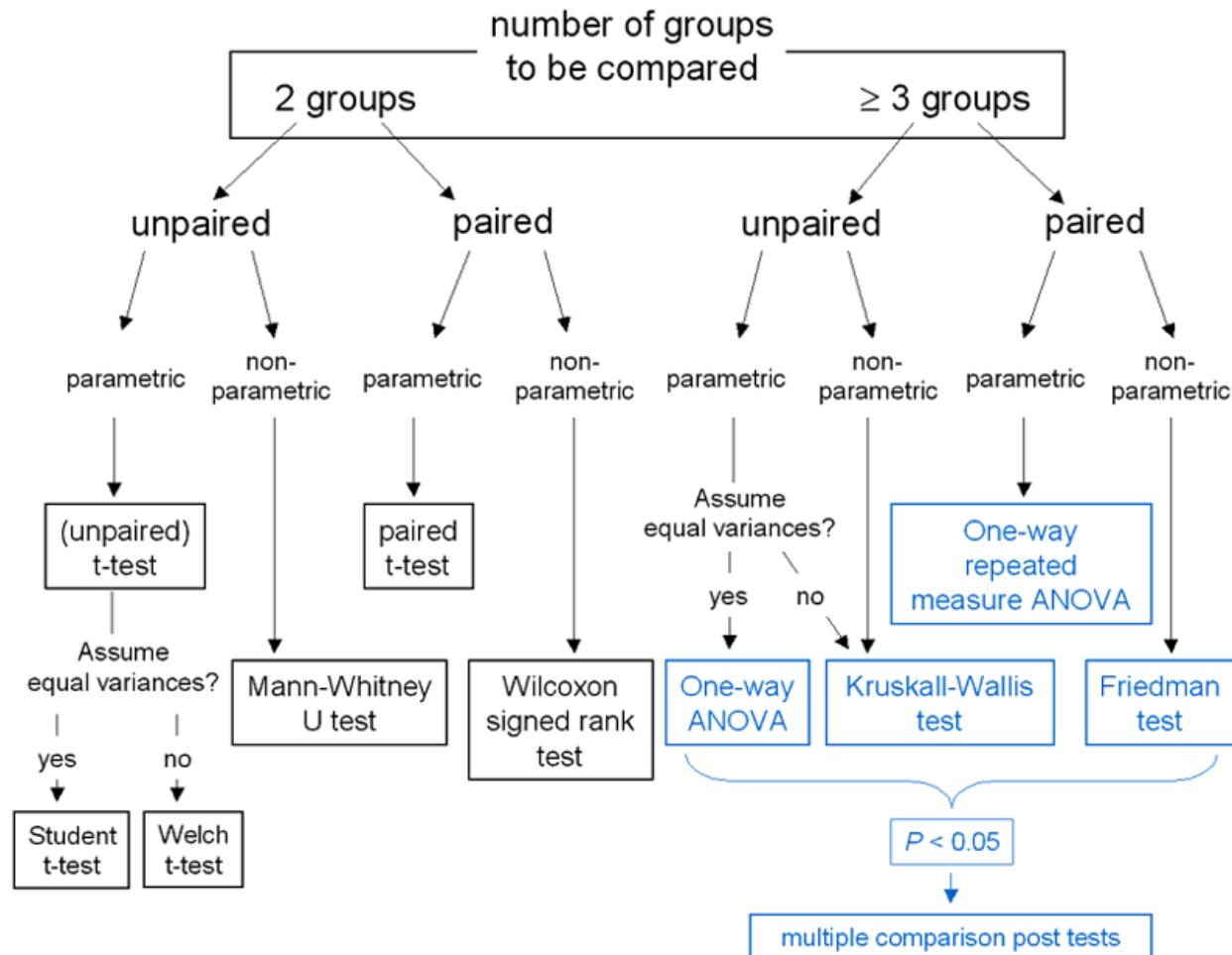
## Mann-Whitney U Test (2)

- MWU-Test rechnet mit den Rängen, nicht mit den Messwerten selbst

ID	Gruppe	Wohlbefinden	Ränge Gruppe 1	Ränge Gruppe 2
5	1	0	1	
6	2	1		2
14	2	2		3
9	2	3		4
18	2	4		5
10	1	5	6	
19	1	5.5	7	
1	2	6		8
8	2	6.5		9
17	1	7	10	
15	2	7.5		11
11	1	8	12	
3	2	8.5		13
2	1	9	14	
20	1	11	15	
12	1	13	16	
16	1	28	17	
4	1	29	18	
7	1	32	19	
13	1	33	20	
Rangsumme			155	55

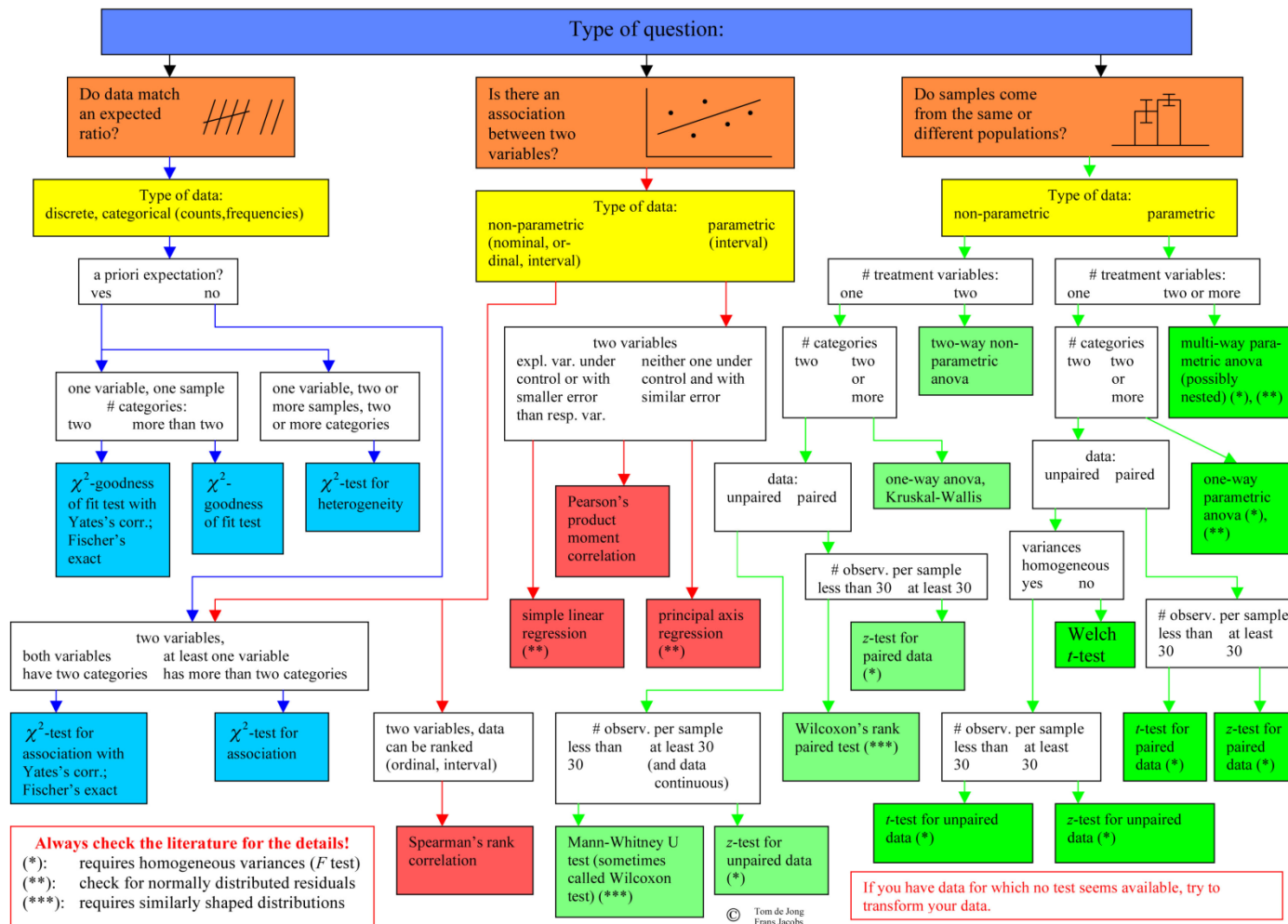
- Berechnung der Rangsummen in beiden Gruppen
- Teststatistik U:  $U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$   
 $n_1$ =Stichprobengröße der Gruppe mit der größeren Rangsumme  
 $n_2$ =Stichprobengröße der Gruppe mit der kleineren Rangsumme  
 $R_1$ =größere der beiden Rangsummen
- Im Bsp.:  $U = 12 \cdot 8 + \frac{12(12+1)}{2} - 155 = 19$
- Stichprobe hinreichend groß ( $n_1+n_2>30$ ): U annähernd normalverteilt
- Im Bsp.:  $z = \frac{U - \mu_U}{\sigma_U} = \frac{19 - \frac{12 \cdot 8}{2}}{\sqrt{12 \cdot 8(12+8+1)}} = -2.34$
- $p < 0.05$ , da  $-2.34 < -1.96$

# Overview of statistical tests (2)





# Overview of statistical tests (3)



# Common statistical tests



	Quantitative outcome variable		Qualitative outcome variable	
	Normal distribution	Any other distribution	Expected frequency in each cell of the crosstable „high“	Expected frequency in each cell of the crosstable „low“
<b>Compare 2 groups</b>	t-test	Wilcoxon-test / Mann-Whitney U-Test	Chi-Square	Fishers exact test
<b>Compare &gt;2 groups</b>	Analysis of Variance (ANOVA)	Kruskal-Wallis-Test	Chi-Square	

**Testing measures of location:**  
Does the mean/median differ between groups

**Testing frequencies in a crosstable:**  
Are the rows and columns independent from each other?

# Wichtige Signifikanztests

		Zielvariable(Outcome)		
		Qualitativ	Quantitativ	
			Normalverteilung	Beliebige Verteilung
Vergleich zweier Gruppen	Unverbunden	<u>Chi-Quadrat Test</u> , Fisher Test	<u>t-Test</u> für unverbundene Stichproben	Mann-Whitney U Test
	Verbunden	McNemar Test	t-Test für verbundene Stichproben	Wilcoxon Test
Vergleich von mehr als zwei Gruppen	Unverbunden	Chi-Quadrat Test	Einfache Varianzanalyse	Kruskal-Wallis Test
	Verbunden	Q-Test von Cochran	Varianzanalyse für Meßwiederholungen	Friedman Test

# Multiples Testen

- „The multiple comparison problem involves the repeated testing of a series of hypotheses and the resultant increasing probability of a type I error.” Van Belle (2002), p. 149.

- P (Fehler 1. Art)

$$\alpha = 1 - (1 - 0.05)^n$$

- z.B. Bonferroni-Korrektur

$$\alpha_i = \frac{0.05}{n}$$



# Sample size estimation

**Question:** How many individuals do you have to include in your study to get a reliable result ?

→ We want to **maximize the probability**  
for rejecting  $H_0$ , if  $H_1$  is true

→ while keeping the **Type I error  $\alpha$**  fixed

**What do you have to know  
to calculate the sample size  
needed?**

1. Power (typically set to 80% or 90%)
2. Type I error  $\alpha$  (typically set to  $\alpha = 0.05$ )
3. The difference you want to find (for t-tests:  
the mean difference between groups)
4. standard deviation / measure of variance



# Sample size estimation

---

## Example

- Hypothesis:  $H_0: \mu_A = \mu_B$  versus  $H_1: \mu_A \neq \mu_B$  → two-sided t-test
  - You consider a difference of 10 as relevant
  - From former studies, you know, that the standard deviation is  $\sim 15$  mmHG
  - So far, you have recruited 20 patients (10 in each treatment arm)
- What is your power?

## Fallzahlschätzung für unverbundene Stichproben und stetige Zielgrößen

- Fallzahlberechnung für vorgegebene Power
- Powerberechnung für vorgegebene Fallzahl
- Entdeckbare Differenz für vorgegebene Fallzahl und Power

Eingabe von  $\mu_1$ :  Eingabe von  $\mu_2$ :

Eingabe von  $\sigma$ :  Differenz Delta:

- Einseitiger Test
- Zweiseitiger Test

Eingabe von  $\alpha$  (Standard ist 0.05):

Eingabe der Power (Standard ist 0.80):

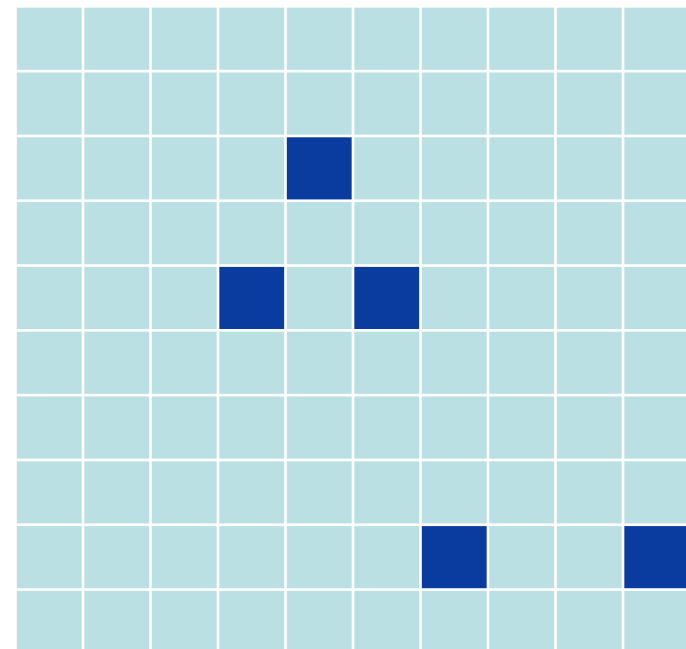
Die Fallzahl für jede Gruppe ist:

# The multiple testing problem

## The situation:

- Consider a dataset with 100 independent parameters, which do not play a role in the etiology of the disease of interest (what you don't know, of course)
- 100 statistical tests are performed with a significance level of  $\alpha=0.05$
- The tests are constructed in that way, that maximum 5 of 100 tests reject the Nullhypothesis, although it is true (which is the case in this example)

→ You expect 5 tests to be significant just by chance





# The multiple testing problem

- The probability to get at least one Type I error increases with increasing number of tests.
- **Family-wise error rate** (the error rate for the complete family of tests performed):  $\alpha^* = 1 - (1 - \alpha)^k$ , with  $\alpha$  being the **comparison-wise error rate**

The probability  
to get one or more  
false discoveries  
(Type I error)

k	$\alpha^*$ ( $\alpha=0.05$ )
1	0.05
5	0.226
10	0.401
100	0.994

→ The significance level has to be modified for multiple testing situations

# The multiple testing problem

## The Bonferroni correction method:

- Control the comparison-wise error rate: Reject  $H_0$ , if  $p < \alpha$
- Control the family-wise error rate (including  $k$  tests): Reject  $H_0$ , if  $p < \alpha/k$   
→ **Advantage: simple**
- Problem: Bonferroni-correction increases the probability of a type II error  
→ the power of detecting a true association is reduced → **Disadvantage: too conservative**

k	$\alpha/k$ ( $\alpha=0.05$ )
1	0.05
5	0.01
10	0.005
100	0.0005



$$0.05/5=0.01$$

# Problem of Multiplicity

Number of tests	$\alpha$ level (in each test)	Global $\alpha$ level (risk of making at least one type 1 error)
1	0.05	0.05
2	0.05	0.10
3	0.05	0.14
5	0.05	0.23
14	0.05	0.51
100	0.05	0.994



Already by 14 tests we have over 50% chance of falsely rejecting  $H_0$  in at least one of the tests.



# Das Problem der Multiplizität

---

ICH E9:

„Multiplicity may arise, for example, from multiple primary variables, ... multiple comparisons of treatments, repeated evaluations over time and/or interim analyses.“

- Mehrere Zielkriterien
- Mehr als 2 Behandlungsgruppen
- Zielkriterium an mehreren Zeitpunkten gemessen
- Zwischenauswertungen



# How to correct for multiplicity?

## **Bonferroni correction:**

Create a **corrected significance level  $\alpha/N$**  and test each of the analyses on this new level.

**Example:** We have 5 tests ( $N=5$ ) and we wish to have an overall  $\alpha=0.05$ . Conduct each test on the corrected significance level  $\alpha=0.05/5=\underline{0.01}$ .

## **Advantages:**

- Easy to implement.
- No order of objectives

## **Disadvantages:**

- Very conservative

# How to correct for multiplicity?

## Gate keeping procedures:

Order objectives/analyses. Test each level at

If significant => move to next level.

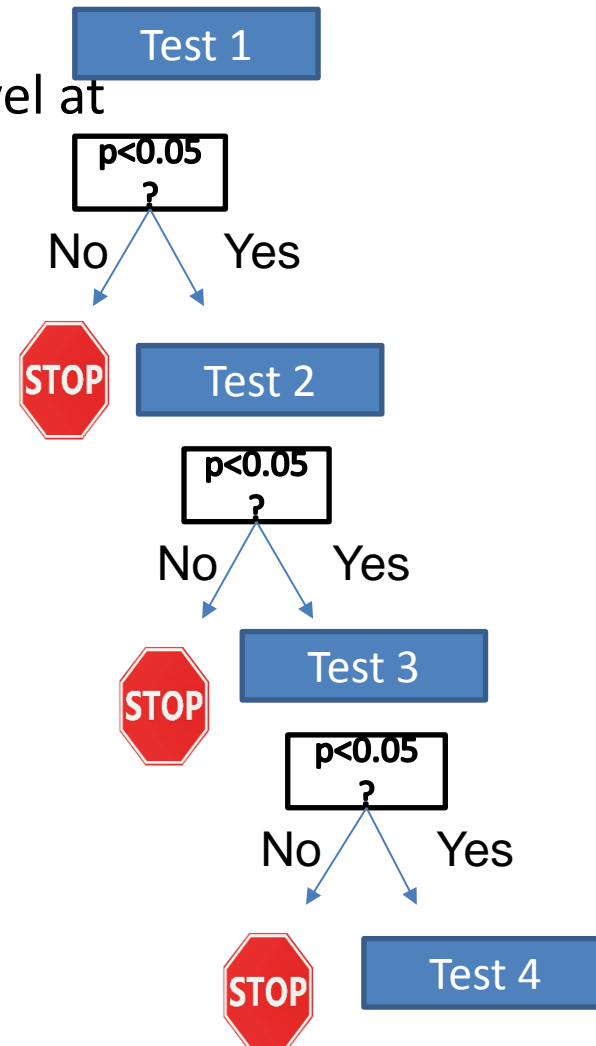
If non significant => STOP!

## Advantages:

- Easy to implement.
- All tests on the same  $\alpha$  level

## Disadvantages:

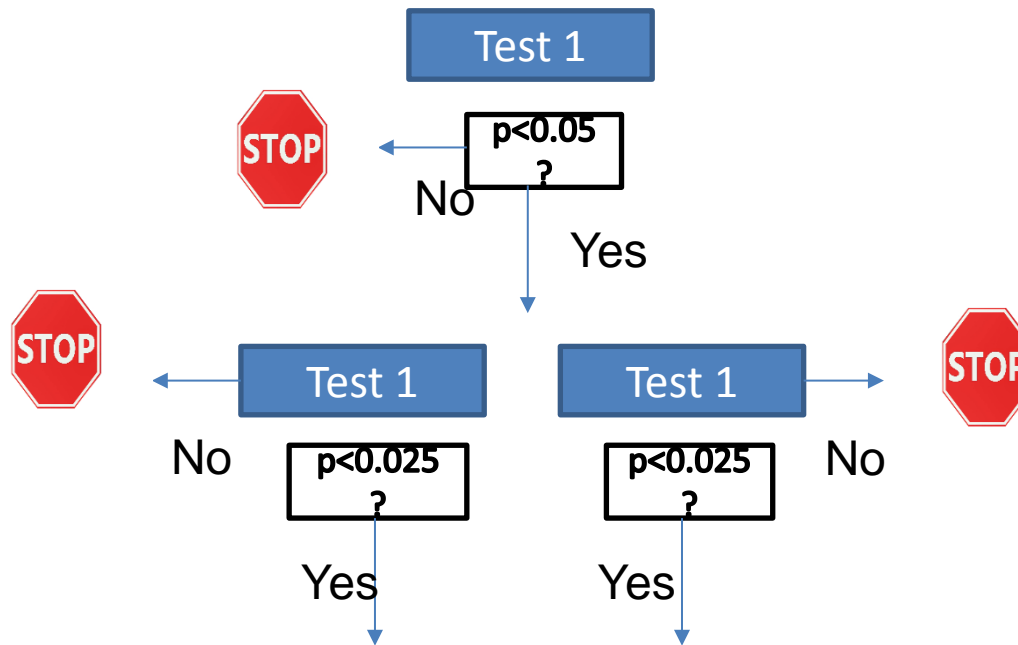
- Must be pre-planned.
- Requires order of analyses.
- Some analyses may not be conducted.



# Combined Bonferroni and Gate keeping



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK





# How to correct for multiplicity?

## Bonferroni-Holm

Conduct all tests (N tests). Order all the p-values from smallest to biggest:  $P_{(1)}, P_{(2)}, \dots, P_{(N)}$

Test the smallest p-value ( $P_{(1)}$ ) on  $P_{(1)} < \alpha/(N+1)$  and the  $m^{\text{th}}$  p-value on  $P_{(m)} < \alpha/(N+1-m)$ . STOP at first non significant p-value (this p-value and all bigger p-values will be considered non-significant).

## Advantages:

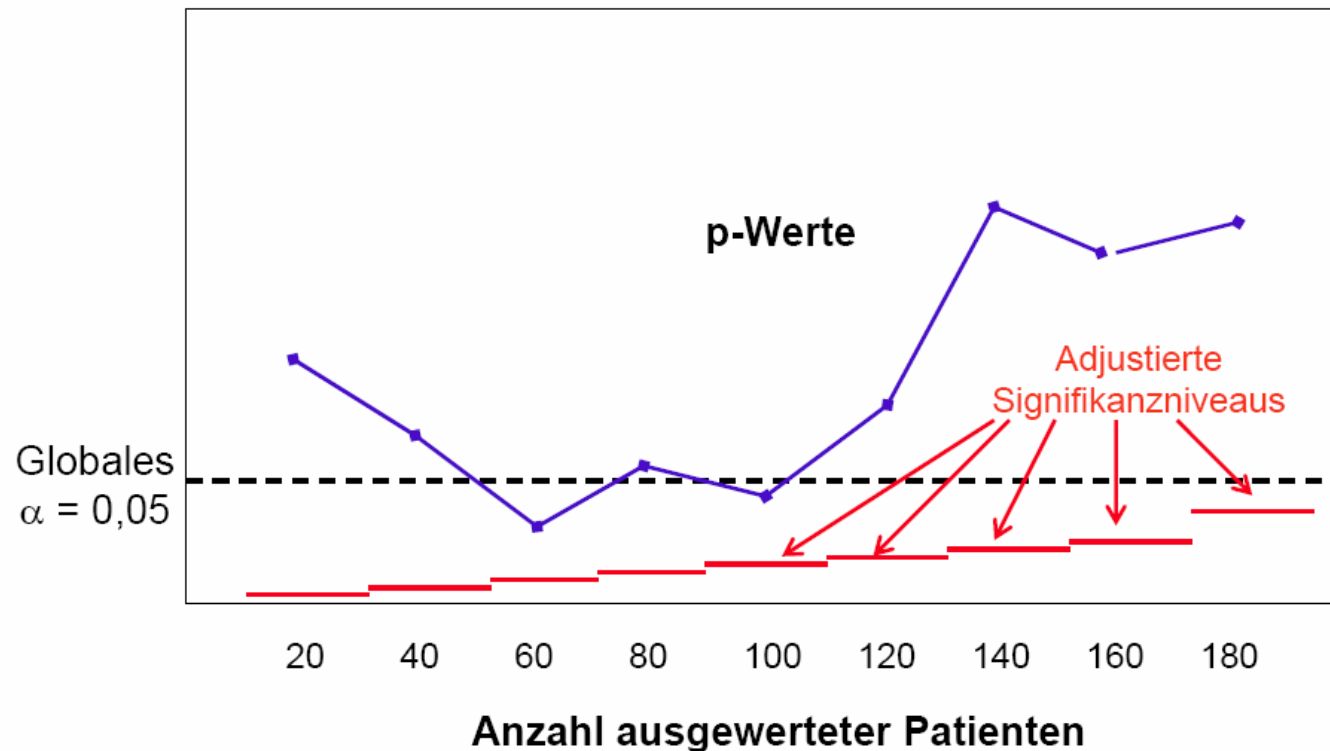
- No order of objectives.
- Not as conservative as Bonferroni.

## Disadvantages:

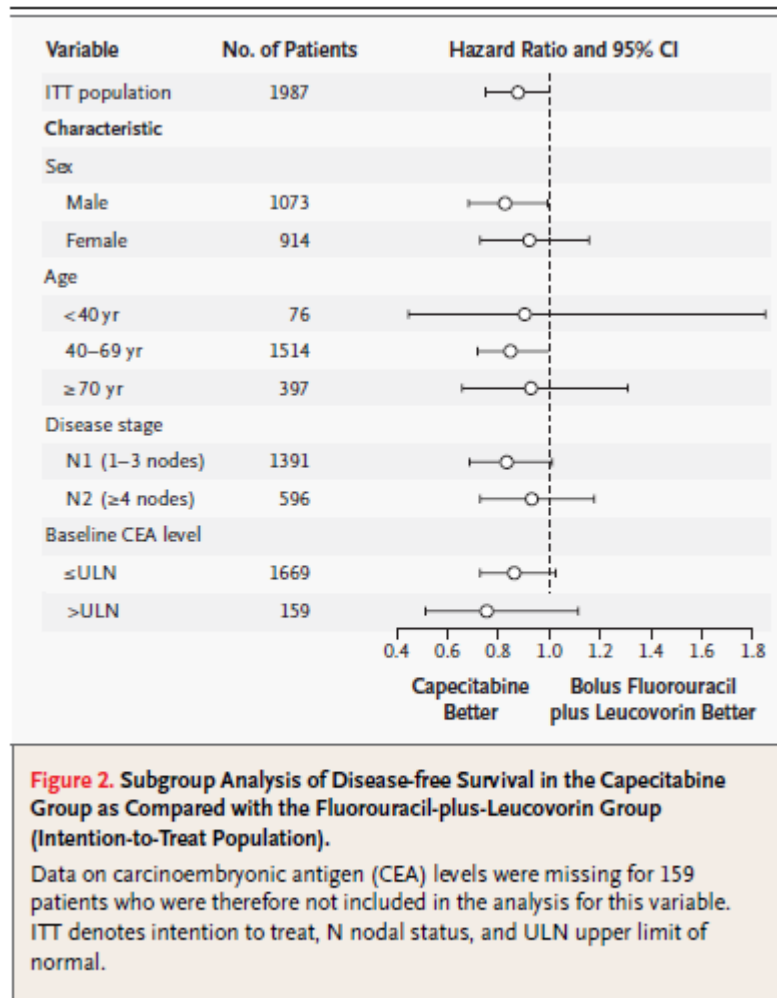
- Hard to implement/explain.



# Beispiel für Adjustierung des Fehlers 1. Art in Zwischenauswertungen



# Subgruppenanalysen



# Lösungen für das Multiplizitätsproblem



- die Studie als explorativ zu definieren – data mining, fishing for significance erlaubt
- 1 primärer Zielparameter wird konfirmatorisch ausgewertet, alle übrigen explorativ
- Alfa-Adjustierung: Aufteilung des Signifikanzniveaus (erhöht Fallzahl)  
Extrembeispiel: Biomarkersuche in Proteomics und Genomics
- A priori Ordnung der Hypothesen (bis zur ersten Nichtsignifikanz mit  $\alpha=0,05$  testen)

- Null Hypothesis:
  - Like a t-test, ANOVA tests the null hypothesis that the means are the same.
- Experimental Hypothesis:
  - The means differ.
- ANOVA is an Omnibus test
  - It test for an overall difference between groups.
  - It tells us that the group means are different.
  - It doesn't tell us exactly which means differ.
- $H_0: \mu_1 = \mu_2 = \dots = \mu_k$
- $H_1: \text{es gibt } i \text{ und } j \text{ mit } \mu_i \neq \mu_j$

# Theory of ANOVA (1)

- Same concept as shown for the F-test for regression
- Decomposition of the total variability ( $SS_T$ ) into (i) variability between groups ( $SS_M$ ) and (ii) variability within groups ( $SS_R$ )

$$SS_T = \sum (x_i - \bar{x}_{grand})^2$$

$$df_T = (N - 1)$$

$$SS_T = SS_M + SS_R$$

$$SS_M = \sum n_i (\bar{x}_i - \bar{x}_{grand})^2$$

$$df_M = (k - 1)$$

$$MS_M = \frac{SS_M}{df_M}$$

$$SS_R = \sum (x_i - \bar{x}_i)^2$$

$$df_R = (n_1 - 1) + \dots + (n_k - 1)$$

$$MS_R = \frac{SS_R}{df_R}$$

- k ... number of groups,  $n_i$  ... size of group i, N ...  $n_1 + \dots + n_k$
- SS ... Sum of squares
- MS ... Mean squares

# Theory of ANOVA (2)

$$F = \frac{MS_M}{MS_R} \text{ follows a } F(k-1, N-k) \text{ distribution}$$

Decision about  $H_0 (\mu_1 = \mu_2 = \dots = \mu_k)$  by comparing  $F$  with the critical value of the  $F(k-1, N-k)$  distribution

Assumptions which have to be fulfilled for the ANOVA (the same as for the t-test)

- data normally distributed
- homogeneity of variances

# ANOVA by hand (1)

- Testing the effects of viagra on libido using three groups:
  - Placebo (Sugar Pill)
  - Low Dose Viagra
  - High Dose Viagra
- The outcome/dependent variable (DV) was an objective measure of libido.

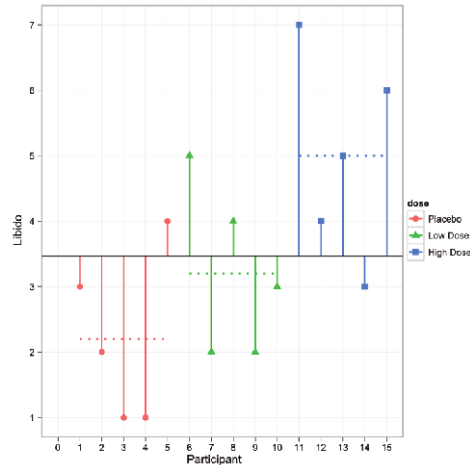
**TABLE 11.1** Data in **Viagra.sav**

	<i>Placebo</i>	<i>Low Dose</i>	<i>High Dose</i>
	3	5	7
	2	2	4
	1	4	5
	1	2	3
	4	3	6
$\bar{X}$	2.20	3.20	5.00
<i>s</i>	1.30	1.30	1.58
$s^2$	1.70	1.70	2.50
	Grand mean = <b>3.467</b> Grand <i>SD</i> = <b>1.767</b>		
	Grand variance = <b>3.124</b>		

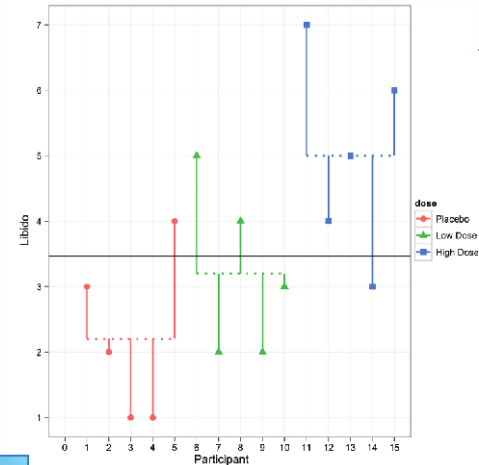
# ANOVA by hand (2)

**FIGURE 11.3**

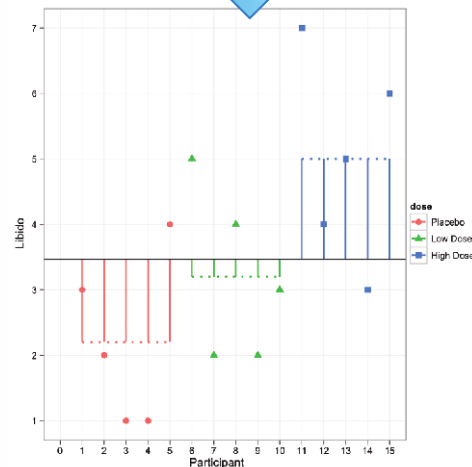
Graphical representation of the different sums of squares in ANOVA designs



$SS_T$  uses the differences between the observed data and the mean value of  $Y$



$SS_R$  uses the differences between the observed data and the model (group means)



$SS_M$  uses the differences between the mean value of  $Y$  and the model (group means)



# ANOVA by hand (3)

## Step 1:

TABLE 11.1 Data in **Viagra.sav**

	Placebo	Low Dose	High Dose
	3	5	7
	2	2	4
	1	4	5
	1	2	3
	4	3	6
$\bar{X}$	2.20	3.20	5.00
s	1.30	1.30	1.58
$s^2$	1.70	1.70	2.50
Grand mean = <b>3.467</b> Grand SD = <b>1.767</b>			
Grand variance = <b>3.124</b>			

## Step 2:

$$SS_M = \sum n_i (\bar{x}_i - \bar{x}_{grand})^2$$



$$\begin{aligned} SS_M &= 5(2.2 - 3.467)^2 + 5(3.2 - 3.467)^2 + 5(5.0 - 3.467)^2 \\ &= 5(-1.267)^2 + 5(-0.267)^2 + 5(1.533)^2 \\ &= 8.025 + 0.355 + 11.755 \\ &= 20.135 \end{aligned}$$

## Step 3:

$$\begin{aligned} SS_R &= s_{group1}^2(n_1 - 1) + s_{group2}^2(n_2 - 1) + s_{group3}^2(n_3 - 1) \\ &= (1.70)(5 - 1) + (1.70)(5 - 1) + (2.50)(5 - 1) \\ &= (1.70 \times 4) + (1.70 \times 4) + (2.50 \times 4) \\ &= 6.8 + 6.8 + 10 \\ &= 23.60 \end{aligned}$$

## Step 4: Double check

$$\begin{aligned} SS_T &= 3.124(15 - 1) \\ &= 43.74 \end{aligned}$$

$$\begin{aligned} SS_T &= SS_M + SS_R \\ 43.74 &= 20.14 + 23.60 \\ 43.74 &= 43.74 \end{aligned}$$

# ANOVA by hand (4)

Step 5: Calculate mean squared errors

$$MS^M = \frac{q\epsilon^M}{22^M} = \frac{5}{50.132} = 10.067$$

$$MS^B = \frac{q\epsilon^B}{22^B} = \frac{15}{53.60} = 1.967$$

Step 6:

$$F = \frac{MS_M}{MS_R} = \frac{10.067}{1.967} = 5.12$$

Step 7: Critical value of F(2,12): 3.89

5.12 > 3.89 – therefore  $H_0$  can be rejected

Summary table:

Source	SS	df	MS	F
Model	20.14	2	10.067	5.12*
Residual	23.60	12	1.967	
Total	43.74	14		

# Follow-up tests

- The F-ratio does not tell us specifically which group means differ from which
  - We need additional tests to find out where the group differences lie
- > **Post Hoc tests**
- Compare each mean against all others.
  - Multiple tests
  - To control the family-wise error rate, stricter criteria to accept an effect as significant must be used
  - Simplest example is the Bonferroni method

$$\text{Bonferroni } \alpha = \frac{\alpha}{\text{Number of Tests}}$$

# Post-hoc tests

- SPSS has 18 types of Post Hoc tests!
  - Bonferroni (conservative option)
  - Bonferroni-Holm
  - Holm-Sidak
  - Tukey HSD
  - Scheffe
  - LSD
  - Dunnett

**Multiple Comparisons**

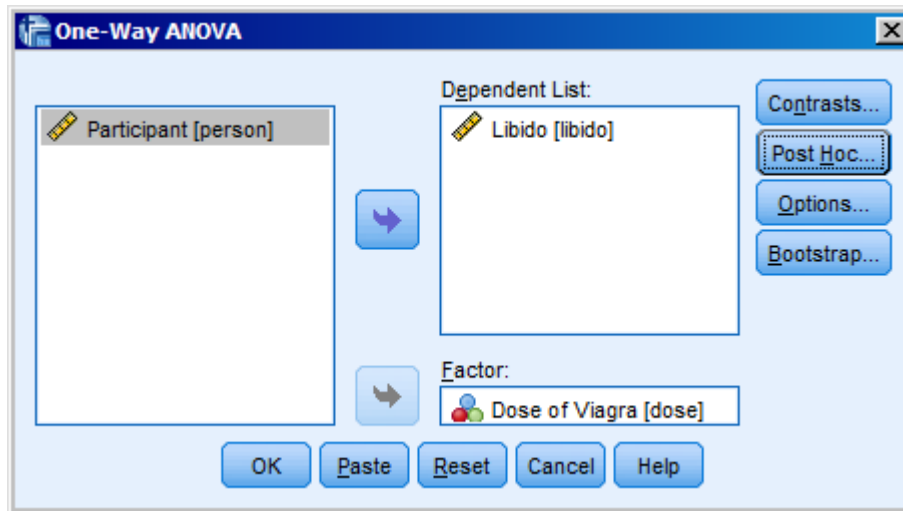
Dependent Variable: Libido

	(I) Dose of Viagra	(J) Dose of Viagra	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	Placebo	Low Dose	-1.000	.887	.516	-3.37	1.37
		High Dose	-2.800 <sup>a</sup>	.887	.021	-5.17	-.43
	Low Dose	Placebo	1.000	.887	.516	-1.37	3.37
		High Dose	-1.800	.887	.147	-4.17	.57
	High Dose	Placebo	2.800 <sup>a</sup>	.887	.021	.43	5.17
		Low Dose	1.800	.887	.147	-.57	4.17
Games-Howell	Placebo	Low Dose	-1.000	.825	.479	-3.36	1.36
		High Dose	-2.800 <sup>a</sup>	.917	.039	-5.44	-.16
	Low Dose	Placebo	1.000	.825	.479	-1.36	3.36
		High Dose	-1.800	.917	.185	-4.44	.84
	High Dose	Placebo	2.800 <sup>a</sup>	.917	.039	.16	5.44
		Low Dose	1.800	.917	.185	-.84	4.44
Dunnett t (-control)	Low Dose	Placebo	1.000	.887	.227	-.87	
	High Dose	Placebo	2.800 <sup>a</sup>	.887	.008	-.93	

<sup>a</sup>. The mean difference is significant at the 0.05 level.

a. Dunnett t-tests treat one group as a control, and compare all other groups against it.

# ANOVA in SPSS



## ANOVA

Libido

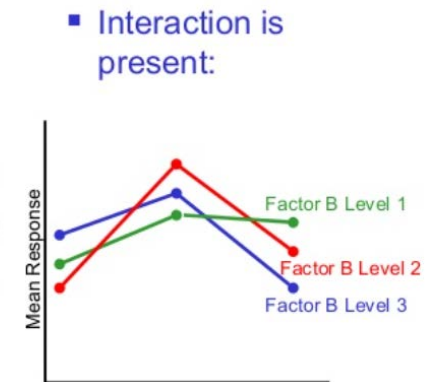
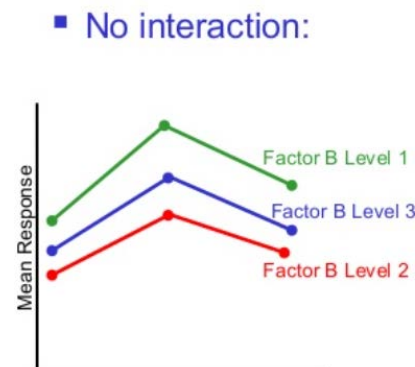
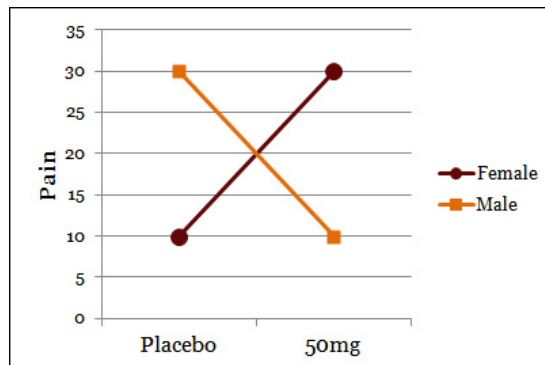
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	20.133	2	10.067	5.119	.025
Within Groups	23.600	12	1.967		
Total	43.733	14			

# ANOVA - Variants

- Till now: only one factor (group variable)
  - Simple ANOVA / **One-way ANOVA**
- Extension: **several factors** each with  $\geq 2$  levels
  - **Two-way/Three-way** etc. **ANOVA**
  - Several Independent Variables is known as a factorial design
- **ANCOVA**: Analysis of Covariance, also continuous/metric independent variables are allowed
- **Multivariate ANOVA**: several dependent variable
- **ANOVA for repeated measurements**: subsequent observations are dependent

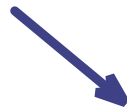
# Two-way (independent) ANOVA

- Two Independent Variables
- Several independent variables is known as a **factorial design**
- **Benefit** of factorial designs
  - We can look at how variables *interact*
- **Interactions**
  - Show how the effects of one IV might depend on the effects of another
  - Are often more interesting than main effects.
- Examples:



# Variance decomposition

---



nd  $B$



# Two-way ANOVA – example (1)

- Field (2013): Testing the effects of Alcohol and Gender on ‘the beer-goggles effect’:
  - IV 1 (Alcohol): None, 2 pints, 4 pints
  - IV 2 (Gender): Male, Female
- Dependent Variable (DV) was an objective measure of the attractiveness of the partner selected at the end of the evening.

