



# Experimentelles Design: Typische Fehler

# Beispiele und typische Fehler

ADVANCES IN THE STUDY OF BEHAVIOR, VOL. 26

---

## How To Avoid Seven Deadly Sins in the ~~Study~~ ~~of Behavior~~ *Experimental Design*

---

MANFRED MILINSKI

ABTEILUNG VERHALTENSÖKOLOGIE

ZOOLOGISCHES INSTITUT

UNIVERSITÄT BERN

HINTERKAPPELEN, SWITZERLAND

## ☠ Deadly sins ☠

TS 1: Kausalität statt Korrelation

TS 2: Pseudoreplikate

TS 3: Behandlungen haben confounder

TS 4: Beobachter hat Bias

TS 5: Verhaltensänderung wegen Experiment-Setting

TS 6: Schlechte / Keine Kontrollen

TS 7: Nullhypothese “beweisen”



## ☠ TS 1: Korrelation und Kausalität

Beobachtungsstudie: “Hormon Replacement Therapy (HRT)” reduziert das Risiko für Erkrankungen der Herzkranzgefässe.


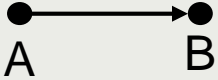
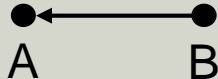
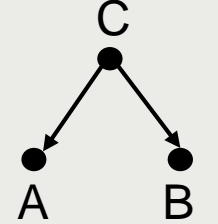
Ärzte sollten daher Frauen nach der Menopause HRT verschreiben.

A: Richtig.

B: Falsch.

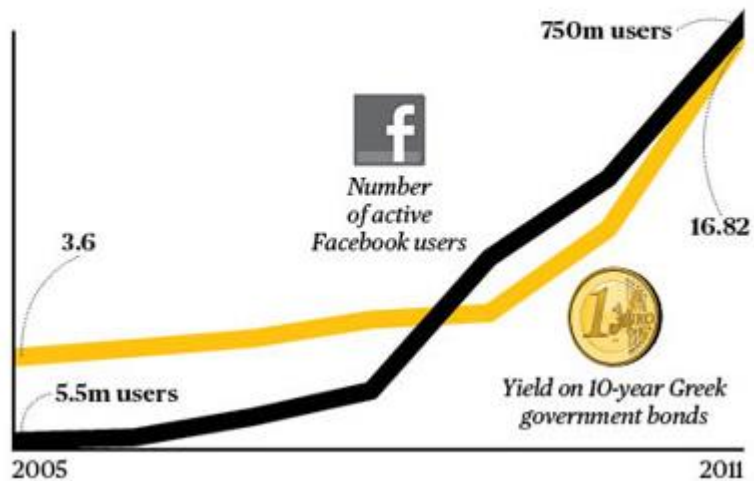
C: Weiss nicht.

# Korrelation und Kausalität - Denkmuster

Muster	Korrelation zw. A und B	A verursacht B
	Nein, aber stat. Test hat evtl. Fehler gemacht	Nein
	Ja	Ja
	Ja	Nein
	Ja	Nein

# Facebook verursacht Schuldenkrise.

Fig. 1  
IS FACEBOOK DRIVING  
THE GREEK DEBT CRISIS?



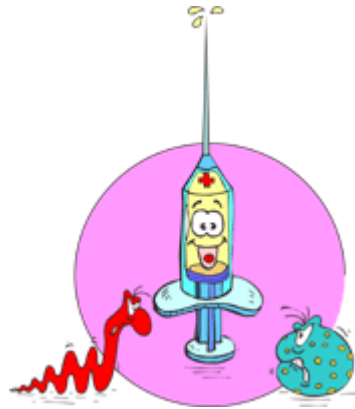
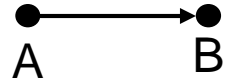
Facebook  
Benutzer



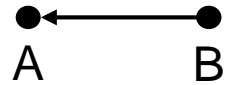
Griechische  
Staatsanleihe

Quelle: <http://www.businessweek.com/magazine/correlation-or-causation-12012011-gfx.html>

# Wer sich gegen Polio impfen lässt, hat ein geringeres Risiko für Polio.

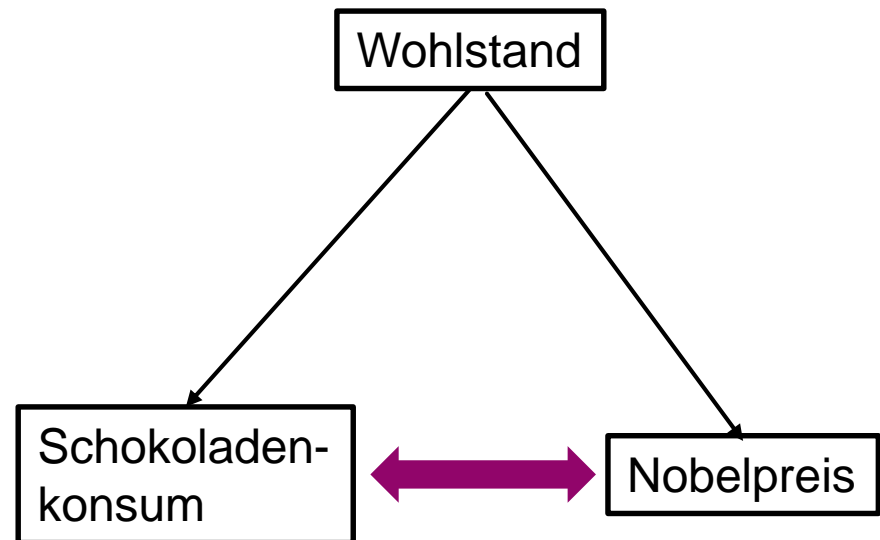
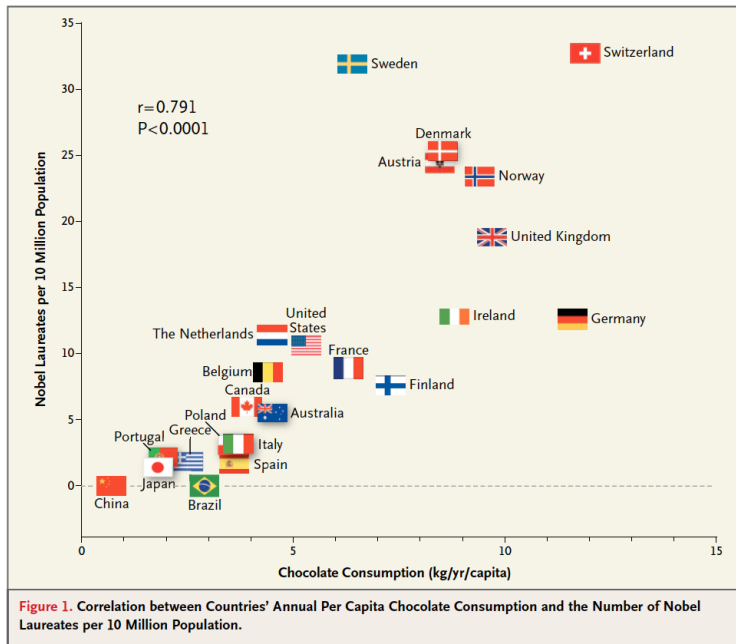
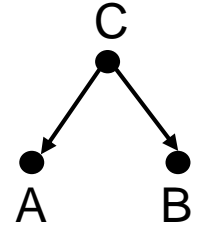


# Wer sich einen Mercedes kauft, wird reich.



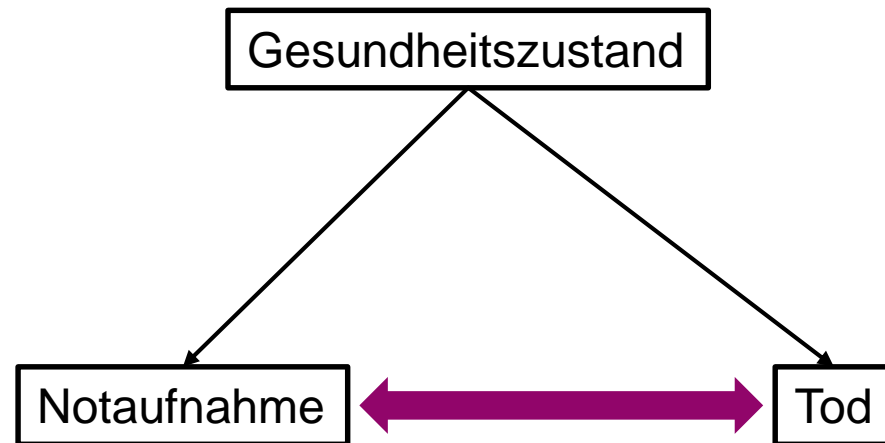
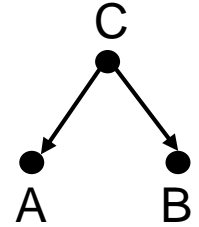


