More Dice less Data

Experimentaldesigns für praktische Fragestellungen

Prof.Dr. Gerrit Hirschfeld

Fachhochschule Bielefeld : FB Wirtschaft : Wirtschaftspsychologie

Was sind Fragen, mit denen ich mich beschäftige

- Ist ein erweitertes Entlassmanagement für Kinder mit chronischen Schmerzen wirklich hilfreich?
- Kann ein neues Berufsbildungsprogramm Jugendliche bei der Jobsuche unterstützen?
- Sind Maßnahmen der Extremismusprävention wirksam?
- Funktioniert eine Präventionskampagne für kritischen Alkoholkonsum bei Jugendlichen?

Vortrag Heute

1. Daten allein machen nicht glücklich: Probleme mit reinen Beobachtungsdaten
2. Dices (Experimente und kontrolliert randomisierten Studien)
3. n=1 Experimente
4. Stepped-wedge Designs

Beispiel 1:Hydroxychloroquine (HCQ) bei Corona?

- Bekannte Patienten berichten von positiven Outcomes.
- Beobachtungsstudien Studien, zeigten, dass HCQ bei Patienten mit positiven Outomes (z.B. Überleben) assoziiert sind.

Abbildung 1: Trump post Covid

Was sagt die Forschung?

- Metaanalysen von Beobachtungsstudien (Castelnuovo et al., 2021) zeigen einen signifikanten Vorteil für HCQ bei Corona.
- Metaanalysen von RCTs (Castelnuovo et al., 2021) zeigen einen nicht-signifikanten Nachteil für HCQ bei Corona.

Welche Verzerrungen führen zu dieser Diskrepanz? (Tleyjeh et al., 2021)

- Treatment Selection Bias: Es erhalten nur die HCQ, denen man noch eine Chance zuspricht und die ggf. auch sonst gut behandelt werden.
- Survival Bias: Patienten, die nach der Aufnahme und vor dem Treatment sterben, werden nicht berücksichtigt.
- ...oder einer der Dutzenden anderer Biases (Sackett, 1979).

Beispiel 2: Serotonin bei Depressionen?

- 1975: Zulassung von Prozac (SSRI) zur Behandlung von Depressionen
- 2022: Studie bringt nachweise, dass Patienten tatsächlich Auffälligkeiten im (Erritzoe et al., 2022)

Abbildung 2: Bsip Sa/Alamy

Ursachen vs. Lösungen

- Die Variablen, die besonders gut etwas vorhersagen sind nicht die, die wir leicht verändern können, wollen, oder dürfen.
- Oft müssen extrem viele Zusatzannahmen getroffen werden, um aus Ursachen Lösungen abzuleiten.
- Viele Variablen (Therapieintensität, Gesprächstechniken), die einen immens großen Einfluss auf die Effektivität haben, haben nichts mit den Ursachen zu tun.

Beispielbefunde

- Mädchen werden öfter Opfer von Cyberbullying.
- Kinder mit Migrationshintergrund haben eine schlechtere Prognose nach Entlassung.
- Einkommen der Eltern limitiert den Zugang zu medizinischer Versorgung.

Far better an approximate answer to the right question, which is often vague, than the exact answer to the wrong question, which can always be made precise.

John W. Tukey

Beispiel 3: Amazon's Algoritmus (Dastin, 2018)

- 2014: Kommen Techniker bei Amazon auf Idee Bewerbungsunterlagen automatisch zu bewerten (1-5 Sterne).
- Bewerbungen der letzten 10 Jahre als Grundlage explizite Geschlechtsinfos werden gelöscht.
- Dennoch gibt es weniger Punkte für weibliche Bewerber.

Abbildung 3: hanse / unsplash

Gesetzmäßigkeiten in sozialen Systemen

- Oft entstehen Daten nur aus bestimmten Praktiken und Regelungen .
- Gerade in sozialen Systemen muss man sich klar sein, dass die entdeckten Gesetzmäßigkeiten sozial konstruiert sind.
- Gelegentlich wollen wir das auch ändern.