TABLE 13.1 Data for the beer-goggles effect

Alcohol	None		2 Pints		4 Pints	
	Female	Male	Female	Male	Female	Male
Gender	65	50	70	45	55	30
	70	55	65	60	65	30
	60	80	60	85	70	30
	60	65	70	65	55	55
	60	70	65	70	55	35
	55	75	60	70	60	20
	60	75	60	80	50	45
	55	65	50	60	50	40
Total	485	535	500	535	460	285
Mean	60.625	66.875	62.50	66.875	57.50	35.625
Variance	24.55	106.70	42.86	156.70	50.00	117.41

Grand mean: 58.33

# Two-way ANOVA – example (2)

- Step 1: Calculate  $SS_T$

$$\begin{aligned}
 SS_T &= s_{\text{grand}}^2 (N - 1) \\
 &= 190.78 (48 - 1) \\
 &= 8966.66
 \end{aligned}$$

- Step 2: Calculate  $SS_M$

$$SS_M = \sum n_i (\bar{x}_i - \bar{x}_{\text{grand}})^2$$

$$\begin{aligned}
 SS_M &= 8(60.625 - 58.33)^2 + 8(66.875 - 58.33)^2 + 8(62.5 - 58.33)^2 \\
 &\quad + 8(66.875 - 58.33)^2 + 8(57.5 - 58.33)^2 + 8(35.625 - 58.33)^2 \\
 &= 8(2.295)^2 + 8(8.545)^2 + 8(4.17)^2 + 8(8.545)^2 + 8(-0.83)^2 + 8(-22.705)^2 \\
 &= 42.1362 + 584.1362 + 139.1112 + 584.1362 + 5.5112 + 4124.1362 \\
 &= 5479.167
 \end{aligned}$$

TABLE 13.1 Data for the beer-goggles effect

Alcohol	None		2 Pints		4 Pints	
	Female	Male	Female	Male	Female	Male
	65	50	70	45	55	30
	70	55	65	60	65	30
	60	80	60	85	70	30
	60	65	70	65	55	55
	60	70	65	70	55	35
	55	75	60	70	60	20
	60	75	60	80	50	45
	55	65	50	60	50	40
Total	485	535	500	535	460	285
Mean	60.625	66.875	62.50	66.875	57.50	35.625
Variance	24.55	106.70	42.86	156.70	50.00	117.41

Grand mean: 58.33

# Two-way ANOVA – example (3)

- Step 2a: Calculate  $SS_A$

$$\begin{aligned}SS_{\text{Gender}} &= 24(60.21 - 58.33)^2 + 24(56.46 - 58.33)^2 \\ &= 24(1.88)^2 + 24(-1.87)^2 \\ &= 84.8256 + 83.9256 \\ &= 168.75\end{aligned}$$

A <sub>1</sub> : Female		
65	70	55
70	65	65
60	60	70
60	70	55
60	65	55
55	60	60
60	60	50
55	50	50

Mean Female = 60.21

A <sub>2</sub> : Male		
50	45	30
55	60	30
80	85	30
65	65	55
70	70	35
75	70	20
75	80	45
65	60	40

Mean Male = 56.46

# Two-way ANOVA – example (4)

- Step 2b: Calculate  $SS_B$

$$\begin{aligned}
 SS_{\text{alcohol}} &= 16(63.75 - 58.33)^2 + 16(64.6875 - 58.33)^2 + 16(46.5625 - 58.33)^2 \\
 &= 16(5.42)^2 + 16(6.3575)^2 + 16(-11.7675)^2 \\
 &= 470.0224 + 646.6849 + 2215.5849 \\
 &= 3332.292
 \end{aligned}$$

- Step 2c: Calculate  $SS_{A \times B}$

$$SS_{A \times B} = SS_M - SS_A - SS_B$$

$$\begin{aligned}
 SS_{A \times B} &= SS_M - SS_A - SS_B \\
 &= 5479.167 - 168.75 - 3332.292 \\
 &= 1978.125
 \end{aligned}$$

$B_1$ : None	
65	50
70	55
60	80
60	65
60	70
55	75
60	75
55	65

Mean None  
= 63.75

$B_2$ : 2 Pints	
70	45
65	60
60	85
70	65
65	70
60	70
60	80
50	60

Mean 2  
Pints =  
64.6875

$B_3$ : 4 Pints	
55	30
65	30
70	30
55	55
55	35
60	20
50	45
50	40

Mean 4  
Pints =  
46.5625

# Two-way ANOVA – example (5)

- Step 3: Calculate  $SS_R$

$$SS_R = s_{\text{group1}}^2 (n_1 - 1) + s_{\text{group2}}^2 (n_2 - 1) + s_{\text{group3}}^2 (n_3 - 1) \dots s_{\text{group n}}^2 (n_n - 1)$$

$$\begin{aligned} SS_R &= s_{\text{group1}}^2 (n_1 - 1) + s_{\text{group2}}^2 (n_2 - 1) + s_{\text{group3}}^2 (n_3 - 1) \\ &\quad + s_{\text{group4}}^2 (n_4 - 1) + s_{\text{group5}}^2 (n_5 - 1) + s_{\text{group6}}^2 (n_6 - 1) \\ &= (24.55 \times 7) + (106.7 \times 7) + (42.86 \times 7) \\ &\quad + (156.7 \times 7) + (50 \times 7) + (117.41 \times 7) \\ &= 171.85 + 746.9 + 300 + 1096.9 + 350 + 821.87 \\ &= 3487.52 \end{aligned}$$

# Two-way ANOVA – example (6)

## Tests of Between-Subjects Effects

Dependent Variable: Attractiveness of Date

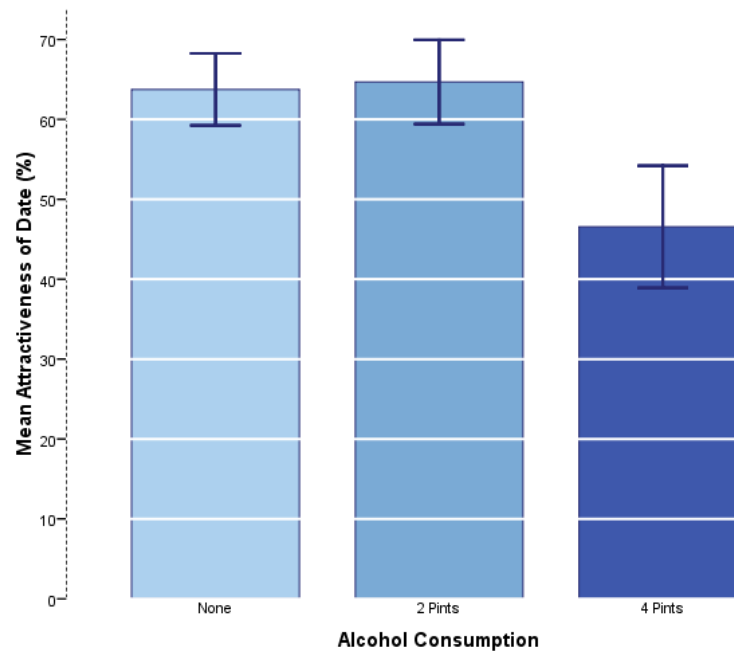
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Gender	168.750	1	168.750	2.032	.161
Alcohol	3332.292	2	1666.146	20.065	.000
Gender * Alcohol	1978.125	2	989.062	11.911	.000
Error	3487.500	42	83.036		

a. R Squared = .611 (Adjusted R Squared = .565)

F-statistic for each factor (gender, alcohol, gender\*alcohol):

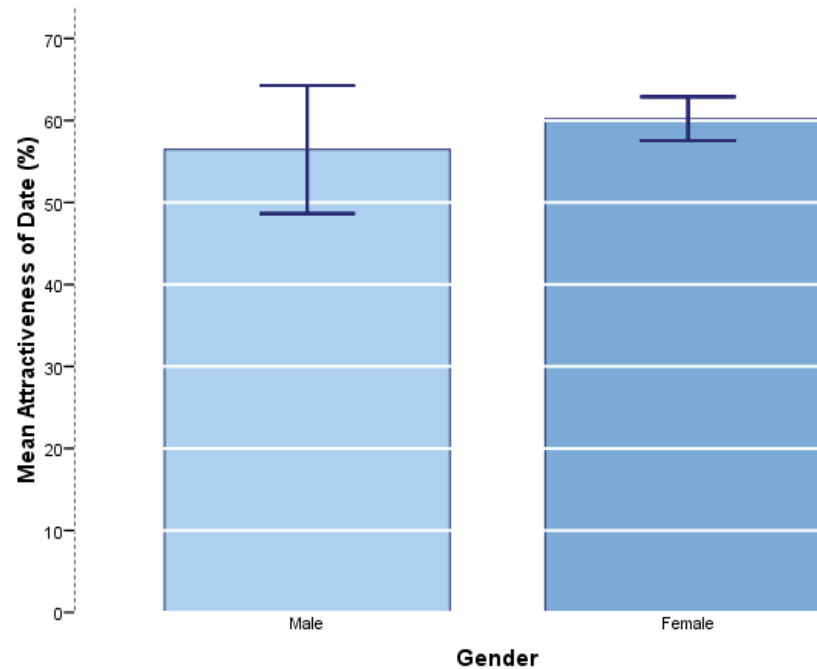
$$\frac{MS_{factor}}{MS_{error}} \quad \text{F-distributed}$$

# Interpretation: Main Effect Alcohol



There was a significant main effect of the amount of alcohol consumed at the night-club, on the attractiveness of the mate that was selected,  $F(2, 42) = 20.07, p < .001$ .

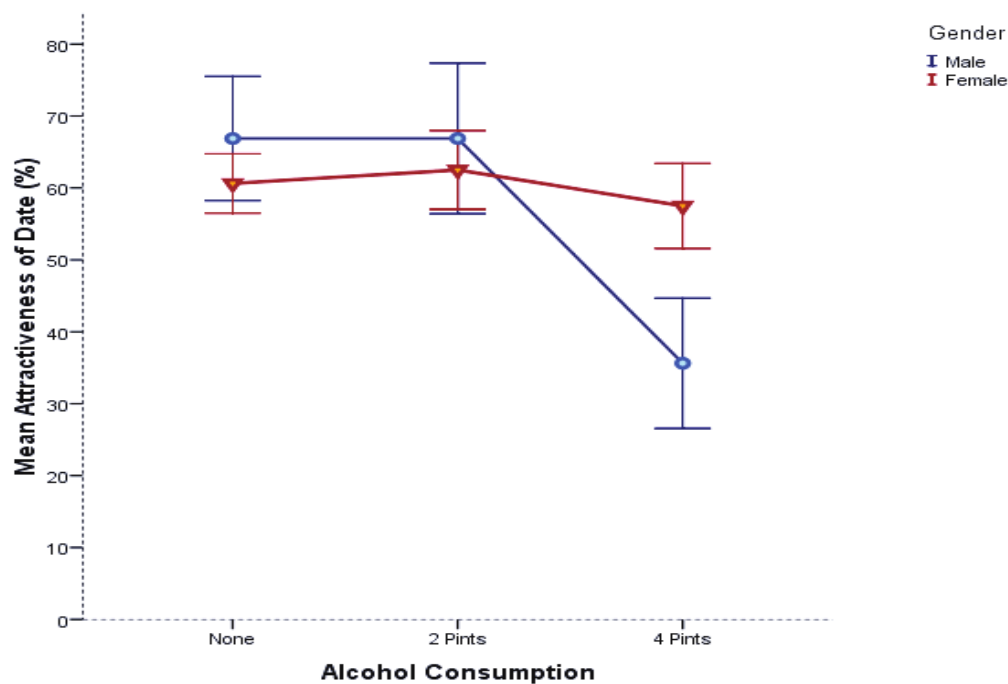
# Interpretation: Main Effect Gender



There was a no significant main effect of gender on the attractiveness of selected mates,  $F(1, 42) = 2.03, p = .161$ .



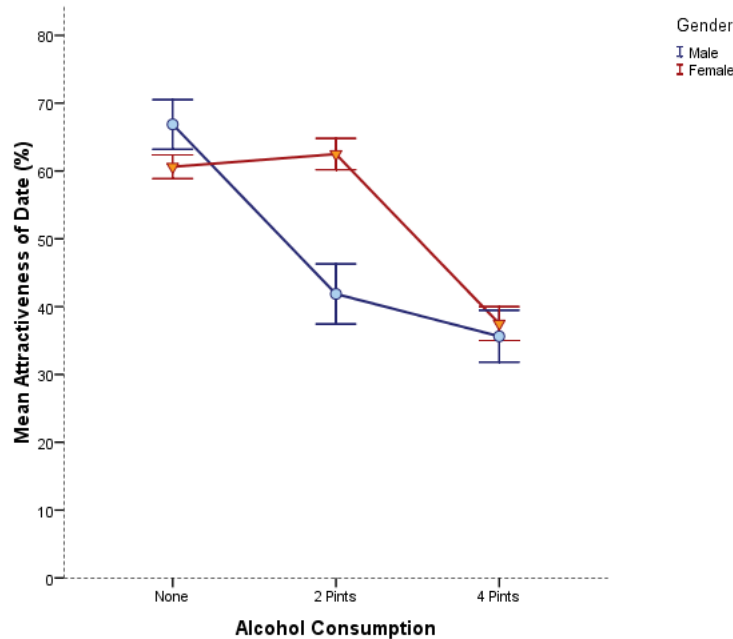
# Interpretation: Interaction



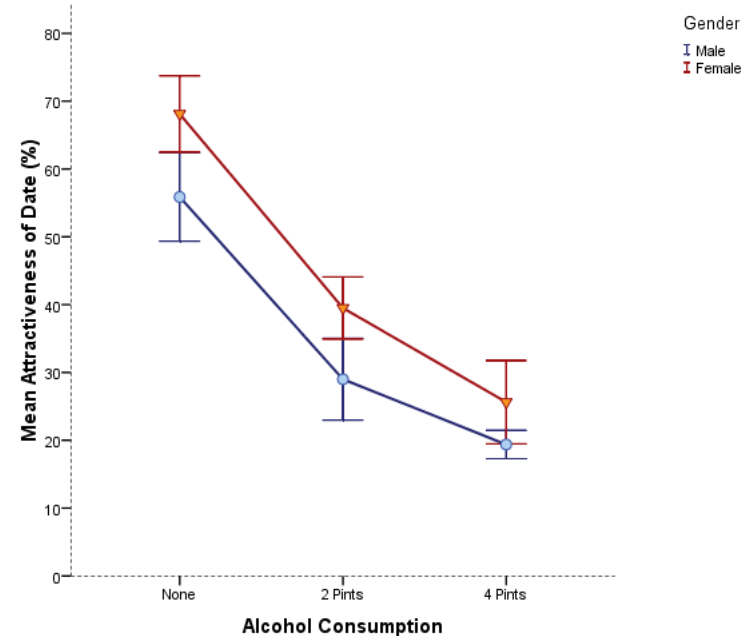
**FIGURE 13.13**  
Graph of the  
interaction  
of gender  
and alcohol  
consumption in  
mate selection

There was a significant interaction between the amount of alcohol consumed and the gender of the person selecting a mate, on the attractiveness of the partner selected,  $F(2, 42) = 11.91, p < .001$ .

# Is there likely to be a significant interaction effect?



Yes



No

- Je größer die Fallzahl desto geringer die Wahrscheinlichkeit für den Fehler 2. Art, desto höher die Präzision der Schätzung
- Fallzahlschätzung bedingt bereits die Definition des Testproblems durch Operationalisierung der Fragestellung, durch Formulierung von Null- und Alternativhypothese
- Fallzahlschätzung bedingt das Festlegen einer speziellen Alternative. „Es gibt keinen Unterschied zwischen den Gruppen“ genügt nicht. Die Größe des zu erwartenden Unterschieds muss festgelegt werden.
- Fallzahlschätzung bedingt bereits die Auswahl des statistischen Tests, die Definition von Fehler 1. und 2. Art



# Sample size estimation

**Question:** How many individuals do you have to include in your study to get a reliable result ?

→ We want to **maximize the probability**  
for rejecting  $H_0$ , if  $H_1$  is true

→ while keeping the **Type I error  $\alpha$**  fixed

**What do you have to know  
to calculate the sample size  
needed?**

1. Power (typically set to 80% or 90%)
2. Type I error  $\alpha$  (typically set to  $\alpha = 0.05$ )
3. The difference you want to find (for t-tests:  
the mean difference between groups)
4. standard deviation / measure of variance



# Sample size estimation

---

## Example

- Hypothesis:  $H_0: \mu_A = \mu_B$  versus  $H_1: \mu_A \neq \mu_B$  → two-sided t-test
  - You consider a difference of 10 as relevant
  - From former studies, you know, that the standard deviation is  $\sim 15$  mmHG
  - So far, you have recruited 20 patients (10 in each treatment arm)
- What is your power?

## Fallzahlschätzung für unverbundene Stichproben und stetige Zielgrößen

- Fallzahlberechnung für vorgegebene Power
- Powerberechnung für vorgegebene Fallzahl
- Entdeckbare Differenz für vorgegebene Fallzahl und Power

Eingabe von  $\mu_1$ :  Eingabe von  $\mu_2$ :

Eingabe von  $\sigma$ :  Differenz Delta:

- Einseitiger Test
- Zweiseitiger Test

Eingabe von  $\alpha$  (Standard ist 0.05):

Eingabe der Power (Standard ist 0.80):

Die Fallzahl für jede Gruppe ist:

# Vergleich von Häufigkeiten zweier unverbundener Stichproben



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

*In einer Karzinomstudie erhofft man sich, die bisher mit der Standardtherapie (S) erreichte Rezidivrate von 50% mit einer neuen Therapie (T) auf 40% zu senken.*

*Für eine kontrollierte klinische Studie mit den beiden Therapiegruppen S und T und dem Zielkriterium 'Rezidivrate' sowie den weiteren Festlegungen (zweiseitiger Test, Irrtumswahrscheinlichkeit 5%, Power 80%) ergeben sich folgende Parameter für die Fallzahlberechnung:  
 $p_1 = 0.5$ ,  $p_2 = 0.4$ ,  $\alpha = 0.05$ ,  $1-\beta = 0.80$ .*

***ergibt eine Fallzahl von 388 Patienten für jede Behandlungsgruppe.***

# Vergleich von Mittelwerten zweier unverbundener Stichproben



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

*In einer Hypertoniestudie erhofft man sich, die bisher mit der Standardtherapie (S) erreichte durchschnittliche Blutdrucksenkung von 15 mm mit einer neuen Therapie (T) auf 20 mm zu senken.*

*Für eine kontrollierte klinische Studie mit den beiden Therapiegruppen S und T und dem Zielkriterium 'Senkung des Blutdrucks' sowie den weiteren Festlegungen (Standardabweichung = 15, zweiseitiger Test, Irrtumswahrscheinlichkeit 5%, Power 80%) ergeben sich folgende Parameter für die Fallzahlberechnung:  
 $\mu_1 = 15$ ,  $\mu_2 = 20$ ,  $\sigma = 15$ ,  $\alpha = 0.05$ ,  $1 - \beta = 0.80$*

***ergibt eine Fallzahl von 142 Patienten für jede Behandlungsgruppe.***



# Vergleich von Mittelwerten zweier verbundener Stichproben



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

*In einer Phase-II-Studie soll überprüft werden, ob sich ein neues Medikament zur Blutdrucksenkung eignet. Geeignet ist das Medikament dann, wenn bei Hypertonikern mit einem durchschnittlichen systolischen Blutdruck von 150 mm eine Senkung um mindestens 10 mm erreicht wird.*

*Für diese klinische Studie lautet das Zielkriterium 'Differenz des Blutdrucks vor und nach Behandlung'. Mit den weiteren Festlegungen (Standardabweichung = 15, zweiseitiger Test, Irrtumswahrscheinlichkeit 5%, Power 80%) ergeben sich folgende Parameter für die Fallzahlberechnung:*

$$\mu_1 = 150, \mu_2 = 140, \sigma = 15, \alpha = 0.05, 1 - \beta = 0.80$$

***ergibt eine Fallzahl von 20 Patienten.***

# Vergleich von Überlebenszeiten zweier unverbundener Stichproben



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

*In einer Karzinomstudie erhofft man sich, die bisher mit der Standardtherapie (S) erreichte mediane Überlebenszeit von 36 Monaten mit einer neuen Therapie (T) auf 48 Monate zu erhöhen.*

*Für eine kontrollierte klinische Studie mit den beiden Therapiegruppen S und T und dem Zielkriterium 'Überlebenszeit' sowie den weiteren Festlegungen (Rekrutierungszeit=24 Monate, Nachbeobachtungszeit = 36 Monate, zweiseitiger Test, Irrtumswahrscheinlichkeit 5%, Power 80%)*

***ergibt eine Fallzahl von 349 Patienten für jede Behandlungsgruppe.***

# Checklist für Fallzahlschätzung (Testproblem)



- Studiendesign mit
- Auswahl des Hauptzielkriteriums
- Zu erwartender Unterschied und Angabe eines Variationsmaßes
- Begründung dafür – Literatur oder Vorstudie
- Fehler 1. Art (üblicherweise 0,05)
- Fehler 2. Art (0,1 oder 0,2)
- Auswahl des statistischen Tests
- Falls mehrere Hypothesen formuliert werden, Korrektur des Fehler 1. Art oder Hierarchisierung der Hypothesen
- Drop-Out Rate berücksichtigen

# Übung: Berechnen Sie die statistische Power



**Table 2. Effect of Cytisine on Smoking Cessation.\***

Outcome	Cytisine (N= 370)	Placebo (N= 370)	Percentage-Point Difference (95% CI)	Relative Rate (95% CI)†
	<i>percent (number)</i>			
Primary outcome: abstinence for 12 mo	8.4 (31)	2.4 (9)	6.0 (2.7–9.2)‡	3.4 (1.7–7.1)
Abstinence for 6 mo	10.0 (37)	3.5 (13)	6.5 (2.9–10.1)‡	2.9 (1.5–5.3)
Point prevalence at 12 mo	13.2 (49)	7.3 (27)	5.9 (1.6–10.3)§	1.8 (1.2–2.8)

# Kategoriale Daten und 4-Felder Tafeln



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

---

Dr. Hanno Ulmer

*hanno.ulmer@imed.ac.at*

*Innsbruck, Oktober 2010*

---

Department für Medizinische Statistik, Informatik und  
Gesundheitsökonomie, Medizinische Universität Innsbruck

# Zusammenhänge qualitativ: Vierfelder Tafel



**Tab.1** Vierfeldertafel zur Untersuchung des Effekts einer intensivierten Insulintherapie auf die Entwicklung einer Neuropathie in 5 Jahren bei 622 Diabetikern (7)

## Interventionsstudie

		Neuropathie		Summe
		ja	nein	
Gruppe	Kontrolle	52	255	307
	Intervention	21	294	315
Summe		73	549	622

**Tab.2** Vierfeldertafel zur Untersuchung des Hämoocult-Test zum Screening auf ein kolorektales Karzinom bei 7493 Personen (1)

## Diagnostische Studie

		kolorektales Karzinom		Summe
		ja	nein	
Hämoocult-Test	+	22	418	440
	-	10	7043	7053
Summe		32	7461	7493

# Therapiestudie, epidemiologische Studie



	Outcome positiv	Outcome negativ	
Exposition positiv	a	b	Rel. Risiko = $= \frac{a}{a+b} / \frac{c}{c+d}$
Exposition negativ	c	d	Odds Ratio $= \frac{a.d}{b.c}$
	Abs. Risiko Reduk. $= \frac{a}{a+b} - \frac{c}{c+d}$	Number needed to treat (NNT) $= 1/ARR$	

**Tab.1** Vierfeldertafel zur Untersuchung des Effekts einer intensivierten Insulintherapie auf die Entwicklung einer Neuropathie in 5 Jahren bei 622 Diabetikern (7)



		Neuropathie		Summe
		ja	nein	
Gruppe	Kontrolle	52	255	307
	Intervention	21	294	315
Summe		73	549	622

	Outcome positiv	Outcome negativ	
Exposition positiv	52 (16,9%)	255	RR = 2,5  $= \frac{a}{a+b} / \frac{c}{c+d}$
Exposition negativ	21 (6,7%)	294	OR = 2,9  $= \frac{a}{b} / \frac{c}{d} = \frac{a.d}{b.c}$
	ARR = 10,2%  $= \frac{a}{a+b} - \frac{c}{c+d}$	NNT = 9,8  NNT = 1/ARR	



# Relative Risiken

- **Relatives Risiko (RR):** Das Verhältnis der Risiken für ein bestimmtes Ereignis (z.B. Neuropathie) in zwei Vergleichsgruppen
- **Odds ratio (OR):** «Odds» entspricht der «Chance», dass ein bestimmtes Ereignis eintritt, bezeichnet also etwas Ähnliches wie ein Risiko, wird aber (wie beim Pferderennen) als Verhältnis (z.B. 1:9) angegeben.
- **Hazard Ratio (HR):** «Hazard» bezeichnet die Wahrscheinlichkeit, dass ein bestimmtes Ereignis eintritt. Dabei wird der Zeitpunkt berücksichtigt, wann das Ereignis eintritt. Im Gegensatz zum relativen Risiko wird also mit einer «hazard ratio» nicht nur ein Ausbleiben, sondern auch ein späteres Eintreffen eines Ereignisses als Effekt erfasst.



# Logistische Regressionsanalyse

---

- Untersuchung / Modellierung von prognostischen Faktoren für binäre Outcomes bei gleichzeitiger Adjustierung für Störfaktoren („confounders“)
- Multivariate Erweiterung zur Kontingenztafelanalyse und Chi-Quadrat Test
- Berechnet wird das adjustierte Relative Risiko oder das adjustierte Odds Ratio

**kurzgefasst:** Mit Hilfe der multiplen logistischen Regression lässt sich der Einfluss erklärender Variablen (Risikofaktoren) auf eine binäre Zielvariable (z.B. Krankheit ja/nein) untersuchen. Aus den Regressionskoeffizienten lassen sich adjustierte Odds Ratios als Maß für die Stärke des Zusammenhangs berechnen.

Tab.2 Multiple logistische Regressionsanalyse für die Entwicklung einer diabetischen Nephropathie nach 6 Jahren bei 480 Typ 1 Diabetikern.

Risikofaktor	Regressionskoeffizient	Standardfehler	p-Wert	Differenz für Odds Ratio	Odds Ratio	95% Konfidenzintervall
Achsenabschnitt	- 8,980	1,736	0,0001			
HbA <sub>1c</sub>	+0,464	0,091	0,0001	1%	1,59	1,33 – 1,90
diast. Blutdruck	+0,048	0,019	0,0148	5 mm Hg	1,27	1,05 – 1,54
Diabetesdauer	+0,004	0,018	0,8220	5 Jahre	1,02	0,85 – 1,22
Geschlecht	- 0,025	0,249	0,9212	männl. vs. weibl.	0,98	0,60 – 1,59

# Diagnostische Studie

	Goldstandard positiv	Goldstandard negativ	
Neues Verfahren positiv	a	b	Pos. Präd. W. $= \frac{a}{a+b} * 100$
Neues Verfahren negativ	c	d	Neg. Präd. W. $= \frac{c}{c+d} * 100$
	Sensitivität $= \frac{a}{a+c} * 100$	Spezifität $= \frac{d}{b+d} * 100$	

**Tab. 2** Vierfeldertafel zur Untersuchung des Hämocult-Test zum Screening auf ein kolorektales Karzinom bei 7493 Personen (1)

		kolorektales Karzinom		
		ja	nein	Summe
Hämocult-Test	+	22	418	440
	-	10	7043	7053
Summe		32	7461	7493



	Goldstandard positiv	Goldstandard negativ	
Neues Verfahren positiv	22	418	PPV = 5% $= \frac{a}{a+b} * 100$
Neues Verfahren negativ	10	7043	NPV = 99,9% $= \frac{c}{c+d} * 100$
	Sensitivität 68,8% $= \frac{a}{a+c} * 100$	Spezifität 94,4% $= \frac{d}{b+d} * 100$	

# Formel von Bayes



Thomas Bayes ~1702 - 1761

$$\text{PPV} = \frac{(\text{Sensitivität} \times \text{Prävalenz})}{(\text{Sensitivität} \times \text{Prävalenz} + (1 - \text{Spezifität}) \times (1 - \text{Prävalenz}))}$$

## Beispiel Mammografie:

Prävalenz: 1%, Sensitivität: 90%, Spezifität: 98%

$$\text{ppV} = \frac{0.90 \cdot 0.01}{0.90 \cdot 0.01 + 0.02 \cdot 0.99} = 0.31$$

# Zur Diskussion

## Brustkrebs-Früherkennung



durch Mammographie-Screening

Zahlen für Frauen ab 50 Jahre, die 10 Jahre oder länger am Screening teilgenommen haben

Nutzen	1000 Frauen ohne Screening	1000 Frauen mit Screening
Wie viele Frauen sind an Brustkrebs gestorben?	5	4
Wie viele sind insgesamt an Krebs gestorben?	21	21
Schaden		
Wie viele Frauen ohne Krebs wurden durch Fehldiagnosen falsch alarmiert oder hatten eine Biopsie?	–	100
Wie viele gesunde Frauen wurden fälschlicherweise mit Brustkrebs diagnostiziert und behandelt?	–	5

Alle Daten aus Gatzsche, PC, Jørgensen, KJ (2013). *Cochrane Database of Systematic Reviews* (6): CD001877. Die Zahlen in der Faktenbox sind gerundet. Wo keine Zahlen für Frauen ab 50 Jahre verfügbar sind, beziehen sie sich auf Frauen ab 40 Jahre. [www.harding-center.mpg.de](http://www.harding-center.mpg.de)

## Prostatakrebs-Früherkennung



durch PSA-Test und Tastuntersuchung der Prostata

Zahlen für Männer ab 50 Jahre, Vergleich Nichtteilnahme mit 11-jähriger Teilnahme

Nutzen	1000 Männer ohne Früherkennung	1000 Männer mit Früherkennung
Wie viele Männer sind an Prostatakrebs gestorben?	7	7*
Wie viele Männer sind insgesamt gestorben?	210	210
Schaden		
Wie viele Männer haben nach einer Biopsie erfahren, dass ihr Testergebnis ein Fehlarbeit war?	–	160
Wie viele gesunde Männer wurden fälschlicherweise mit Prostatakrebs diagnostiziert und behandelt**?	–	20

\* Das bedeutet: Von 1000 Männern (Alter: 50+) mit Früherkennung sind innerhalb von 11 Jahren etwa 7 an Prostatakrebs gestorben.

\*\* Z.B. operative Entfernung der Prostata oder Strahlentherapie, was zu Inkontinenz oder Impotenz führen kann.

Quelle: Ilic et al. (2013) *Cochrane Database of Systematic Reviews*, Art. No.:CD004720.

## Gebärmutterhalskrebs-Früherkennung



durch den Pap-Test (auch „Abstrich“ genannt) für Frauen ab 20 Jahre.

Alle Angaben beziehen sich auf den Nutzen und Schaden pro Jahr.

Nutzen	100.000 ohne Screening	100.000 mit Screening
Wie viele Frauen erkrankten an Gebärmutterhalskrebs?	40	15
Wie viele Frauen verstarben an Gebärmutterhalskrebs?	6	3
Wie viele Frauen verstarben insgesamt an Krebs?	230	230
Schaden		
Wie viele gesunde Frauen wurden durch das Screening fälschlicherweise mit Gebärmutterhalskrebs diagnostiziert*?	–	5000

\* Dies führt zu Testwiederholungen, Biopsien, Operation/Konisation (Herausschneiden eines Kegels am Gebärmutterhals, was später zu Schwangerschaftskomplikationen führen kann), psychische Belastungen wie Angst.

Quellen: Gesundheitsberichterstattung 2009 des Statistischen Bundesamtes. Siebert, Muth, Sroczynski et al. (2003) [http://portal.dimdi.de/de/hta/hta\\_berichte/hta067\\_bericht\\_de.pdf](http://portal.dimdi.de/de/hta/hta_berichte/hta067_bericht_de.pdf). Siebert, Sroczynski, Hillmanns et al. (2006) *Eur J Public Health* 16.

Es liegen keine randomisiert-kontrollierten Studien vor. Die Zahlen stammen aus Bevölkerungsstatistiken, die seit Einführung des Pap-Tests erhebt wurden.

## Gebärmutterhalskrebs-Prävention



durch die HPV-Impfung mit Gardasil. Zahlen für Frauen von 12-25 Jahren, die noch keinen sexuellen Kontakt hatten. Die Angaben beziehen sich auf ein Jahr.

Alle Angaben beziehen sich auf den Nutzen und Schaden pro Jahr.

Nutzen für Frauen, die neben der HPV-Impfung auch am Pap-Test teilnehmen	100.000 ohne Impfung	100.000 mit Impfung
Wie viele Frauen erkrankten an Gebärmutterhalskrebs?	15	11
Wie viele Frauen verstarben an Gebärmutterhalskrebs?	3	2
Nebenwirkungen		
Bei wie vielen Frauen traten Fieber und Empfindlichkeit der Injektionsstelle auf?	–	1.000 - 10.000
Bei wie vielen Frauen traten unspezifische Gelenkentzündungen oder Nesselsucht auf?	–	10 - 1.000
Bei wie vielen Frauen trat eine Verengung der Atemwege mit schwerer Luftnot auf?	–	1 - 10

Quellen: Gesundheitsberichterstattung 2009 des Statistischen Bundesamtes. FUTURE II Study Group (2007) *N Engl J Med* 356. European Medicines Agency (2008) 31/10/2008 Gardasil, H.C. 203-II-13. Center for Disease Control and Prevention (CDC) (2008)

# Übung: Berechnen Sie das relative und absolute Risiko



**Table 2. Effect of Cytisine on Smoking Cessation.\***

Outcome	Cytisine (N= 370)	Placebo (N= 370)	Percentage-Point Difference (95% CI)	Relative Rate (95% CI)†
	<i>percent (number)</i>			
Primary outcome: abstinence for 12 mo	8.4 (31)	2.4 (9)	6.0 (2.7–9.2)‡	3.4 (1.7–7.1)
Abstinence for 6 mo	10.0 (37)	3.5 (13)	6.5 (2.9–10.1)‡	2.9 (1.5–5.3)
Point prevalence at 12 mo	13.2 (49)	7.3 (27)	5.9 (1.6–10.3)§	1.8 (1.2–2.8)



# Überlebenszeitanalyse (Survival Analysis)



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

e 36 Statistik | Statistics

## Überlebenszeitanalyse: Eigenschaften und Kaplan-Meier Methode

– Artikel Nr. 15 der Statistik-Serie in der DMW –

Survival analysis: Properties and Kaplan-Meier method

**Autoren** A. Ziegler<sup>1</sup> S. Lange<sup>2</sup> R. Bender<sup>2</sup>

**Institut** <sup>1</sup> Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Schleswig-Campus Lübeck, Universität zu Lübeck  
<sup>2</sup> Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln

Statistik | Statistics e 39

## Überlebenszeitanalyse: Der Log-Rang-Test

– Artikel Nr. 16 der Statistik-Serie in der DMW –

Survival analysis: Log rank test

**Autoren** A. Ziegler<sup>1</sup> S. Lange<sup>2</sup> R. Bender<sup>2</sup>

**Institut** <sup>1</sup> Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Universität zu Lübeck  
<sup>2</sup> Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln

e 42 Statistik | Statistics

## Überlebenszeitanalyse: Die Cox-Regression

– Artikel Nr. 17 der Statistik-Serie in der DMW –

Survival analysis: Cox regression

**Autoren** A. Ziegler<sup>1</sup> S. Lange<sup>2</sup> R. Bender<sup>2</sup>

**Institut** <sup>1</sup> Institut für Medizinische Biometrie und Statistik, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Universität zu Lübeck  
<sup>2</sup> Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen, Köln

**Literatur:**  
**Ziegler A., Lange S., Bender R.:**  
**Statistik-Supplement DMW**

# Überlebenszeitanalyse (Survival Analysis)

---

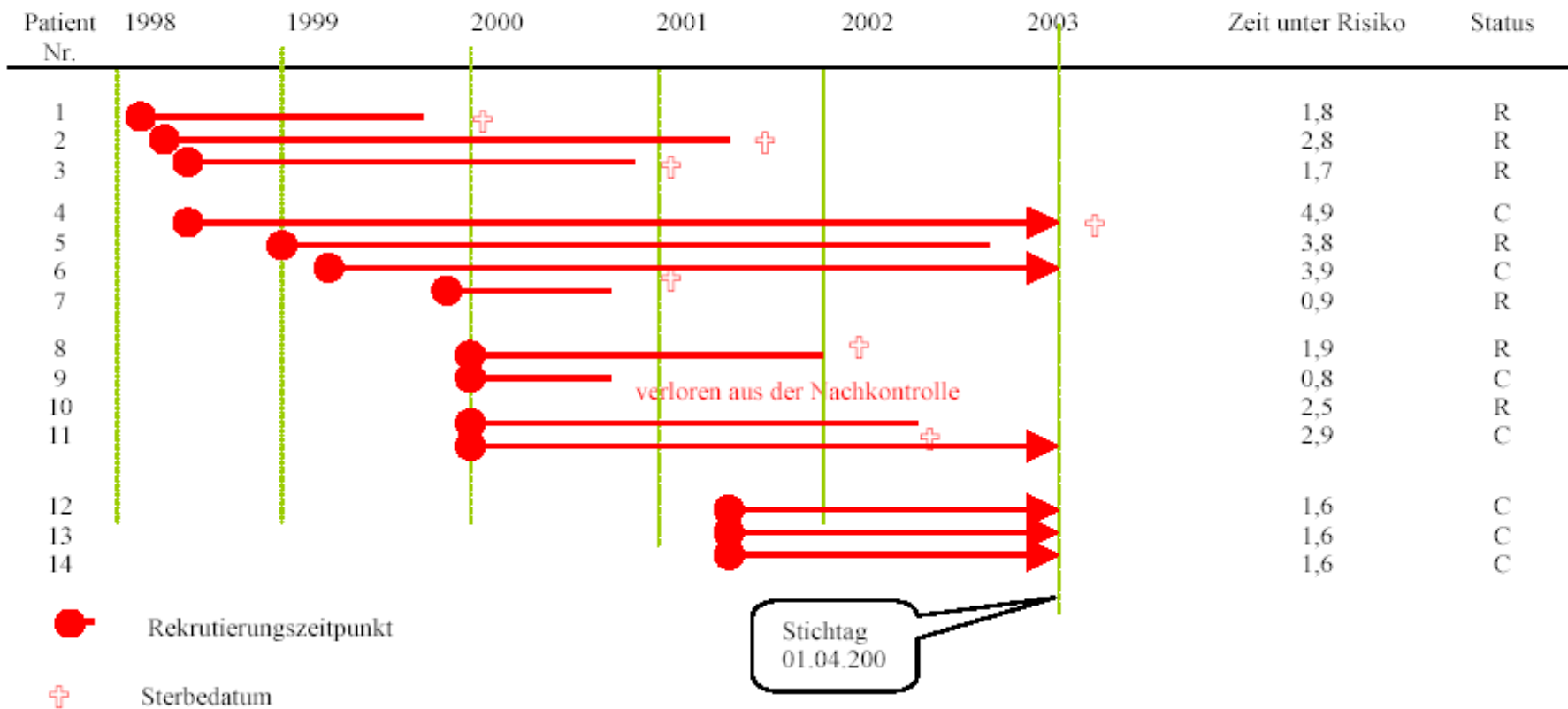


- Time-to-Event Analyse - Ereigniszeitanalyse
- Zwei Variablen:
  - Zeit
  - Ereignis: bereits eingetreten  
(noch) nicht eingetreten = zensiert

# Zensierte Daten

## Kaplan-Meier-Schätzung der Überlebensfunktion an einem Beispiel mit 14 Patienten

Lebenslinien der Patienten:



# Überlebenszeitanalyse (Survival Analysis)



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

- Kaplan-Meier Methode

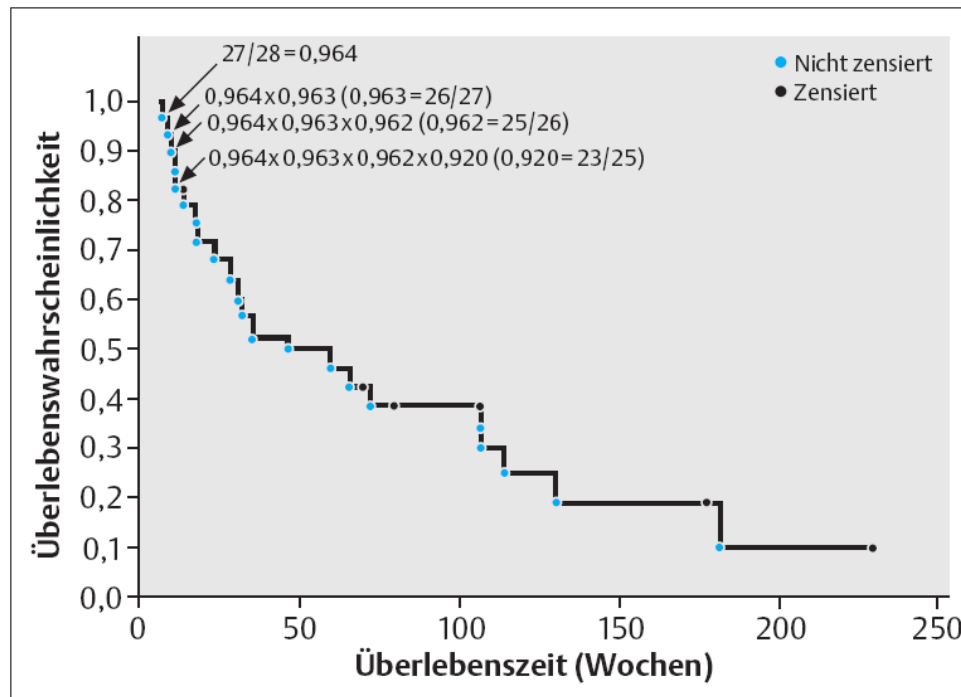
KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc. 53 457-481.

