Hochschule Niederrhein
University of Applied Sciences
Gesundheitswesen
Faculty of Health Care

Bewertung von RCT-Studienpublikationen.
Hinweise zur Krefelder Ampel.

Jörg große Schlarmann[1], Matthias Mertin[1]
1 Hochschule Niederrhein, Fachbereich Gesundheitswesen
Kontakt: Jörg große Schlarmann, joerg.grosseschlarmann@hs-niederrhein.de

Modul 10 EBN1 13.10.2023

Abstract: In diesem Text werden Hinweise zur Bewertung von Studienpublikationen gegeben. Dabei wird der Schwerpunkt auf randomisiert kontrollierten Studien (RCTs) gelegt. Anhand der Fragen der Krefelder Ampel wird ausgeführt, welche Informationen in der Pubikation enthalten sein sollten, und welche Auswirkungen diese Informationen auf die Studienergebnisse sowie deren Glaubwürdigkeit haben.

! Niemals graue Literatur zitieren!!!

Achtung, bei diesem Artikel handelt es sich um *graue Literatur*. Er ist für Ihre Vor- und Nachbereitung gedacht. Dieser Artikel sollte **auf keinen Fall** zitiert werden, da er für andere (außerhalb des Studiengangs) nicht zugänglich ist und kein Peer-Review-Verfahren durchlaufen hat.

1 Einleitung

Wenn ein Forschungsprojekt abgeschlossen ist, werden die Ergebnisse in Form von Fachpublikationen (*papers*) in Fachzeitschriften (*journals*) veröffentlicht. Diese Artikel sind für Außenstehende der einzige Zugang zu den durchgeführten Erhebungen, Analysen und Ergebnissen. Die Bewertung der Glaubwürdigkeit solcher Studienergebnisse stellt eine wesentliche Kompetenz in der Wissenschaft dar. Sicherlich ist es begrüßenswert, wenn englische Texte verstanden und Inhalte wiedergeben werden können. Die große Kunst liegt aber darin zu *verstehen*, mit welcher Methode die Ergebnisse erzielt wurden, was die dargestellten Tabellen und Abbildungen bedeuten, und ob die bereitgestellten Informationen plausibel, glaubwürdig und vollständig sind. Diese Bewertungen gestalten sich abhängig vom Studiendesign (Fall-Kontroll-Studie, Kohortenstudie, randomisiert-kontrollierte Studie, usw.) unterschiedlich.

Dieser Artikel konzentriert sich auf die Bewertung von Publikationen randomisiert-kontrollierter Studien (*randomised controlled trials*, RCT)[1].

Für Autor_innen von RCT-Studien steht das Publikationsstatement CONSORT zur Verfügung (Schulz et al. (2010)), in welchem formuliert wurde, welche Informationen an welcher Stelle publiziert werden sollten, so dass sich Lesende einen umfassenden Überblick über die Forschung, deren Durchführung sowie deren Ergebnisse verschaffen können. Für die Studienbewertung ist es zur Überprüfung der Informationsvollständigkeit (siehe Abschnitt 2.17) hilfreich.

Für Lesende gibt es darüber hinaus eigene Beurteilungshilfen zur Studienbewertung (z.B bei Behrens & Langer (2022)).

2 Fragen der Krefelder Ampel

An der Hochschule Niederrhein verwenden wir im Pflegestudiengang die *Krefelder Ampel*. Sie besteht aus 18 Fragen an die Studie, die - gemäß einer Ampel - mit grün, gelb oder rot markiert werden können, je nachdem, wie *gut* die Antworten ausfallen. So erhalten Sie - nach dem Ausfüllen der Ampel - eine Übersicht über die Güte der Studie.

Im Folgenden wird für jede Fragen der Krefelder Ampel ausgeführt, welche Informationen in der Pubikation enthalten sein sollten, und welche Auswirkungen diese Informationen auf die Studienergebnisse sowie deren Glaubwürdigkeit haben.