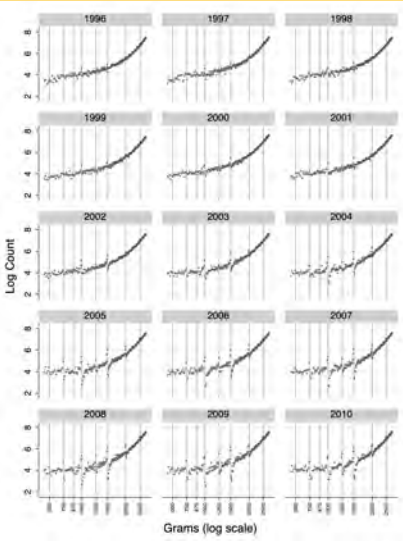


Abbildung 4: (Jürges & Köberlein, 2015)

Zwischenfazit:

Wenn wir an Veränderung interessiert sind, reichen Daten alleine oft nicht aus

- Viele Biases verzerren die Ergebnisse
- Fokussieren nicht auf interventionsrelevante Variablen
- Gesetzmäßigkeiten in sozialen Systemen

Echte Experimente (aka kontrolliert-randomisierte Studien) helfen diese Probleme zu addressieren.

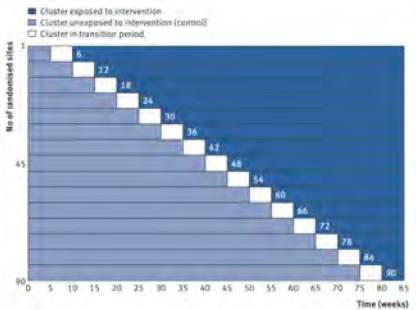
- Kontrolle durch Randomisierung verringert viele - aber nicht alle - Biases.
- Durch Manipulation fokussieren wir auf die Variablen, die wirklich veränderbar sind.

Warum macht man das nicht öfter?

- **Praktische Umsetzbarkeit**:
  - Oft können einzelne Personen gar nicht randomisiert werden (Gruppeninterventionen, bestehende Cluster).
  - Oft gibt es nur sehr wenige Betroffene.
  - Follow-ups sind teuer.
- **Ethische Bedenken**:
  - Aber dann bekommen ja einige der Studienteilnehmer eine weniger effektive Behandlung.

Stepped-Wedge Designs ...

- sind eine Sonderform der Clusterrandomisierten Trials.
- Erlauben eine Randomisierung während der Roll-out Periode.
- Ermöglichen es auch zeitliche Effekte abzuschätzen.
- Sind oft effizienter als parallele Cluster designs. (Hemming et al., 2015)



Beispielstudie: Depressionsmanagement in Altenheimen (Leontjevas et al., 2013)

- Act in Depression = Komplexe Intervention.
- Relativ lange Vorlaufzeit, bis man das umsetzen kann.
- Praktisch schwer auf einzelne Patienten runterzubrechen.
- Ethisch schwierig wenn man das umsetzt

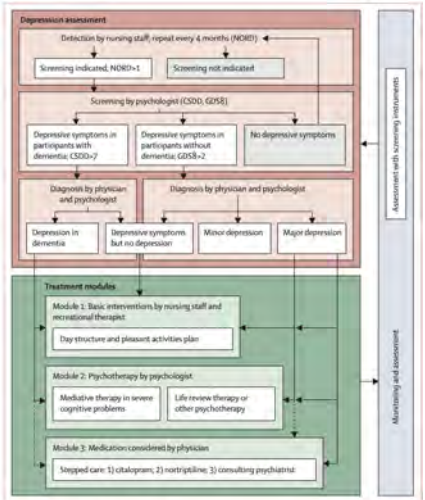


Figure 2: Act in Case of Depression pathways
Dotted arrows indicate that the treatment was to be considered if symptoms are severe, or when a psychosocial treatment (ie, module 1 and module 2) was not effective. NORD=Nijmegen observer rated depression scale.[44] CSDD=Cornell scale for depression in dementia.[45] GDS8=eight-item geriatric depression scale.[46]

Beispielstudie: Depressionsmanagement in Altenheimen (Leontjevas et al., 2013)