- Log-Rank Test

- Cox Proportional Hazards Modelle

Cox, D. R. (1972). Regression Models and Life Tables (with discussion). J. R. Statistic. Soc. 34: 187-220.

# Kaplan-Meier Methode



Tab. 1 Überlebenszeit (Wochen) von 28 Männern mit Zungenkrebs mit diploidem DNA-Tumorprofil – Daten aus Sickle-Santanello et al.

	Verstorben	Zensiert
1	18	69
3	23	104
4	26	104
5	27	112
5	30	129
8	42	181
12	56	
13	62	

Abb. 1 Kaplan-Meier Kurve für die Überlebenszeit der 28 Zungenkrebspatienten mit diploidem Tumor. Es wird die Wahrscheinlichkeit gezeigt, dass ein Patient eine Zeit (in Wochen) überlebt.

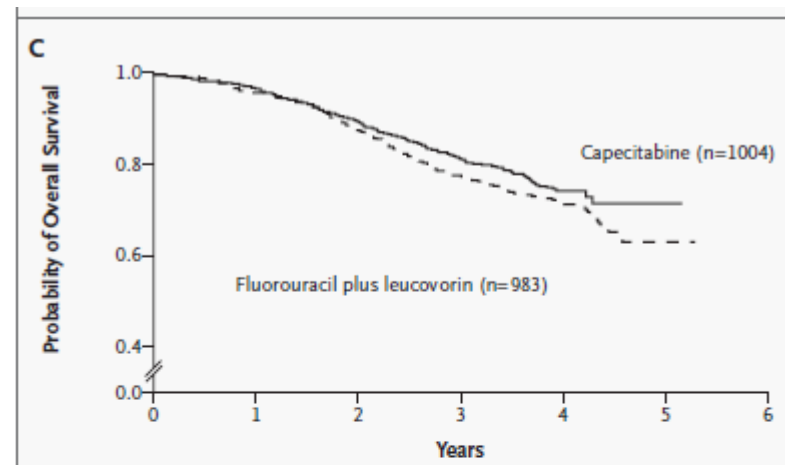
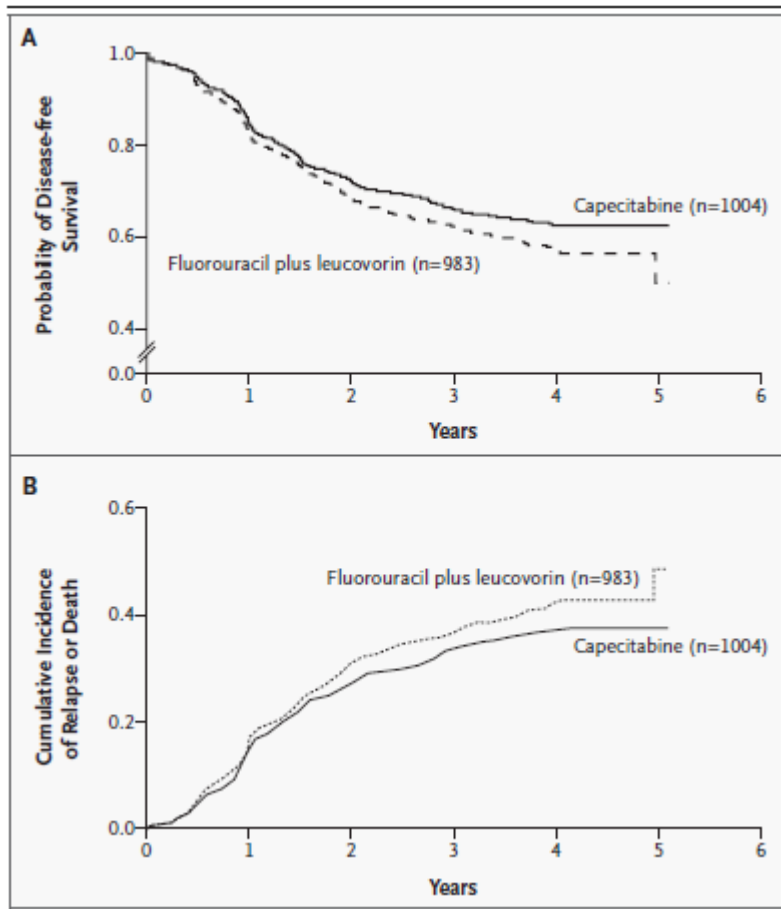


# Log-Rank Test

---

- **Vergleich von Überlebenszeiten**
  - $H_0$ : Es besteht eine Gleichverteilung der Überlebenszeiten
  - $H_1$ : Die Überlebenszeiten sind unterschiedlich verteilt
- **Einschränkungen**
  - Probleme bei Überschneidungen der Kurven
  - Alternative Peto-Test

# Disease-free/Overall Survival Incidence of Relapse/Death



**Figure 1. Disease-free Survival, Incidence of Relapse or Death, and Overall Survival among Patients Receiving Fluorouracil plus Leucovorin or Capecitabine (Intention-to-Treat Population).**

Panel A shows Kaplan–Meier estimates of disease-free survival. The upper limit of the confidence interval of the hazard ratio was significantly below both the predefined margins, 1.25 and 1.20, for equivalence ( $P < 0.001$  in both cases). The analysis for superiority showed a trend favoring capecitabine (hazard ratio, 0.87 [95 percent confidence interval, 0.75 to 1.00];  $P = 0.05$ ). Panel B shows the cumulative incidence of relapse or death; only deaths related to colon cancer or the study treatment were included. A Cox proportional-hazards model showed that relapse-free survival in the capecitabine group was statistically superior to that in the fluorouracil-plus-leucovorin group ( $P = 0.04$ ; hazard ratio, 0.86; 95 percent confidence interval, 0.74 to 0.99). Panel C shows Kaplan–Meier estimates of overall survival. The analysis for survival showed a trend favoring capecitabine (hazard ratio, 0.84 [95 percent confidence interval, 0.69 to 1.01];  $P = 0.07$ ).

# Cox Proportional Hazards Regressionsanalyse

---



- Untersuchung / Modellierung von prognostischen Faktoren für Ereigniszeiten bei gleichzeitiger Adjustierung für Störfaktoren („Confounder“)
- Multivariate Erweiterung zu Kaplan-Meier und Log-Rank Test
- Berechnet wird die (adjustierte) Hazard Ratio als Maß für das relative Risiko



# COX Modell für Versuche bis zur Lebendgeburt

Variablen	Signifikanz	Hazard Ratio	(95% CI)
Alter 30 – 34,9 Jahre	0,008	0,82	(0,70 - 0,95) vs. <30J
Alter 35 - < 39,9 Jahre	<,001	0,58	(0,46 - 0,65)
Alter 40 +	<,001	0,15	(0,11 - 0,22)
Blastozystentransfer	<0,01	2,13	(1,90 - 2,40)
Endometriose	0,51	0,95	(0,80 - 1,11)
PCO	0,24	0,92	(0,80 - 1,06)

Mit Hilfe des Cox-Modells läßt sich der Einfluß von erklärenden Variablen auf eine Ereigniszeit untersuchen. Aus den Regressionskoeffizienten lassen sich adjustierte Hazard Ratios für die Stärke des Zusammenhangs berechnen.

# Design of Experiments



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

---

Dr. Hanno Ulmer

*hanno.ulmer@imed.ac.at*

*Innsbruck, Oktober 2010*

---

Department für Medizinische Statistik, Informatik und  
Gesundheitsökonomie, Medizinische Universität Innsbruck

# DOE Grundlagen

- DOE = Design of Experiments
- Idee:
  - Simultane Bewertung von mehreren Einflussfaktoren
  - Auswahl der Zielvariable
  - Planung des Experiments, Anzahl der Durchläufe
  - Verwendung von ANOVA/Regressionsanalyse
- Anwendungen: Naturwissenschaftliche Experimente, Marktforschung, Simulationsexperimente
- Software (MODDE, SAS JMP,...) erlaubt
  - Auswahl des experimentellen Designs
  - Analyse der Ergebnisse des Experiments

# DOE Grundlagen

- Input: Beeinflussbare Faktoren  $x_1, x_2, x_3$
- Output  $Y$ : Response, wird gemessen
- Noise: bekannte oder unbekannte Störungen
- Bestimmung der Koeffizienten  $c_1, c_2, c_3$  mittels Varianzanalyse (ANOVA)

$$\text{Modell } y = f(x_1, \dots, x_n) = c_1 \cdot x_1 + c_2 \cdot x_2 + \dots + c_{12} \cdot x_1 \cdot x_2 + \dots + c_{11} \cdot x_1^2$$



- The **agricultural** origins, 1918 – 1940s
  - R. A. Fisher & his co-workers
  - Profound impact on agricultural science
  - Factorial designs, ANOVA
- The **first industrial** era, 1951 – late 1970s
  - Box & Wilson, response surfaces
  - Applications in the chemical & process industries
- The **second industrial** era, late 1970s – 1990
  - Quality improvement initiatives in many companies
  - Taguchi and robust parameter design, process robustness
- The **modern** era, beginning circa 1990
  - Wide use of computer technology in DOE
  - Expanded use of DOE in Six-Sigma and in business
  - Use of DOE in computer experiments

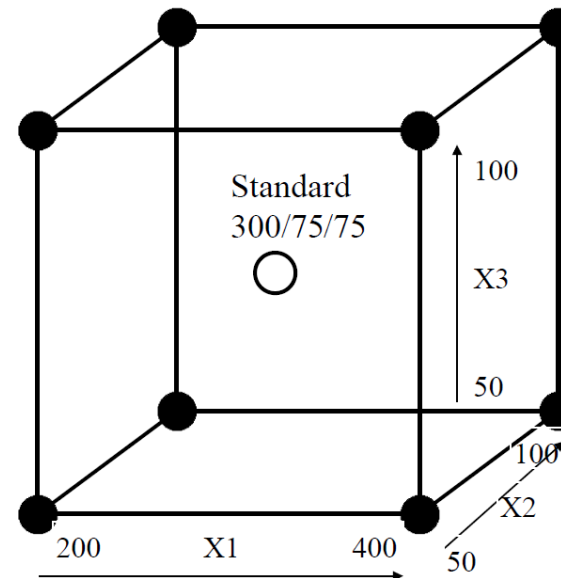
# CakeMix Beispiel mit MODDE

## Overview of DOE - CakeMix application

- Three factors varied: Flour (200-400g), Shortening (50-100g), and Eggpowder (50-100g)
- Response: Taste of resulting cake

Cake Mix Experimental Plan

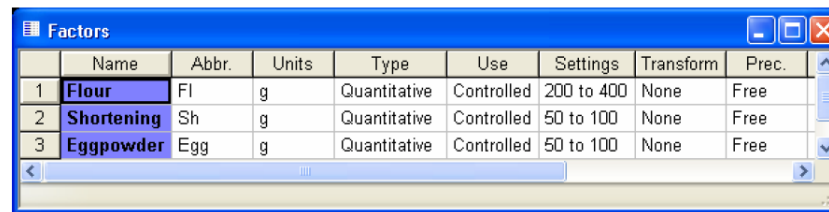
Cake No	Flour	Shortening	Egg Powder	Taste
1	200	50	50	3.52
2	400	50	50	3.66
3	200	100	50	4.74
4	400	100	50	5.20
5	200	50	100	5.38
6	400	50	100	5.90
7	200	100	100	4.36
8	400	100	100	4.86
9	300	75	75	4.73
10	300	75	75	4.61
11	300	75	75	4.68



# CakeMix Beispiel mit MODDE

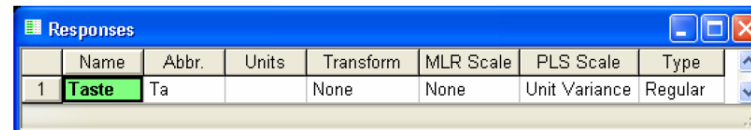
## Overview of steps in DOE - part I

### 1. Define Factors



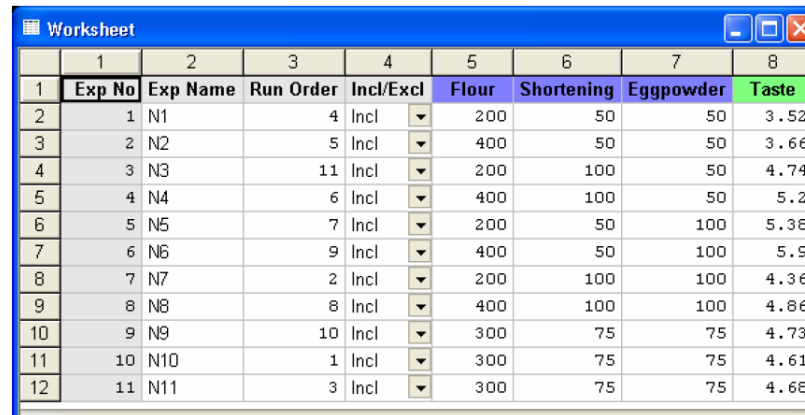
	Name	Abbr.	Units	Type	Use	Settings	Transform	Prec.
1	Flour	Fl	g	Quantitative	Controlled	200 to 400	None	Free
2	Shortening	Sh	g	Quantitative	Controlled	50 to 100	None	Free
3	Eggpowder	Egg	g	Quantitative	Controlled	50 to 100	None	Free

### 2. Define Response(s)



	Name	Abbr.	Units	Transform	MLR Scale	PLS Scale	Type
1	Taste	Ta		None	None	Unit Variance	Regular

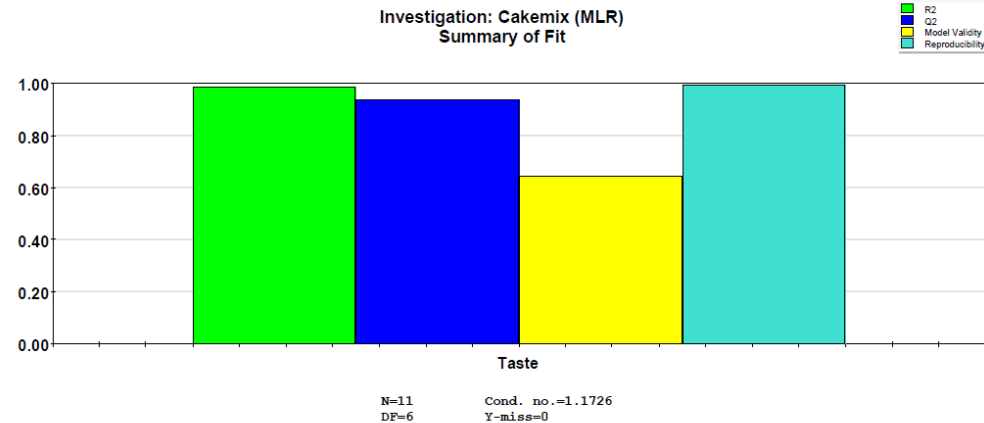
### 3. Create Design (Make experiments)



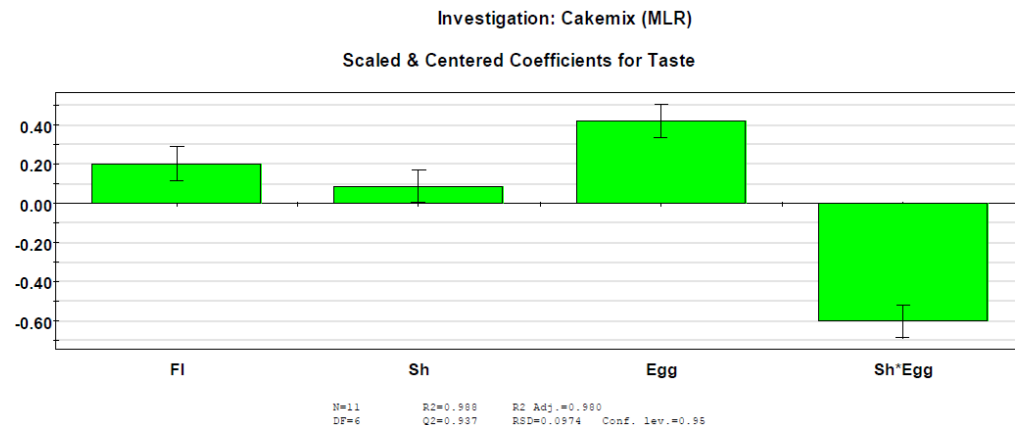
	1	2	3	4	5	6	7	8
1	Exp No	Exp Name	Run Order	Incl/Excl	Flour	Shortening	Eggpowder	Taste
2	1	N1	4	Incl	200	50	50	3.52
3	2	N2	5	Incl	400	50	50	3.66
4	3	N3	11	Incl	200	100	50	4.74
5	4	N4	6	Incl	400	100	50	5.2
6	5	N5	7	Incl	200	50	100	5.38
7	6	N6	9	Incl	400	50	100	5.9
8	7	N7	2	Incl	200	100	100	4.36
9	8	N8	8	Incl	400	100	100	4.86
10	9	N9	10	Incl	300	75	75	4.73
11	10	N10	1	Incl	300	75	75	4.61
12	11	N11	3	Incl	300	75	75	4.68

## Overview of steps in DOE - part II

### 4. Make Model



### 5. Interpret Model





- Full Factorial Design:  
Alle Level Kombinationen werden getestet
- Latin Square Design:  
nur ein zufällige Teilmenge aller Level-Kombinationen werden getestet
- Plackett-Burman Designs:  
sehr effizient, große Anzahl von Hauptfaktoren wird mit so wenig Durchgängen wie möglich untersucht
- Box-Behnken, Central Composite, Orthogonal Arrays, Center Point Designs, etc.

## Beispiel Crash-Test

Design Of Experiments



- Jeder Faktor hat mehrere Level
  - Stetige (z.B. Geschwindigkeit: 0 – 100 km/h)
  - Diskrete (z.B. Airbag: an/aus)
- Einteilung der stetigen Level in sinnvolle Stufen
  - Z.B. Geschwindigkeit niedrig, mittel, hoch
- Zielgröße: Kraftwirkung auf Dummy.
- Faktoren, Level, gewünschte Genauigkeit werden im Versuchsplan festgehalten

# Statistik und Gender Medicine



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

---

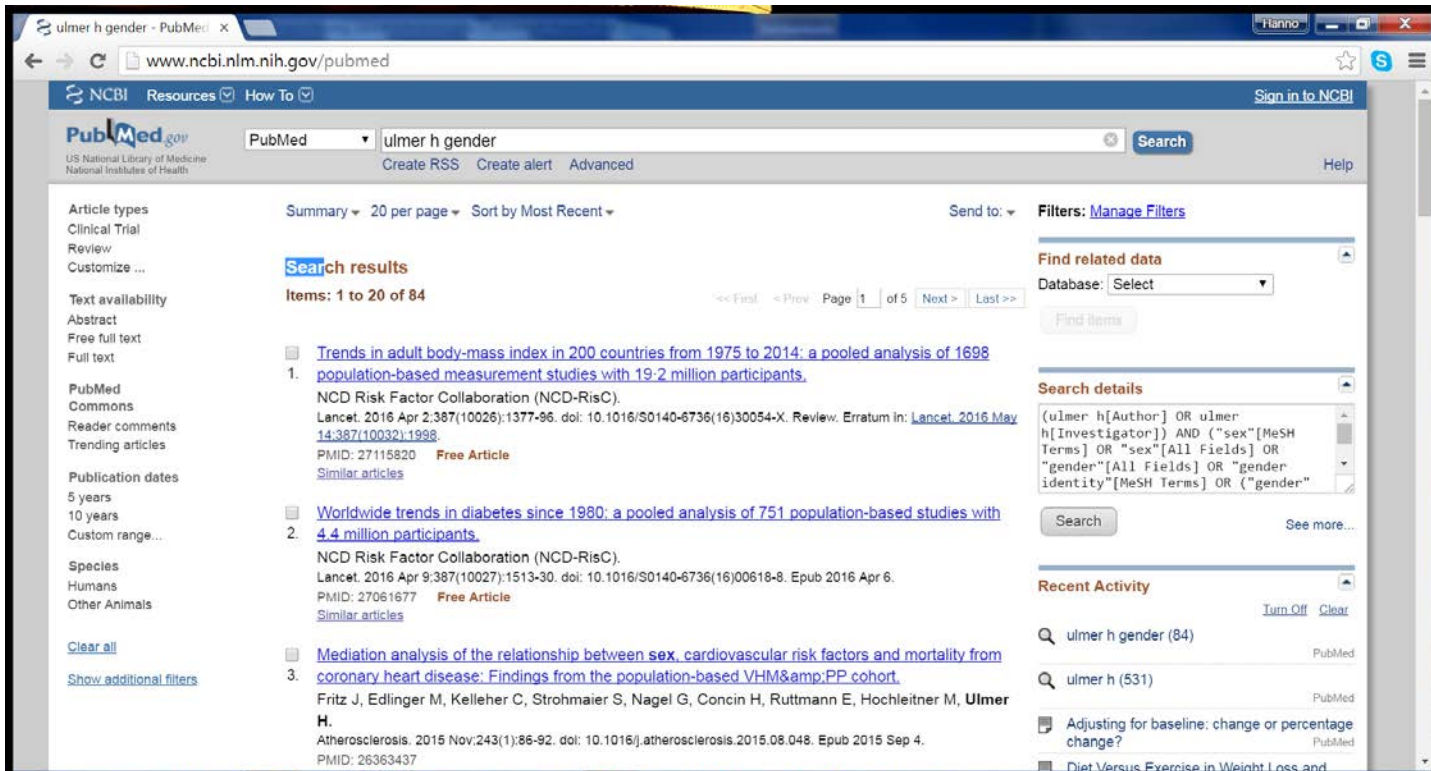
Dr. Hanno Ulmer

*hanno.ulmer@imed.ac.at*

---

Department für Medizinische Statistik, Informatik und  
Gesundheitsökonomie, Medizinische Universität Innsbruck

(ulmer h[Author] OR ulmer h[Investigator]) AND ("sex"[MeSH Terms] OR "sex"[All Fields] OR "gender"[All Fields] OR "gender identity"[MeSH Terms] OR ("gender"[All Fields] AND "identity"[All Fields]) OR "gender identity"[All Fields])



The screenshot shows a web browser window displaying the PubMed search results for the query "ulmer h gender". The search results are displayed in a list format, showing the first three items. The first item is "Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19.2 million participants", published in Lancet in 2016. The second item is "Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants", also published in Lancet in 2016. The third item is "Mediation analysis of the relationship between sex, cardiovascular risk factors and mortality from coronary heart disease: Findings from the population-based VHM&PP cohort", published in Atherosclerosis in 2015. The search results are displayed in a list format, showing the first three items. The search query is visible in the search bar, and the search results are displayed in a list format. The search results are displayed in a list format, showing the first three items. The search results are displayed in a list format, showing the first three items.

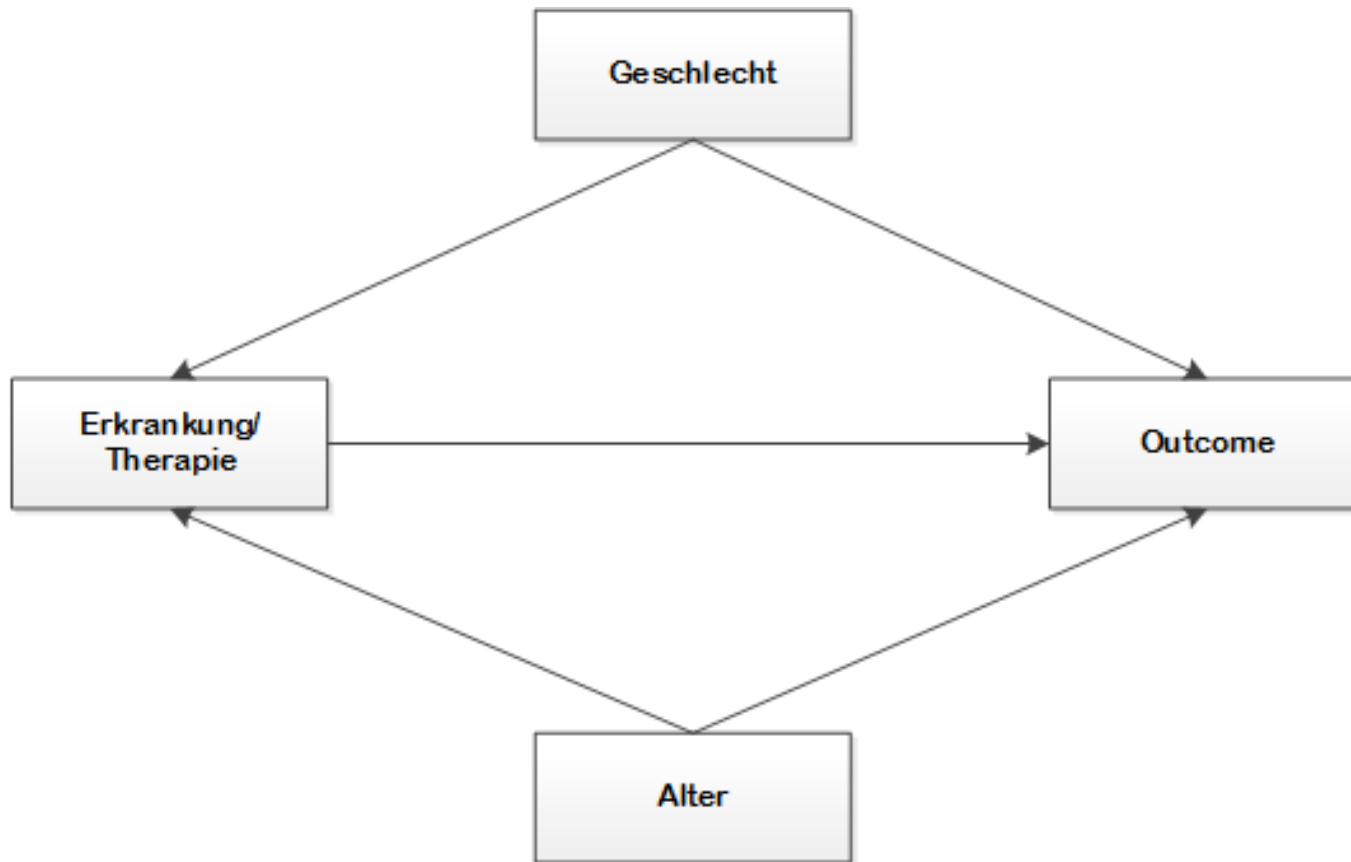
- sex-specific
- by sex
- age-sex-groups
- Independent of sex
- sex-matched
- after adjustment for sex, adjusting for sex
- stratified by sex
- age- and sex standardized

- Auswertung separat für Männer und Frauen, Unterschiede zwischen Männer und Frauen  
sex-specific, by sex, sex-groups
- Geschlecht als Störvariable (Confounder) und Berücksichtigung dieses Einflusses  
adjusting for sex, independent of sex, stratified by sex, sex standardized
- Im Studiendesign (z.B. Fall-Kontroll Studie) wird jedem männlichen Fall eine männliche Kontrolle und jedem weiblichen Fall eine weibliche Kontrolle zugewiesen  
sex-matched
- Randomisierung führt zu gleicher Anzahl von Männern und Frauen

# Für den Statistiker

- Ist Geschlecht in der Regel oft ein Störfaktor (Confounder) oder Effektmodifikator (effect modifier)
- Alter und Geschlecht sind die wichtigsten Einflussfaktoren in der Medizin,
- deren Einfluss muss entweder im Studiendesign durch Randomisierung und Matching oder nachträglich in der Analyse durch separate Auswertung, Standardisierung, Adjustierung oder Stratifizierung (Gewichtung) berücksichtigt werden
- Voraussetzung: Variable Geschlecht muss vorhanden sein!

# DAG: Geschlecht und Alter als Confounder





- Epidemiologen sprechen von Verschleierung oder **Confounding**, wenn die Assoziation zwischen einer Exposition und einem Outcome durch eine Störgröße überlagert oder verzerrt wird. Diese Störgröße heißt Confounder (Zitat aus Razum O et al. Epidemiologie für Dummies).

Zu unterscheiden von:

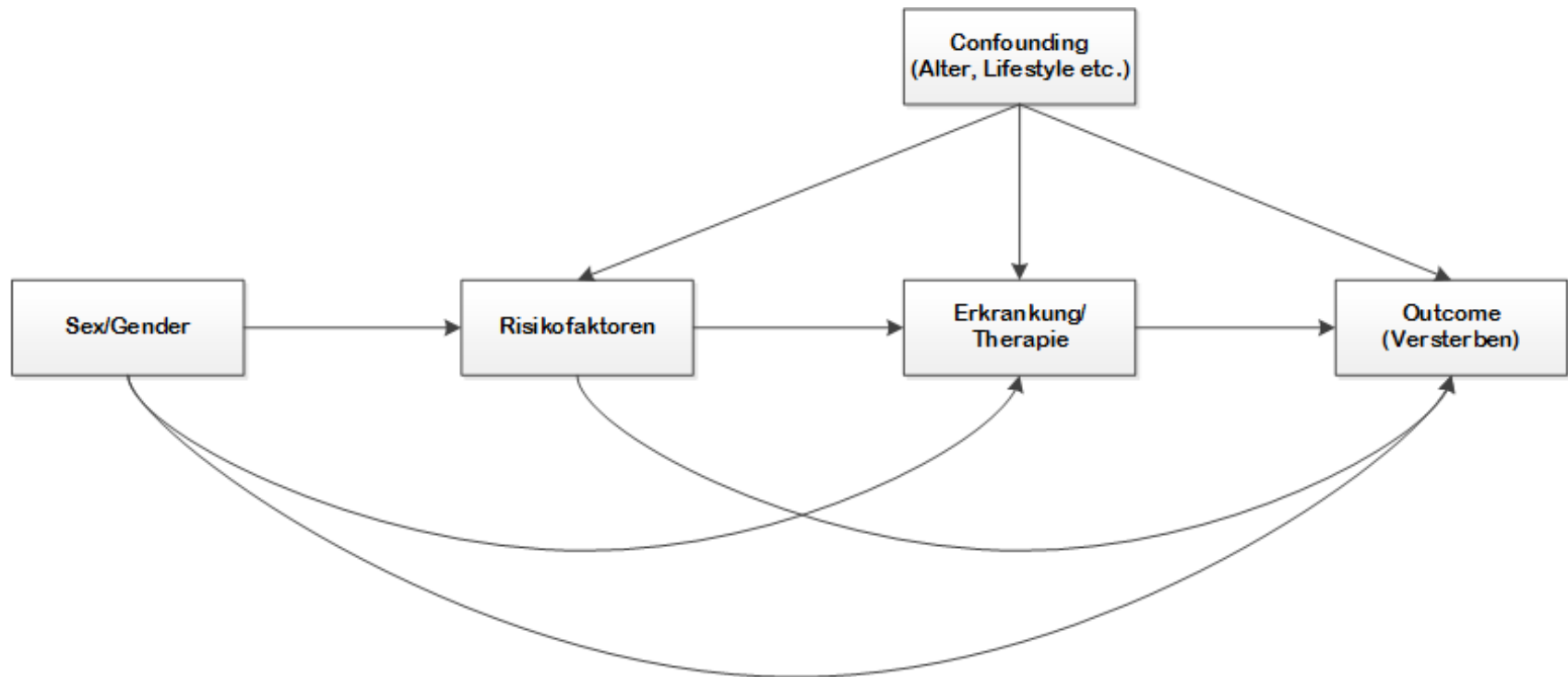
- Wenn sich die Stärke einer Exposition und einem Outcome verändert wenn eine oder mehrere weitere Variablen hinzukommen, dann spricht man von **Effektmodifikation** (Cholesterin Beispiel).
- Intermediärvariablen (**Mediatoren**) sind Zwischenstufen in der Kausalkette Exposition und Outcome (KHK Beispiel).

# Geschlechtsunterschiede und Kausalität



- Geschlechtsunterschiede können nur beobachtet werden
- Geschlechtsunterschiede können zwar in einem RCT beobachtet werden, aber nicht per se durch einen RCT untersucht werden
- Der Faktor Geschlecht kann nicht randomisiert werden
- Aber: Geschlecht ist ab Geburt defacto natürlich randomisiert
- Für die Gender Medizin Forschung gelten die Limitationen der Beobachtungsstudien ,  
Cave: Selection Bias und Informations Bias!
- Confounding ist per Definition nicht möglich
- Außer Frage stehende Fakten: Männer Prostatakrebs, etc.

# DAG: Geschlecht als Untersuchungsobjekt



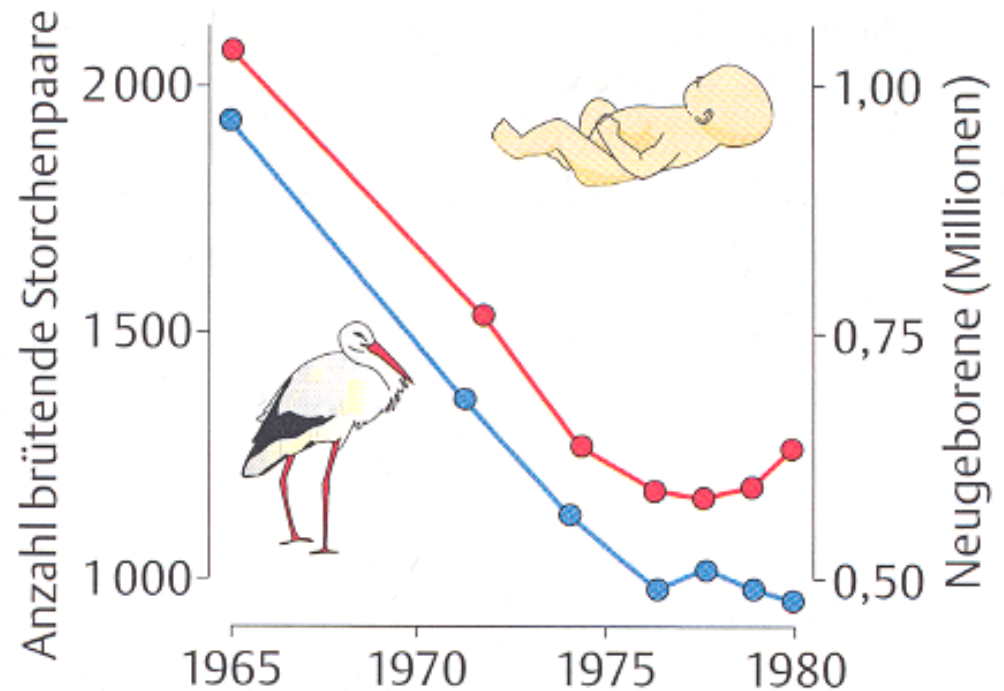
# Glauben Sie den Ergebnissen einer Studie?



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

- **Konzepte zur Kausalität z.B. von** Henle&Koch (1880), Hill (1965) oder Rothman (1976)
- **Assoziation** (negativer oder positiver Zusammenhang)
- **Kausalität** (Risikofaktor als (Mit-)Ursache einer Krankheit)

Korrelation Abnahme brütender Storchpaare/  
Geburtenrückgang in der BRD 1965 – 1980



# Statistische Assoziation oder Kausalität ??



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

- Kausalitätskriterien nach
- Sir Austin Bradford-Hill:
  - **Temporalität**
  - **Konsistenz (Meta-Analysen, Systematische Reviews)**
  - **Biologischer Gradient**
  - **Stärke des Effekts (z.B. doppeltes Risiko)**
  - u.a.



# Systematische Fehler, Confounding

---

- **Selection Bias**
  - Nichterreichen von Berufstätigen bei Telefonumfragen
- **Information Bias**
  - Fehlklassifikation / Fehldiagnosen,
  - Messfehler
- **Confounding**
  - Mangelnde Berücksichtigung von Störgrößen

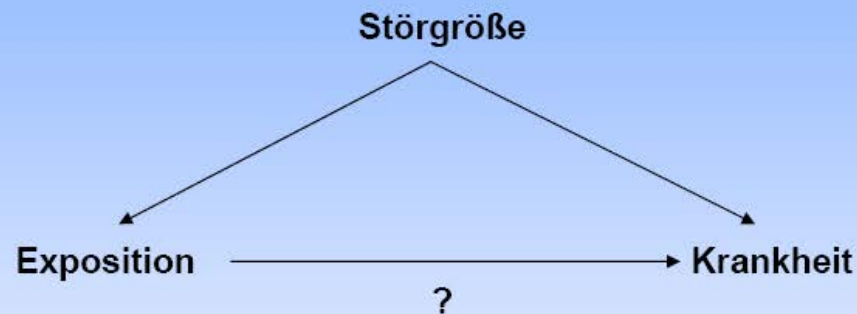
Scheinassoziation von Alkohol und Lungenkarzinom über Rauchen erklärbar

# Systematische versus zufällige Fehler



- Systematische Fehler führen zu einer Verzerrung (Bias) der Effektschätzer (RR, OR etc.)
- Zufällige Fehler (durch zu geringe Fallzahl, siehe Kapitel Fallzahlschätzung) erniedrigen die Präzision
- Systematische Fehler erniedrigen die Validität
- Die Validität hat Priorität vor der Präzision.

## Störgrößen (Confounder)



z.B. unterschiedliches Alter in den Therapiegruppen  
Alter beeinflusst den Blutdruck und womöglich die Therapie.

Mögliche Lösung: Adjustierung für Alter  
mittels multivariater Analyse



# Simpson's Paradoxon

## Reserpin-Beispiel

		Brustkrebs		
		ja	nein	
Reserpin	ja	32	57	$32/89 = 36 \%$
	nein	149	351	$149 / 500 = 30 \%$

**OR =  $(32/57) / (149/351)$   
=  $0,56 / 0,42 = 1,3$**

**Alter ≤ 50**

		<b>Brustkrebs</b>		
		<b>ja</b>	<b>nein</b>	
<b>Reserpin</b>	<b>ja</b>	2	14	$2/16 = 13 \%$
	<b>nein</b>	42	221	$42 / 263 = 16 \%$
				<b>OR = 0,14 / 0,19 = 0,75</b>

**Alter > 50**

		<b>Brustkrebs</b>		
		<b>ja</b>	<b>nein</b>	
<b>Reserpin</b>	<b>ja</b>	30	43	$30/73 = 41 \%$
	<b>nein</b>	107	130	$107 / 237 = 45 \%$
				<b>OR = 0,70 / 0,82 = 0,85</b>



# Logistische Regressionsanalyse

---

- Untersuchung / Modellierung von prognostischen Faktoren für binäre Outcomes bei gleichzeitiger Adjustierung für Störfaktoren („confounders“)
- Multivariate Erweiterung zur Kontingenztafelanalyse und Chi-Quadrat Test
- Berechnet wird das adjustierte Odds Ratio
- Cytisine for smoking cessation: Adjustment for all baseline characteristics shown in Table 1 had a negligible effect.