[1] zu RCTs siehe Mad et al. (2008) und Kabisch et al. (2011)

2.1 Ist die Forschungsfrage (oder das Forschungsziel) klar formuliert?

Der Artikel sollte die konkrete Forschungsfrage enthalten, möglichst nach dem PICO-Schema (Schardt et al. (2007)) aufgebaut, so dass die Elemente

- **P**opulation (*bei wem*)
- **I**ntervention (*wird* was *gemacht*)
- **C**ontrol (*im Vergleich zu wem oder was*)
- **O**utcome (*hinsichtlich was*)

hinreichend exakt beschrieben sind. Mindestens sollte das Forschungsziel beschrieben sein, welches ebenfalls die genannten Elemente enthalten sollte. Nur wenn Forschungsziel oder -frage bekannt sind, können die Ergebnisse sinnvoll interpretiert werden. Da die Forschungsfrage das Forschungsdesign und die -methoden bestimmt, sollten die Leser bereits an dieser Stelle eine eigene Vorstellung davon haben, wie man vorgehen sollte, um die Frage exakt beantworten zu können.

2.2 Sind die gewählten Messwerte (Variablen) geeignet, um die Forschungsfrage zu beantworten?

Grundsätzlich sollte man sich fragen, ob die gemessenen Variablen geeignet sind, die Forschungsfrage zu beantworten. Eine Intervention, die Jugendliche über die Anwendung der Ernährungsampel schulen möchte, sollte beispielsweise nicht mittels der Variable `Gewichtsverlust` evaluiert werden. Die gewählten Variablen müssen geeignet sein, das gewählte Outcome der Fragestellung (Abschnitt 2.1) abzubilden.

2.3 Ist die Intervention ausreichend beschrieben?

Die Studie muss ausführlich und nachvollziehbar beschreiben, welche Art von Maßnahmen wann und wie oft und von wem in den gebildeten Gruppen durchgeführt werden, und wann und wie welche Variablen (Daten) gemessen werden.

Wie genau unterscheiden sich die Maßnahmen in der Interventionsgruppe von solchen in der Kontrollgruppe? Sind diese Maßnahmen geeignet, um die Forschungsfrage zu beantworten? Kann es sein, dass Komponenten der Intervention als Confounder (Störvariablen) zum Ergebnis oder zum Drop-Out (siehe Abschnitt 2.10) beitragen? Vor vielen Jahren wurden Dekubiti mit “Eisen und Föhnen” behandelt, und die Wunden heilten besser. Hierbei wurde übersehen, dass die Patienten zum “Eisen oder Föhnen” auf die Seite gedreht werden mussten, und dass dieser Lagewechsel den Heilungsprozess maßgeblich beeinflusst.

Eine detaillierte Beschreibung der Intervention ist entscheidend, um bewerten zu können, ob die Maßnahmen konsistent und standardisiert durchgeführt wurden. Dies ermöglicht es, die Ergebnisse hinsichtlich des Kausalzusammenhangs zwischen Maßnahmen und Ergebnissen zu bewerten.

2.4 Ist ausreichend beschrieben, wo das Forschungsvorhaben an wem durchgeführt wurde?

Kulturen sind verschieden, Gesundheitssysteme auch, und die Ergebnisse einer Studie können je nach geografischer Lage und Bevölkerungsgruppe variieren. Es ist wichtig zu wissen, an welchem Ort (in welchem Land, welcher Region, welchem Setting) die Studie durchgeführt wurde, und aus welcher Population die Stichprobe gezogen wurde. Dies hilft bei der Bewertung, ob die Ergebnisse nur in einer bestimmten Umgebung oder bei einer spezifischen Bevölkerungsgruppe gültig sind, oder auch auf andere Kontexte übertragbar sein könnten.

Die Beschreibung des Studienorts und der Population ermöglicht es den Lesern zudem, potenzielle kontextuelle Unterschiede zu ihrem eigenen Standort zu berücksichtigen und zu verstehen, wie sich diese auf die Ergebnisübertragung auswirken könnten (siehe Abschnitt 2.18).

Davon abgesehen stellt sich die Frage, wie plausibel es ist, dass die Interventionen in der beschriebenen Form wirklich an diesem Ort (in diesem Setting) durchgeführt worden sind.