# Schokolade macht Nobelpreisträger.

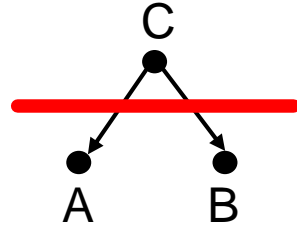
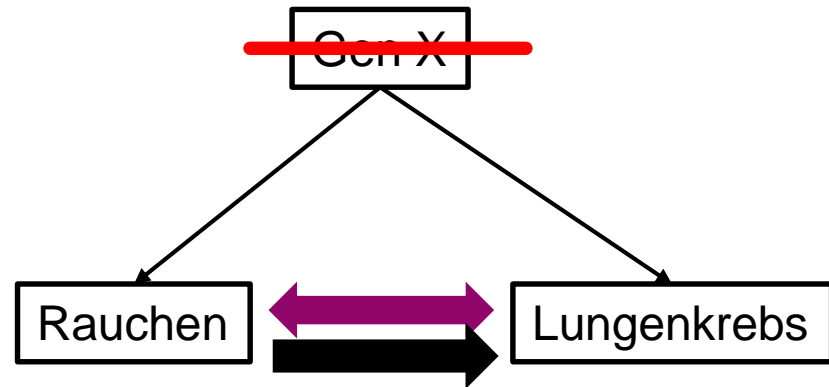


Quelle: <http://www.nejm.org/doi/full/10.1056/NEJMon1211064>

# Notaufnahmen bringen Patienten um.



# Rauchen verursacht Lungenkrebs.

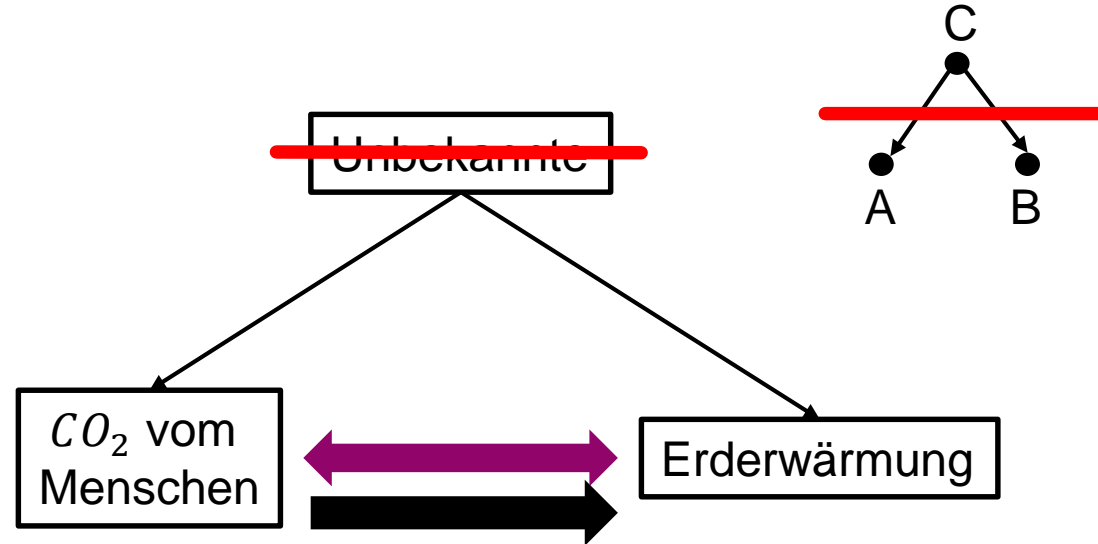


“cigarette smoking has been **causally linked to diseases** of nearly all organs of the body” (“50 Years of Progress”, Ch. 1, “Major conclusions from the Report, Punkt 3)

<http://www.surgeongeneral.gov/library/reports/50-years-of-progress/exec-summary.pdf>

Gegenbeispiel: ☠ TS 1

# $CO_2$ vom Menschen verursacht Erderwärmung.



“It is extremely likely that **human influence** has been **the dominant cause** of the observed warming since the mid-20th century. “ (AR5, D.3)

[http://www.ipcc.ch/pdf/assessment-report/ar5/wg1/WG1AR5\\_SPM\\_FINAL.pdf](http://www.ipcc.ch/pdf/assessment-report/ar5/wg1/WG1AR5_SPM_FINAL.pdf)

Gegenbeispiel: ☠ TS 1

## ☠ TS 1: Fazit

- Korrelation beweist in der Regel keine Kausalität
- Die **Beweislast** für Kausalität ist enorm hoch  
Experimente beweisen Kausalität
- Kausalität durch Korrelation zu beweisen braucht **extremen** Zusatzaufwand  
(z.B. Rauchen, Klimawandel)

## ☠ TS 2: Pseudoreplikate

- Sind Männer im Schnitt grösser als Frauen?
- Wähle zufällig einen Mann und eine Frau und vergleiche.
- Replikate: Messe den Mann und die Frau 100 mal und vergleiche mit t-Test.



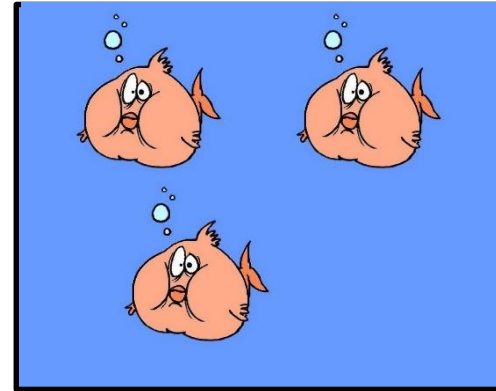
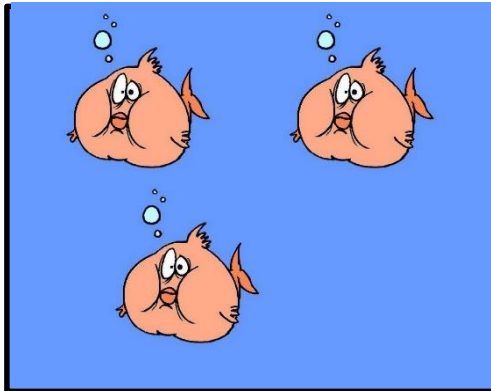
**Problem:** Messungen sind  
nicht unabhängig  
→ 100 **verschiedene**  
Männer und Frauen  
messen



Replikate sind  
**unabhängige**  
Versuchseinheiten

# Fische im Aquarium

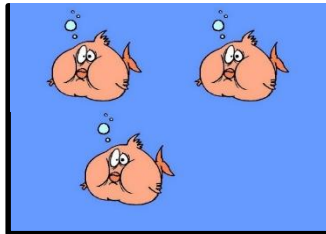
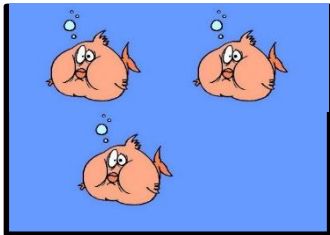
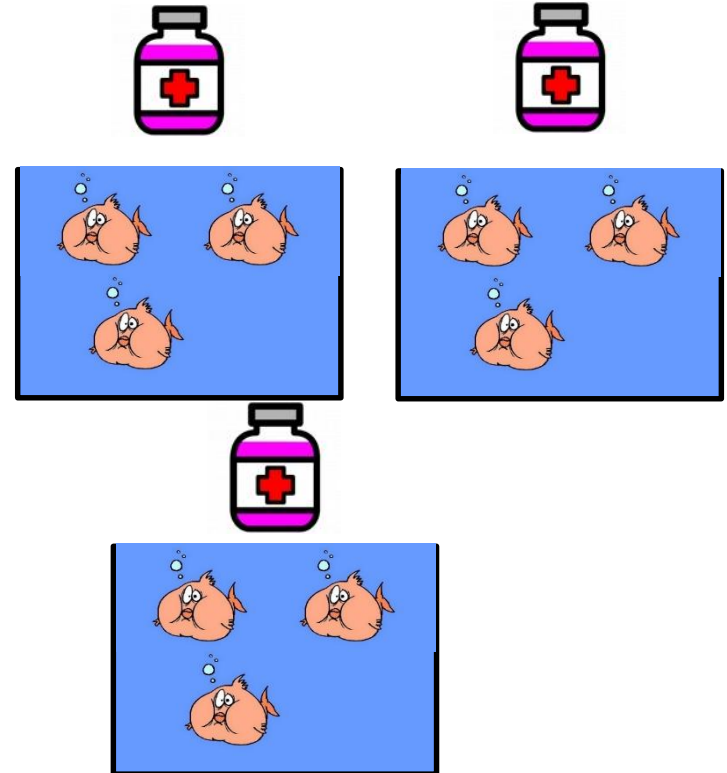
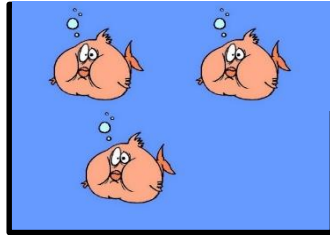
$n = 2$  oder  $n = 6$  ?