- Evaluation Mit Hilfe eines Stepped-Wedge Designs
- Randomisierung der einzelnen Altenheime (n= 33; jeweils eine somatische und eine Demenz-Station) in 6 Gruppen, die zu unterschiedlichen Zeitpunkten mit dem Programm beginnen.
- Primäres Outcome: Anteil der Depressiven Bewohner (anhand des CSSD Scores)

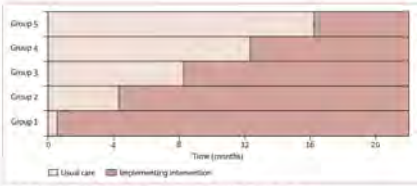
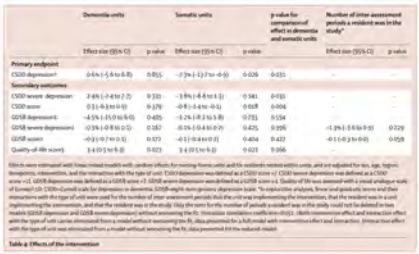


Figure 1: Stepped-wedge design with six measurements and five cluster groups
No clusters receive the intervention at baseline. They are randomly assigned to five groups that crossover to receive the intervention after measurements at 4-month intervals. How long a group has been receiving the intervention at 20 months varies from about 4 months (group 5) to 20 months (group 1).

Abbildung 5: (Leontjevas et al., 2013)

Beispielstudie: Depressionsmanagement in Altenheimen (Leontjevas et al., 2013)

- Ergebnisse zeigen einen Vorteil der Intervention in Altenheimen nach dem Wechsel in die Interventionsphase.
- Dies gilt aber nur für somatische Stationen.



	Dementia units		Somatic units		p value for comparison of effect in dementia and somatic units	Number of inter-assessment periods a resident was in the study*	
	Effect size (95% CI)	p value	Effect size (95% CI)	p value		Effect size (95% CI)	p value
Primary endpoint							
CSDD depression†	0·44 (−1·6 to 0·8)	0·815	−7·74 (−13·7 to −0·9)	0·026	0·031	..	..
Secondary outcomes							
CSDD severe depression	2·48 (−2·41 to 7·7)	0·331	−3·84 (−8·4 to 1·1)	0·141	0·031	..	..
CSDD score	0·31 (−0·3 to 0·9)	0·379	−0·81 (−1·4 to −0·2)	0·018	0·004	..	..
GDSR depression‡	−4·5% (−15·0 to 6·0)	0·405	−1·2% (−9·2 to 5·8)	0·731	0·554	..	..
GDSR severe depression‡	−2·9% (−9·8 to 2·5)	0·257	−0·1% (−0·4 to 0·2)	0·425	0·295	−1·3% (−3·6 to 0·9)	0·229
GDSR score‡	−0·6 (−0·7 to 0·5)	0·972	−0·1 (−0·4 to 0·2)	0·404	0·422	−0·1 (−0·3 to 0·0)	0·059
Quality-of-life score§	1·4 (0·5 to 5·3)	0·023	1·4 (0·5 to 5·3)	0·023	0·996	..	..

Effects were estimated with linear mixed models with random effects for nursing home units and for residents nested within units, and were adjusted for sex, age, hypothesis, intervention, and the interaction with the type of unit. CSDD depression was defined as a CSDD score >7. CSDD severe depression was defined as a CSDD score >12. GDSR depression was defined as a GDSR score >7. GDSR severe depression was defined as a GDSR score >14. Quality of life was assessed with a visual analogue scale of Cornell-CD. CSDD=Cornell scale for depression in dementia. GDSR=eight-item geriatric depression scale. *In explorative analyses, linear and quadratic terms and their interactions with the type of unit were used for the number of inter-assessment periods that the unit was implementing the intervention, that the resident was in a unit implementing the intervention, and that the resident was in the study. Only the term for the number of periods a resident was in the study could not be deleted in two models (GDSR depression and GDSR severe depression) without worsening the fit. †Intraclass correlation coefficient=0·051. ‡Both intervention effect and interaction effect with the type of unit can be eliminated from a model without worsening the fit; data presented for a full model with intervention effect and interaction. §Interaction effect with the type of unit was eliminated from a model without worsening the fit; data presented for the reduced model.