# Logistische Regressionsanalyse

## Modell mit Reserpin:

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
reserpin	,280	,242	1,339	1	,247	1,323	,824	2,123

## Modell mit Reserpin und Alter:

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
reserpin	-,179	,255	,491	1	,483	,836	,507	1,379
alter50	1,474	,205	51,497	1	,000	4,367	2,920	6,532

# Logistische Regression



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

**Untersuchungsziel: hat eine/mehrere (unabhängige) Variable/n X  
einen Einfluss auf eine andere (abhängige) Variable Y?**

---

- 🚧 Abhängige Variable ist binär (Ausprägungen z. B. ja/nein)
- 🚧 Unabhängige Variablen sind intervallskaliert oder als Dummy-Variablen codiert
- 🚧 Unterschied zur linearen Regression: Y kann nur die Werte 0 oder 1 annehmen
- 🚧 Hintergrund der logistischen Regression: Untersuchung des Zusammenhang  $p=P(Y=1)$  und der unabhängigen Variable(n) X
- 🚧 Es wird nicht der Wert der abhängigen Variablen vorhergesagt, sondern die Wahrscheinlichkeit, dass die abhängige Variable den Wert 1 annimmt

**Cave: erwarteter kausaler Zusammenhang zwischen unabhängiger und abhängiger Variable muss theoretisch erklärbar sein**

# Logistische Regression – statistischer Hintergrund



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

- 🚩 Logistische Regressionsanalyse basiert auf der Maximum-Likelihood-Schätzung (MLE) und unterscheidet sich von der Methode der kleinsten Quadrate (lineare Regressionsanalyse)
- 🚩 Ziel der Analyse: Identifikation einer Funktionskurve zu finden, die möglichst gut zu den Daten passt
- 🚩 Funktion ist eine logistische Funktion (bei linearen Regressionsanalyse eine Gerade)
- 🚩 Werte der logistischen Funktion werden als Wahrscheinlichkeit interpretiert (dass die abhängige Variable  $y$  den Wert 1 annimmt – gegeben die unabhängigen Variablen  $x_k$ )
- 🚩 Wert nahe bei 0 bedeutet, dass das Eintreten von  $y$  ( $y = 1$ ) sehr unwahrscheinlich ist; Wert nahe bei 1, dass das Eintreten von  $y$  sehr wahr

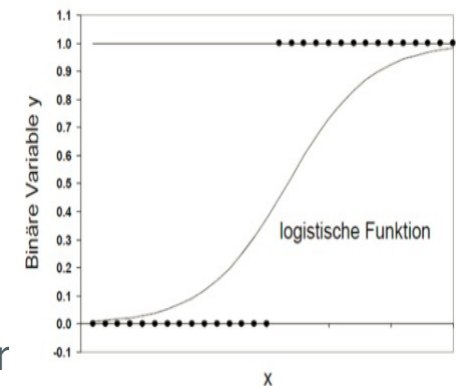


Abbildung 2: Logistische Funktion

# Logistische Regression – statistischer Hintergrund



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

Logistische Regressionsfunktion:

$$P(y=1) = \frac{1}{1+e^{-z}}$$

$P(y=1)$  = Wahrscheinlichkeit, dass  $y = 1$

$e$  = Basis des natürlichen Logarithmus, Eulersche Zahl

$z$  = Logit (lineares Regressionsmodell der unabhängigen Variablen)

$z$ , der sogenannte "Logit", stellt dabei ein lineares Regressionsmodell dar:

$$z = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \dots + \beta_k \cdot x_k + \varepsilon$$

$x_k$  = unabhängige Variablen

$\beta_k$  = Regressionskoeffizienten

$\varepsilon$  = Fehlerwert

$$P(y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x_1+\beta_2 \cdot x_2+\beta_3 \cdot x_3+\dots+\beta_k \cdot x_k+\varepsilon)}}$$

# Logistische Regression - Interpretation

- Zusammenhang zwischen unabhängigen Variablen und abhängiger Variable wird mittels sogenannter "Odds" interpretiert
- Odds: Wahrscheinlichkeit, dass das Ereignis eintritt, in Relation zum Nichteintreffen des Ereignisses

$$Odds = \frac{P(y \text{ trifft ein})}{P(y \text{ trifft nicht ein})} = \frac{P(y \text{ trifft ein})}{1 - P(y \text{ trifft ein})}$$

$$Odds \text{ Ratio} = \text{Exp}(B) = e^{\beta} = \frac{\text{Odds nach dem Anstieg von } x \text{ um eine Einheit}}{\text{Odds vor dem Anstieg von } x \text{ um eine Einheit}}$$

$$= \frac{Odds_{\text{nach}}}{Odds_{\text{vor}}}$$



# Logistische Regression - Interpretation

- Odds Ratio einer unabhängigen Variablen = **Veränderung der relativen Wahrscheinlichkeit** von  $y = 1$  an, wenn diese unabhängige Variable um eine Einheit steigt
- Odds Ratio einer unabhängigen Variablen ist der **Faktor**, um den sich die Odds verändern, wenn diese Variable um eine Einheit ansteigt
- Beträgt eine Odds Ratio ( $\text{Exp}(B)$ ) = 1 ergibt sich keine Veränderung ( $\text{Odds}_{\text{nach}} = \text{Odds}_{\text{vor}}$ )
- Odds Ratio  $> 1$  ergibt eine Zunahme der Odds ( $\text{Odds}_{\text{nach}} > \text{Odds}_{\text{vor}}$ )
- Odds Ratio  $< 1$  ergibt eine Abnahme der Odds ( $\text{Odds}_{\text{nach}} < \text{Odds}_{\text{vor}}$ )
- Zusammenhang Odds Ratios und Regressionskoeffizienten: Odds Ratio =  $\text{Exp}(B) = e^{\beta}$
- Odds Ratio = 1 wenn Regressionskoeffizient = 0,  $> 1$  wenn Regressionskoeffizient positiv ist,  $< 1$  wenn Regressionskoeffizient negativ ist

# Logistische Regression – statistische Signifikanz

- 🚩 1. Schritt: Überprüfen ob das Regressionsmodell insgesamt signifikant ist: Chi-Quadrat-Test
- 🚩 Prüft ob das Modell insgesamt einen Erklärungsbeitrag leistet
- 🚩 Modellgüte: Passung zwischen Modell und Daten ("Goodness of fit"): Analog zum R-Quadrat der linearen Regression – verschiedene Pseudo-R-Quadrate
- 🚩 2. Schritt: Überprüfen ob Regressionskoeffizienten (Betas) ebenfalls signifikant sind  
Wald-Test für jeden der Regressionskoeffizienten

Variablen in der Gleichung

	Regressionskoeffizient B	Standardfehler	Wald	df	Sig.	Exp(B)	95% Konfidenzintervall für EXP (B)	
							Unterer Wert	Oberer Wert
Schritt 1 <sup>a</sup>								
Einkommen	-.022	.006	14.651	1	.000	.979	.968	.990
Risikobereitschaft	.348	.088	15.541	1	.000	1.416	1.191	1.683
Interesse	.085	.018	23.036	1	.000	1.089	1.052	1.127
Konstante	-1.668	.279	35.731	1	.000	.189		

a. In Schritt 1 eingegebene Variablen: Einkommen, Risikobereitschaft, Interesse.

- 🚩 Exp(B) = entlogarithmierter logit-Koeffizienten – wenn Konfidenzintervall von Exp(B) den Wert 1 nicht einschließt: signifikanter Einfluss
- 🚩 *Risikobereitschaft* Exp(B) > 1 positiver Zusammenhang: steigt Risikobereitschaft um eine Einheit steigt die relative Wahrscheinlichkeit, dass eine Person bereits einmal Aktien gekauft hat, um 41.6% (1.416 – 1 = .416)

# Vergleichbarkeit: Patientenflussdiagramm

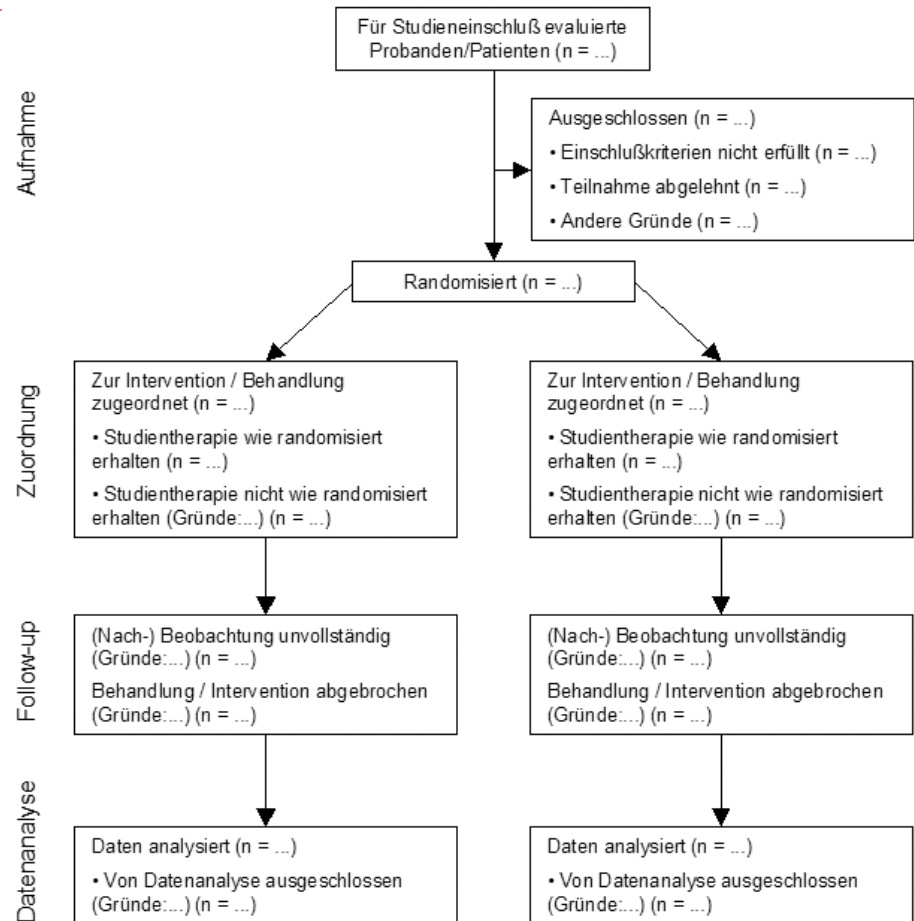


MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

- Intention-To-Treat

versus

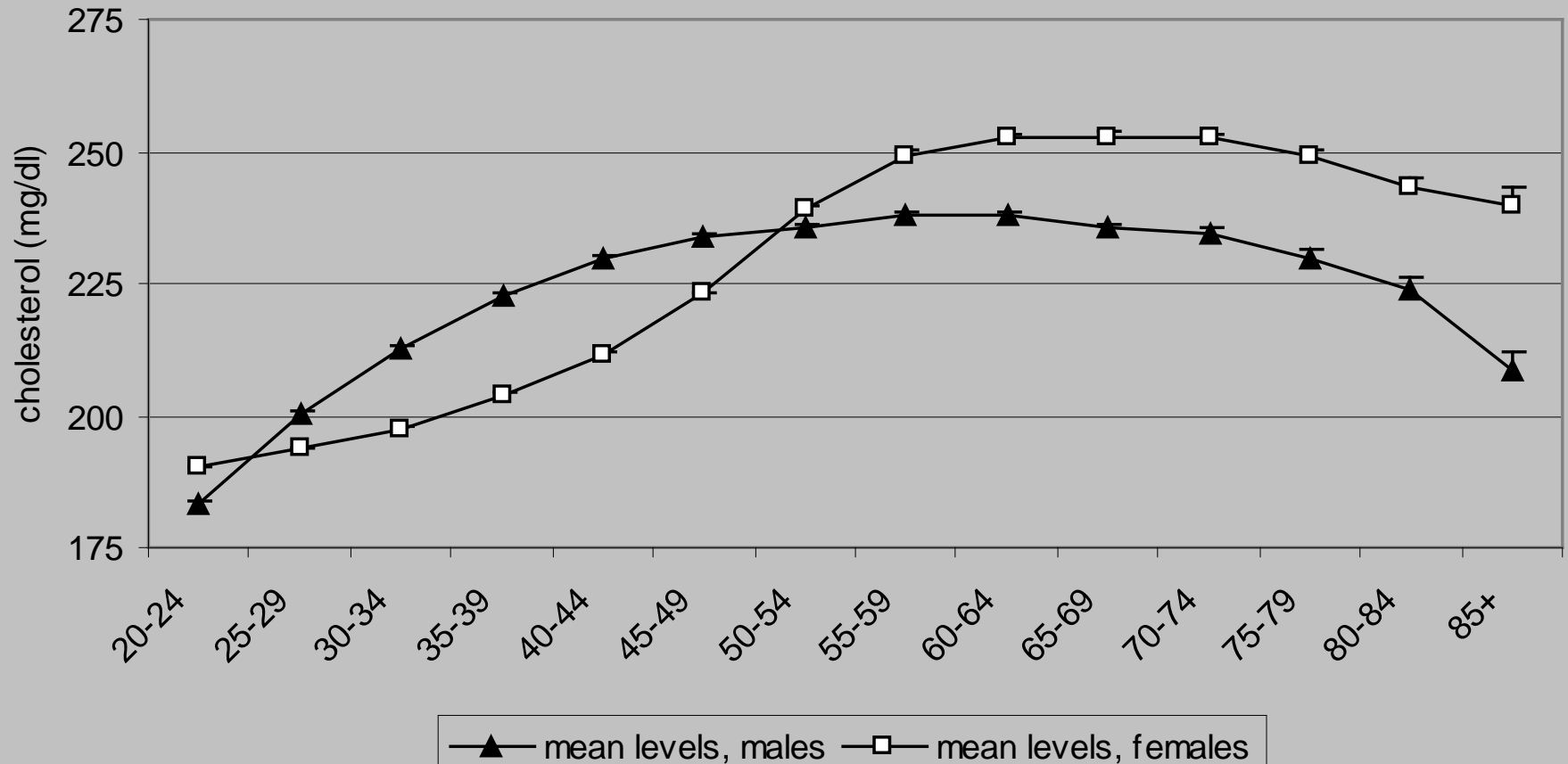
- Per-Protocol
- Vergleichbarkeit der Gruppen:
  - Strukturgleichheit
  - Beobachtungsgleichheit
  - Behandlungsgleichheit



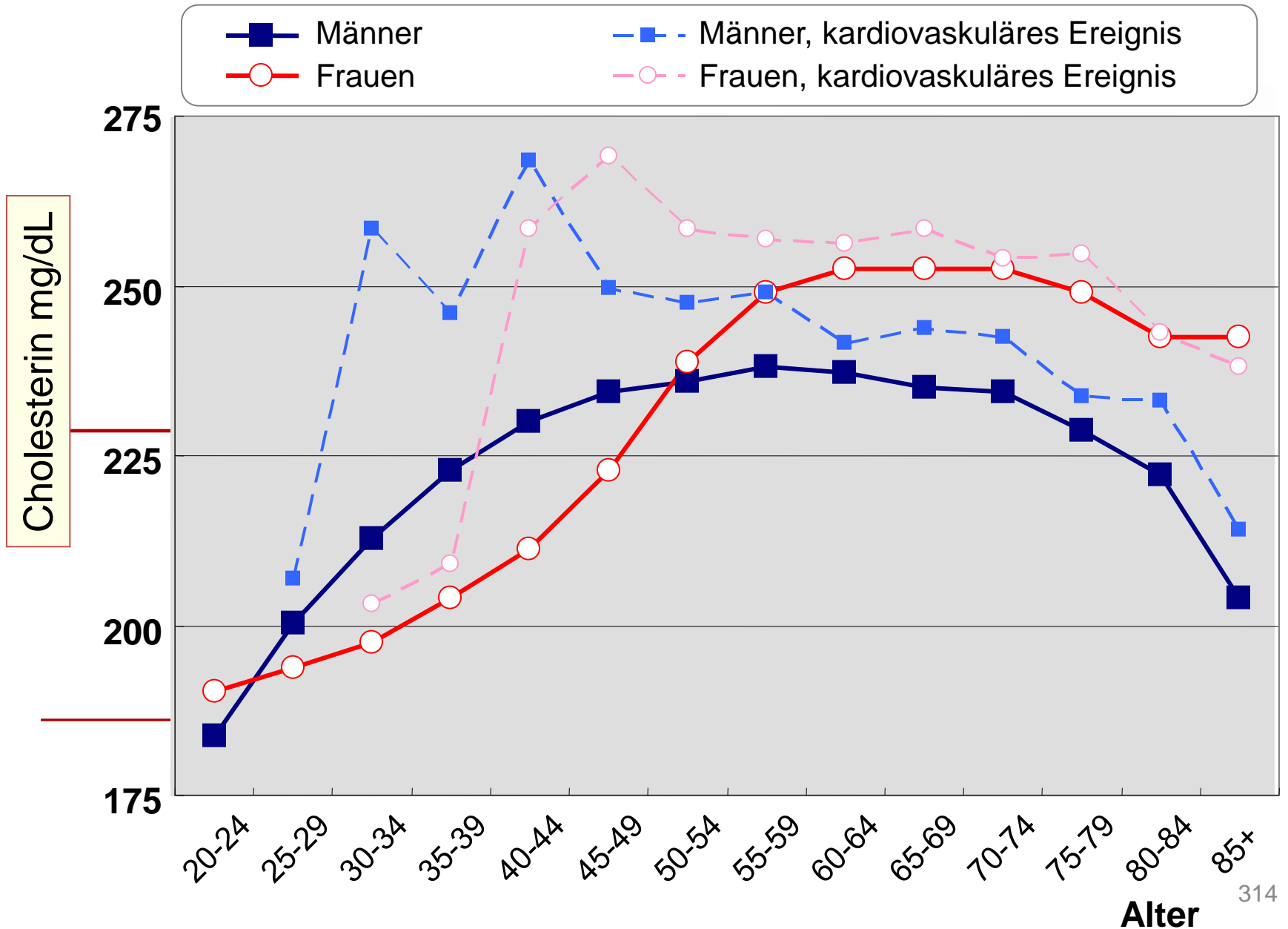
# Effektmodifikation (Interaktion)

- Existieren in verschiedenen Strata (Schichten) einer Variablen unterschiedliche Effektschätzer,
- so spricht man von **Effektmodifikation bzw. Interaktion**.
- **Die Schichtvariable** wird als Effektmodifikator (effect modifier) bezeichnet
- Reine Effektmodifikation führt nicht zu einer Verzerrung des Effektmaßes und gehört damit nicht zu den Fehlern in epidemiologischen Studien
- Modellierung durch Aufnahme von multiplikativen Termen in Regressionsmodellen

# Mittleres Gesamtcholesterin für Männer und Frauen nach Alter bei der Erstuntersuchung



# Mittleres Gesamtcholesterin für Männer und Frauen mit kardiovaskulärer Mortalität



# Studiendesign und Ethikeinreichung



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

---

Dr. Hanno Ulmer

*hanno.ulmer@imed.ac.at*

---

Department für Medizinische Statistik, Informatik und  
Gesundheitsökonomie, Medizinische Universität Innsbruck

# Hierarchie von Medizinischen Studien



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK





# RCT: Randomisierte kontrollierte Studie



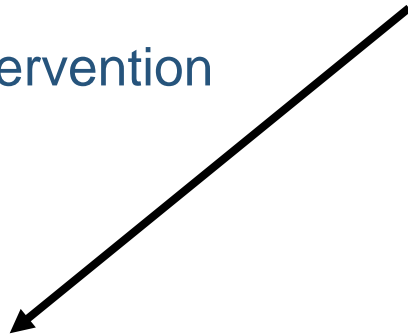
MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

- **Klinische Studien sind ein Experiment**
  - Einflussfaktor wird gesteuert
  - Alle anderen Faktoren sollen möglichst konstant gehalten werden
- **Randomisierung**
  - Zufällige Behandlungszuteilung
  - Ausschluss von Verzerrungen (Bias) durch Selektion
- **Verblindung**
  - Ausschluss von Verzerrungen (Bias) durch Information

# Experiment vs. Erhebung

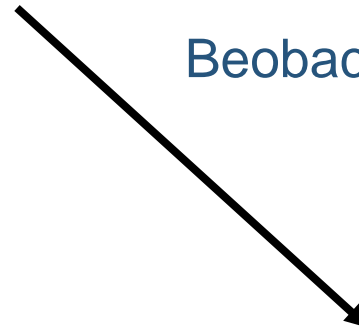
- Zu untersuchender Faktor  
(z.B. Wirksamkeit einer Therapie oder eines Medikaments,  
schädlicher Einfluss von Rauchen oder Übergewicht,  
Schützender Einfluss von Obst/Gemüse, Sport...)

Intervention



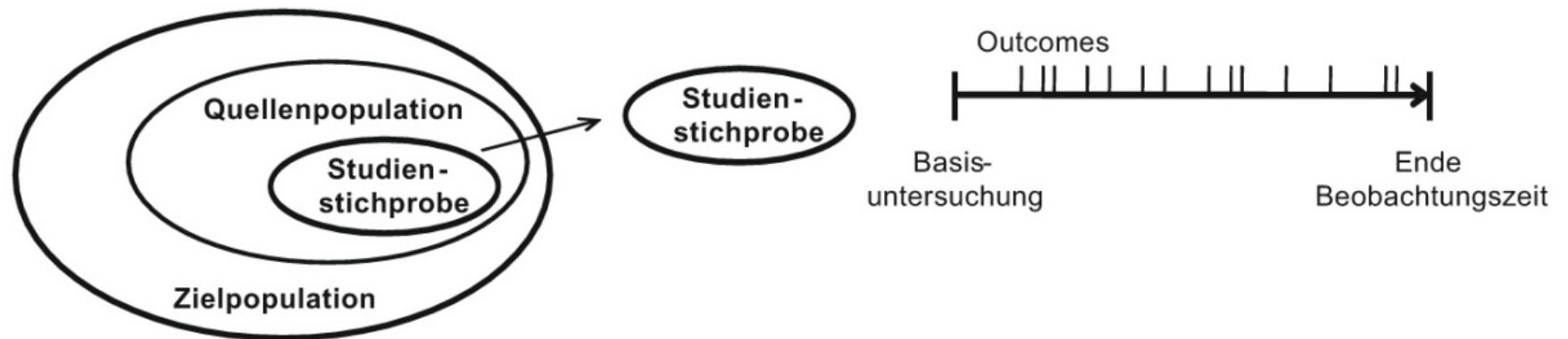
Der Faktor wird gezielt eingesetzt und vorgegeben. Wirksamkeit (und Verträglichkeit) in Hinblick auf die Zielerkrankung werden geprüft.

Beobachtung



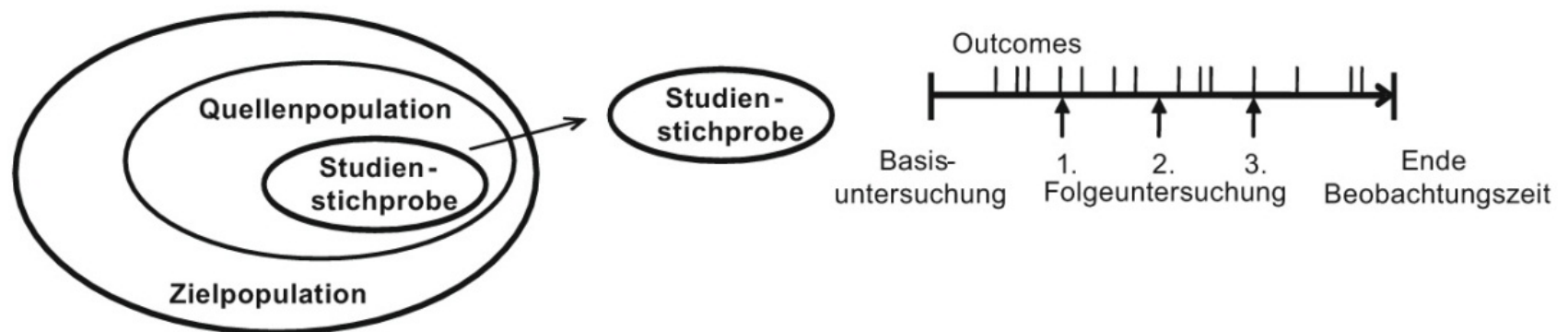
Der Faktor wird beobachtet und Zusammenhänge mit bzw. sein Einfluss auf das Auftreten von Krankheit geprüft.

## Kohortenstudie 1



**Abbildung 14.2:** Grundstruktur einer Kohortenstudie.

## Kohortenstudie 2



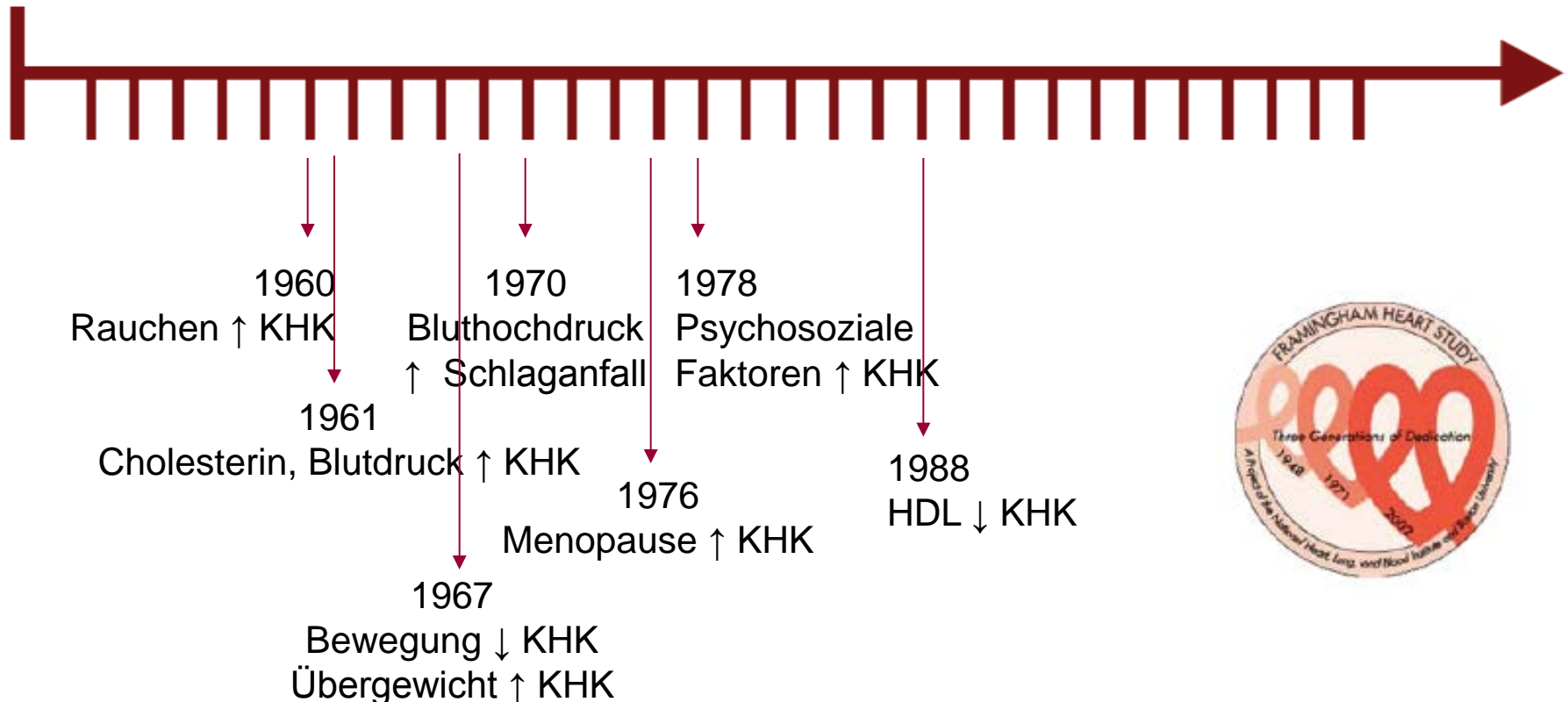
**Abbildung 14.3:** Eine Kohortenstudie mit wiederholten Untersuchungen.

# Kohorten Studie: Framingham Heart Study



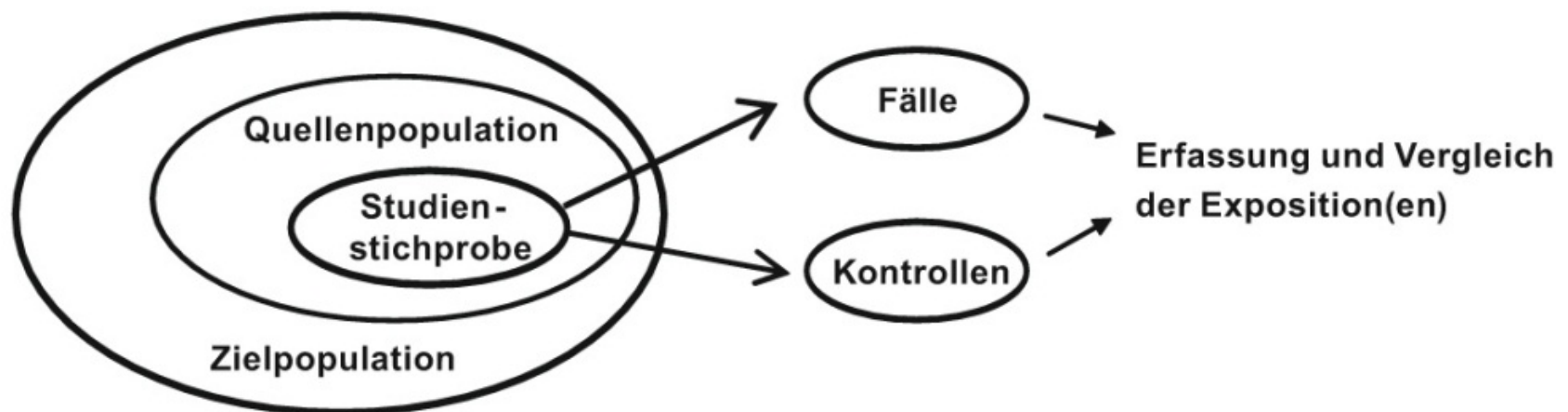
MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

1948 Rekrutierung von 5209 gesunden Probanden (30-62J)



# Fall-Kontroll Studie

## Fall Kontroll Studie

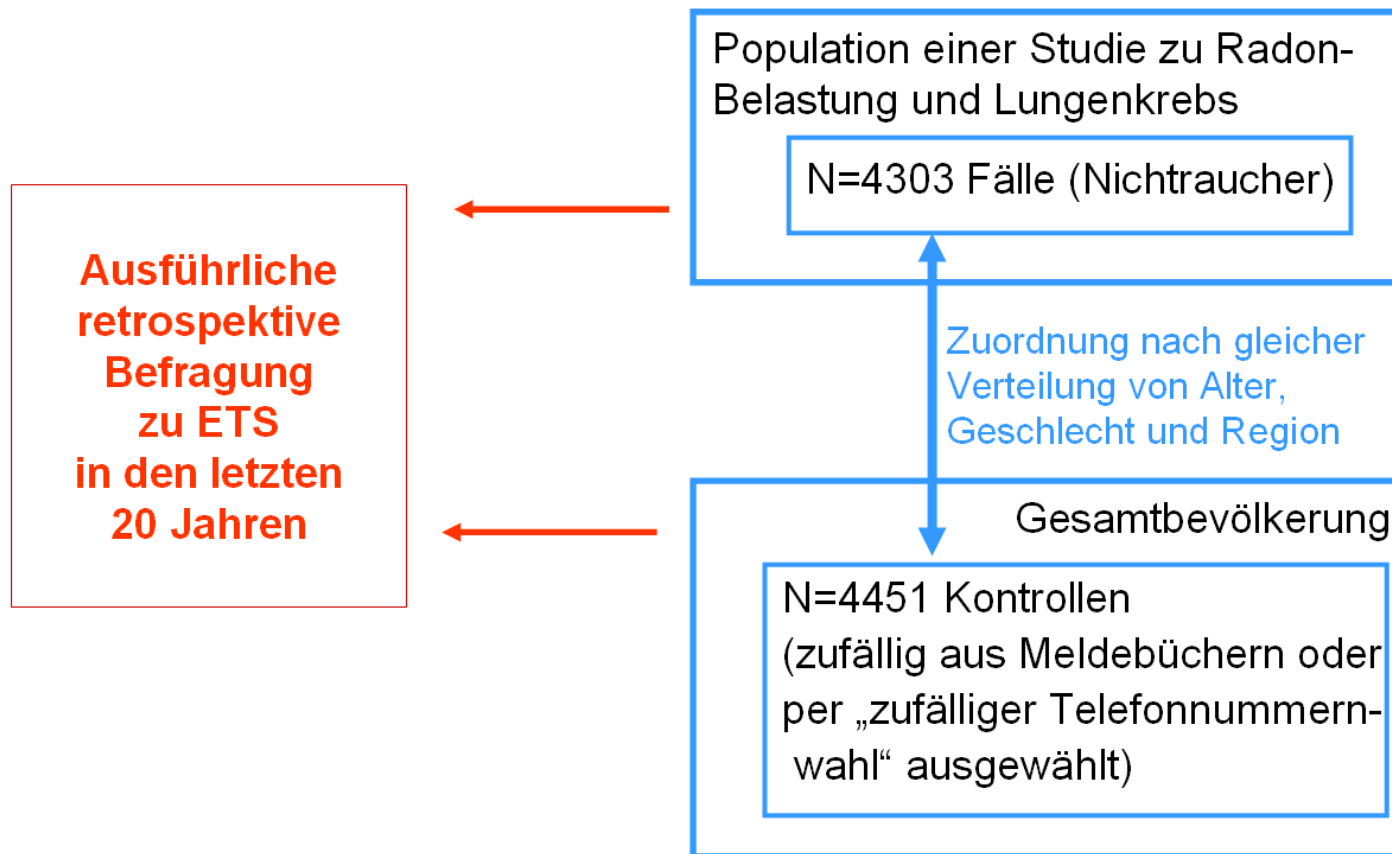


**Abbildung 14.4:** Grundstruktur einer Fall-Kontrollstudie.

# Fall-Kontroll Studie: Passivrauchen (ETS) und Lungenkrebs



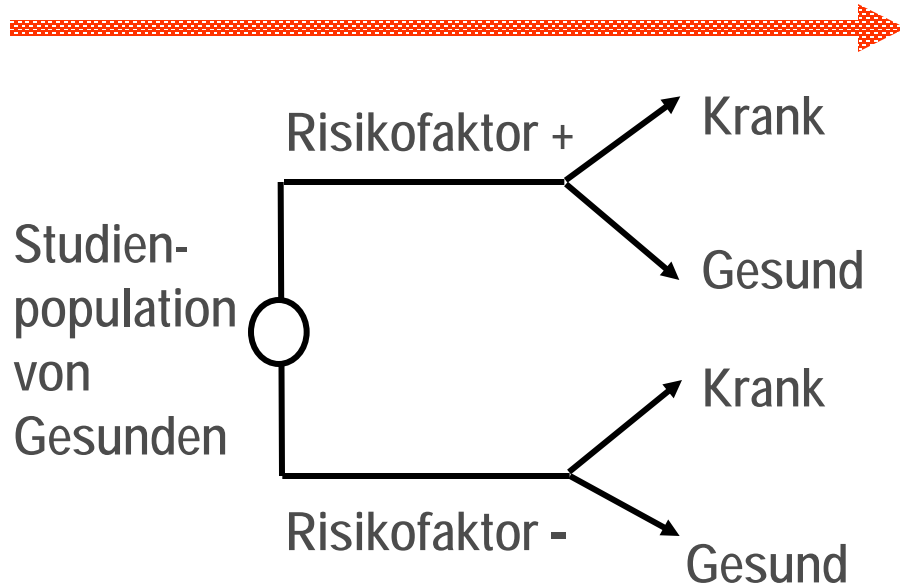
MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK



Kreuzer, Krauss, Kreienbrock, Jöckel, Wichmann in AJE, 1999

# Kohortenstudie

prospektiv



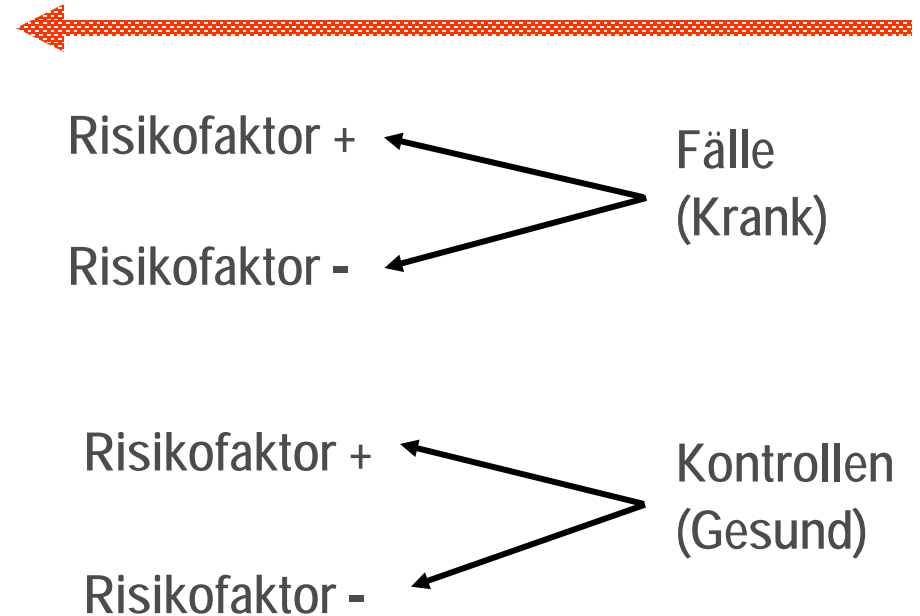
Kranke    Gesunde

Risikofaktor +	A	B	A + B
Risikofaktor -	C	D	C + D

**Relatives Risiko** =  $\frac{A / A+B}{C / C+D}$

# Fall-Kontroll-Studie

retrospektiv



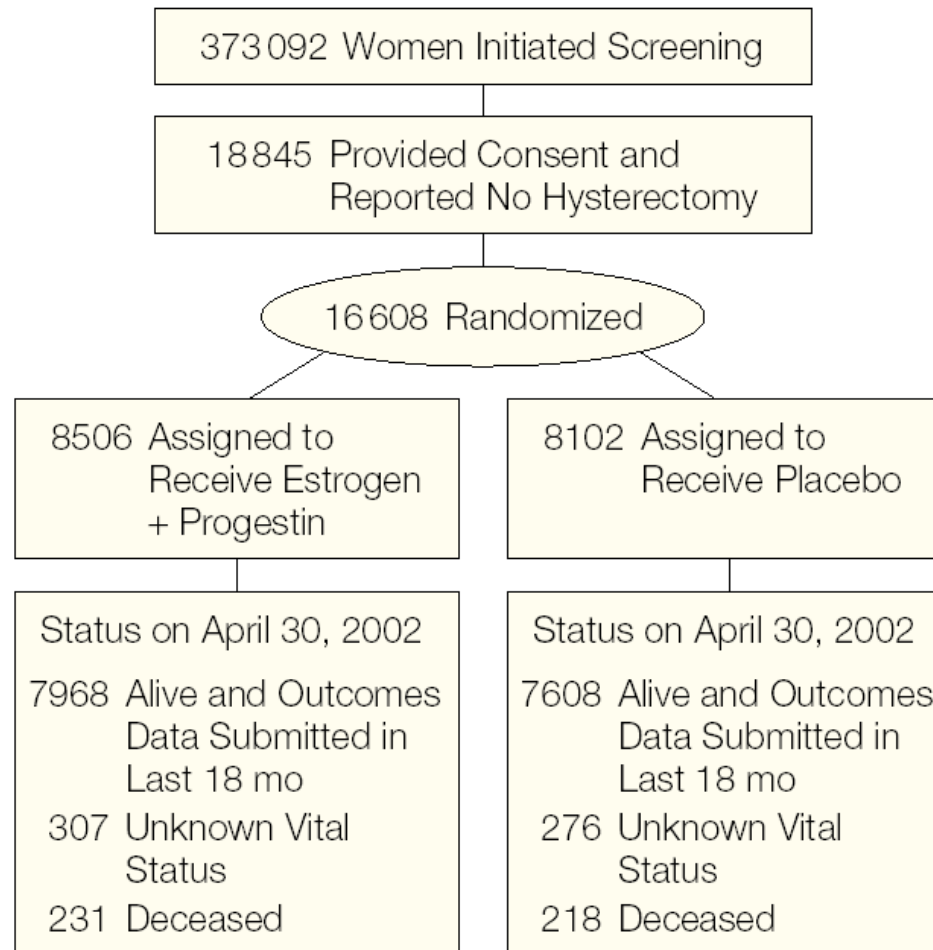
Kranke    Gesunde

Risikofaktor +	A	B	A + B
Risikofaktor -	C	D	C + D

**Odds Ratio** =  $\frac{(A / A+B) / (B / A+B)}{(C / C+D) / (D / C+D)} = \frac{A \times D}{B \times C}$



# Interventionsstudie: Women`s Health Initiative



# Meta-Analyse



---

MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

---

Dr. Hanno Ulmer

*hanno.ulmer@i-med.ac.at*

---

Department für Medizinische Statistik, Informatik und  
Gesundheitsökonomie, Medizinische Universität Innsbruck

- Allgemeiner Teil
- Praktische Meta-Analyse mit MedCalc
- Übung Lionheart-Levorep



# Literaturhinweise, Quellen

---

- Marcus Müllner - Erfolgreich wissenschaftlich arbeiten in der Klinik; Evidence Based Medicine-Springer Wien 2005, S.125 ff.
- George Davey Smith, Matthias Egger – Meta-Analysen, pharma-kritik Jahrgang 14 , Nummer 14, 1992
- Hansueli Stamm, Thomas M. Schwarb: Meta-analyse- Eine Einführung; ZfP 1995
- A. Ziegler, S. Lange, R. Bender; Systematische Übersichten und Meta-Analysen, Deutsche Medizinische Wochenschrift, 129. Ausgabe, 2004.
- Leo Held, Burkhardt Seifert, Kaspar Rufibach Medizinische Statistik, Pearson 2013.
- <https://training.cochrane.org/interactivelearning>

# Definition und Idee



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

- Die Meta-Analyse ist die quantitative Kombination der Resultate mehrerer Einzelstudien
- Idee ca. 100 Jahre alt, erste Analyse 1955 (Becher et al JAMA 1955)
- Begriff erstmals 1979 durch Psychologen Gene V. Glass verwendet
- Mit der Entwicklung von geeigneteren statistischen Methoden haben Meta-Analysen seit dem Anfang der achtziger Jahre zunehmend an Bedeutung gewonnen. Der meta-analytische Ansatz blieb jedoch von heftiger Kritik nicht verschont. Während die einen die Meta-Analyse als «objektive, quantitative Methode» rühmen, wird das Verfahren von anderen als «statistischer Trick» bezeichnet, der «ungerechtfertigte Annahmen macht und zu unzulässigen Verallgemeinerungen führt».