Fische in einem Tank sind nicht unabhängig (z.B. schlechte Belüftung, Sonnenschein, Lärm, ...)  $\rightarrow n = 2$

# Fische im Aquarium

$n = 6$  oder  $n = 18$  ?



Fische in einem Tank sind nicht unabhängig (z.B. schlechte Belüftung, Sonnenschein, Lärm, ...)  $\rightarrow n = 6$



# Grenzfälle: “Hinreichend” unabhängig ?

- Bsp: Klimakammer



KK 1  
 $T = -40\text{ C}$

Rep 1

Rep 2

Rep 3

KK 2  
 $T = 15\text{ C}$

Rep 1

Rep 2

Rep 3

Klimakammer kann sehr genau kontrolliert werden  
Aber: Gibt es evtl. einen Fehler in einer Kammer ?

# Grenzfälle: “Hinreichend” unabhängig ?

- Mögliche Lösung



KK 1  
 $T = -40\text{ C}$

Rep 1

Rep 2

Rep 3

KK 1  
 $T = 15\text{ C}$

Rep 1

Rep 2

Rep 3

KK 2  
 $T = 15\text{ C}$

Rep 1

Rep 2

Rep 3

KK 2  
 $T = -40\text{ C}$

Rep 1

Rep 2

Rep 3

## ☠ TS 2: Fazit

- Replikate müssen **unabhängige** Versuchseinheiten sein
- Sonst spricht man von Pseudoreplikaten
- Die Grenze zwischen Replikat und Pseudoreplikat ist fliegend
- Auch bei kontrollierten Umgebungen (z.B. Klimakammer) können Fehler passieren
- Mixed Effects Models oder Blockfaktoren können in solchen Situationen weiterhelfen

## ☠ TS 3: Behandlung ist mit anderer Variable vermisch (confounder)



- Wirkstoff gegen leichte Depression wird getestet
- Behandlungsgruppe: 1 h Gespräch mit Arzt und dann Abgabe von Wirkstoff (Tablette)
- Kontrollgruppe: Kein Gespräch und kein Wirkstoff
- Stimmung wird nach 2 Stunden mit Fragebogen abgefragt
- Behandlungsgruppe schneidet signifikant besser ab:

Ist der Wirkstoff ein Erfolg ?

# ☠ TS 3: Behandlung ist mit anderer Variable vermischt (confounder)

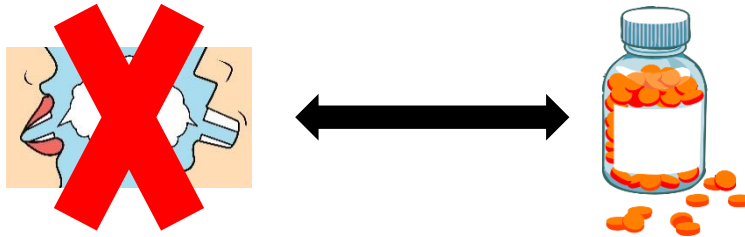


- Wirkstoff gegen leichte Depression wird getestet
- Behandlungsgruppe: 1 h Gespräch mit Arzt und dann Abgabe von Wirkstoff (Tablette)
- Kontrollgruppe: Kein Gespräch und kein Wirkstoff
- Stimmung wird nach 2 Stunden mit Fragebogen abgefragt
- Falls Behandlungsgruppe besser abschneidet: Lag das am Gespräch, dem Wirkstoff oder beidem ?
- Wirkstoff und Gespräch sind vermischt → können Effekte nicht trennen → Studie ist nutzlos

- Wirkstoff gegen leichte Depression wird getestet
- Behandlungsgruppe: 1 h Gespräch mit Arzt und dann Abgabe von Wirkstoff (Tablette)
- Kontrollgruppe: Kein Gespräch und kein Wirkstoff
- Stimmung wird nach 2 Stunden mit Fragebogen abgefragt
- Falls Behandlungsgruppe besser abschneidet: Lag das am Gespräch, dem Wirkstoff oder beidem ?
- Wirkstoff und Gespräch sind vermischt → können Effekte nicht trennen → Studie ist nutzlos

## ☠ TS 3: Behandlung ist mit anderer Variable vermischt (confounder)

- Mögliche Lösung:



Gleiche Bedingungen für Behandlungs- und Kontrollgruppe;  
einziger Unterschied ist Wirkstoffabgabe



# Reaktionszeittest revisited

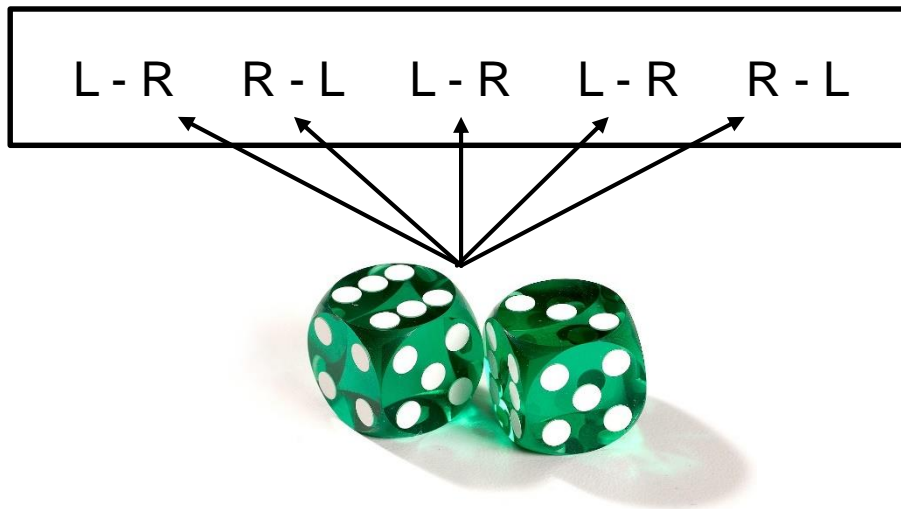
- Ist die rechte Hand schneller als die linke Hand ?
- Jeder Proband: 5 mal hintereinander mit linker Hand, dann 5 mal hintereinander mit rechter Hand
- Falls rechte Hand schneller ist: Gab es Lerneffekte ?
- Zeit und Dominanz der Hand sind vermischt → können Effekte nicht trennen → Studie ist nutzlos



# Reaktionszeittest revisited

## Mögliche Lösung

Zufällige Sequenz der Hände:  
Die Reihenfolge in jedem L-R Block ist zufällig





## ☠ TS 3: Fazit

- Die erklärende Variable von Interesse darf nicht mit einer anderen Variable perfekt korreliert (= vermischt) sein
- Der Effekt kann nicht mehr getrennt werden