Table 2: Effects of the intervention

Abbildung 6: (Leontjevas et al., 2013)

Stepped Wedge Designs

- Beispiel für Randomisierung, die nicht auf Ebene einzelner Personen stattfindet.
- Alle Zentren bekommen die gute Intervention.
- Etwas aufwändig zu planen und auszuwerten.

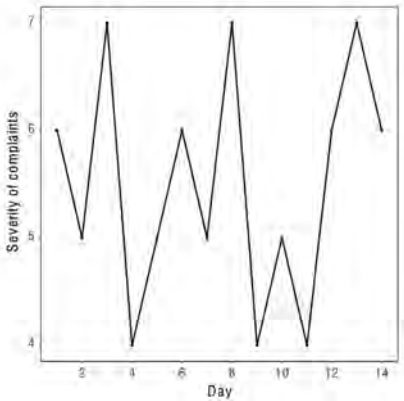
Abbildung 7: Geht das auch etwas kleiner?

Einzelfallexperimente (Onghena, 2020)

- Experimente, bei denen eine Beobachtungseinheit wiederholt in verschiedenen Bedingungen mindestens einer manipulierter Variable beobachtet wird.

Vorgehen

1. Festlegen eines Designs (Anzahl der Beobachtungen und Art der Randomisierung: Komplett zufällig, Randomized Block)
2. Messung der AV
3. Vergleich der Teststatistik mit allen möglichen Ergebnissen



Einzelfallexperimente (Onghena, 2020)

- Experimente, bei denen eine Beobachtungseinheit wiederholt in verschiedenen Bedingungen mindestens einer manipulierter Variable beobachtet wird.

Vorgehen

1. Festlegen eines Designs (Anzahl der Beobachtungen und Art der Randomisierung: Komplett zufällig, Randomized Block)
2. Messung der AV
3. Vergleich der Teststatistik mit allen möglichen Ergebnissen

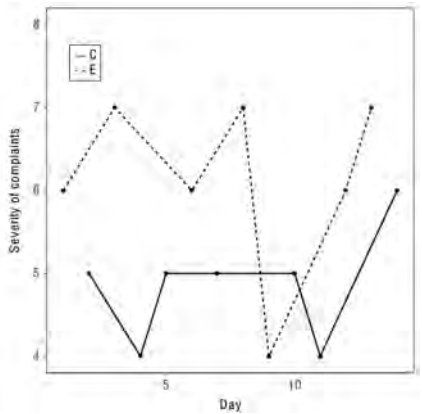


Abbildung 8: (Onghena, 2020)

Einzelfallexperimente (Onghena, 2020)

- Experimente, bei denen eine Beobachtungseinheit wiederholt in verschiedenen Bedingungen mindestens einer manipulierter Variable beobachtet wird.

Vorgehen

1. Festlegen eines Designs (Anzahl der Beobachtungen und Art der Randomisierung: Komplett zufällig, Randomized Block)
2. Messung der AV
3. Vergleich der Teststatistik mit allen möglichen Ergebnissen

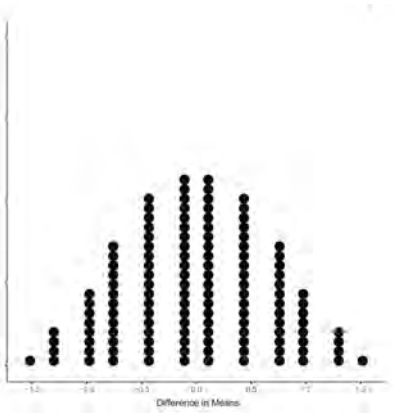


Abbildung 9: (Onghena, 2020)

Wichtige Designs

- Completely Randomized Design
- Randomized Block Design: Paarweise wird randomisiert (AB; BA)
- Phase Designs: AB Phase Design, ABA Phase Design, ABAB Phase Design
- Alternating Treatments Design (Maximalanzahl der wiederholten Bedingungen)
- Multiple Baseline Design: Randomisierung des Startzeitpunkts über verschiedene Probanden.