- Einzelergebnisse inhaltlich homogener Primärstudien werden zusammengefasst und empirisch ausgewertet.
- Ziel ist eine Effektgrößenschätzung. Es soll untersucht werden, ob ein Effekt vorliegt und wie groß dieser ist.
- Effektgrößen können sein: Mittelwerte, Differenzen, relative oder absolute Risiken, Odds Ratio, etc.
- Mit Hilfe der Meta-Analyse lassen sich mehrere geeignete Einzelstudien statistisch zusammenfassen, diese können dabei verschieden gewichtet werden.
- Gewinn neuer Erkenntnisse aus alten Daten

- Voraussetzung, es wurden zwei oder mehr Studien mit ähnlicher Fragestellung durchgeführt und/oder publiziert
- Stichprobenumfang der Einzelstudien zu klein
- Vorhandene Studien liefern inkonsistente Ergebnisse
- Untersuchung von Einflüssen und ihre Stärke auf das Ergebnis
- Grundlage für künftige Forschungstätigkeit
- Ermittlung des Publikationsbias

# Beispiel

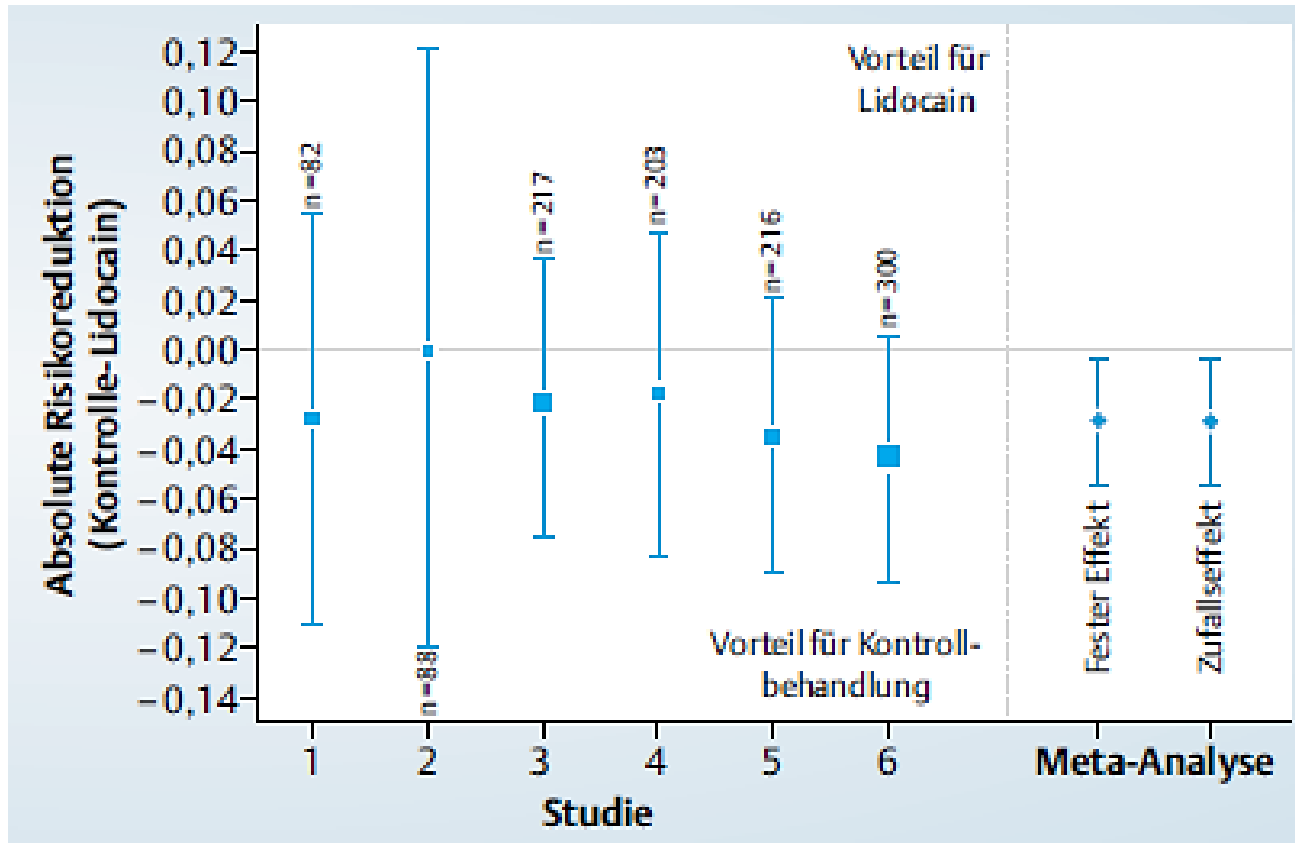


**Tab. 1** Untersuchung der Mortalität durch Prophylaxe mit Lidocain im akuten Myokardinfarkt (Quelle: Referenz [18], siehe auch [20]).

	Quelle	Anzahl randomisierter Patienten		Anzahl verstorbener Patienten	
		Lidocain	Kontrolle	Lidocain	Kontrolle
1	Chopra et al.	39	43	2	1
2	Mogensen	44	44	4	4
3	Pitt et al.	107	110	6	4
4	Darby et al.	103	100	7	5
5	Bennett et al.	110	106	7	3
6	O'Brian et al.	154	146	11	4
<i>Gesamt</i>		557	549	37	21



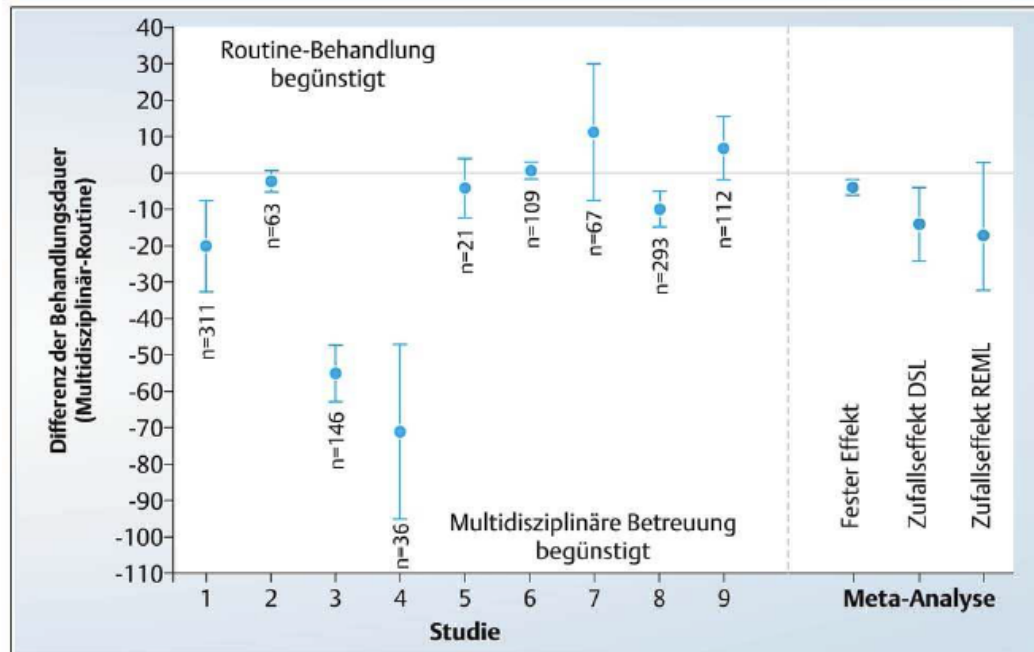
# Beispiel



Wirksamkeit von Lidocain zur Reduktion von Mortalität im akuten Myokardinfarkt

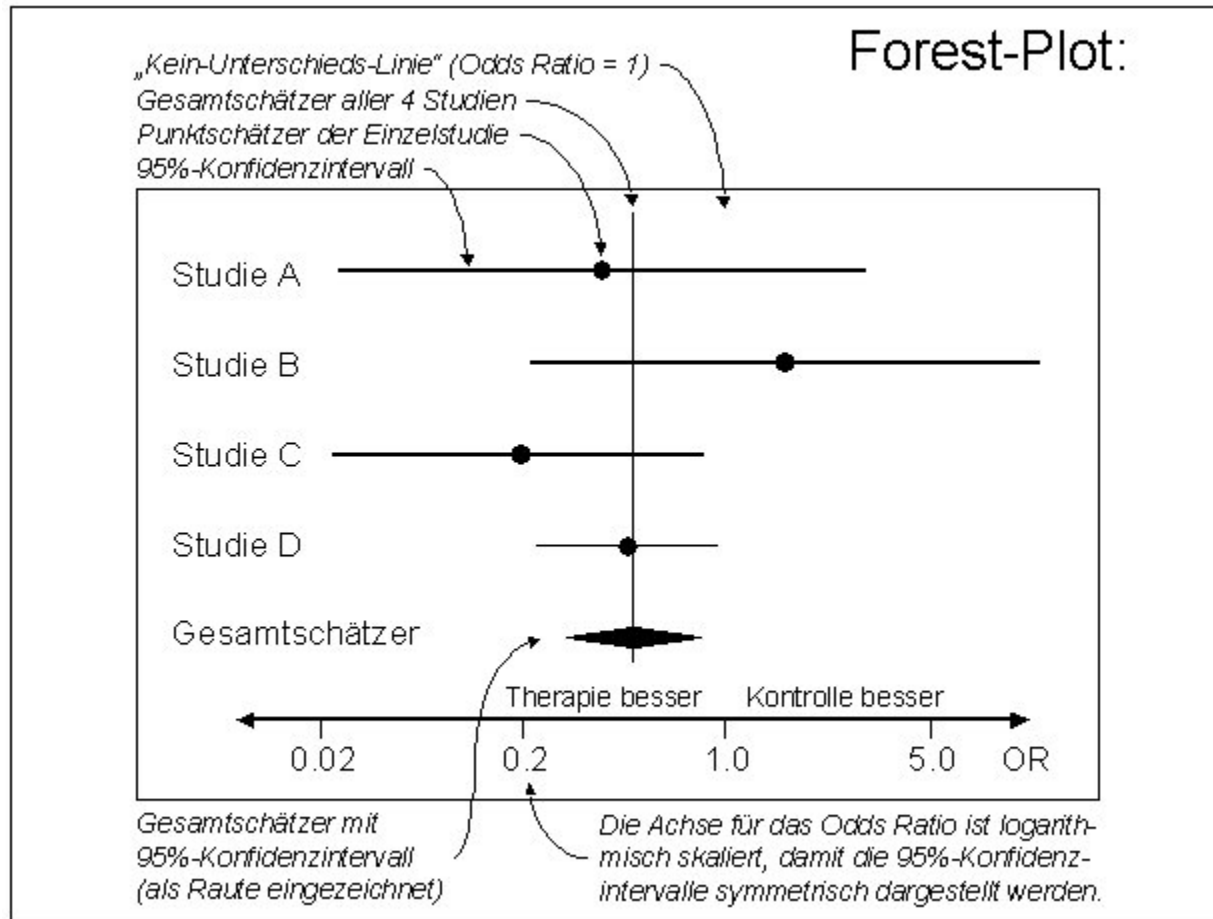
[A. Ziegler, S. Lange, R. Bender; Systematische Übersichten und Meta-Analysen.]

# Beispiel, Heterogenität



**Abb. 2** Betreuung von Schlaganfallpatienten durch Team von Spezialisten mehrerer Disziplinen im Vergleich zum Routinemanagement: x-Achse stellt die einzelnen Studien sowie die Meta-Analysen dar; y-Achse Dauer des Krankenhausaufenthalts in Tagen (LOS). Für jede einzelne Studie und die Meta-Analysen sind die geschätzte LOS (Punkt) mit dem dazugehörigen 95% Konfidenzintervall (Schnurrbärte) dargestellt.

# Meta-Analyse



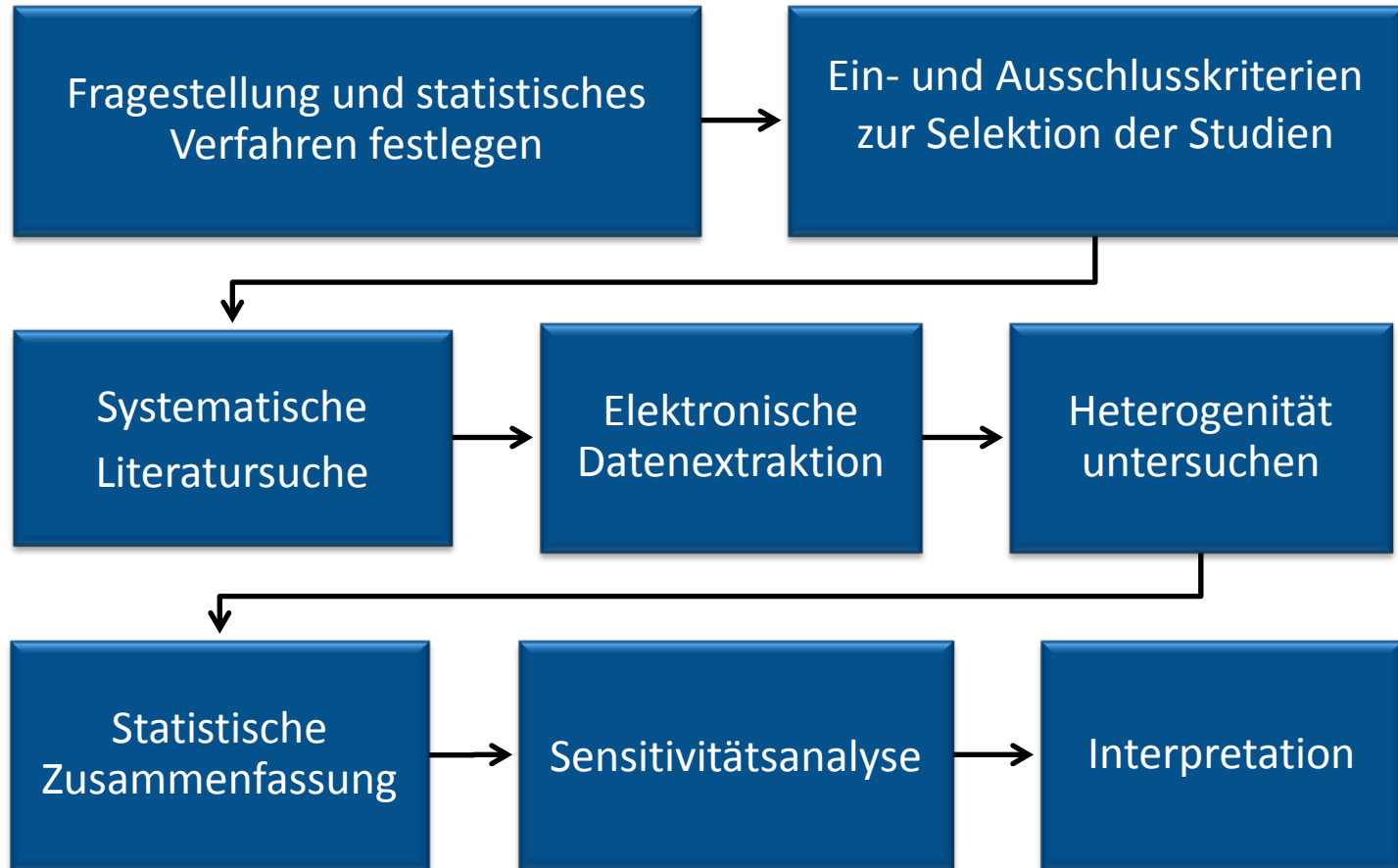


# Stufen einer Meta-Analyse

---

- Definition der Hypothese
  - Definition der abhängigen und unabhängigen Variablen
  - Standardisierung der Terminologie
- Definition der Keywords für Literatursuche und Datensammlung
  - MeSH
- Definition von Ein- und Ausschlusskriterien für Studien
- Analyse und Validierung der Ergebnisse
  - Mit geeigneter Software

# Ablauf der Meta-Analyse



# Meta-Analysis Protocol (FDA guidance)



- The planned purpose of the meta-analysis
- The background information available at the time of protocol development that motivated the meta-analysis
- The design features of the meta-analysis, including outcome definition and ascertainment, exposure periods and assessment, comparator drugs, and target subject population
- A description of the search strategy that will be used to identify candidate trials and the criteria that will be applied for trial selection
- The analysis strategy for conducting the meta-analysis, including planned subgroup analyses and sensitivity analyses

# Methods: Search strategy

---

## Search strategy

Key terms:

“physical activity”, “exercise”,  
“increase”, “brief intervention”,  
“counselling”, “systematic  
review”, “meta analysis”.

Period covered: 1854 - October  
2011.

## Databases

- CINAHL
- Cochrane Database of Systematic Reviews
- Database of Abstracts of Reviews of Effects (DARE) on Cochrane Library and Centre for Reviews and Dissemination (CRD)
- Health Technology Assessment database on Cochrane Library and Centre for Reviews and Dissemination (CRD)
- Embase
- MEDLINE
- PsycINFO
- SCI-Expanded
- SSCI SIGN
- Hand search of first authors’(LL) personal collections of articles

# Hierarchie von Medizinischen Studien

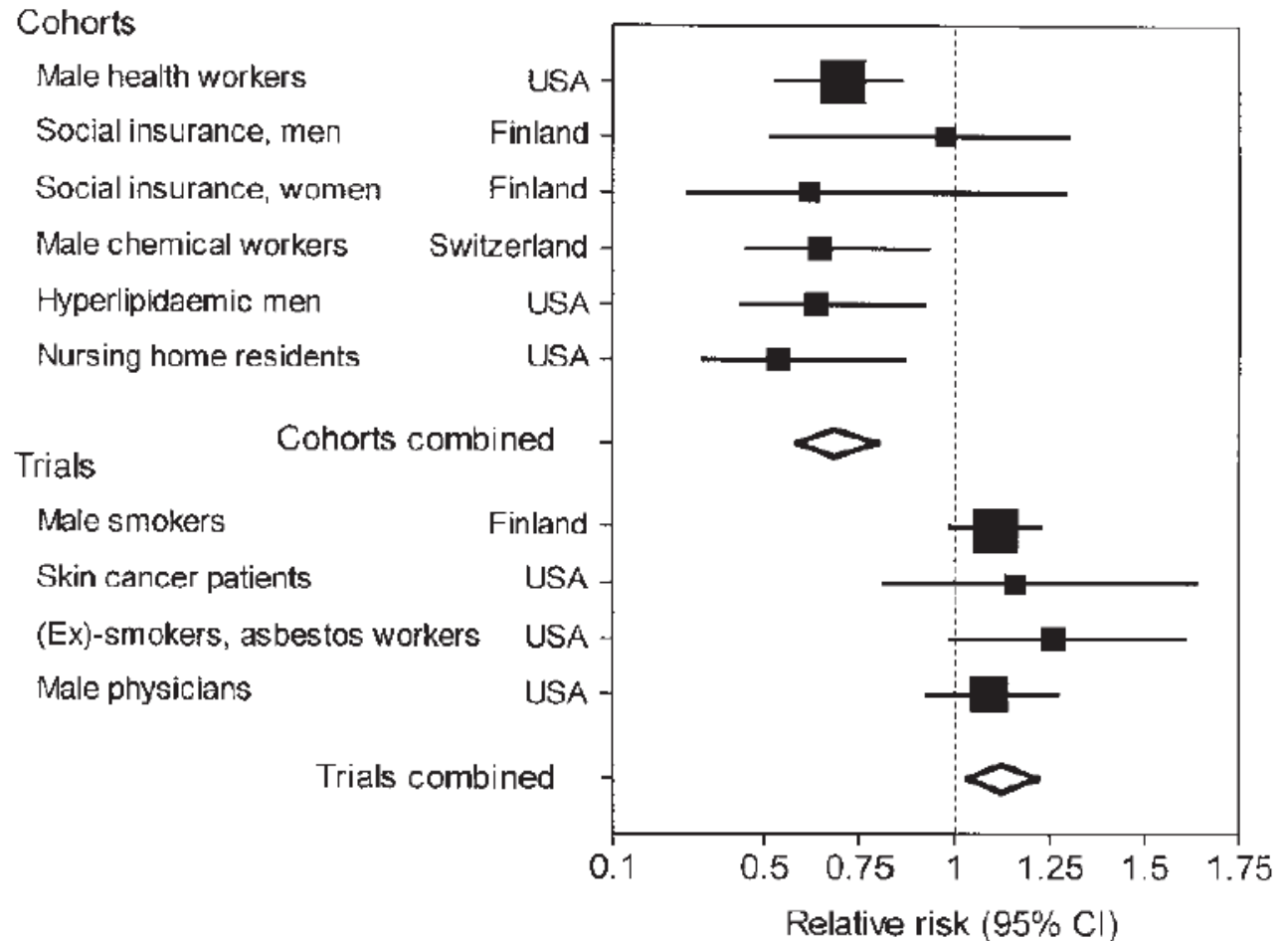




# Konsistenz

## Beta-Carotin und kardio-vaskuläre Mortalität

Davey Smith & Ebrahim, Int J Epidemiol 2001



# Evidenzgrade von Studien

Stufe	Für Therapie, Prävention, Ätiologie, Nebenwirkungen
1a	Systematisches Review mit homogenen RCTs
1b	Einzelne RCT (mit engen Konfidenzintervallen)
1c	Sonderfälle
2a	Systematisches Review mit homogenen Kohortenstudien
2b	Einzelne Kohortenstudie oder methodisch schwache RCT (z.B. <80% Nachbeobachtung)
2c	"Outcome"-Forschung, Ökologische Studien
3a	SR mit homogenen Fall-Kontroll-Studien
3b	Einzelne Fall-Kontroll-Studie
4	Fallserien (+ qualitativ schlechte Kohorten- u. Fall-Kontroll-Studien)
5	Expertenmeinung ohne explizite Bewertung der Evidenz oder basierend auf physiologischen Grundlagenmodellen

- *Ist die Qualität der berücksichtigten Studien zufriedenstellend?*

Zwischen verschiedenen Studien bestehen zum Teil erhebliche Qualitätsunterschiede, die in einer Meta-Analyse zunächst nicht berücksichtigt werden. Als Mindestanforderung gilt, dass nur korrekt *randomisierte Studien* mit vollständigen Angaben über alle am Anfang in die Studie aufgenommenen Personen meta-analysiert werden. Der Randomisierungsvorgang, der leider oft ungenügend beschrieben ist, verdient deshalb besondere Aufmerksamkeit.

Daneben ist es auch von Bedeutung, dass in den verschiedenen Studien tatsächlich gleiche oder vergleichbare Therapien und Beurteilungsverfahren eingesetzt wurden. Bei vielen Medikamentenstudien muss z.B. gefordert werden, dass die Bewertung der Therapieergebnisse *blind* erfolgt.

- *Wurden alle relevanten Studien berücksichtigt?*

Da «positive» Resultate mit grösserer Wahrscheinlichkeit veröffentlicht werden als Studien, die keinen Effekt zeigen, sollten auch unpublizierte Studien berücksichtigt werden. Dies gilt natürlich nur, soweit die erwähnten Qualitätskriterien erfüllt sind. Wenn allenfalls in Frage kommende Studien nicht in eine Meta-Analyse aufgenommen werden, so sollten sie erwähnt und die entsprechende Entscheidung begründet werden.

# Fragen zur Meta-Analyse

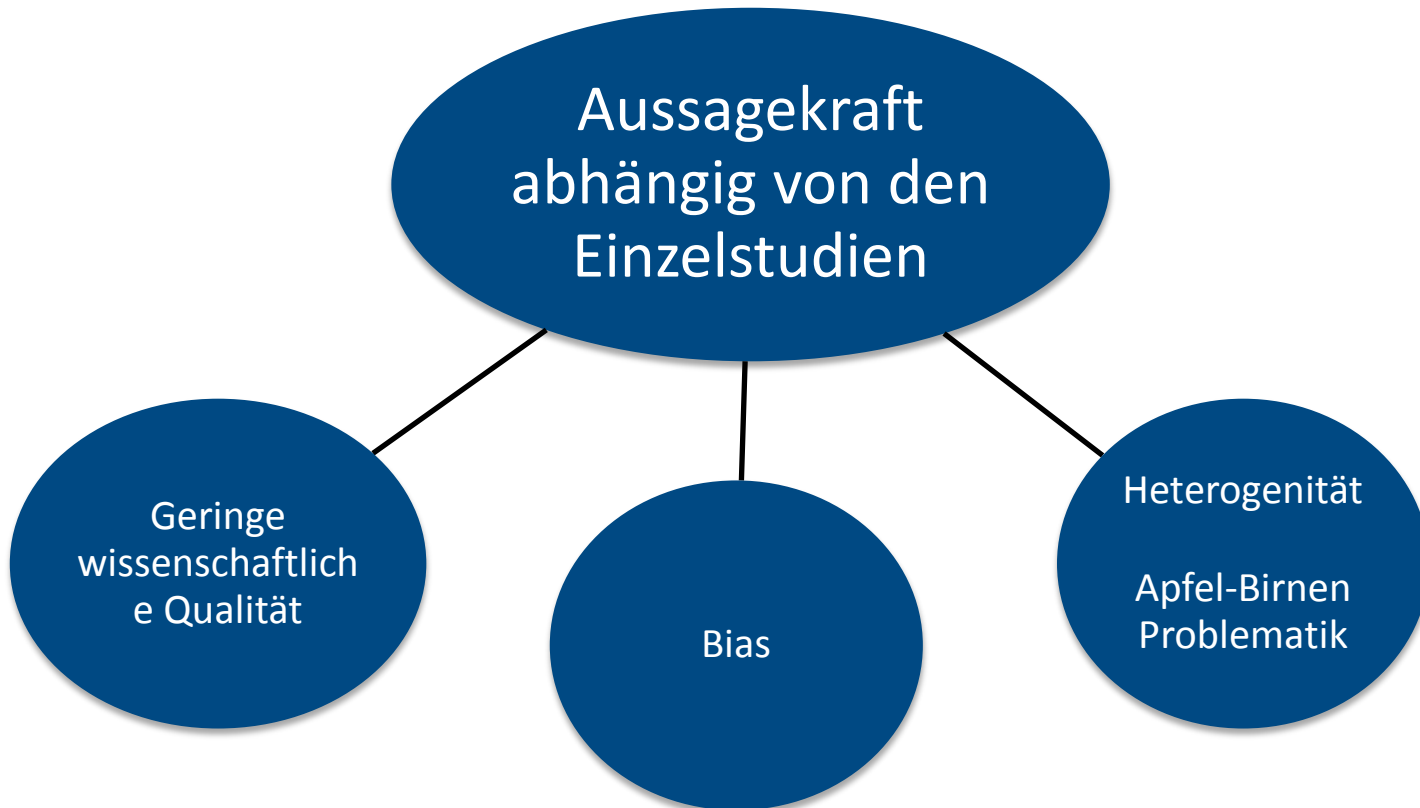
- *Haben die berücksichtigten Studien teilweise entgegengesetzte Resultate erbracht?*

Der Variabilität der Ergebnisse wird oft nicht genügend Beachtung geschenkt. Das Problem der Heterogenität von Studienresultaten kann nicht durch die Anwendung eines statistischen Tests gelöst werden. Wenn stark voneinander abweichende Resultate vorliegen, geben klinische und biologische Überlegungen den Ausschlag, ob überhaupt eine Meta-Analyse sinnvoll ist.

- *Wie «robust» sind die Ergebnisse der Meta-Analyse?*

Bei der Durchführung einer Meta-Analyse gibt es eine Reihe von teilweise *arbiträren* Entscheidungen (Ausschluss von Studien, Wahl der statistischen Methoden, Interpretation bestimmter Resultate). Das Resultat einer Meta-Analyse sollte von solchen Entscheidungen einigermaßen unabhängig sein. Wenn also die gleiche Meta-Analyse z.B. mit einer etwas anderen Studienausswahl oder mit einem anderen statistischen Verfahren durchgeführt wird, sollte sie ungefähr dasselbe Resultat ergeben.

- Mathematische-statistische Zusammenfassung von Einzelstudien
- Im Gegensatz dazu steht der narrative Review
- Meta-Analyse objektiver durch ihre Festlegung von Kriterien für die Auswahl von Primärstudien (gegebenenfalls sinkt aber die Studienanzahl)
- Relativ kostengünstig
- Erhöhung der Validität und Trennschärfe
- Ermittlung welche Eigenschaften zu welchen Effektstärken führen



# Meta-Analyse

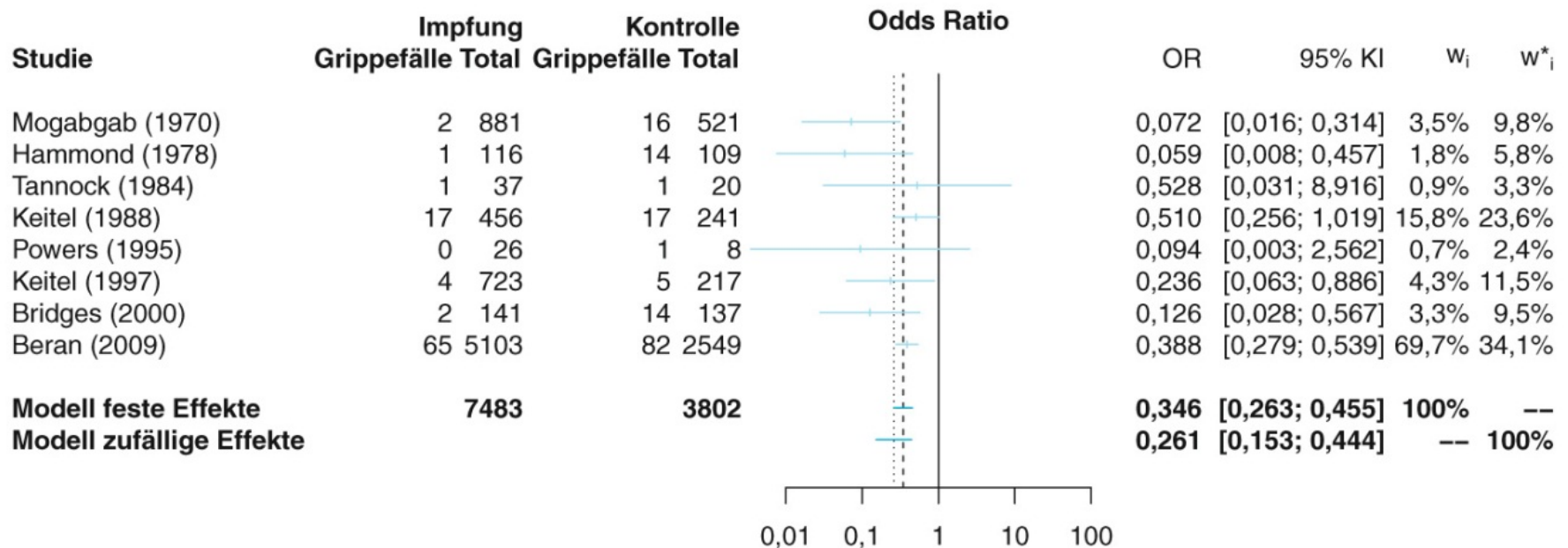
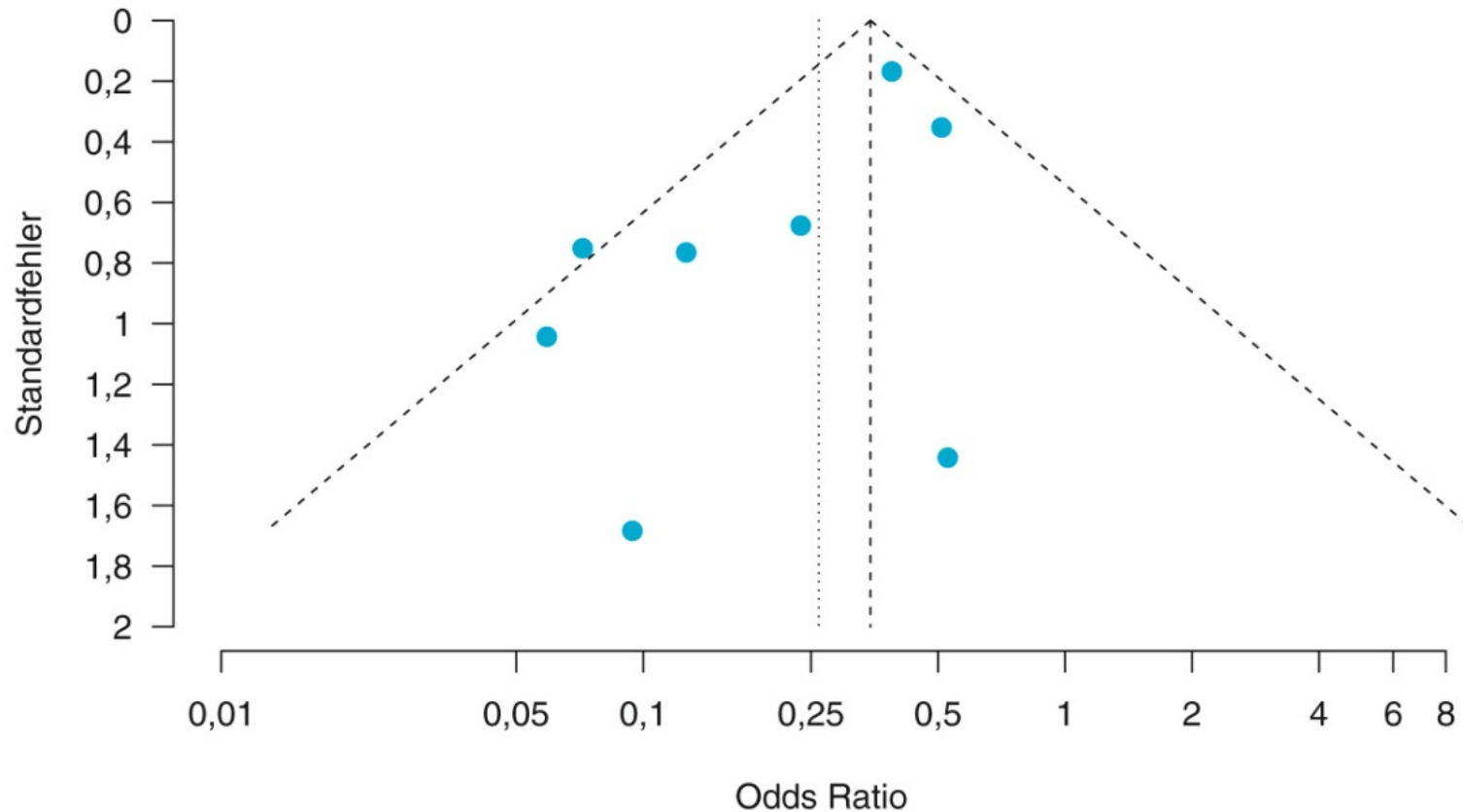


Abbildung 15.1: Forest-Plot der Meta-Analyse zur Grippeimpfung.

Geimpft	Grippenerkrankung	
	Ja	Nein
Ja	0 (0,5)	26 (26,5)
Nein	1 (1,5)	7 (7,5)

# Publications Bias



**Abbildung 15.3:** Funnel-Plot zur Beurteilung von Publikations-Bias.



# Stratifying the analysis by trial

- An important principle involved in estimating risk from a meta-analysis is that the randomized comparisons of the individual trials should be maintained when analyzing the combined data.
- In other words, when comparing drug A to drug B, subjects randomly assigned to drug A in a single trial are compared to subjects assigned to drug B from the same trial and not to subjects from other trials. In the statistics literature, this is referred to as stratifying the analysis by trial.

# Simpson's Paradox

**Table 1. An Illustration of Simpson's Paradox from Incorrect Pooling of Data**

Trial	<i>Drug A</i>			<i>Drug B</i>		
	Events	Patients	Risk	Events	Patients	Risk
1	1	100	1.0%	2	200	1.0%
2	1	100	1.0%	2	200	1.0%
3	200	1200	16.7%	50	300	16.7%
4	2	200	1.0%	2	200	1.0%
Total	204	1600	12.8%	56	900	6.2%

The hypothetical example in Table 1 illustrates an extreme example of Simpson's paradox in which, for each of four trials, the estimated risk of a safety event is identical for both Drug A and Drug B. With simple pooling, however, the risk for Drug A appears to be more than twice as high as that for Drug B (12.8 percent vs. 6.2 percent)

# Mantel-Haenszel Methode – stratifizierte Analyse

	Erfolgreich		Total
	Ja	Nein	
Therapie A	273 (78 %)	77	350
Therapie B	289 (83 %)	61	350

**Tabelle 7.3:** Analyse des Erfolgs zweier Therapien zur Nierensteinentfernung.

	Kleine Nierensteine			Große Nierensteine		
	Erfolgreich		Total	Erfolgreich		Total
	Ja	Nein		Ja	Nein	
Therapie A	81 (93 %)	6	87	192 (73 %)	71	263
Therapie B	234 (87 %)	36	270	55 (69 %)	25	80

**Tabelle 7.4:** Subgruppenanalyse des Erfolgs zweier Therapien zur Nierensteinentfernung in Abhängigkeit von der Nierensteingröße.

$$Odds\ Ratio_{Mantel - Haenszel} = \frac{\frac{81 \times 36}{357} + \frac{192 \times 25}{343}}{\frac{6 \times 234}{357} + \frac{71 \times 55}{343}} = 1,45 \quad (p=0,12)$$

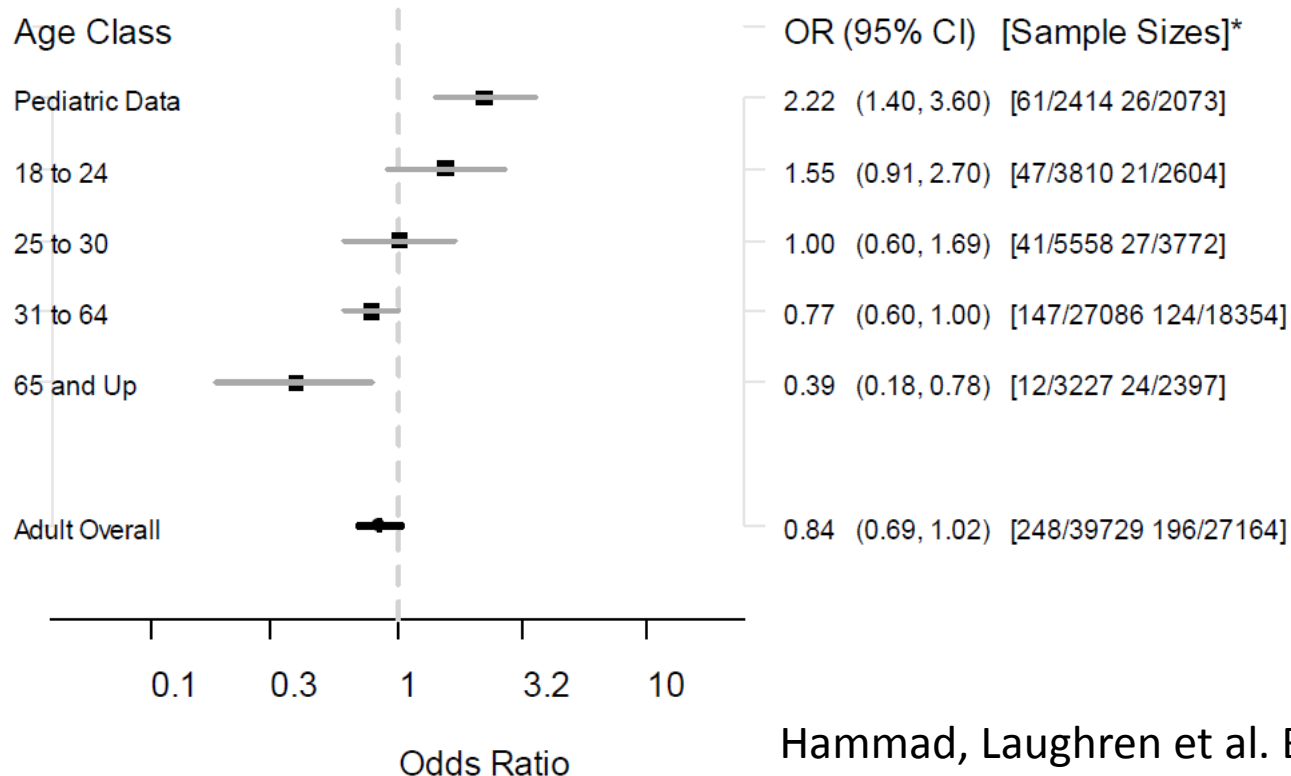
# Sensitivity Analysis

- The goal of any sensitivity analysis should not be to search for additional findings, but to support and understand the primary findings of the meta-analysis.
- For example, a meta-analysis that included one very large study contributing a large proportion of subjects and events could raise a concern that it was overly influencing the meta-analytic results. A sensitivity analysis that excluded that study would have reduced numbers of subjects and events and lower power to yield a significant finding, but a risk estimate that was consistent with the original estimate would add to the weight of evidence of the finding

# FDA Meta-Analysis of Antidepressants and Suicidal Behaviour



## Suicidal Behavior and Ideation Psychiatric Indications



Hammad, Laughren et al. BMJ 2009



Das Cochrane-Logo spiegelt die Ergebnisse eines Systematischen Cochrane Reviews mit Kultcharakter wider. In dem Review von 1989 ging es um die Frage, ob die Reifung der Lungen bei Frühgeborenen durch die Gabe von Kortikosteroiden unterstützt werden kann. Der Cartoonist David Mostyn kreierte aus den Studienergebnissen ein prägnantes Logo.

# MedCalc statistical software: Meta-analysis Introduction



- A meta-analysis integrates the quantitative findings from separate but similar studies and provides a numerical estimate of the overall effect of interest (Petrie et al., 2003).
- Different weights are assigned to the different studies for calculating the summary or pooled effect. The weighting is related with the inverse of the standard error (and therefore indirectly to the sample size) reported in the studies. Studies with smaller standard error and larger sample size are given more weight in the calculation of the pooled effect size.