## ☠ TS 4: “Wunschdenken”

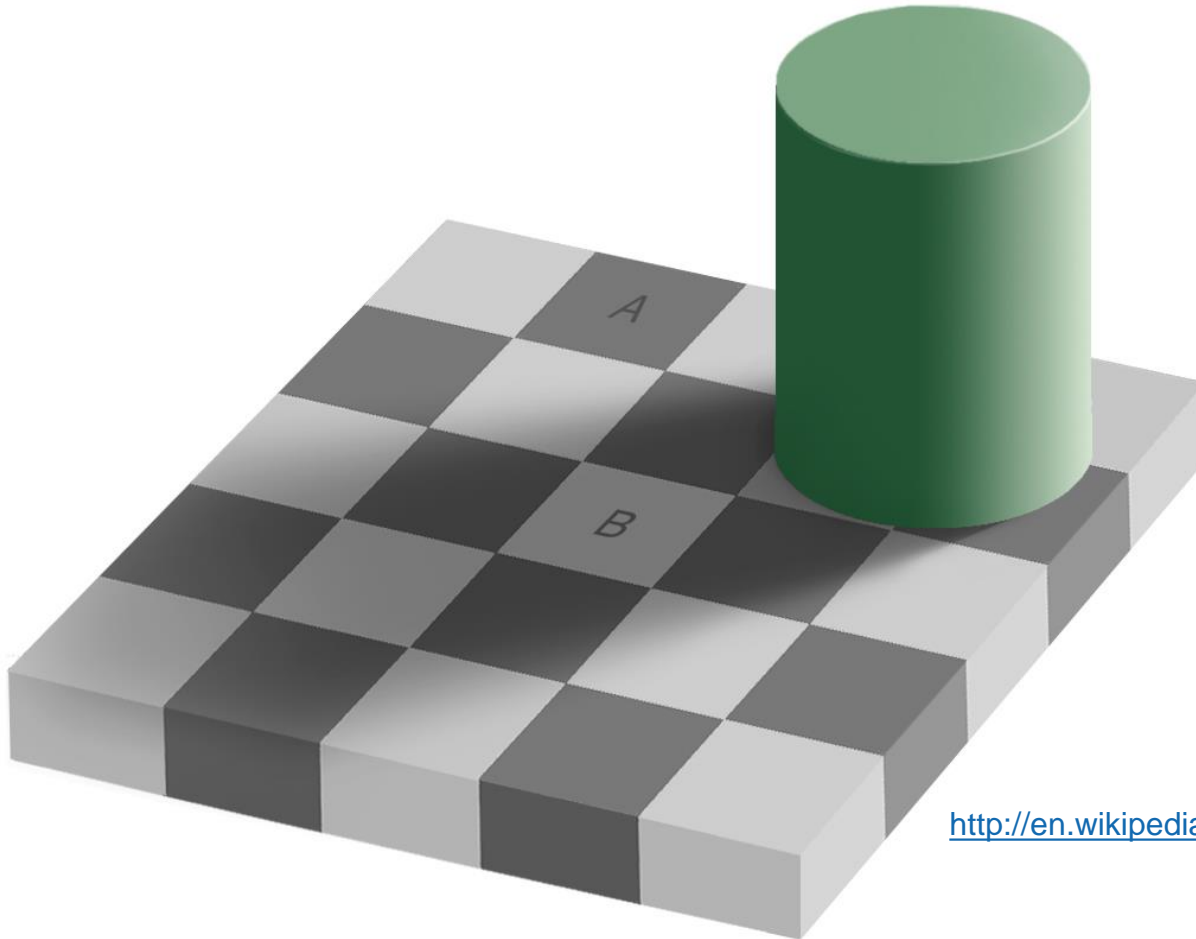
Erwartungen prägen unsere Wahrnehmung

Das betrifft sowohl

- Studienleiter als auch
- Studienteilnehmer



# Auch Studienleiter sind nur Menschen...

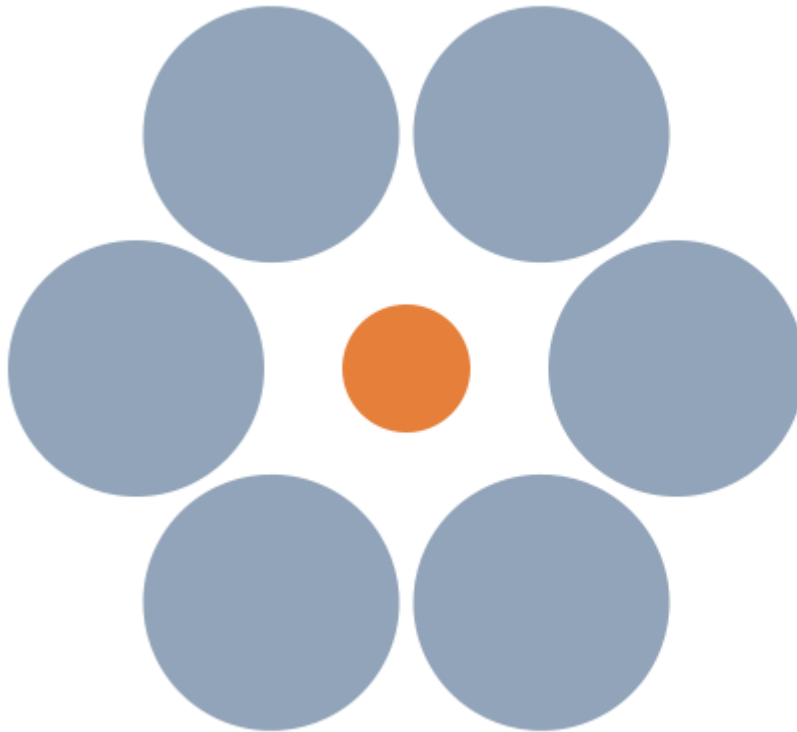


Ist B heller als A ?

Möglichst  
objektiv  
vorgehen !

[http://en.wikipedia.org/wiki/Checker\\_shadow\\_illusion](http://en.wikipedia.org/wiki/Checker_shadow_illusion)

# Auch Studienleiter sind nur Menschen...

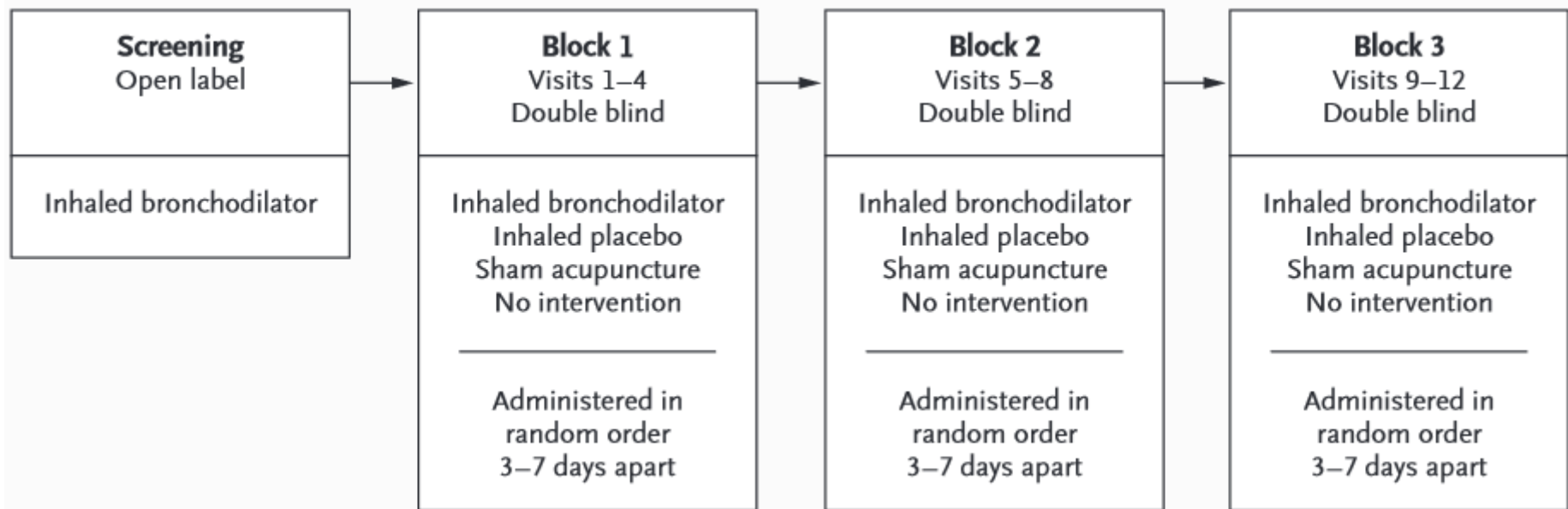


Welcher orangefarbener Kreis  
ist grösser ?