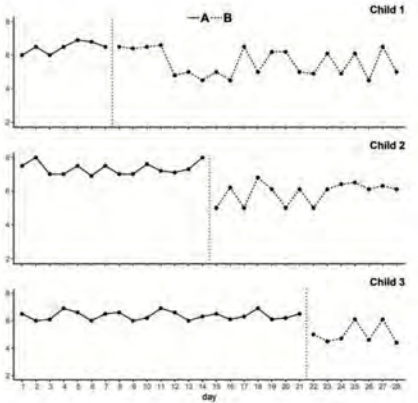


FIGURE 6.4 Hypothetical data for a single-case experiment using a four-week multiple-baseline design across three children for evaluating the effect of a behavioral intervention on the general distress level of the children as assessed daily by the staff at a daycare center

Abbildung 10: (Onghena, 2020)

Single Case Studies

- Möglichkeit an Einzelfällen
- Mindestens 5 Beobachtungen pro Bedingung
- Soviele Wechsel wie möglich
- Integration der Ergebnisse verschiedener Probanden über Metaanalysen
- Leicht umzusetzen: https://tamalkd.shinyapps.io/scda/

Vielen Dank für Ihre Aufmerksamkeit

Gerrit.Hirschfeld@fh-bielefeld.de

Wer will, dass die Welt so bleibt, wie sie ist, der will nicht, dass Sie bleibt.
Erich Fried

Referenzen I

Castelnuovo, A. D., Costanzo, S., Cassone, A., Cauda, R., Gaetano, G. D., & Iacoviello, L. (2021). Hydroxychloroquine and mortality in COVID-19 patients: A systematic review and a meta-analysis of observational studies and randomized controlled trials. *Pathogens and Global Health*, *115*(7-8), 456–466. https://doi.org/10.1080/20477724.2021.1936818

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics* (pp. 296–299). Auerbach Publications.

Erritzoe, D., Godlewska, B. R., Rizzo, G., Searle, G. E., Agnorelli, C., Lewis, Y., Ashok, A. H., Colasanti, A., Boura, I., Farrell, C., Parfit, H., Howes, O., Passchier, J., Gunn, R. N., Nutt, D. J., Cowen, P. J., Knudsen, G., & Rabiner, E. A. (2022). BRAIN SEROTONIN RELEASE IS REDUCED IN PATIENTS WITH DEPRESSION: A [11C]cimbi-36 PET STUDY WITH a d-AMPHETAMINE CHALLENGE. *Biological Psychiatry*. https://doi.org/10.1016/j.biopsych.2022.10.012

Hemming, K., Haines, T. P., Chilton, P. J., Girling, A. J., & Lilford, R. J. (2015). The stepped wedge cluster randomised trial: Rationale, design, analysis, and reporting. *Bmj*, *350*, h391.

Jürges, H., & Köberlein, J. (2015). What explains DRG upcoding in neonatology? The roles of financial incentives and infant health. *Journal of Health Economics*, *43*, 13–26. https://doi.org/10.1016/j.jhealeco.2015.06.001

Leontjevas, R., Gerritsen, D. L., Smalbrugge, M., Teerenstra, S., Vernooij-Dassen, M. J., & Koopmans, R. T. (2013). A structural multidisciplinary approach to depression management in nursing-home residents: A multicentre, stepped-wedge cluster-randomised trial. *The Lancet*, *381*(9885), 2255–2264. https://doi.org/10.1016/s0140-6736(13)60590-5

Onghena, P. (2020). *SMALL SAMPLE SIZE SOLUTIONS* (R. van de Schoot & M. Miocevic, Eds.). Routledge.

Sackett, D. L. (1979). Bias in analytic research. *Journal of Chronic Diseases*, *32*(1-2), 51–63. https://doi.org/10.1016/0021-9681(79)90012-2

Tleyjeh, I. M., Kashour, T., Mandrekar, J., & Petitti, D. B. (2021). Overlooked shortcomings of observational studies of interventions in coronavirus disease 2019: An illustrated review for the clinician. *Open Forum Infectious Diseases*, *8*(8). https://doi.org/10.1093/ofid/ofab317