# The effect of interest can be

- The effect of interest can be:
- an average of a continuous variable
- a correlation between two variables
- an odds ratio, suitable for analyzing retrospective studies
- a relative risk (risk ratio) or risk difference, suitable for analyzing prospective studies
- a proportion
- the area under the ROC curve



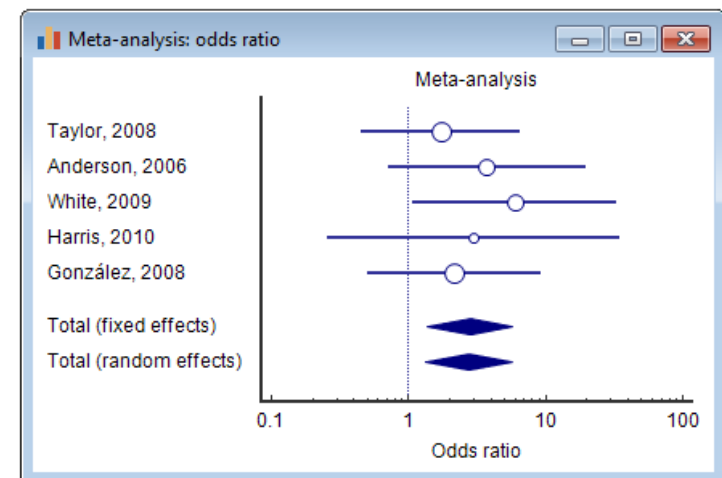
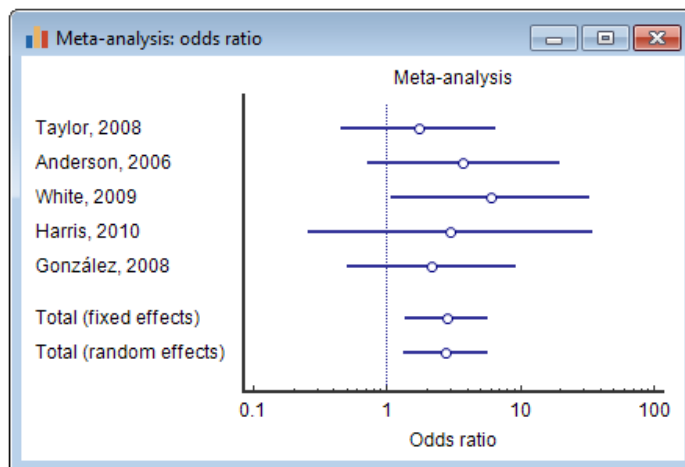
# Fixed and random effects model

- Under the fixed effects model, it is assumed that the studies share a common true effect, and the summary effect is an estimate of the common effect size.
- Under the random effects model the true effects in the studies are assumed to vary between studies and the summary effect is the weighted average of the effects reported in the different studies (Borenstein et al., 2009).
- The random effects model will tend to give a more conservative estimate (i.e. with wider confidence interval), but the results from the two models usually agree when there is no heterogeneity.
- When heterogeneity is present (see below) the random effects model should be the preferred model.

- Cochran's Q is the weighted sum of squares on a standardized scale. It is reported with a P value with low P-values indicating presence of heterogeneity. This test however is known to have low power to detect heterogeneity and it is suggested to use a value of 0.10 as a cut-off for significance (Higgins et al., 2003). Conversely, Q has too much power as a test of heterogeneity if the number of studies is large.
- $I^2$  statistics is the percentage of observed total variation across studies that is due to real heterogeneity rather than chance. It is calculated as  $I^2 = 100\% \times (Q - df)/Q$ , where Q is Cochran's heterogeneity statistic and df the degrees of freedom. Negative values of  $I^2$  are put equal to zero so that  $I^2$  lies between 0% and 100%. A value of 0% indicates no observed heterogeneity, and larger values show increasing heterogeneity (Higgins et al., 2003).

# Forest plot

The results of the different studies, with 95% CI, and the overall effect (under the fixed and random effects model) with 95% CI are illustrated in a graph called "forest plot", e.g.:

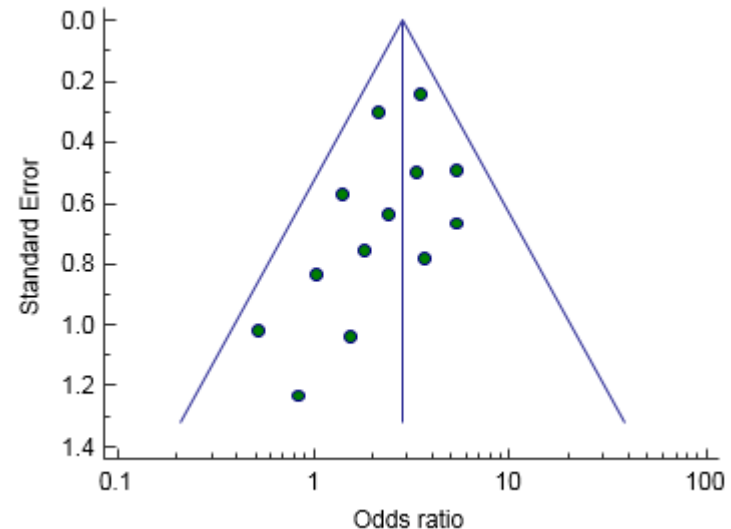
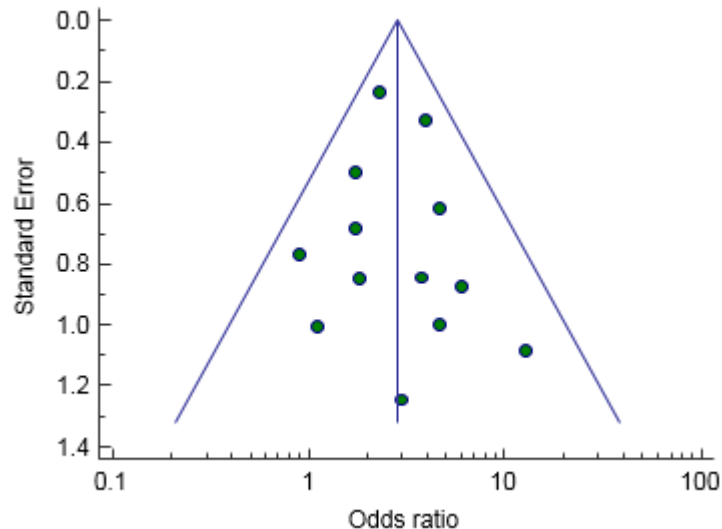


In this example the markers representing the effect size all have the same size. Optionally, the marker size may vary in size according to the weights assigned to the different studies. In addition, the pooled effects can be represented using a diamond. The location of the diamond represents the estimated effect size and the width of the diamond reflects the precision of the estimate,

# Funnel plot

- A funnel plot (Egger et al., 1997) is a graphical tool for detecting bias in meta-analysis.
- In a funnel plot treatment effect is plotted on the horizontal axis and MedCalc plots the standard error on the vertical axis (Sterne & Egger, 2001).
- The vertical line represents the summary estimated derived using fixed-effect meta-analysis.
- Two diagonal lines represent (pseudo) 95% confidence limits (effect  $\pm 1.96$  SE) around the summary effect for each standard error on the vertical axis. These show the expected distribution of studies in the absence of heterogeneity or of selection bias. In the absence of heterogeneity, 95% of the studies should lie within the funnel defined by these diagonal lines.

# Funnel plot



Publication bias results in asymmetry of the funnel plot. If publication bias is present, the smaller studies will show the larger effects. See Sterne et al. (2011) for interpreting funnel plot asymmetry. The funnel plot may not always be a reliable tool, in particular when the number of studies included in the analysis is small.

# References

- Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2009) Introduction to meta-analysis. Chichester, UK: Wiley.
- Egger M, Smith GD, Schneider M, Minder C (1997) Bias in meta-analysis detected by a simple, graphical test. *BMJ* 315: 629–634.
- Higgins JP, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. *BMJ* 327:557-560.
- Petrie A, Bulman JS, Osborn JF (2003) Further statistics in dentistry. Part 8: systematic reviews and meta-analyses. *British Dental Journal* 194:73-78.
- Sterne JA, Egger E (2001) Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of Clinical Epidemiology* 54:1046–1055.
- Sterne JA, Sutton AJ, Ioannidis JP et al. (2011) Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011;343:d4002.
- DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. *Controlled Clinical Trials* 7:177-188
- Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from the retrospective analysis of disease. *Journal of the National Cancer Institute* 22: 719-748.
- Hedges LV, Olkin I (1985) Statistical methods for meta-analysis. London: Academic Press.
- Zhou XH, NA Obuchowski, DK McClish (2002) Statistical methods in diagnostic medicine. New York: Wiley.



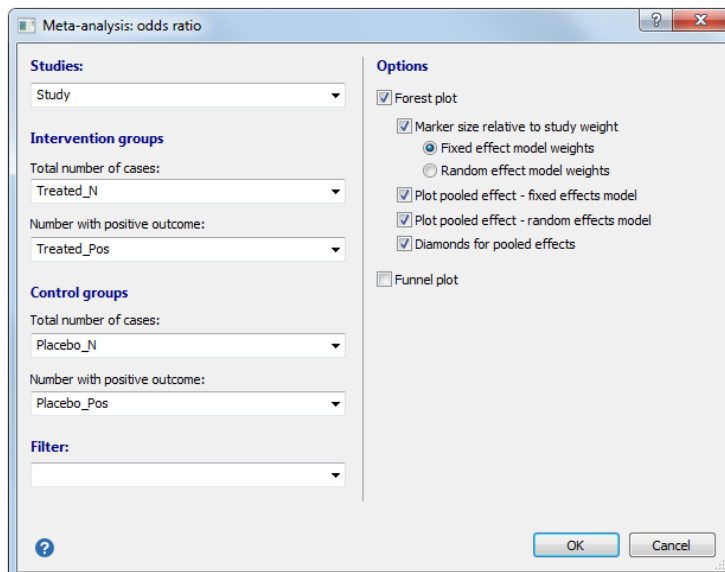
# Meta-analysis in MedCalc

---

- Continuous measure
- Correlation
- Proportion
- Relative risk
- Risk difference
- Odds ratio
- Area under ROC curve
- Generic inverse variance method

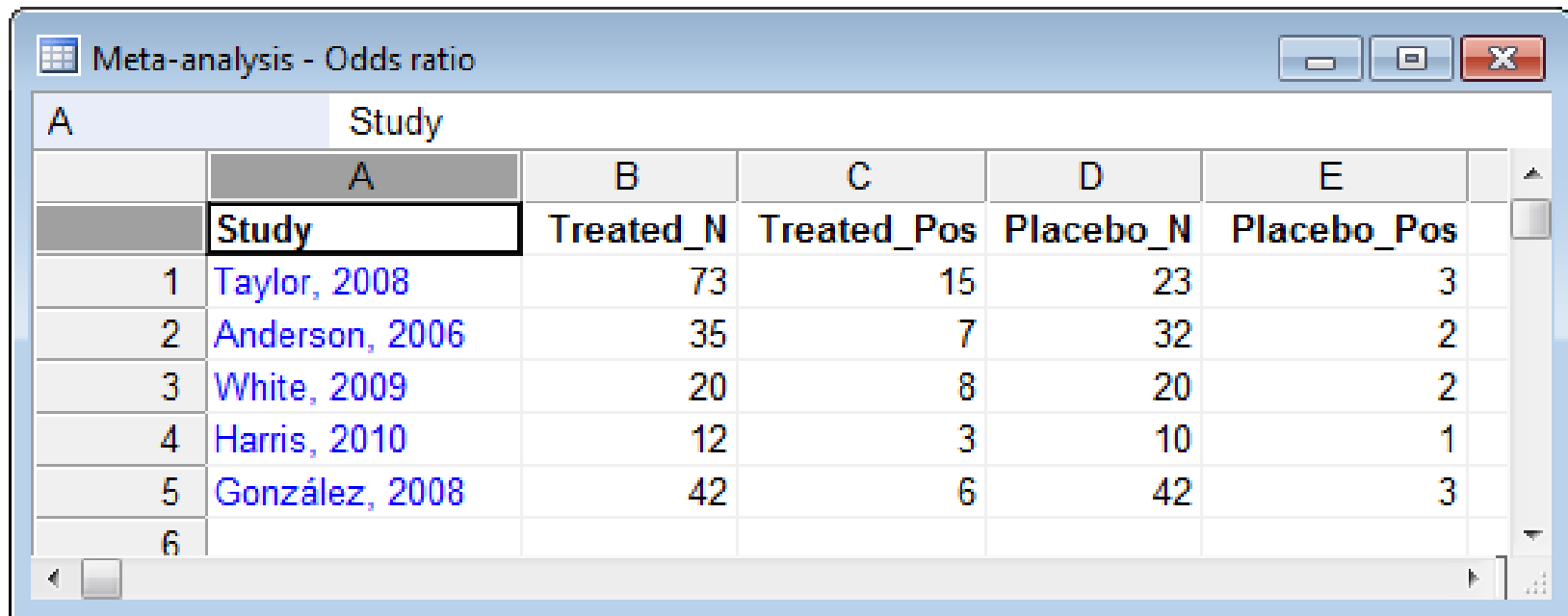
# Meta-analysis: odds ratio

MedCalc uses the Mantel-Haenszel method (Mantel & Haenszel, 1959) for calculating the weighted pooled odds ratio under the fixed effects model. Next the heterogeneity statistic is incorporated to calculate the summary odds ratio under the random effects model (DerSimonian & Laird, 1986).





# Meta-analysis: odds ratio



A	Study	B	C	D	E
	Study	Treated_N	Treated_Pos	Placebo_N	Placebo_Pos
1	Taylor, 2008	73	15	23	3
2	Anderson, 2006	35	7	32	2
3	White, 2009	20	8	20	2
4	Harris, 2010	12	3	10	1
5	González, 2008	42	6	42	3
6					

# Meta-analysis: relative risk and risk difference



MedCalc uses the Mantel-Haenszel method (based on Mantel & Haenszel, 1959) for calculating the weighted pooled relative risk and risk difference under the fixed effects model. Next the heterogeneity statistic is incorporated to calculate the summary relative risk under the random effects model (DerSimonian & Laird, 1986).

The screenshot shows the 'Meta-analysis: relative risk' dialog box. It is divided into two main sections: 'Studies' and 'Options'.

**Studies:**

- Intervention groups:**
  - Total number of cases: Treated\_Total
  - Number with positive outcome: Treated\_positive
- Control groups:**
  - Total number of cases: Controls\_total
  - Number with positive outcome: Controls\_positive
- Filter:** (Empty dropdown)

**Options:**

- Forest plot
  - Marker size relative to study weight
    - Fixed effect model weights
    - Random effect model weights
  - Plot pooled effect - fixed effects model
  - Plot pooled effect - random effects model
  - Diamonds for pooled effects
- Funnel plot

Buttons: OK, Cancel

# Meta-analysis: relative risk



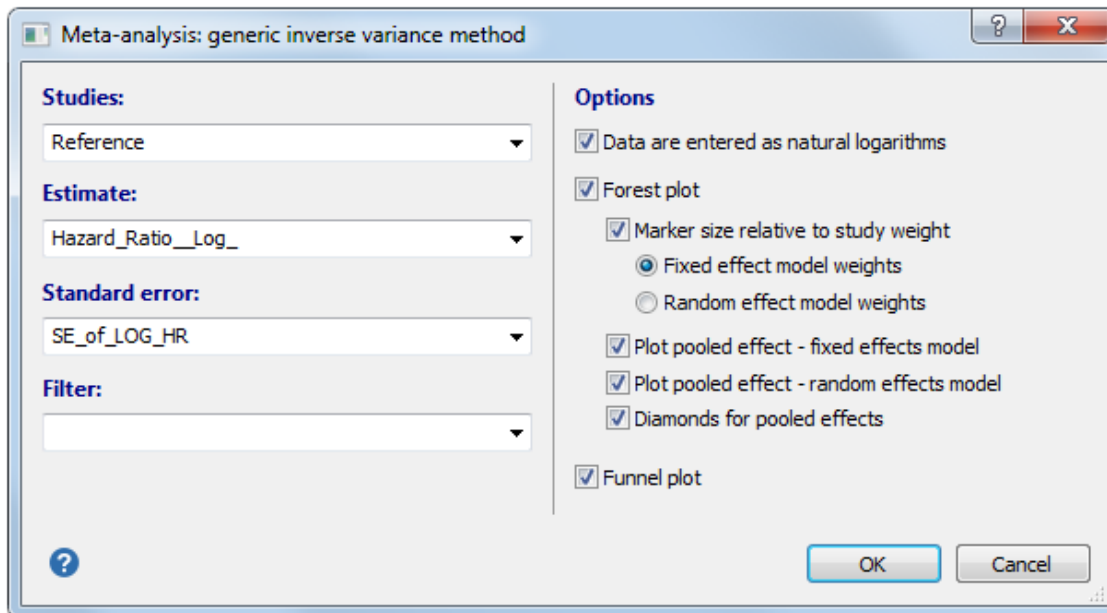
Meta-analysis - Risk ratio & difference

A	Study	B	C	D	E
	Study	Treated_positive	Treated_Total	Controls_positive	Controls_total
1	佐藤, 2012	95	103	47	104
2	渡辺, 2008	119	127	34	129
3	山本, 2013	51	223	12	76
4	長谷川, 2006	122	139	61	142
5	田村, 2006	47	53	10	51
6	小野, 2004	121	135	29	68
7	菊地, 2008	337	378	170	376
8					

# Meta-analysis: hazard ratio (generic inverse variance method)

Estimates and their standard errors are entered directly. For ratio measures of intervention effect, the data should be entered as natural logarithms (for example as a log Hazard ratio and the standard error of the log Hazard ratio).

In the inverse variance method the weight given to each study is the inverse of the variance of the effect estimate (i.e. one over the square of its standard error). Thus larger studies are given more weight than smaller studies, which have larger standard errors.



# Meta-analysis: hazard ratio (generic inverse variance method)



Meta-analysis - Generic

A		Reference	
	A	B	C
	Reference	Hazard_Ratio_Log	SE_of_LOG_HR
1	Study 1	-0.077	0.212
2	Study 2	0.012	0.221
3	Study 3	0.323	0.426
4	Study 4	0.154	0.230
5	Study 5	0.051	0.348
6	Study 6	-0.661	0.232
7	Study 7	-0.199	0.337
8	Study 8	0.040	0.245
9	Study 9	0.305	0.432
10			

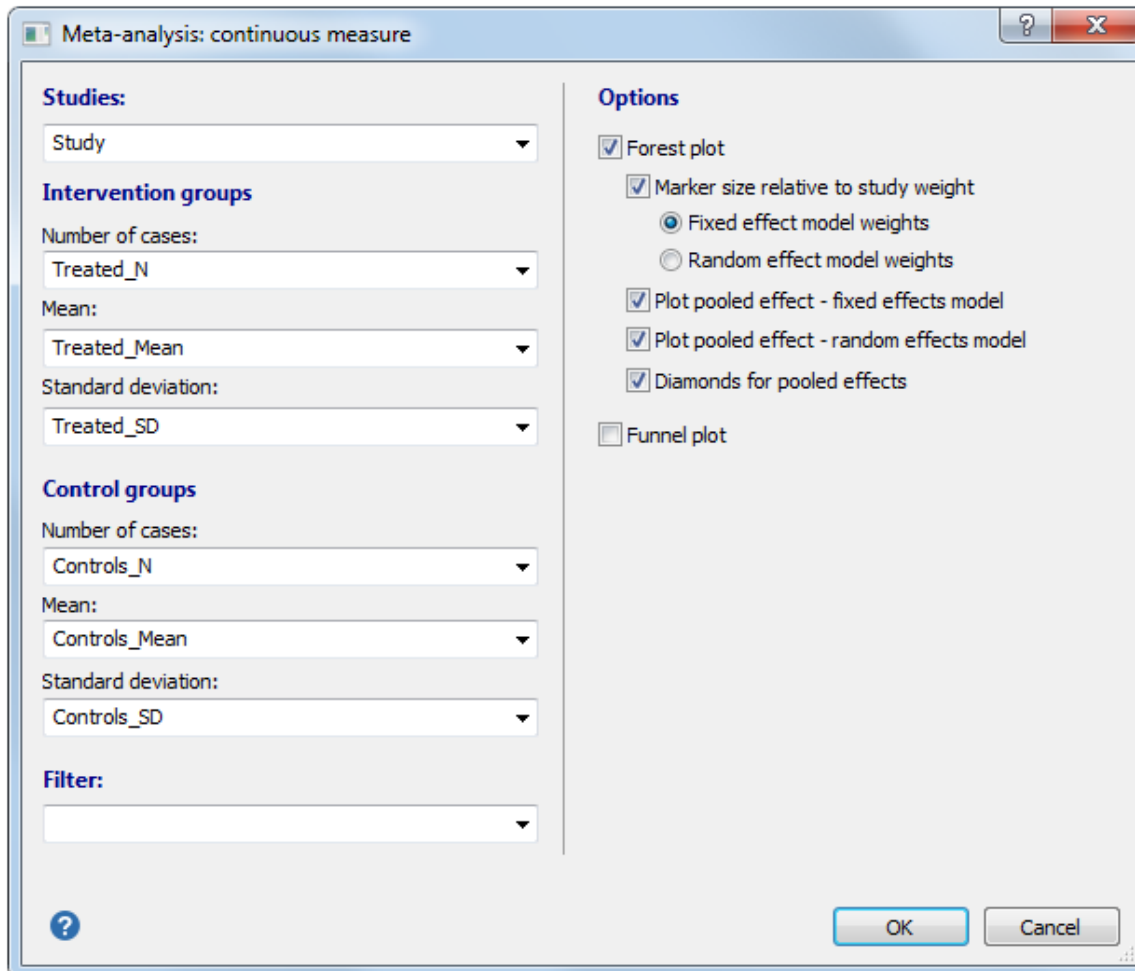
# Meta-analysis: continuous measure

---

For meta-analysis of studies with comparison of means between treated cases and controls, MedCalc uses the Hedges  $g$  statistic as a formulation for the standardized mean difference under the fixed effects model. Next, the heterogeneity statistic is incorporated to calculate the summary standardized mean difference under the random effects model (DerSimonian & Laird, 1986).

The standardized mean difference Hedges  $g$  is the difference between the two means divided by the pooled standard deviation, with a correction for small sample bias.

# Meta-analysis: continuous measure



The screenshot shows a dialog box titled "Meta-analysis: continuous measure". It is divided into two main sections: "Studies:" and "Options".

**Studies:**

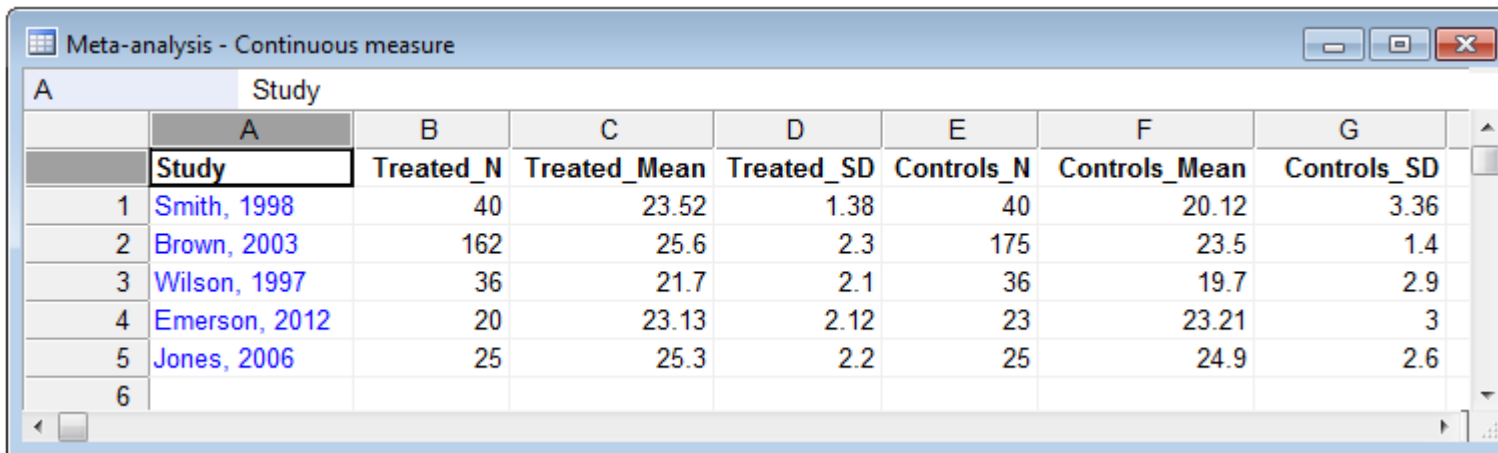
- Study:** A dropdown menu with "Study" selected.
- Intervention groups:**
  - Number of cases: "Treated\_N"
  - Mean: "Treated\_Mean"
  - Standard deviation: "Treated\_SD"
- Control groups:**
  - Number of cases: "Controls\_N"
  - Mean: "Controls\_Mean"
  - Standard deviation: "Controls\_SD"
- Filter:** An empty dropdown menu.

**Options:**

- Forest plot
  - Marker size relative to study weight
    - Fixed effect model weights
    - Random effect model weights
  - Plot pooled effect - fixed effects model
  - Plot pooled effect - random effects model
  - Diamonds for pooled effects
- Funnel plot

Buttons: "OK" and "Cancel".

# Meta-analysis: continuous measure



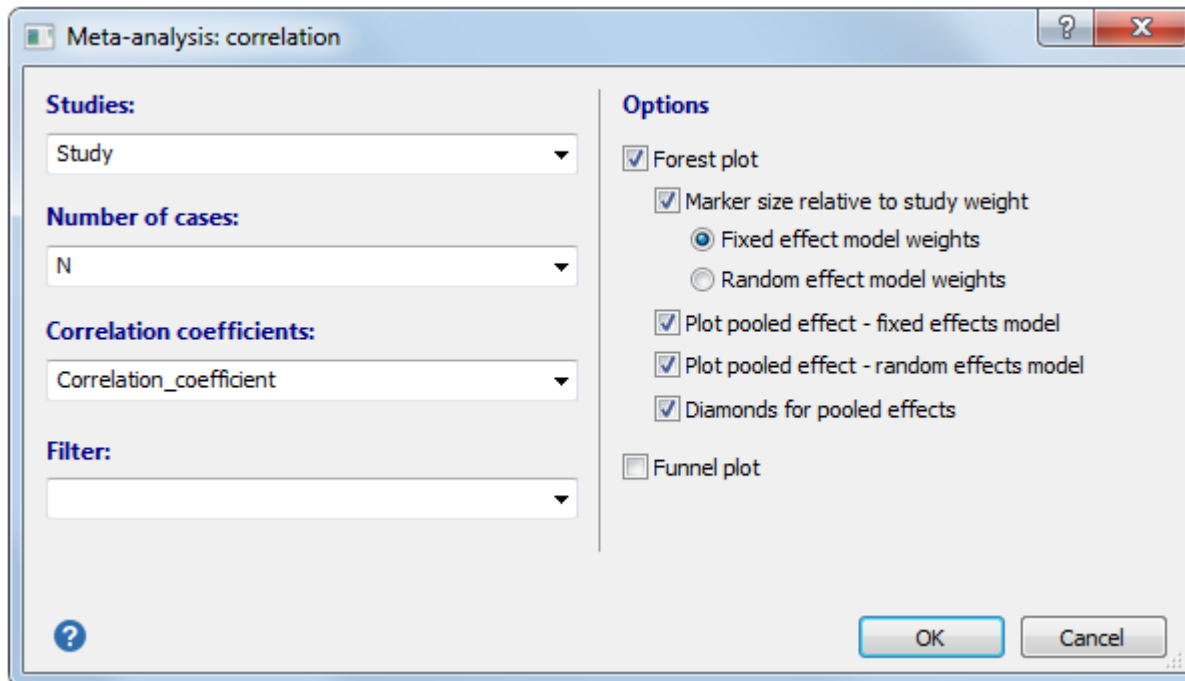
The screenshot shows a software window titled "Meta-analysis - Continuous measure" with a table of study data. The table has columns for Study, Treated\_N, Treated\_Mean, Treated\_SD, Controls\_N, Controls\_Mean, and Controls\_SD. The data is as follows:

	A	B	C	D	E	F	G
	Study	Treated_N	Treated_Mean	Treated_SD	Controls_N	Controls_Mean	Controls_SD
1	Smith, 1998	40	23.52	1.38	40	20.12	3.36
2	Brown, 2003	162	25.6	2.3	175	23.5	1.4
3	Wilson, 1997	36	21.7	2.1	36	19.7	2.9
4	Emerson, 2012	20	23.13	2.12	23	23.21	3
5	Jones, 2006	25	25.3	2.2	25	24.9	2.6
6							

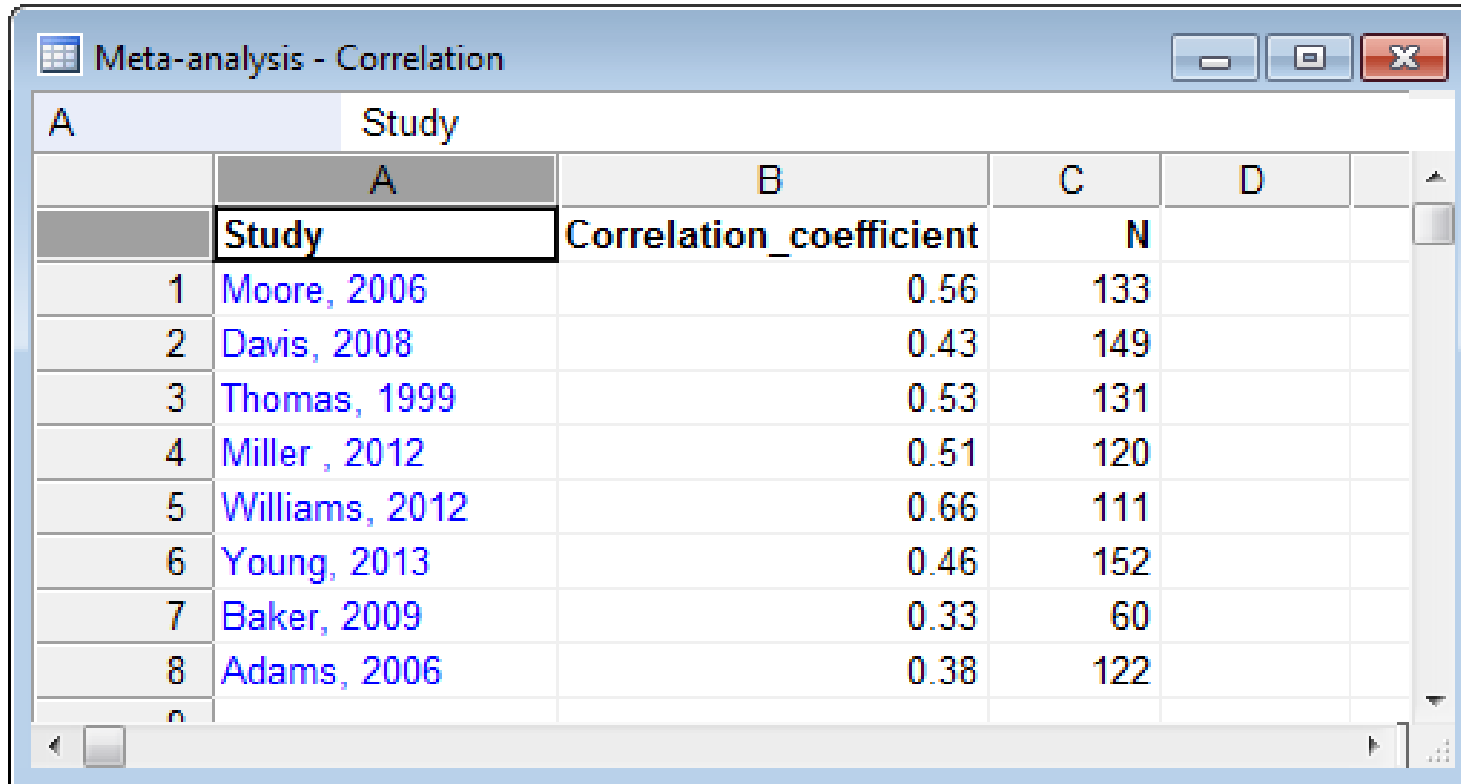


# Meta-analysis: correlation

MedCalc uses the Hedges-Olkin (1985) method for calculating the weighted summary Correlation coefficient under the fixed effects model, using a Fisher Z transformation of the correlation coefficients. Next, the heterogeneity statistic is incorporated to calculate the summary Correlation coefficient under the random effects model (DerSimonian and Laird, 1986).



# Meta-analysis: correlation

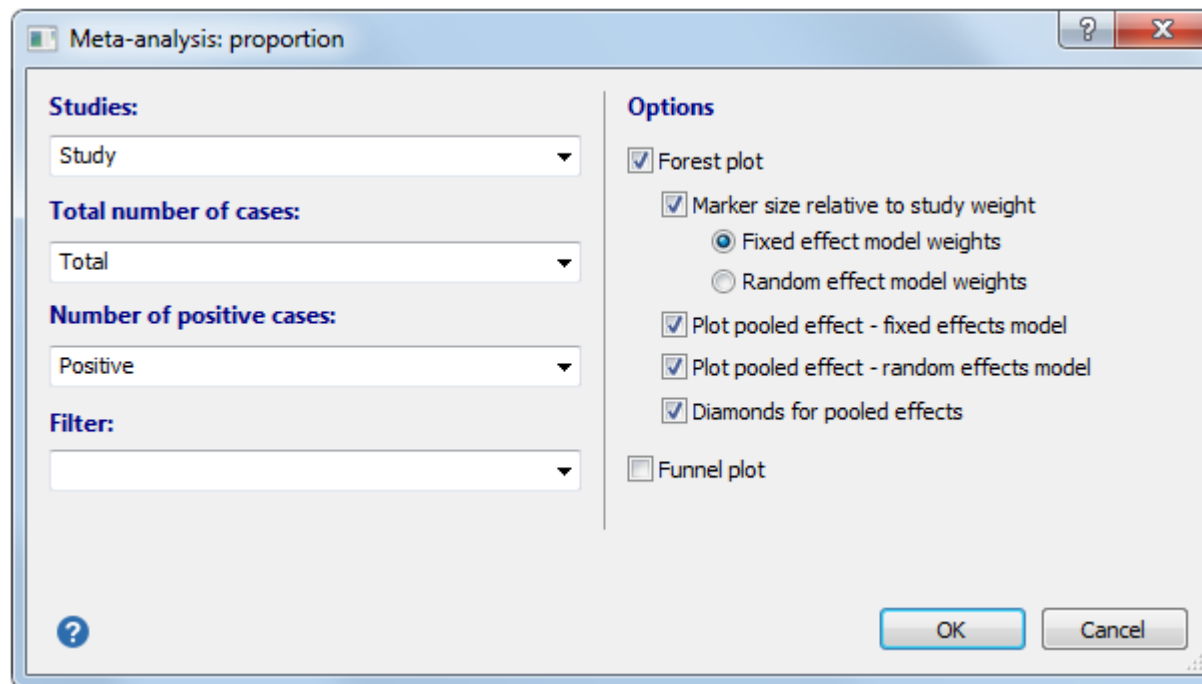


The screenshot shows a spreadsheet window titled "Meta-analysis - Correlation". The spreadsheet contains a table with the following data:

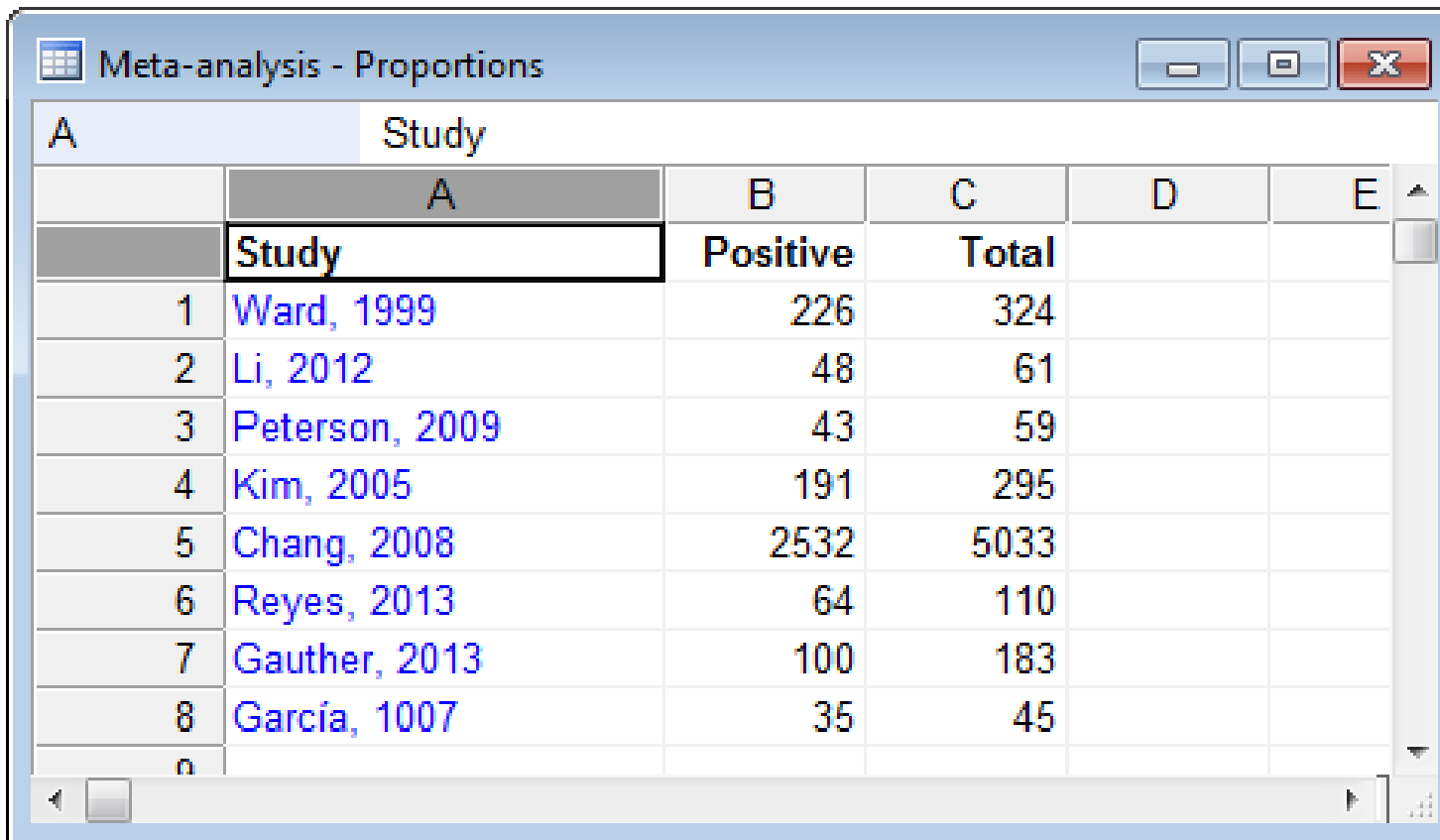
A	Study	B	C	D
	Study	Correlation_coefficient	N	
1	Moore, 2006	0.56	133	
2	Davis, 2008	0.43	149	
3	Thomas, 1999	0.53	131	
4	Miller, 2012	0.51	120	
5	Williams, 2012	0.66	111	
6	Young, 2013	0.46	152	
7	Baker, 2009	0.33	60	
8	Adams, 2006	0.38	122	

# Meta-analysis: proportions

MedCalc uses a Freeman-Tukey transformation (arcsine square root transformation; Freeman and Tukey, 1950) to calculate the weighted summary proportion under the fixed and random effects model (DerSimonian & Laird, 1986).