[http://en.wikipedia.org/wiki/Ebbinghaus\\_illusion](http://en.wikipedia.org/wiki/Ebbinghaus_illusion)

# Studienteilnehmer: Placebo

Active Albuterol or Placebo, Sham Acupuncture, or No Intervention in Asthma



[M. Wechsler et.al., N Engl J Med 2011;365:119-26](#)

# Studienteilnehmer: Placebo

Active Albuterol or Placebo, Sham  
Acupuncture, or No Intervention in Asthma

## RESULTS

Among the 39 patients who completed the study, albuterol resulted in a 20% increase in  $FEV_1$ , as compared with approximately 7% with each of the other three interventions ( $P < 0.001$ ). However, patients' reports of improvement after the intervention did not differ significantly for the albuterol inhaler (50% improvement), placebo inhaler (45%), or sham acupuncture (46%), but the subjective improvement with all three of these interventions was significantly greater than that with the no-intervention control (21%) ( $P < 0.001$ ).

[M. Wechsler et.al., N Engl J Med 2011;365:119-26](#)

## ☠ TS 4: Fazit

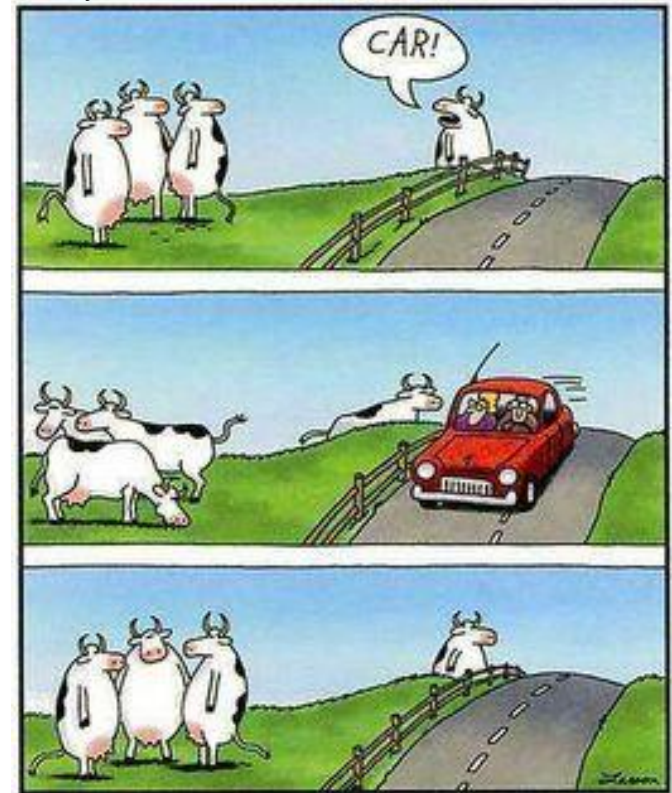
- Erwartungen prägen unsere Wahrnehmung
- Möglichst **objektive Messmethoden**
- “**Doppelblinde**” Studien machen, falls Menschen beteiligt sind



# ☠ TS 5: Verhalten verändert sich durch experimentelles Setting

- Experiment verändert Umfeld
- Daher reagieren Tiere evtl. anders als sie natürlicherweise reagieren würden

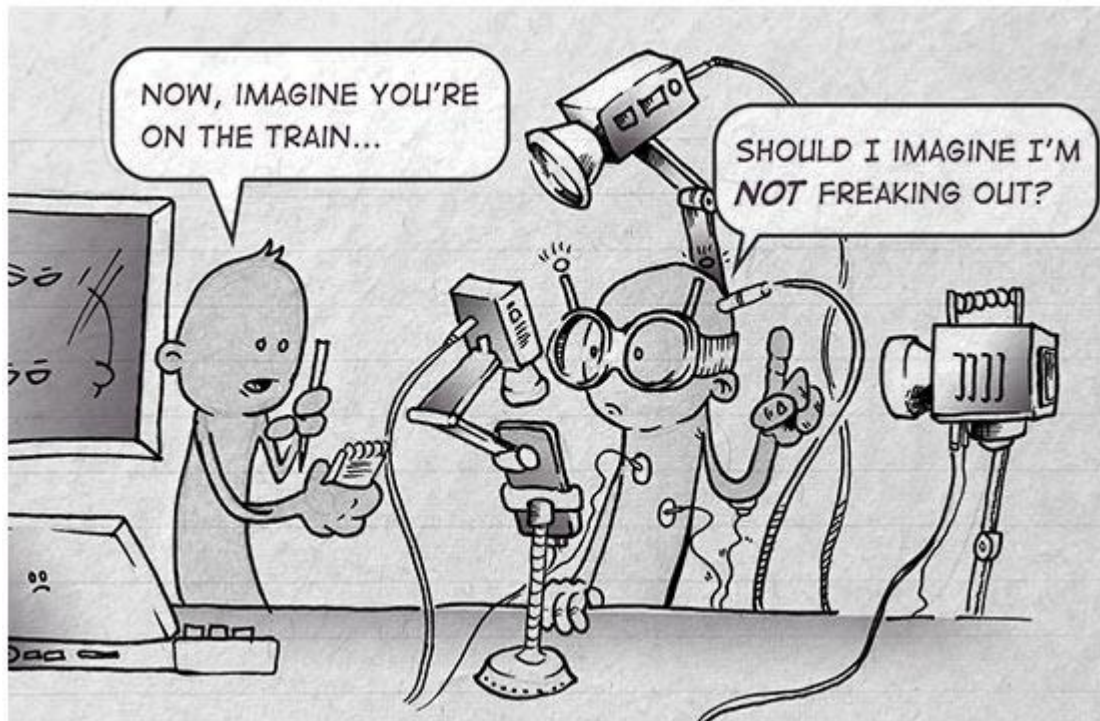
Gary Larson



☠ TS 5



# Umstritten: Hawthorne Effekt



## ☠ TS 5: Fazit

- Das experimentelle Setting könnte das Verhalten der Subjekte verändern
- Störung minimieren: Tarnung, Gewöhnung, etc.

# Exkurs: Metastudie

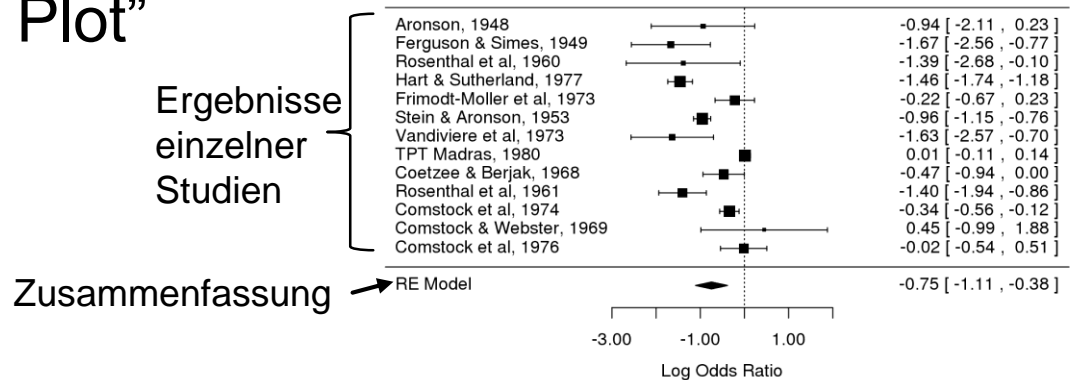
- RCT gilt als Goldstandard um etwas nachzuweisen
  - Metastudie: Fasst viele andere Studien zusammen
- «Diamant»standard: Metastudie von RCTs

- Beispiele:

Medizin: <http://www.cochranelibrary.com/>

Entwicklungshilfe: <http://www.givewell.org/>

- Ergebnis als “Forest Plot”



## ☠ TS 6: Schlechte Kontrollen

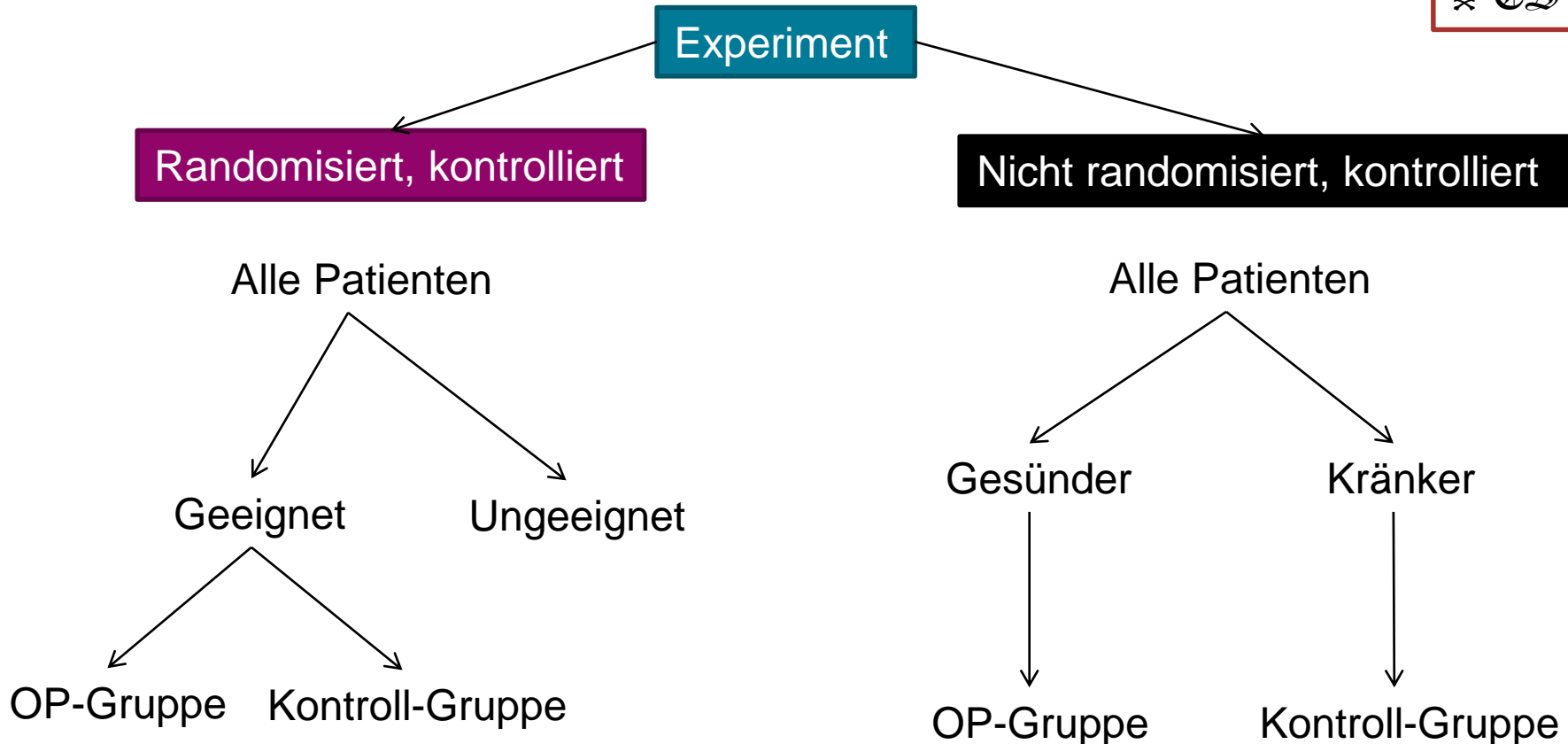
- Leberzirrhose – Shunt (Blutumleitung)
- 51 klinische Studien untersucht: “Bringt die (riskante) Operation einen Vorteil?”

	Ja, sehr	Etwas	Nein
Keine Kontrollgruppe	24	7	1
Kontrollgruppe, nicht randomisiert	10	3	3
Kontrollgruppe, randomisiert	0	1	3

[aus “Statistics”, D. Freedman et.al., 4<sup>th</sup> ed., Kap. 1.2]

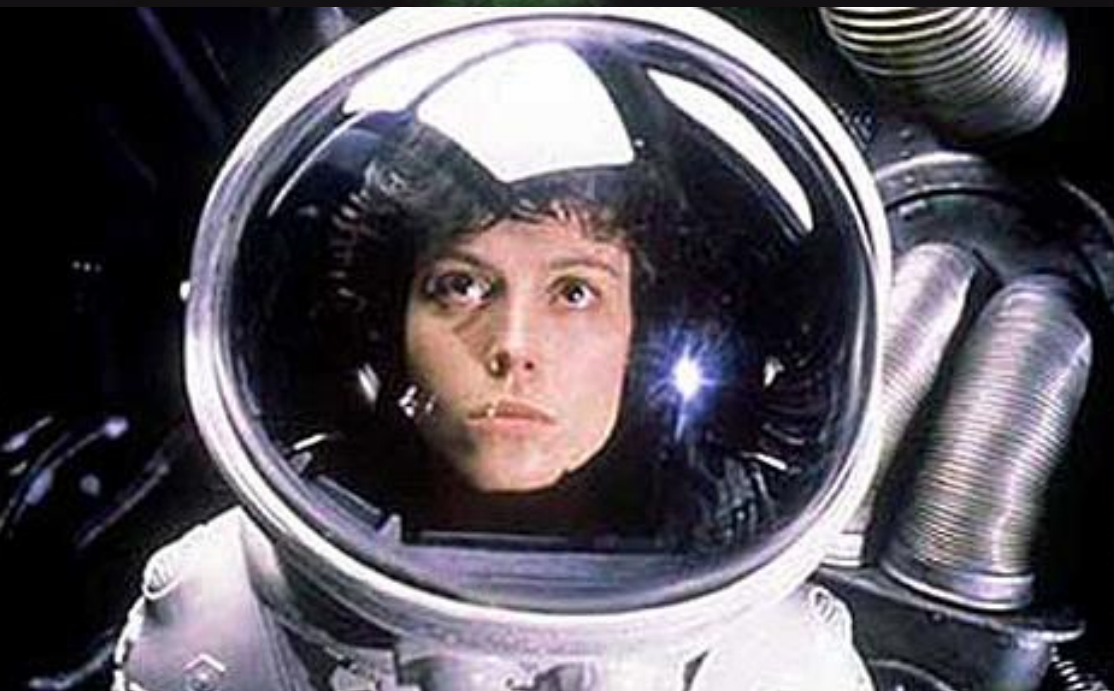
# Problem: Gesundere Patienten werden eher operiert

☠ TS 6



Gesündere Patienten in der OP-Gruppe !