# Meta-analysis: proportions

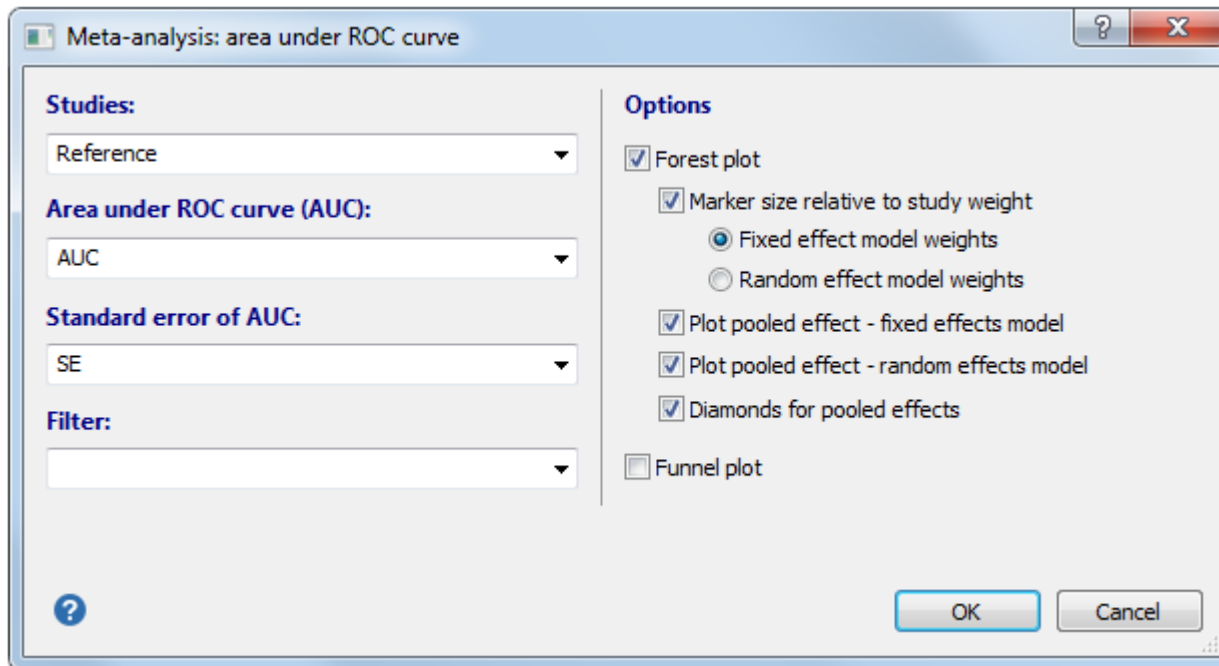


The screenshot shows a software window titled "Meta-analysis - Proportions" with a table of study data. The table has columns labeled A, B, C, D, and E. The data is as follows:

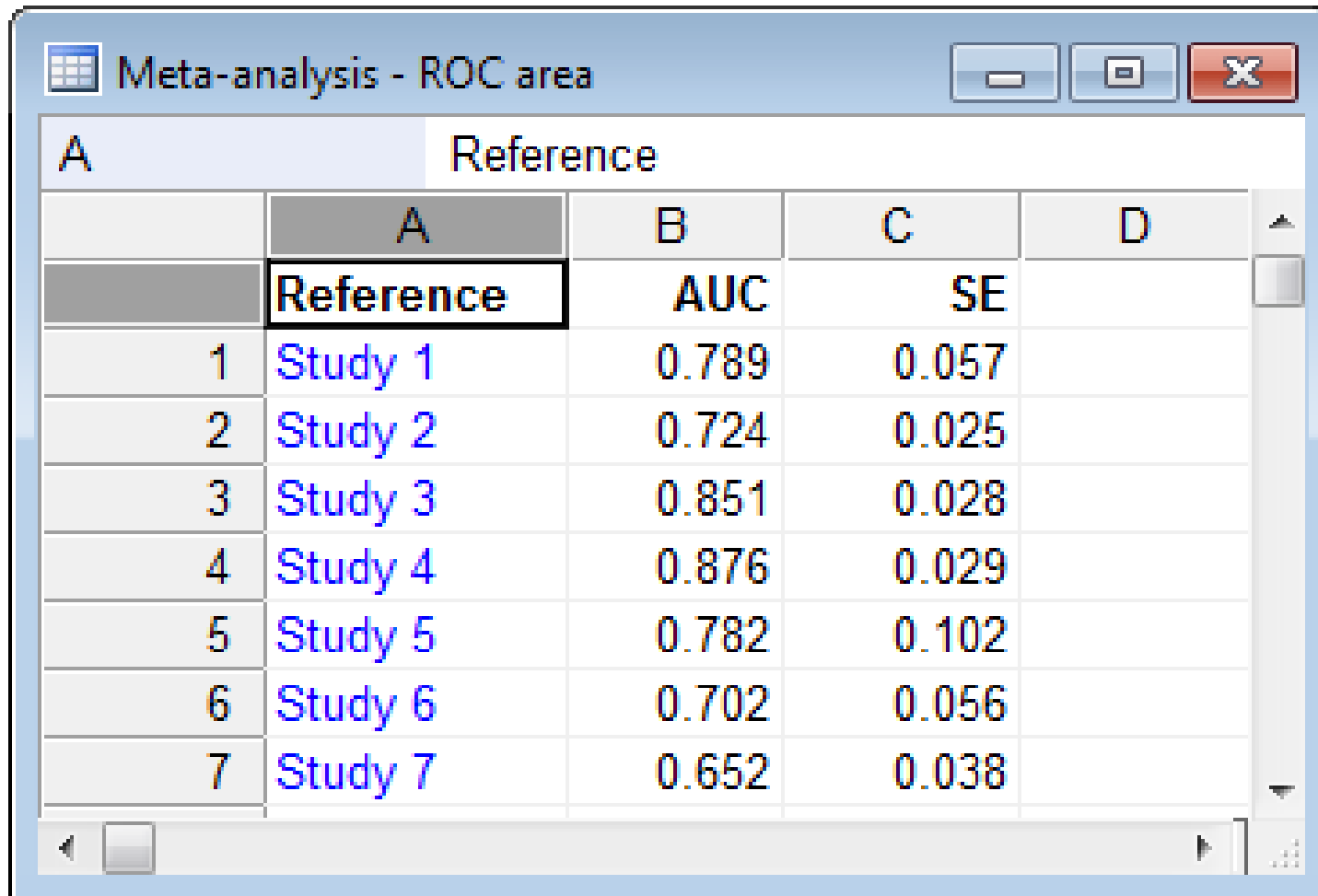
	A	B	C	D	E
Study	Positive	Total			
1	Ward, 1999	226	324		
2	Li, 2012	48	61		
3	Peterson, 2009	43	59		
4	Kim, 2005	191	295		
5	Chang, 2008	2532	5033		
6	Reyes, 2013	64	110		
7	Gauthier, 2013	100	183		
8	García, 1007	35	45		
9					

# Meta-analysis: AUC

MedCalc uses the methods described by Zhou et al. (2002) for calculating the weighted summary Area under the ROC curve under the fixed effects model and random effects model.



# Meta-analysis: AUC



The screenshot shows a software window titled "Meta-analysis - ROC area". The window contains a table with the following data:

A		Reference		
	A	B	C	D
	Reference	AUC	SE	
1	Study 1	0.789	0.057	
2	Study 2	0.724	0.025	
3	Study 3	0.851	0.028	
4	Study 4	0.876	0.029	
5	Study 5	0.782	0.102	
6	Study 6	0.702	0.056	
7	Study 7	0.652	0.038	

# Lionheart+Levorep



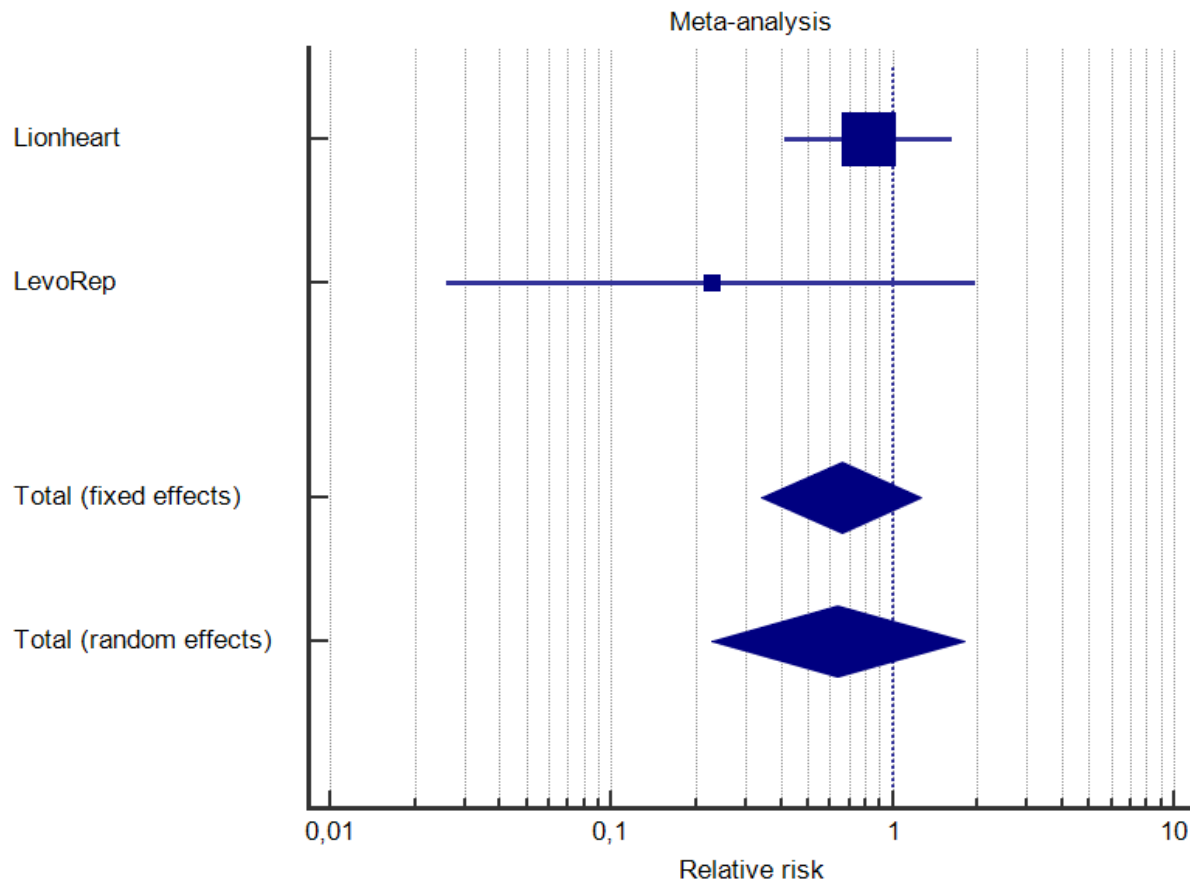
## Meta-analysis: relative risk

Variable for studies	study								
1. Intervention groups									
Variable for total number of cases	Levosimendan_N								
Variable for number of positive cases	Levosimendan_Deaths								
2. Control groups									
Variable for total number of cases	Placebo_N								
Variable for number of positive cases	Placebo_Deaths								
Study	Intervention	Controls	Relative risk	95% CI	z	P	Weight (%)		
							Fixed	Random	
Lionheart	15/48	8/21	0,820	0,412 to 1,632			90,80	80,61	
LevoRep	1/63	4/57	0,226	0,0260 to 1,965			9,20	19,39	
Total (fixed effects)	16/111	12/78	0,658	0,341 to 1,269	-1,250	0,211	100,00	100,00	
Total (random effects)	16/111	12/78	0,639	0,227 to 1,800	-0,848	0,397	100,00	100,00	

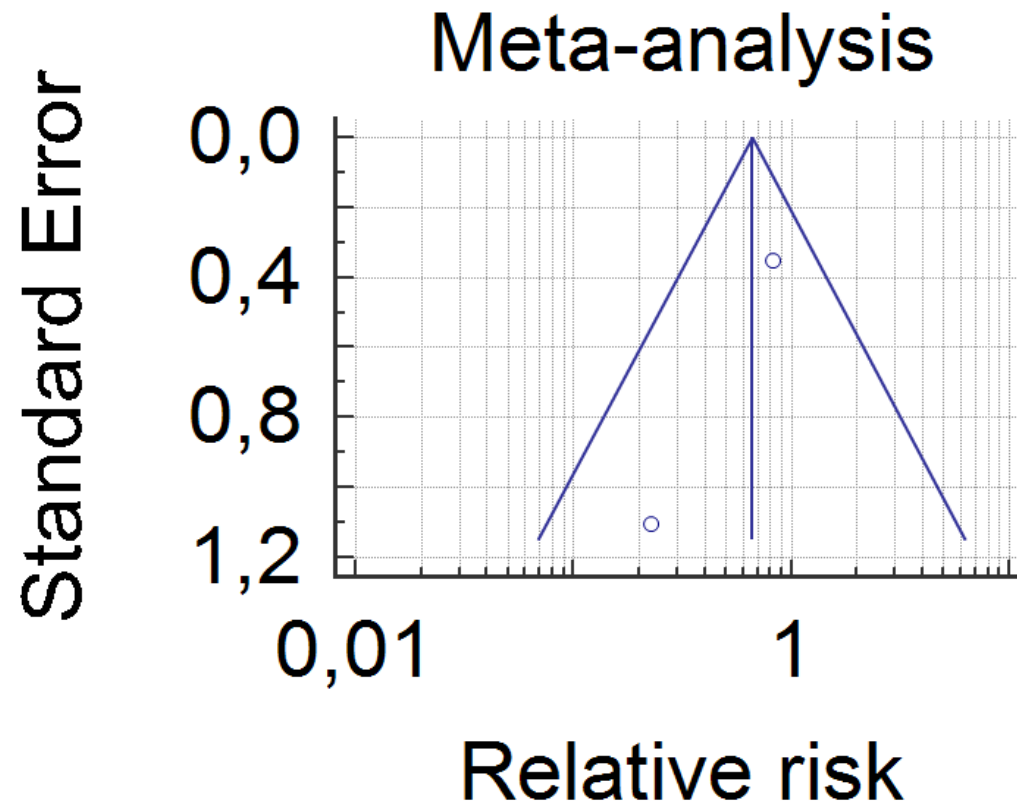
## Test for heterogeneity

Q	1,3331
DF	1
Significance level	P = 0,2483
I <sup>2</sup> (inconsistency)	24,98%
95% CI for I <sup>2</sup>	0,00 to 0,00

# Lionheart+Levorep







# Lionheart+Levorep



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK



European Journal of Heart Failure (2018) 20, 1120–1135  
doi:10.1093/ehj/ehf1145

RESEARCH ARTICLE

## Efficacy and safety of intermittent intravenous outpatient administration of levosimendan in patients with advanced heart failure: the LION-HEART multicentre randomised trial

Josep Comin-Colet<sup>1,2\*</sup>, Nicolás Manito<sup>3</sup>, Javier Segovia-Cubero<sup>3</sup>, Juan Delgado<sup>4</sup>, José Manuel García Pinilla<sup>5</sup>, Luis Almenar<sup>6</sup>, María G. Crespo-Leiro<sup>7</sup>, Alessandro Sionis<sup>8</sup>, Teresa Blasco<sup>9</sup>, Domingo Pascual-Figal<sup>10</sup>, Francisco Gonzalez-Vilchez<sup>11</sup>, José Luis Lambert-Rodriguez<sup>12</sup>, María Grau<sup>13</sup>, and Jordi Bruguera<sup>1</sup>, on behalf of the LION-HEART Study Investigators

<sup>1</sup>Heart Disease Biomedical Research Group, IMH (Hospital del Mar Medical Research Institute), and Universitat Autònoma de Barcelona, Barcelona, Spain; <sup>2</sup>Heart Disease Institute, Hospital Universitari de Bellvitge, IDIBELL, University of Barcelona, L'Hospitalet del Llobregat, Spain; <sup>3</sup>Hospital Universitario Reina del Mar, Madrid, Spain; <sup>4</sup>Unidad de Tratamiento Coronario y Preload, Servicio de Cardiología, Hospital 12 de Octubre, Madrid, Spain; <sup>5</sup>Hospital Universitario Virgen de la Victoria, Málaga, Spain; <sup>6</sup>Hospital Universitario y Politécnico La Fe, Valencia, Spain; <sup>7</sup>Complejo Hospitalario Universitario de A Coruña (CHUAC) e Instituto de Investigación Biomédica de A Coruña (IBIAC), Universidad de A Coruña (UCA), A Coruña, Spain; <sup>8</sup>Hospital de la Santa Cruz (HSC), San Rafael Research Institute III, San Rafael, Universidad Autónoma de Barcelona, Barcelona, Spain; <sup>9</sup>Hospital Universitario Miguel Serres, Zaragoza, Spain; <sup>10</sup>Hospital Universitario Virgen de la Arrixaca, Murcia, Spain; <sup>11</sup>Hospital Universitario Hospital de Valme, Universidad de Córdoba, Córdoba, Spain; <sup>12</sup>Unidad de Tratamiento Coronario e Insuficiencia Cardíaca, Hospital Universitario Central de Asturias, Oviedo, Spain; and <sup>13</sup>Cardiovascular Epidemiology & Care Unit, IMH (Hospital del Mar Medical Research Institute) and University of Barcelona, Barcelona, Spain

Received 7 November 2016; revised 4 January 2017; accepted 1 January 2018; online publication date 1 February 2018

**Aims** The LION-HEART study was a multicentre, double-blind, randomised, parallel-group, placebo-controlled trial evaluating the efficacy and safety of intravenous administration of intermittent doses of levosimendan in outpatients with advanced chronic heart failure.

**Methods and results** Sixty-nine patients from 12 centres were randomly assigned at a 2:1 ratio to levosimendan or placebo groups, receiving treatment by a 6-hour intravenous infusion (0.2 µg/kg/min without bolus) every 2 weeks for 12 weeks. The primary endpoint was the effect on serum concentrations of N-terminal pro-B-type natriuretic peptide (NT-proBNP) throughout the treatment period in comparison with placebo. Secondary endpoints included evaluation of safety, clinical events and health-related quality of life (HRQL). The area under the curve (AUC<sub>0-6h</sub>, pg day/mL) of the levels of NT-proBNP over time for patients who received levosimendan was significantly lower than for the placebo group [ $344 \times 10^3$  (95% confidence interval (CI) 283–404) vs.  $535 \times 10^3$  (443–626) (P = 0.003)]. In comparison with the placebo group, the patients on levosimendan experienced a reduction in the rate of heart failure hospitalisation (hazard ratio 0.25; 95% CI 0.11–0.56; P = 0.001). Patients on levosimendan were less likely to experience a clinically significant decline in HRQL over time (P = 0.022). Adverse event rates were similar in the two treatment groups.

**Conclusions** In this small pilot study, intermittent administration of levosimendan to ambulatory patients with advanced systolic heart failure reduced plasma concentrations of NT-proBNP, worsening of HRQL, and hospitalisation for heart failure. The efficacy and safety of this intervention should be confirmed in larger trials.

**Keywords** Levosimendan • Pulsed infusions • Outpatient setting • Advanced heart failure • Safety • Natriuretic peptide

\*Corresponding author: IMH (Hospital del Mar Medical Research Institute), Carrer Doctor Aiguader, 38–39, 08003 Barcelona, Spain. Tel: +34 93 3463136; Fax: +34 93 3463396; Email: jcomin@imh.cat

© 2018 The Authors  
European Journal of Heart Failure © 2018 European Society of Cardiology



European Journal of Heart Failure (2014) 16, 898–906  
doi:10.1093/ehj/ehf118

## Efficacy and safety of the pulsed infusions of levosimendan in outpatients with advanced heart failure (LevoRep) study: a multicentre randomized trial

Johann Altenberger<sup>1</sup>, John T. Parissis<sup>2</sup>, Angelika Costard-Jaeckle<sup>3</sup>, Andreas Winter<sup>4</sup>, Christian Ebner<sup>5</sup>, Apostolos Karavadas<sup>6</sup>, Kurt Silbersch<sup>7</sup>, Ekaterini Avgeropoulou<sup>8</sup>, Thomas Weber<sup>9</sup>, Lida Dimopoulos<sup>10</sup>, Hanno Ulmer<sup>11</sup>, and Gerhard Poelzl<sup>12\*</sup>

<sup>1</sup>Cardiac Rehabilitation Center Griesganger, Paroissensanatsambulanz and Department of Cardiology, Paroisse Medical (PH), Salzburg, Austria; <sup>2</sup>Second Cardiology Department and Heart Failure Unit, University of Athens Medical School, Athens University Hospital, Athens, Greece; <sup>3</sup>Department of Cardiology, University Heart Center Hamburg, Hamburg, Germany; <sup>4</sup>Department of Cardiology, Hospital of the State of Carinthia, Lienz, Austria; <sup>5</sup>Department of Cardiology, St. Elizabeth Hospital, Lienz, Austria; <sup>6</sup>Cardiology Department, G. Gerovassiliou General Hospital, Athens, Greece; <sup>7</sup>Department of Cardiology, General Hospital Lienz, Austria; <sup>8</sup>Department of Cardiology, Epitaphion General Hospital, Athens, Greece; <sup>9</sup>Department of Cardiology, Graz-Karl-Feld-Klinik, Austria; <sup>10</sup>Department of Cardiology, Danube Hospital, Vienna, Austria; <sup>11</sup>Department of Medical Statistics, Informatics and Health Economics, Innsbruck Medical University, Innsbruck, Austria; and <sup>12</sup>Department of Cardiology and Angiology, Innsbruck Medical University, Innsbruck, Austria  
Received 28 February 2014; revised 12 April 2014; accepted 20 April 2014; online publication date 17 June 2014

**Aims** The aim of this study was to determine whether intermittent ambulatory treatment with levosimendan would improve functional capacity, quality of life, and event-free survival in patients with advanced heart failure.

**Methods and results** This was a prospective, randomized, double-blind, placebo-controlled, multicentre, parallel-group trial of pulsed infusions of levosimendan in 120 outpatients with advanced heart failure (EF <35%, NYHA class III or IV). The study was conducted at 11 centres in Austria, Greece, and Germany. Levosimendan (0.2 µg/kg/min) or placebo was administered for 6 h at 2-week intervals over 6 weeks, in addition to standard care therapy. The primary outcome was the proportion of patients with a >20% improvement in the 6-min walk test and a >15% score increase on the Kansas City Cardiomyopathy Questionnaire at the end of the 24-week study period. Secondary outcomes included event-free survival after 24 weeks. Analyses were performed on an intention-to-treat basis. The primary endpoint was reached in 19% of patients receiving levosimendan and 11.8% of patients receiving placebo (odds ratio 1.28; 95% confidence interval 0.44–3.59; P = 0.670). Cardiac death (four vs. one), heart transplants (two vs. one), and acute heart failure (14 vs. nine) were more frequent with placebo as compared with levosimendan. The incidence of side effects was comparable between groups.

**Conclusion** Intermittent ambulatory treatment with levosimendan in patients with advanced heart failure did not improve significantly functional capacity or quality of life as compared with placebo. An adequately powered, event-driven trial is warranted to enlarge on our findings.

**Trial registration:** NCT01061194

**Keywords** Levosimendan • Pulsed infusions • Advanced heart failure • Outcome • Safety • Outpatient setting

\*Corresponding author: Tel: +43 512 8713136; Fax: +43 512 23204; Email: gerhard.poelzl@i-med.ac.at

© 2014 The Authors  
European Journal of Heart Failure © 2014 European Society of Cardiology

Study	Levo_N	Levo_Deaths	Placebo_N	Placebo_Deaths
Lionheart	48	15	21	8
LevoRep	63	1	57	4

# Lionheart+Levorep



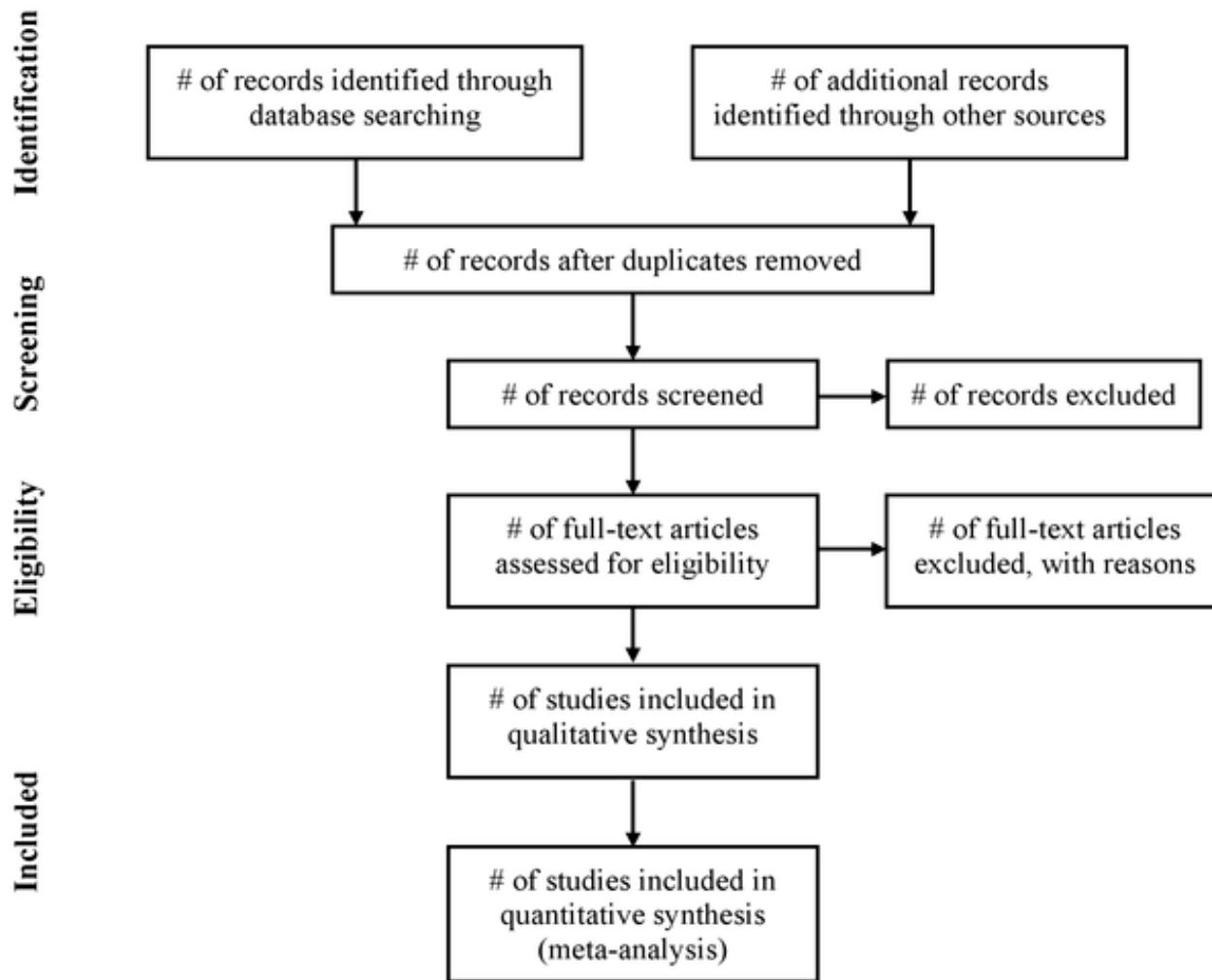
## Meta-analysis: risk difference

Variable for studies	study							
1. Intervention groups								
Variable for total number of cases	Levosimendan_N							
Variable for number of positive cases	Levosimendan_Deaths							
2. Control groups								
Variable for total number of cases	Placebo_N							
Variable for number of positive cases	Placebo_Deaths							
Study	Intervention	Controls	Risk Difference	95% CI	z	P	Weight (%)	
							Fixed	Random
Lionheart	15/48	8/21	-0,0685	-0,314 to 0,177			8,15	8,15
LevoRep	1/63	4/57	-0,0543	-0,127 to 0,0188			91,85	91,85
Total (fixed effects)	16/111	12/78	-0,0589	-0,153 to 0,0354	-1,224	0,221	100,00	100,00
Total (random effects)	16/111	12/78	-0,0555	-0,126 to 0,0146	-1,550	0,121	100,00	100,00

## Test for heterogeneity

Q	0,02123
DF	1
Significance level	P = 0,8842
I <sup>2</sup> (inconsistency)	0,00%
95% CI for I <sup>2</sup>	0,00 to 0,00

Figure 1. Flow of information through the different phases of a systematic review.



Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009) Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Medicine 6(7): e1000097. <https://doi.org/10.1371/journal.pmed.1000097>  
<https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000097>

**Table 1. Checklist of items to include when reporting a systematic review or meta-analysis.**

Section/Topic	#	Checklist Item	Reported on Page #
<b>TITLE</b>			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	
<b>ABSTRACT</b>			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria; participants; and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	
<b>INTRODUCTION</b>			
Rationale	3	Describe the rationale for the review in the context of what is already known.	
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	
<b>METHODS</b>			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., $I^2$ ) for each meta-analysis.	
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	
<b>RESULTS</b>			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome-level assessment (see Item 12).	
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group and (b) effect estimates and confidence intervals, ideally with a forest plot.	
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	
<b>DISCUSSION</b>			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., health care providers, users, and policy makers).	
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review level (e.g., incomplete retrieval of identified research, reporting bias).	
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	
<b>FUNDING</b>			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	

doi:10.1371/journal.pmed.1000097.t001

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009) Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Medicine 6(7): e1000097. <https://doi.org/10.1371/journal.pmed.1000097>  
<https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000097>

# PRISMA for Individual Patient Data



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

PRISMA for Individual Patient Data systematic reviews (PRISMA-IPD)  
PRISMA-IPD was published in 2015 and provides guidelines for reporting systematic reviews and meta-analyses of IPD. Systematic reviews and meta-analyses of IPD aim to collect, check, and reanalyze individual-level data from all studies addressing a particular research question.

Statement paper:

Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, Stewart G, Tierney JF; PRISMA-IPD Development Group. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. *JAMA*. 2015;313(16):1657-1665.

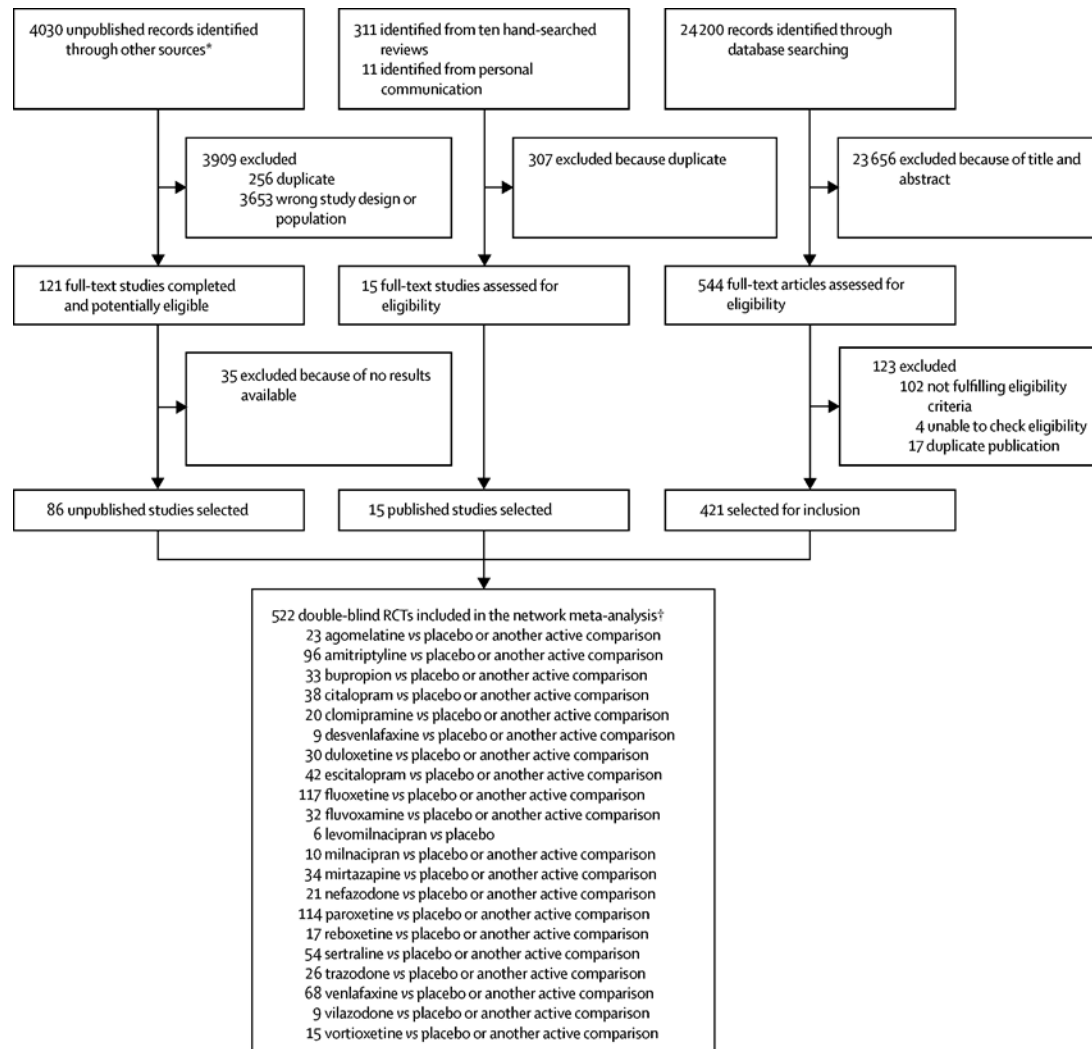


# Effect size calculator

Practical Meta-Analysis Effect Size Calculator David B. Wilson,  
Ph.D., George Mason University:

<http://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php>

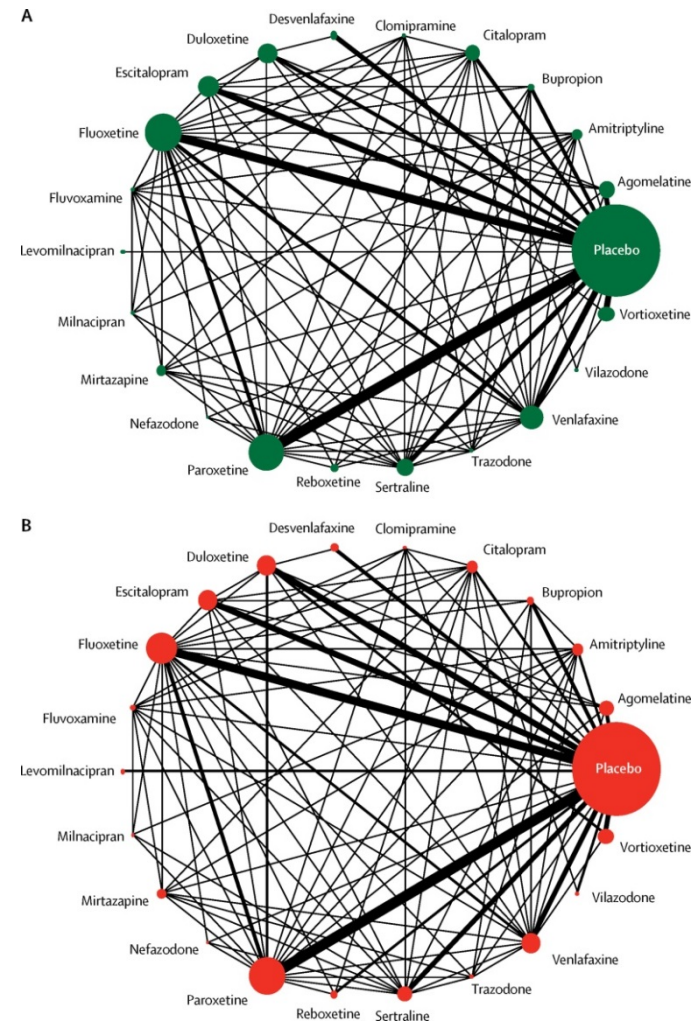
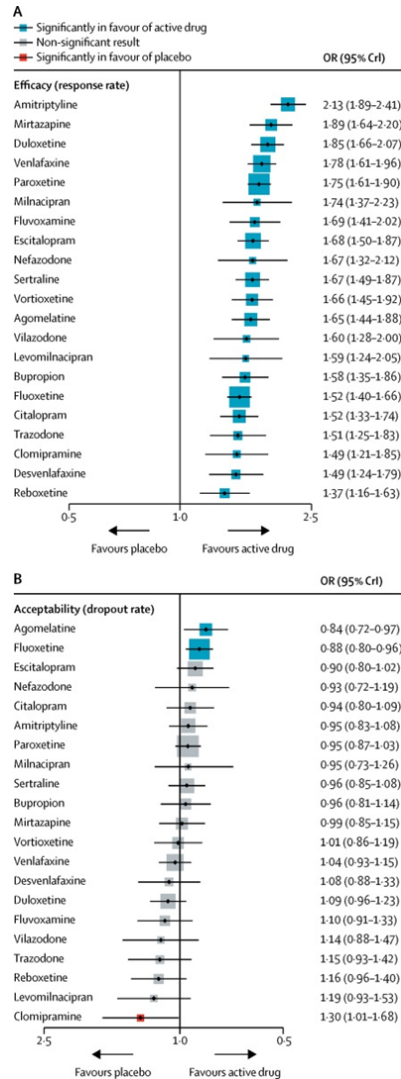
# Network Meta-Analysis



Cipriani A et al.  
Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis.  
Lancet 2018



# Network Meta-Analysis





# PRISMA for Network Meta-Analyses

## PRISMA for Network Meta-Analyses (PRISMA-NMA)

The PRISMA-NMA extension was published in 2015. It provides guidance for reporting systematic reviews comparing multiple treatments using direct and indirect evidence in network meta-analyses. In addition to providing guidance It also highlights educational information related to key considerations in the practice of network meta-analysis.

### Statement/Explanatory paper:

Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, Ioannidis JP, Straus S, Thorlund K, Jansen JP, Mulrow C, Catalá-López F, Gøtzsche PC, Dickersin K, Boutron I, Altman DG, Moher D. The PRISMA Extension Statement for Reporting of Systematic Reviews Incorporating Network Meta-analyses of Health Care Interventions: Checklist and Explanations. *Ann Intern Med.* 2015;162(11):777-784.



# Richtlinien und Empfehlungen

---

- **Deklaration von Helsinki (World Medical Association)**
- Ethische Gesichtspunkte
- **GCP ... Good Clinical Practice**
  - International anerkannte formale Kriterien
- **ICH ... International Committee on Harmonization**
  - Ergänzt durch nationale Gesetzgebung (z.B. Arzneimittelgesetz)
- **Consort-Richtlinien**
  - Studienberichterstellung

# Deklaration von Helsinki

- Regelt seit 1964 ethische Prinzipien für die medizinische Forschung am Menschen
- Revision 1983, 1989, 1996, 2000
  - Seit 2000 (52. Generalversammlung, Edinburgh)
    - Einleitung (Abschnitt 1-9)
    - Allgemeine Grundsätze für jede Art von medizinischer Forschung (Abschnitt 10-27)
    - Grundsätze für die medizinische Forschung in Verbindung mit ärztlicher Versorgung (Abschnitt 28-32)

Download: World Medical Association: <http://www.wma.net>

Inoffizielle Übersetzung: <http://www.bundesaerztekammer.de>

# Deklaration von Helsinki

## Auszug

- Ethische Grundsätze als Leitlinie für Personen, die in der medizinischen Forschung am Menschen tätig sind, incl identifizierbaren menschlichen Daten und Materialien
- Forschung muss allgemein anerkannten wissenschaftlichen Grundsätzen entsprechen
- Anlegen eines Versuchsprotokoll, der Ethikkommission vorlegen
- Überwachung der laufenden Versuche durch die Ethikkommission

# Deklaration von Helsinki

## Auszug

- Forscher muss Ethikkommission informieren über
  - Finanzierung
  - Sponsoren
  - institutionelle Verbindungen
  - potentielle Interessenskonflikte
  - Anreize für Versuchspersonen

# International Conference of Harmonisation (ICH)



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

## Mitglieder:

- Kommission der Europäischen Gemeinschaft
- European Federation of Pharmaceutical Industries and Associations (EFPIA)
- Ministry of Health, Labor and Welfare (MHLW)
- Japan Pharmaceutical Manufacturers Association (JPMA)
- US Food and Drug Administration (FDA)
- Pharmaceutical Research and Manufacturers of America (PhRMA)

# Richtlinie ICH E9

## (Statistical Principles for Clinical Trials, 1998)



- Harmonisierung statistischer Vorgehensweisen
  - Biostatistical Methodology in Clinical Trials in Application for Marketing Authorisations (Committee for Proprietary Medicinal Products, 1994)
  - Guidelines on Statistical Analysis of Clinical Studies (Japanese Ministry of Health and Welfare, 1992)
  - Guideline for the Format and Content of Clinical and Statistical Sections of a New Drug Application (U.S. Food and Drug Administration, 1988)





# Weitere wichtige Richtlinien

---

- ICH E6 Guideline for good clinical practice
  - Beschreibung der wichtigsten Elemente
  - Richtlinie muß für Zulassungsstudien befolgt werden
  - Bezieht sich auf Aspekte zur Durchführung klinischer Studien
  - Beschreibt die Rolle der Ethikkommission
  - Beschreibt Aufgaben des Prüfarztes, des Sponsors und des Monitors

# CONSORT-Richtlinien

- **CONSORT (Consolidated Standards of Reporting Trials)**
- 22 internationale Standards und Richtlinien zur Darstellung von Ergebnissen in wissenschaftlichen Arbeiten
  - Studienplanung
  - Durchführung
  - Statistische Analyse
  - Interpretation
- Flussdiagramm über die Rekrutierung der Patienten
- Transparenz einer Studie gewährleisten
- [www.consort-statement.org/](http://www.consort-statement.org/)
  - Revidierte Version 2001

# Randomisierte Klinische Studien (Randomised Clinical Trials)

---



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

Dr. Hanno Ulmer

*hanno.ulmer@imed.ac.at*

*Innsbruck, Oktober 2010*

---

Department für Medizinische Statistik, Informatik und  
Gesundheitsökonomie, Medizinische Universität Innsbruck

- **Randomisiert**
  - Zufällige Zuteilung zu einer Therapieform
    - Stratifizierung
      - Unterteilung in Subgruppen nach bestimmten Merkmalen, wie Alter, Ausgangswerte, ...
- **Kontrolliert**
  - Mindestens eine Kontrollgruppe
    - z.B.: Vergleich mit Standardtherapie, Placebo-Kontrolle
    - Statistische Vergleichbarkeit der wesentlichen Merkmale zwischen Gruppen
- **Verblindet – wenn möglich**
  - Offen
  - Einfach verblindet: Patient weiß nicht, welche Therapie er bekommt
  - Doppel-blind: Patient und Arzt wissen nicht, welche Therapie Patient bekommt

# Einige einfache Studienpläne

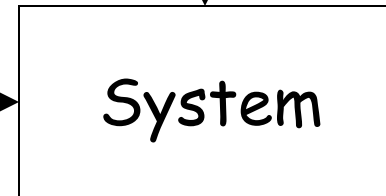
- **Parallelgruppen**
  - Zwei oder mehrere unabhängige, aber vergleichbare Gruppen
    - Werden zeitlich parallel behandelt
    - Mit unterschiedlichen Therapieformen
- **Cross-Over**
  - Eine Gruppe mit zwei oder mehreren aufeinander folgenden Therapien behandeln
  - Jeder Proband erhält jede Therapie aber in unterschiedlicher Reihenfolge
    - Randomisierung der Reihenfolge
      - z.B.: ABC, ACB, BAC, BCA, CAB, CBA
      - Carry-over-Effekte beachten
- **Faktorielle Pläne**
  - Kombination von zwei oder mehreren Einflussfaktoren
    - Wechselwirkungen überprüfbar

# Parallelgruppenstudie

- z.B. Untersuchung der Wirksamkeit eines neuen blutdrucksenkenden Medikaments im Vergleich zu einer Standardtherapie (=Kontrolle)
  - Manchmal Placebo
- **Nullhypothese  $H_0$** : die beiden Therapien sind im Mittel gleich wirksam
  - z.B. die Änderung des diastolischen Blutdrucks ist im Mittel in beiden Gruppen gleich hoch
  - **Die Ungültigkeit der Nullhypothese ist zu beweisen**
- **Alternativhypothese  $H_1$** : die beiden Therapien sind im Mittel unterschiedlich stark wirksam
- **Voraussetzung**: Die beiden Gruppen stimmen in den wesentlichen Merkmalen überein – siehe Randomisierung
  - Nur die Therapien sind unterschiedlich
  - Strukturgleichheit der Gruppen

# Was muss man messen?

**Einflussgröße**



**System**

**Zielgröße**



- **Zielgröße(n)**

- Haupt- und Nebenzielkriterien

- In engem Zusammenhang mit dem Studienziel
    - Klinisch relevant

- **Einflussgröße(n)**

- Können Auswirkungen auf die Zielgröße(n) haben

- **Störgröße(n)**

- Nicht von primären Studieninteresse
  - Können jedoch das Studienergebnis wesentlich beeinflussen bzw. verzerren („Bias“!)