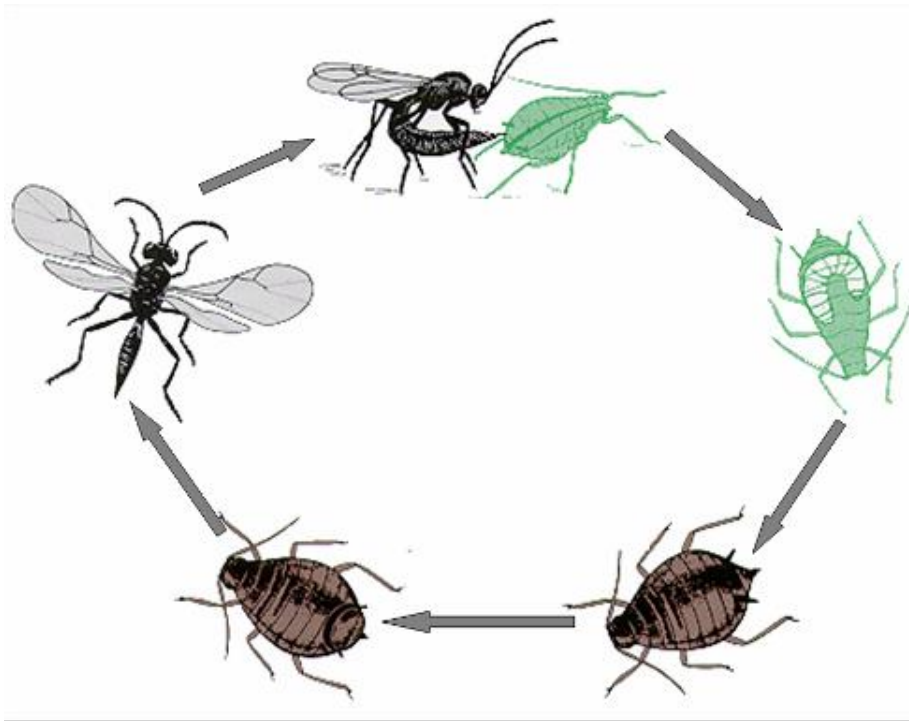
# A L I E N





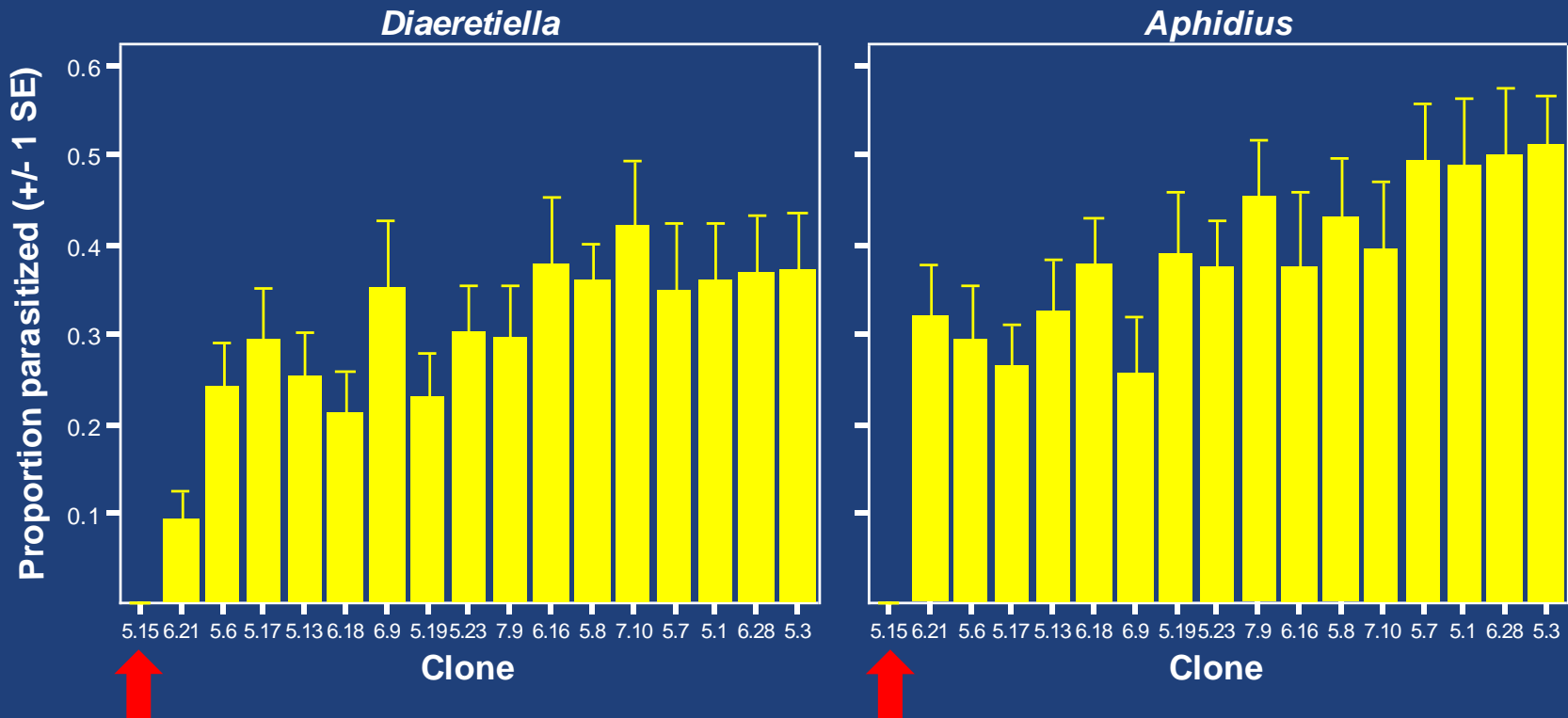
# Gegenbeispiel: Blattläuse und Schlupfwespen

(mit Erlaubnis aus Vortrag von Prof. Dr. Christoph Vorburger, ETHZ)



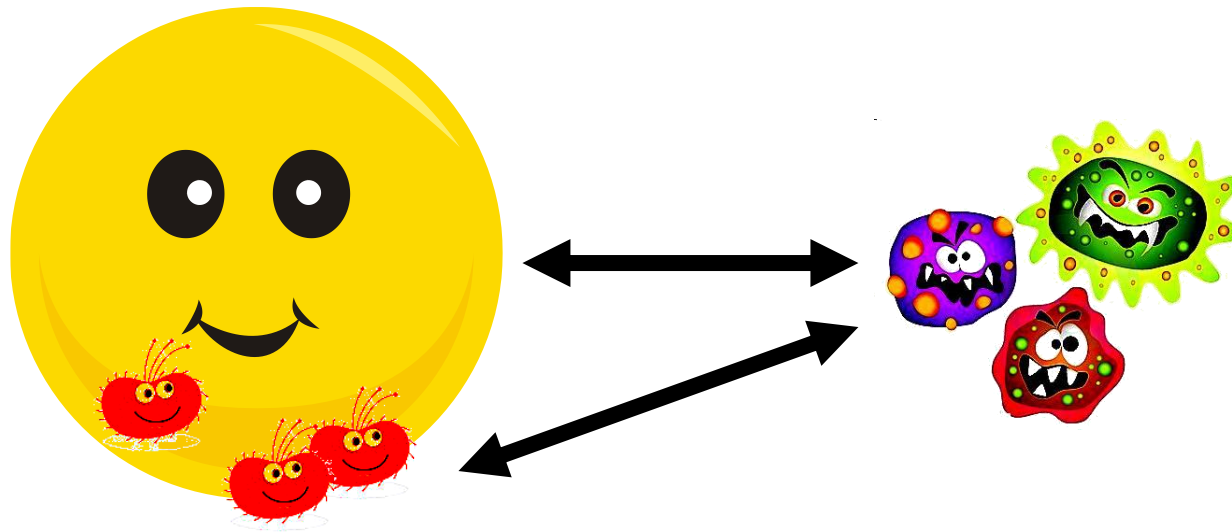
# Resistenter Klon ?

Von Burg *et al.* (2008)  
*ProcB* 275: 1089-1094





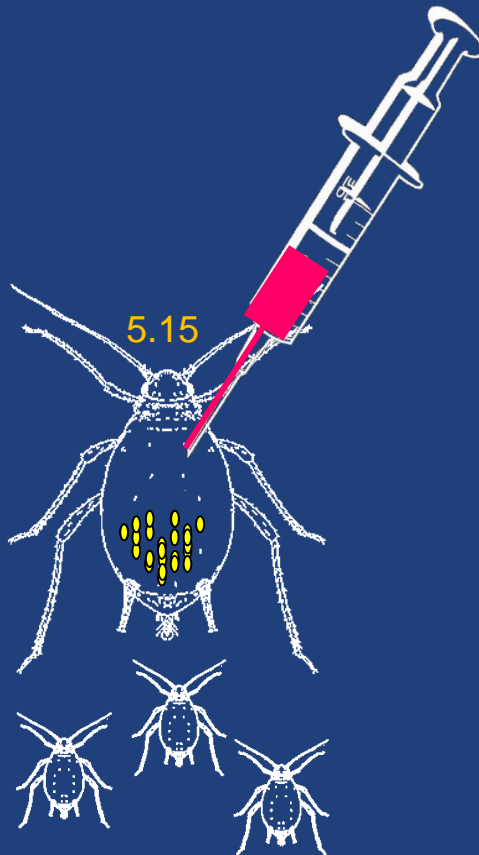
# Vermutung: Endosymbionten (Bakterien) schützen



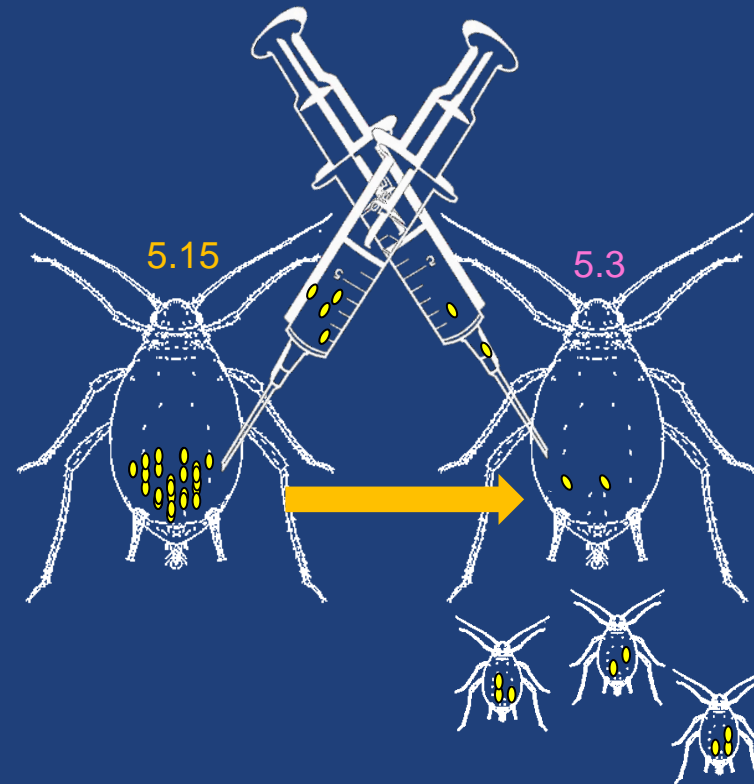
Wie testet man diese Vermutung ?

# Zwei Arten von Kontrolle: Positiv und Negativ

“Heilung”



“Infektion”



# Behandlung und Kontrollen

Aphid clone

*Myzus persicae*

*Aphis fabae*

*Regiella*

5.15

5.3

7.9

A06-405

absent

$n = 10$

$n = 10$

$n = 10$

$n = 10$

present

$n = 10$

$n = 10$

$n = 10$

$n = 10$

“geheilt”

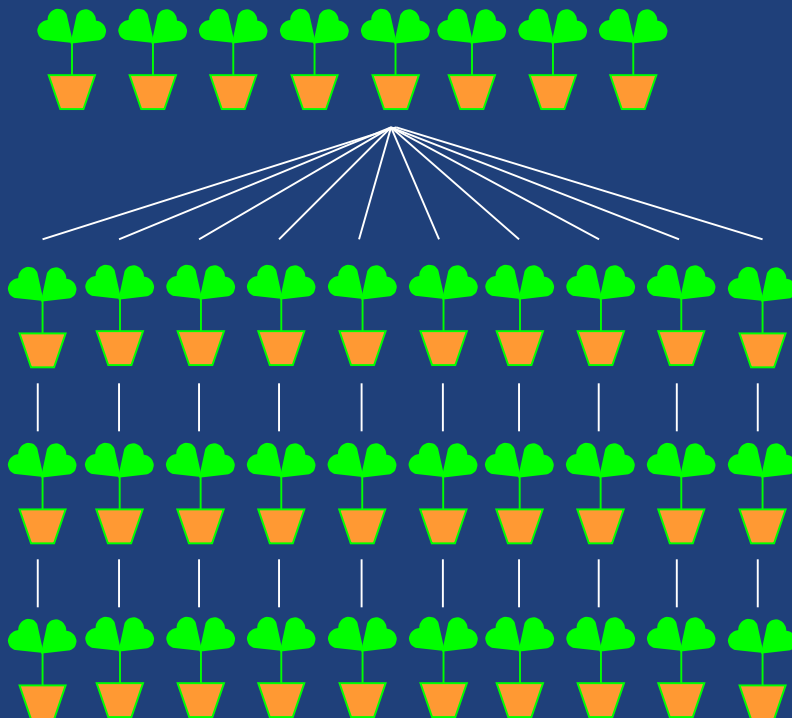
“infiziert”

Gruppen von je ~ 30 Blattläusen für 24h mit Schlupfwespen  
Gemessen: Anteil Verpuppungen



# Wo kommen die 10 Replikate her ?

8 Gruppen



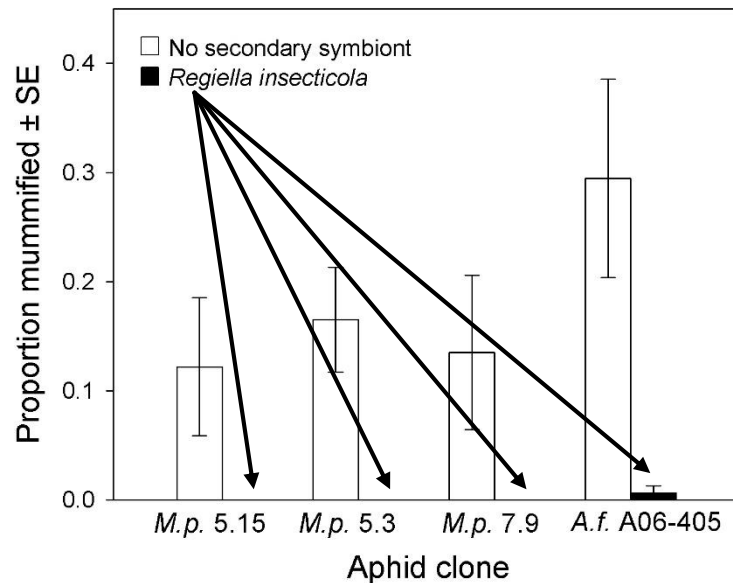
Je 10 Replikate  
3rd = test generation



x 10

10 Blöcke mit  
zufälliger Anordnung

# Ergebnis



Endosymbiont hat einen offensichtlichen Einfluss  
(Auswertung im Paper mit Logistischer Regression)

biology  
**letters**

*Biol. Lett.* (2010) 6, 109–111

doi:10.1098/rsbl.2009.0642

Published online 23 September 2009

Evolutionary biology

## A strain of the bacterial symbiont *Regiella insecticola* protects aphids against parasitoids

Christoph Vorburger\*, Lukas Gehrler  
and Paula Rodriguez

## ☠ TS 6: Fazit

- Sorgfältige Kontrollen sind notwendig für eine überzeugende Schlussfolgerung



Faire Münze ?

## ☠ TS 7: Nullhypothese beweisen



Faire Münze ?



## Faire Münze ? Ja !

- Riddler “beweist” mit 2-seitigem Binomialtest:
- 5 Würfe, 4 mal Kopf  
p-Wert: 0.375  
→  $H_0$  konnte nicht verworfen werden, also muss  $H_0$  stimmen
- **FALSCH:** Falls  $H_0$  nicht verworfen wird war entweder
  - $H_0$  richtig oder
  - wir hatten **zu wenig Macht um eine Abweichung festzustellen**

# Faire Münze ?

- Mit nur 5 Würfeln könnten wir  $H_0: p = 0.5$  niemals verwerfen:

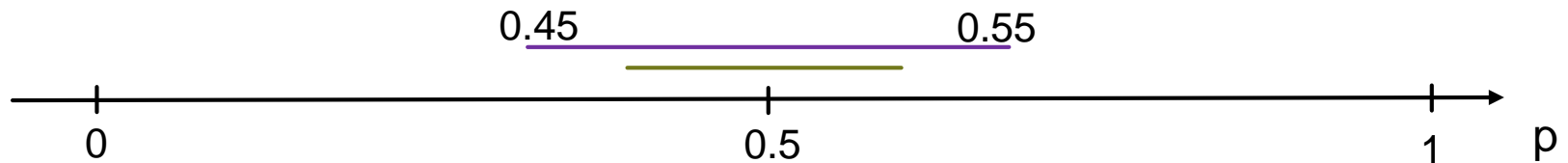
Anzahl Kopf	P-Wert	95%-VI
$x = 0$	0.0625	[0.00; 0.52]
$x = 1$	0.375	[0.01; 0.72]
$x = 2$	1	[0.05; 0.85]
$x = 3$	1	[0.15; 0.95]
$x = 4$	0.375	[0.28; 0.99]
$x = 5$	0.0625	[0.48; 1.00]

95%-VI liefert mehr Information



## Besserer Vorschlag

- Angenommen, für unsere Zwecke ist eine Münze “fair genug”, wenn die Gewinnwa. zw. 0.45 und 0.55 liegt.
- Schwankungen in diesem Bereich sind also **irrelevant**; diese Info kann nur Kontextwissen liefern.
- Falls 95%-VI komplett im irrelevanten Bereich liegt, gilt Münze als fair (Medizin: s. “Äquivalenztest”)



## ☠ TS 7: Fazit

- Wenn man die  $H_0$  nicht verwerfen kann, hat man sie noch lange nicht bewiesen
- Praxis: Vertrauensintervall und irrelevanten Bereich vergleichen

## 💀 Deadly sins 💀

TS 1: Kausalität statt Korrelation

TS 2: Pseudoreplikate

TS 3: Behandlungen haben confounder

TS 4: Beobachter hat Bias

TS 5: Verhaltensänderung wegen Experiment-Setting

TS 6: Schlechte / Keine Kontrollen

TS 7: Nullhypothese “beweisen”



Studien immer auf diese Punkte prüfen !