# Primäre und sekundäre Zielgrößen



- Wenn möglich nur ein einziges Hauptzielkriterium (=primäre Zielgröße)
  - Blutdrucksenkung, Überlebenszeit, ...
  - Kann multidimensional sein
  - Fragebogen, QALY, DALY, ...
- Im Studienprotokoll operationalisieren
  - Messbar machen
- Mehrere sekundäre Zielvariablen sind zulässig
- Surrogatkriterien





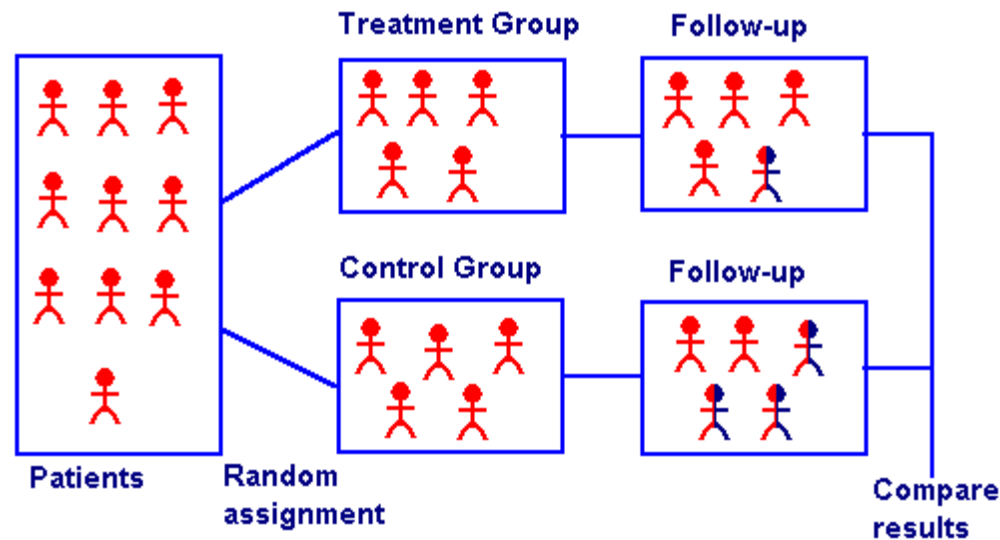
# Vergleichbarkeit der Gruppen

---

- Bezüglich der
  - Gemessenen Merkmale
    - Dokumentieren
    - Bei der Auswertung berücksichtigen
  - Nicht-gemessenen Merkmale
  - Unbekannten Merkmale
    - Können möglicherweise zu Verzerrungen führen
- Randomisierung
- Stratifizierung
- Paarbildung
  - Zwillinge, Matching

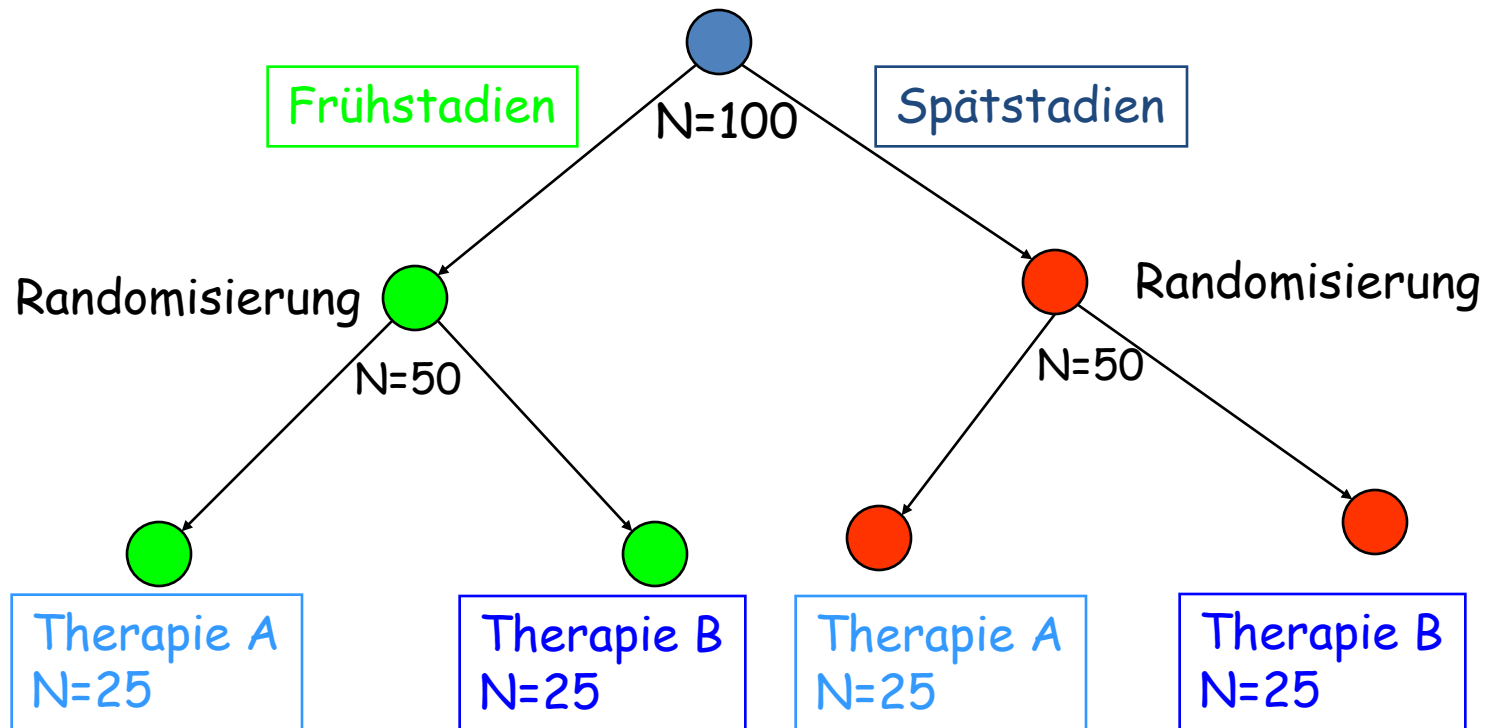
# Randomisierung

- **Zufällige Zuordnung eines Subjektes/Objektes einer Stichprobe zu einer der Gruppen des Einflussfaktors**
  - z.B. mit Hilfe eines Zufallsgenerators
- **ZIEL:**
  - Vermeidung eines Selektions-Bias
    - Jedes Objekt hat die gleiche Chance zufällig einer Gruppe zugeteilt zu werden
  - Strukturgleichheit (Zusätzliche Stratifizierung ist möglich)



# Randomisierung / Stratifizierung

## Stratifizierung nach Stadium



# Verblindung



- **Ausschaltung von Verzerrungen/systematischen Fehlern (Bias) aufgrund von Vorinformationen**
  - Selektions-Bias, Beobachter-Bias
- Ausprägungen des Einflussfaktors sind für den Beobachter unbekannt
  - **Doppel-Blindstudie**
    - Patient und Arzt wissen nicht, welche Therapieform angewendet wird
    - Nicht immer möglich!!!
- Beschreibung der Art der Verblindung
  - äußere Form, Geschmack, Farbe
  - sichtbare Unterschiede bei Behandlung (z.B. Farbe des Urins)
- Beurteilung des Outcomes
  - unabhängige Beurteilung durch Dritte
- **Was passiert, wenn Code gebrochen/entziffert wird?**

- Fixe, berechnete Fallzahl
  - Basierend auf Annahmen über Unterschiede/Behandlungseffekte
  - Fehler 1. und 2. Art (z.B. Signifikanzniveau=0.05, Power=80%)
- Sequentielle Pläne
  - 1 bis K geplante Zwischenauswertungen
    - **Interimsanalysen, Adaptive Studienpläne**
  - Möglicherweise frühere Entscheidung
- Berücksichtigung der Ausfallrate bei statistischen Analyse
  - **Intention-to-treat** Analyse: alle **eingeschlossenen** Fälle
  - **Per Protokoll** Analyse: alle **abgeschlossenen** Fälle

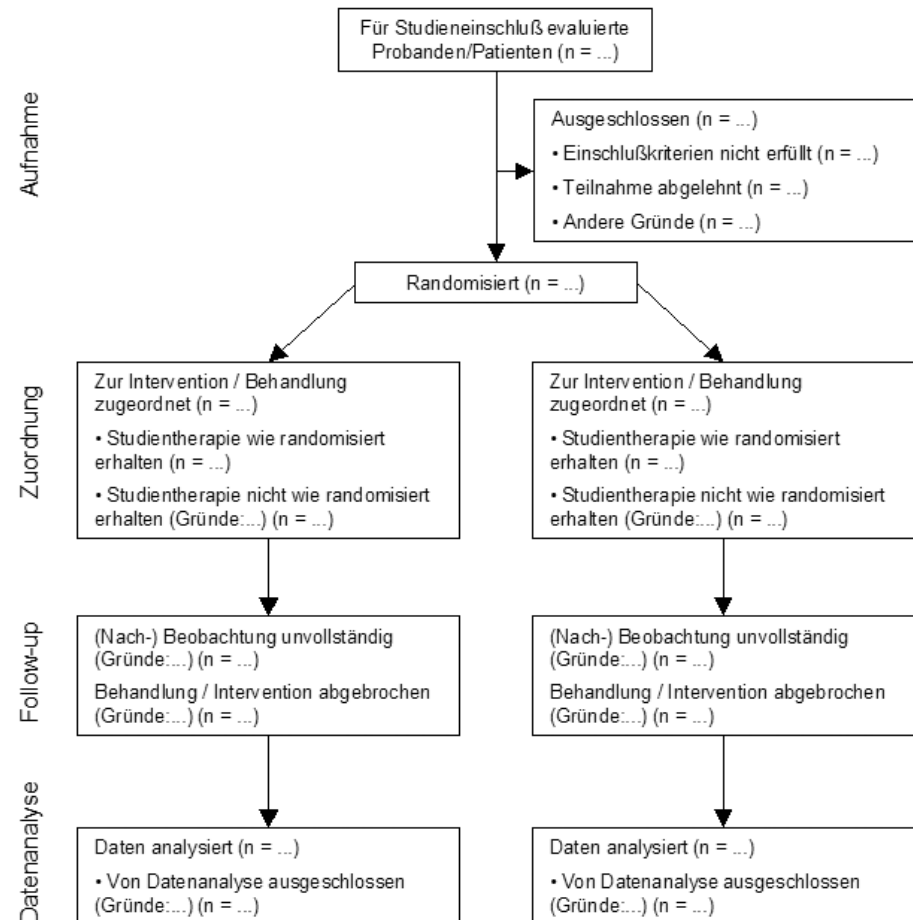
# Patientenflussdiagramm

- Intention-To-Treat

versus

- Per-Protocol

- Vergleichbarkeit der Gruppen:
  - Strukturgleichheit
  - Beobachtungsgleichheit
  - Behandlungsgleichheit





# Checklist für Fallzahlschätzung (Testproblem)

- Fehler 1. Art (üblicherweise 0,05)
- Fehler 2. Art (0,1 oder 0,2)
- Auswahl des Hauptzielkriteriums
- Zu erwartender Unterschied und Angabe eines Variationsmaßes
- Begründung dafür – Literatur oder Vorstudie
- Auswahl des statistischen Tests
- Falls mehrere Hypothesen formuliert werden, Korrektur des Fehler 1. Art oder Hierarchisierung der Hypothesen
- Drop-Out Rate berücksichtigen

# Vorschlag „Studienprotokoll – Statistik“

---

## ad 9 Statistik

### 9.1 Fallzahlplanung

*Die Fallzahl wird in der Regel berechnet aus der primären Zielvariablen, dem klinisch relevanten Unterschied, der in der Studie nachgewiesen werden soll, dem statistischen Verfahren, das dazu verwendet wird, sowie aus dem Fehler 1. Art  $\alpha$  und der Power  $1-\beta$ , die erzielt werden soll. Die Fallzahl ist außerdem abhängig von der Anzahl der Gruppen, die verglichen werden sollen. Angaben zu diesen Parametern sowie zum Verfahren der Fallzahlplanung sind erforderlich. .*

### 9.2 Randomisierung

*Bei der Beschreibung der Randomisierung sind Angaben erforderlich zur Art der Randomisierung, zur Anzahl der Gruppen und der erforderlichen Strata und ggf. zum organisatorischen Ablauf der Randomisierung. Dabei sind besonders bei verblindeten Studien Maßnahmen aufzulisten, die die Verblindung gewährleisten.*

### 9.3 Statistische Methoden

*Zielgrößen, Definition von Auswertungskollektiven, Datenanalyse, Zwischenauswertungen, Verweis auf ICH-GCP E9: Statistical Principles for Clinical Trials*



# Übung: EK1 Statistik ausfüllen



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

Antragsformular EK1 Seite 5 von 7

<http://www.i-med.ac.at/ethikkommission/>

## **Statistik:**

Angaben zum Studiendesign und Statistik

### **Design:**

- unkontrolliert     kontrolliert     doppelblind     placebokontrolliert     Meßwiederholungen  
 cross over     andere
- multizentrisch     ja     nein  
konfirmatorisch     ja     explorativ  
Randomisierung:     ja     nein  
Verblindung:     ja     nein

### **Haupt- und Nebenzielkriterien:**

### **Null- und Alternativhypothese:**

verbale und formale Formulierung

multiples Testen     ja     nein

Fehler 1. Art :

### **Angaben zur Fallzahlberechnung – Stichprobenumfang**

Fehler 2. Art :

### **Statistische Analyse**

Intent to treat /  per protocol

Zwischenauswertungen     ja     nein

- wenn ja, welche Abbruchkriterien?

Verwendete statistische Verfahren:

Behandlung der Nebenzielkriterien:

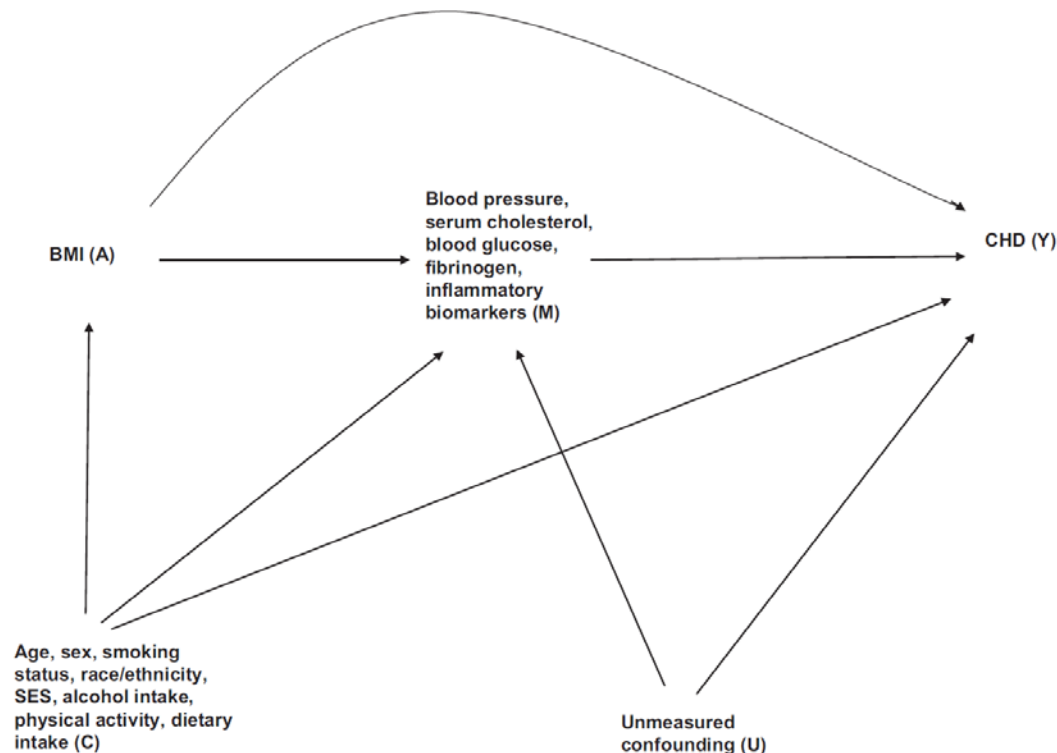
Wer wird die statistische Analyse durchführen?

### **Dokumentationsbogen (CRF)**

Angaben zur Datenqualitätsprüfung:

Angaben zum Datenmanagement und Datenschutz:

# Confounding, Moderation, Mediation anhand einer Fallstudie erklärt



**FIGURE 1.** Causal diagram of the relation among BMI (A), metabolic risk factors, prothrombotic and inflammatory biomarkers (M), and CHD (Y) with measured confounders (C)\* and unmeasured confounding for BMI, mediators, and CHD (U). Measured confounders were pre-baseline variables.

Regressionsanalyse als statistische Methode um Zusammenhänge zu beschreiben :

Exposition (z.B. Risikofaktor, Therapie) -- > Outcome (z.B. Erkrankung)

Multivariable Analyse:

k unabhängige Variable (Prädiktoren) -- > 1 abhängige Variable (Outcome)

Verschiedene Arten und Berechnungsweisen von Regressionsanalysen:

Bereits 1805 wurde beispielsweise die Methode der kleinsten Quadrate von Legendre publiziert.

# Die wichtigsten Regressionsanalysen



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

Multivariable Analyse: Regression von  $k$  unabhängigen auf 1 abhängige Variable

Abhängige Variable ist metrisch (stetig): Linear Regression

z.B. Geschlecht, Alter, BMI -> systolischer Blutdruck

Geschätzt (berechnet) wird das (standardisierte) Beta

Abhängige Variable ist kategoriell : Logistische Regression

z.B. Geschlecht, Alter, BMI -> KHK in den nächsten 10 Jahren

Geschätzt wird das Odds Ratio

Abhängige Variable ist eine Ereigniszeit: Cox proportional hazards Regression

z.B. Geschlecht, Alter, BMI -> Zeit bis KHK (survival analysis)

Geschätzt wird das Hazard Ratio

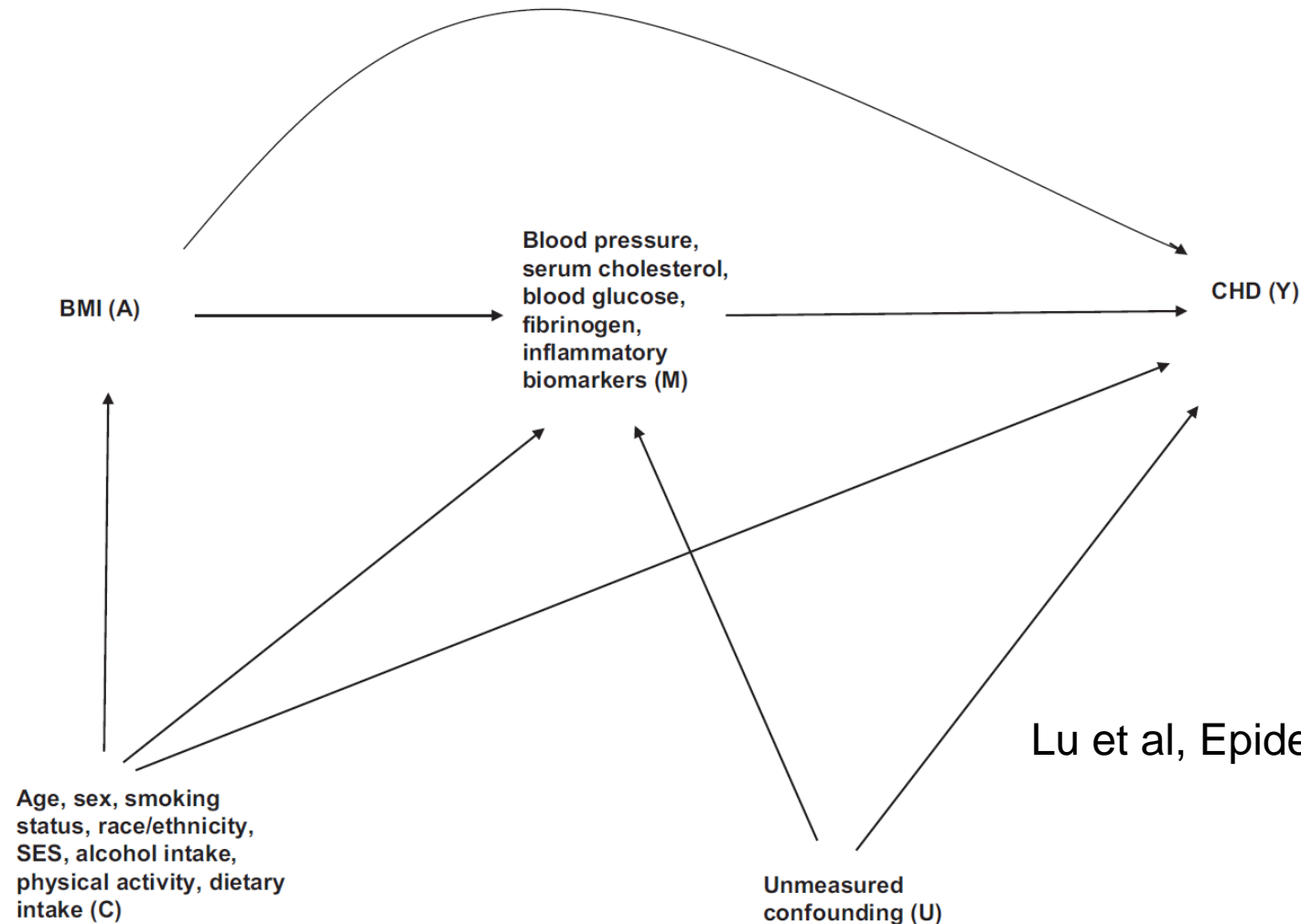
# Confounding, Moderation, Mediation



Die Regressionsanalyse ermöglicht es, den Effekt einer Prädiktorvariable (z.B. Adipositas) auf eine Zielvariable (z.B. KHK) unter Berücksichtigung von **‘dritten Faktoren’** (z.B. Geschlecht, Alter, Rauchen, Blutdruck, Cholesterin, Diabetes, etc.) abzuschätzen.

Diese **‘dritten Faktoren’** können als Confounder, Moderatoren, oder Mediatoren agieren, je nach angenommenen kausalen Wirkzusammenhang.

Die drei Konzepte werden nun am Beispiel (engl.) BMI --- > KHK illustriert.



Lu et al, Epidemiology 2015

**FIGURE 1.** Causal diagram of the relation among BMI (A), metabolic risk factors, prothrombotic and inflammatory biomarkers (M), and CHD (Y) with measured confounders (C)\* and unmeasured confounding for BMI, mediators, and CHD (U). Measured confounders were pre-baseline variables.



# Example data

---

## Vorarlberg Health Examinations (VHM&PP)

Sex (male, female)	categorical
Age in years	continuous
Year of examination	continuous
Body mass index in kg/m <sup>2</sup>	continuous
Systolic blood pressure in mmHG	continuous
Total cholesterol in mg/dl	continuous
Fasting glucose in mg/dl	continuous
Smoking (current or past, never)	categorical
Coronary heart disease mortality (ICD-10: I20-I25)	time to event continuous and categorical

# Confounding



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

Confounding:

A “mixing of the effect” of the exposure-disease relationship with a third (or more) factors

BMI ----- > CHD

< ----- sex, age, smoking ----- >



# Example



Relationship between BMI and CHD incidence:

Crude Hazard Ratio

Obesity (30+ kg/m<sup>2</sup>) versus normal weight (20-25 kg/m<sup>2</sup>)

HR = 2.54 95%CI (2.32-2.78)

Sex, age and smoking adjusting hazard ratio:

HR = 1.60 95%CI (1.46-1.75)

Adjusted = controlled for confounding

Calculated with Cox proportional hazards regression analysis

# Confounding



Three essential characteristics:

The confounder is associated with the exposure of interest (BMI)

The confounder is associated with the disease (CHD)

The confounder is not in the causal pathway leading from the exposure of interest (BMI) to the disease of interest (CHD)

# Methods for Preventing Confounding in Study Designs



MEDIZINISCHE UNIVERSITÄT  
INNSBRUCK

1. Stringent inclusion criteria to narrow the variability between study participants
2. Randomization (intervention/RCT only)  
In an optimal RCT, study groups only differ regarding the intervention
3. Matching (observational studies):

Simple Matching e.g. for age and sex in case-controls studies  
versus

Propensity Score Matching (involves logistic regression analysis)

Very popular in clinical research:

Blackstone EH. Comparing apples and oranges. *J Thoracic and Cardiovascular Surgery* 2002; 1: 8-15.

An example:

Ruttman E et al. Second internal thoracic artery versus radial artery in coronary artery bypass grafting: a long-term, propensity score-matched follow-up study. *Circulation*. 2011 20;124(12):1321-9.

# Effect Modification/Moderation

Effect modification occurs when the association between the exposure (BMI) and the disease (CHD) varies by levels of a third factor.

How to assess: include interaction terms into the regression model

Interaction age\*obesity  $p < 0.001$

Young: BMI ----- > CHD

Old: BMI ----- > CHD

# Example

---

Relationship between BMI and CHD incidence moderated by age:

Interaction age\*obesity  $p < 0.001$

Obesity (30+ kg/m<sup>2</sup>) versus normal weight (20-25 kg/m<sup>2</sup>)

Sex, age and smoking adjusting hazard ratio:

<50 years of age:

HR = 3.13 95%CI (2.27-4.31)

50+ years of age:

HR = 1.51 95%CI (1.37- 1.66)



# Mediation

Mediation occurs if factors, like confounders, are associated with the exposure of interest (BMI) and the disease (CHD), but are **in the causal pathway** leading from the exposure to the disease.

These factors are called mediators:

BMI ---- > blood Pressure, cholesterol, diabetes ---- > CHD

# Example



Mediators in the relationship between BMI and CHD incidence:

Sex, age and smoking adjusting hazard ratio:

Total effect of BMI (obesity versus normal weight) on CHD:

HR = 1.70 95%CI (1.57-1.85)

Direct effect of BMI on CHD

HR = 1.30 95%CI (1.15–1.47)

Indirect effect mediated by blood pressure, cholesterol and glucose

HR = 1.31 95%CI ( 1.16-1.48) (95%CIs estimated by Bootstrap)

HRs ... multiplicative, do not add

# Example



Mediators in the relationship between BMI and CHD incidence:

Effect of BMI on CHD mediated by blood pressure, cholesterol and glucose

PERM (Percentage of excess risk mediated) =  
 $(1.70 - 1.30) / (1.70 - 1) * 100$   
= 57% (approximative formula)

Global Burden of Metabolic Risk Factors for Chronic Diseases  
Collaboration. Metabolic mediators of the effects of body-mass index,  
overweight, and obesity on coronary heart disease and stroke: a  
pooled analysis of 97 prospective cohorts with 1.8 million participants.  
Lancet. 2014 Mar 15;383(9921):970-83





# Mediation Techniques

---

Traditional approach:

Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol.* 1986 Dec;51(6):1173-82

New approaches:

Lange T, Rasmussen M, Thygesen LC. Assessing natural direct and indirect effects through multiple pathways. *Am J Epidemiol.* 2014 Feb 15;179(4):513-8.

VanderWeele T. *Explanation in Causal Inference: Methods for Mediation and Interaction.* Oxford University Press 2015.

New approaches applied on BMI --- > CHD problem:

Lu Y, Hajifathalian K, Rimm EB, Ezzati M, Danaei G. Mediators of the effect of body mass index on coronary heart disease: decomposing direct and indirect effects. *Epidemiology.* 2015 Mar;26(2):153-62.



# Do risk factors explain the sex/gender gap in mortality from coronary heart disease?

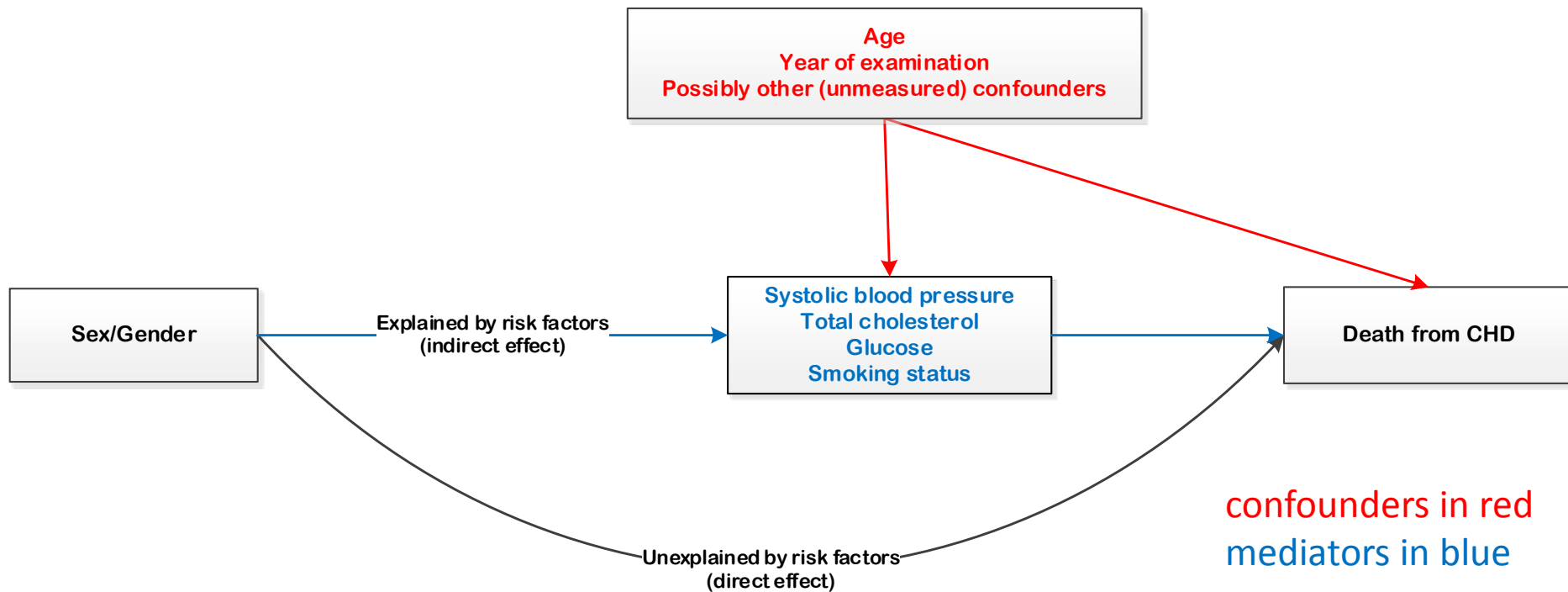
Josef Fritz, Michael Edlinger, Cecily Kelleher, Susanne Strohmaier, Gabriele Nagel, Hans Concin, Elfriede Ruttmann, Margarethe Hochleitner, Hanno Ulmer

Medical University of Innsbruck, Austria, University College Dublin, Ireland  
University of Oslo, Norway, University of Ulm, Germany  
Agency for Preventive and Social Medicine, Bregenz, Austria

# Purpose

- Age and sex are the strongest predictors of coronary heart disease mortality
- Premature CHD (I20-I25) deaths – before age 65 - in Europe: 330,000 death cases, of which 77% in males, 23% in females (Nichols et al, Eur Heart J 2014)
- Aim of study:  
to estimate, how much of this large sex difference is explained by the major risk factors (RFs):
  - systolic blood pressure
  - total cholesterol
  - fasting glucose
  - smoking

# Figure 1. Underlying model



We assume that the 4 RFs are in the causal chain between sex and mortality, mediating the total sex effect, e.g. male sex causes hypertension, and hypertension causes CHD

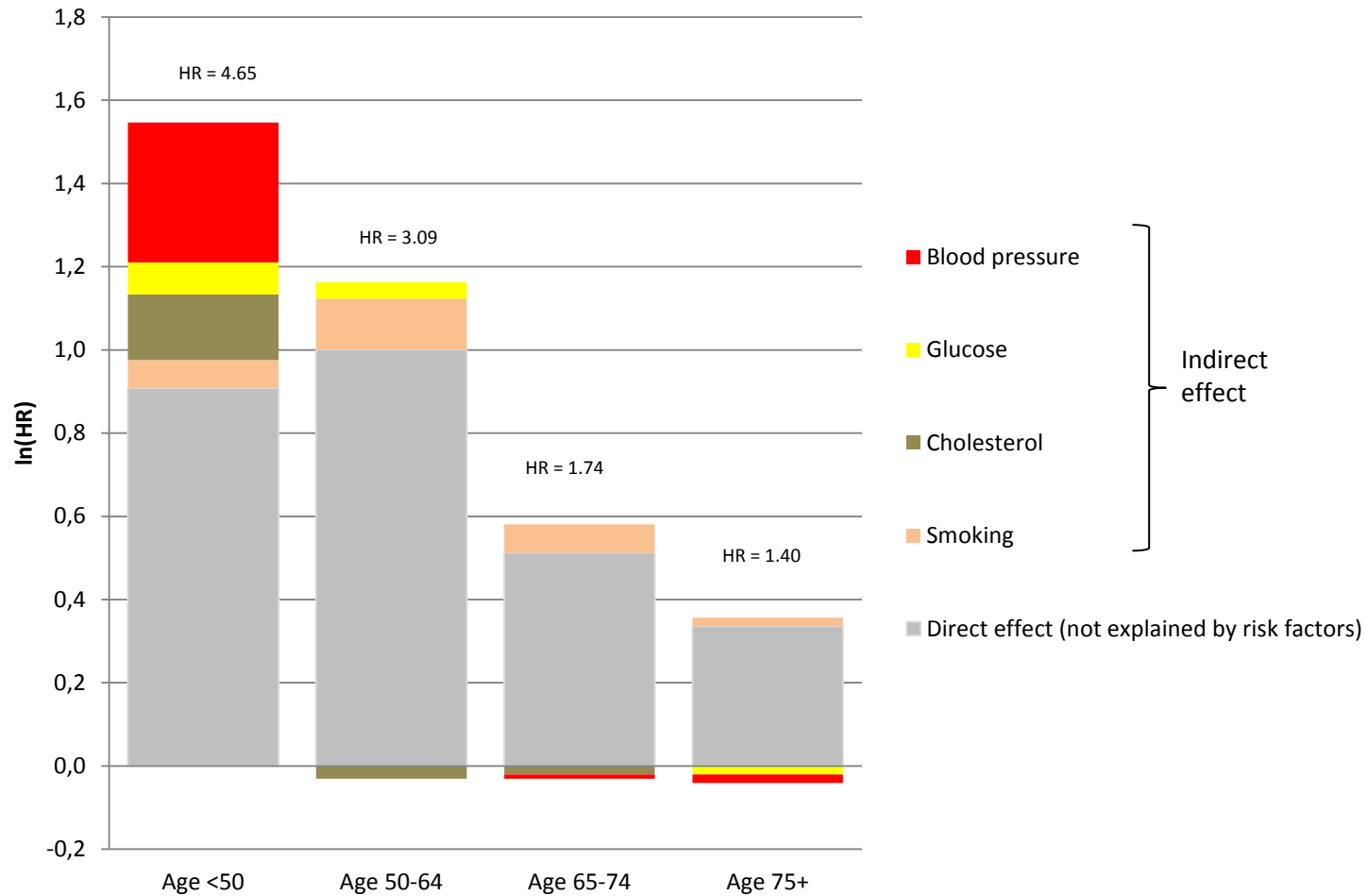
# Material and methods

- We used prospective cohort data from the Vorarlberg Health Monitoring and Promotion Programme (VHM&PP), Austria
  - A total of 172,262 individuals underwent baseline health examinations with fasting measurements of RFs
  - There were 3,892 CHD deaths during a follow-up of 14.6 years
- 
- For data analysis, we used a recently developed statistical mediation method (Lange et al, Am J Epidemiol 2014)
  - Designed for survival data
  - Allowing breakdown into single components of the indirect sex effect (that is explained by the RFs)

# Results and conclusions

- The mortality difference between sexes decreased with age
  - <50 years: HR=4.7 (95%CI 3.5-6.1)  
≥50 years: HR=1.9 (95%CI 1.7-2.1)
- The extent to which risk factors contributed varied with age
  - <50 years: the 4 RF explained 41% (95%CI 27-54%) of sex effect  
≥50 years: the 4 RF explained 8% (95%CI 4-12%) of sex effect
- In younger individuals, the female survival advantage was explained to a substantial part through the pathways of the 4 major risk factors
  - As blood pressure and cholesterol were the strongest factors, our results correspond to the oestrogen/testosterone thesis

# Figure 2. What risk factors explain





---

SWISS MED WKLY 2007;137:44–49 · [www.smw.ch](http://www.smw.ch)

---

## Statistical errors in medical research – a review of common pitfalls

---

*Alexander M. Strasak, Qamruz Zaman, Karl P. Pfeiffer, Georg Göbel, Hanno Ulmer*



# Appendix (SMW-Artikel)



**Table 1**

Statistical errors and deficiencies related to the design of a study.

Study aims and primary outcome measures not clearly stated or unclear
Failure to report number of participants or observations (sample size)
Failure to report withdrawals from the study
No a priori sample size calculation/effect-size estimation (power calculation)
No clear a priori statement or description of the Null-Hypothesis under investigation
Failure to use and report randomisation
Method of randomisation not clearly stated
Failure to use and report blinding if possible
Failure to report initial equality of baseline characteristics and comparability of study groups
Use of an inappropriate control group
Inappropriate testing for equality of baseline characteristics

**Table 2**

Statistical errors and deficiencies related to data analysis.

Use of wrong statistical tests
Incompatibility of statistical test with type of data examined
Unpaired tests for paired data or vice versa
Inappropriate use of parametric methods
Use of an inappropriate test for the hypothesis under investigation
Inflation of Type I error
Failure to include a multiple-comparison correction
Inappropriate post-hoc Subgroup analysis
Typical errors with Student's t-test
Failure to prove test assumptions
Unequal sample sizes for paired t-test
Improper multiple pair-wise comparisons of more than two groups
Use of an unpaired t-test for paired data or vice versa
Typical errors with $\chi^2$ -tests
No Yates-continuity correction reported if small numbers
Use of chi-square when expected numbers in a cell are $<5$
No explicit statement of the tested Null-Hypotheses
Failure to use multivariate techniques to adjust for confounding factors

**Table 3**

Errors related to the documentation of statistical methods applied.

Failure to specify/define all tests used clear and correctly

---

Failure to state number of tails

---

Failure to state if test was paired or unpaired

---

Wrong names for statistical tests

---

Referring to unusual or obscure methods without explanation or reference

---

Failure to specify which test was applied on a given set of data if more than one test was done

---

“Where appropriate” statement

---

**Table 4**

Statistical errors and deficiencies related to the presentation of study data.

---

**Inadequate graphical or numerical description of basic data**

---

Mean but no indication of variability of the data

---

Giving SE instead of SD to describe data

---

Use of mean (SD) to describe non-normal data

---

Failure to define  $\pm$  notion for describing variability or use of unlabeled error bars

---

**Inappropriate and poor reporting of results**

---

Results given only as p-values, no confidence intervals given

---

Confidence intervals given for each group rather than for contrasts

---

“p = NS”, “p < 0.05” or other arbitrary thresholds instead of reporting exact p-values

---

Numerical information given to an unrealistic level of precision

---

ÄT

—

**Table 5**

Statistical errors and deficiencies related to the interpretation of study findings.

---

**Wrong interpretation of results**

---

“non significant” interpreted as “no effect”,  
or “no difference”

---

Drawing conclusions not supported by the study data

---

Significance claimed without data analysis or statistical test  
mentioned

---

**Poor interpretation of results**

---

Disregard for Type II error when reporting non-significant  
results

---

Missing discussion of the problem of multiple significance  
testing if done

---

Failure to discuss sources of potential bias and confounding  
factors

---