2.5 Ist ausreichend beschrieben, wie die Probanden rekrutiert wurden?

Mit der Stichprobe steht und fällt die Güte der Forschung. Daher muss exakt beschrieben sein, wie die Stichprobe zustande gekommen ist. Bei einer Zufallsstichprobe wurden aus der Gesamtpopulation (z.B. alle adipösen Männer in Deutschland) zufällig x Personen ausgewählt, wobei jede Person dieselbe

Chance hatte, für die Studie gezogen zu werden. Solche Stichproben sind sehr selten. Ihre Ergebnisse lassen sich dafür besonders gut auf die Allgemeinheit übertragen. Bei Gelegenheitsstichproben "nimmmt man, was gerade kommt", also z.B. solche Patienten, die während eines bestimmten Zeitraums mit Magenblutung in die Notaufnahme unserer Klinik kommen. Solche Stichproben sind viel einfacher zu realisieren als echte Zufallsstichproben, jedoch sind ihre Ergebnisse schlechter verallgemeinerbar und gelten erst einmal nur für die gezogene Stichprobe.

Die Ein- und Ausschlusskriterien für die Studienteilnahme müssen klar beschrieben werden, damit ersichtlich wird, ob diese zur Beantwortung der Fragestellung sinnvoll gewählt wurden (oder zu lasch, oder zu restriktiv). Ebenso sollte beschrieben sein, wer diese Kriterien auf potentielle Probanden anwendet. Hätte diese Person ein potentielles Interesse sowie die Möglichkeit, bestimmte Menschen als Probanden aufzunehmen (oder abzulehnen), obwohl die Ausschluss- (oder Einschlusskriterien) klar dagegen (oder dafür) gesprochen haben?

2.6 War die Stichprobe ausreichend groß?

Wenn Forschende einen Unterschied "zeigen" möchten, dann hängt es von der Größe des erwarteten Unterschieds (Effektgröße) ab, wie viele Probanden untersucht werden müssen. Bei großen Unterschieden (großen Effekten) werden weniger Probanden benötigt, als bei kleinen. Möchte man zeigen, dass Basketballspieler größer sind als Kindergartenkinder, so werden nur wenige Basketballer und Kindergartenkinder benötigt (eben weil der Größenunterschied so deutlich ist). Möchte man den Größenunterschied von Studierenden der Hochschule Niederrhein mit jenen der Universität Bochum vergleichen, so wird man sehr viele Studierende benötigen, da davon auszugehen ist, dass der Unterschied - wenn er denn wirklich da ist - sehr klein ist. Aus forschungsethischer Sicht dürfen keine unnötigen Daten erhoben werden, es sollten also nicht prinzipiell große Stichproben angestrebt werden, nur weil damit sehr kleine Unterschiede aufgezeigt werden können.

Forschende müssen daher möglichst präzise die Anzahl an Probanden errechnen, die benötigt wird, um den gewünschten Effekt "sehen" zu können (sofern er denn da ist). Dies nennt man Fallzahlkalkulation oder Poweranalyse[2]. Standardmäßig wird in der Wissenschaft eine Power (die Wahrscheinlichkeit, den Effekt zu sehen, sofern er denn da ist; dargestellt durch β) von 80% als akzeptabel angesehen. Die benötigte Fallzahl kann mit Hilfe von Computerprogrammen wie `G*Power`[3] oder `R`[4] berechnet werden. Anschließend wird der zu erwartende Drop-Out oben aufgeschlagen, also die erwartete Anzahl an Probanden, die während der Studie herausfallen werden, z.B. weil sie versterben oder aus anderen Gründen ihre Teilnahme widerrufen. Die kalkulierte Fallzahl + der erwartete Drop-Out ergeben so die angestrebte Stichprobengröße.

Wenn eine Studie die benötigte Fallzahl nicht erreicht, z.B. durch einen unerwartet hohen Drop-Out, dann darf man sich nicht wundern, wenn der Effekt nicht gefunden werden kann (dass "nichts dabei herauskommt").

Studien, die keine (nachvollziehbare) Fallzahl kalkuliert haben, "schießen ins Blaue" und sollten eher als Pretest für größere Studien angesehen werden.

2.7 Kann ausreichend nachvollzogen werden, wie die Probanden in Interventions- und Kontrollgruppe(n) eingeteilt wurden?

Ein Alleinstellungsmerkmal des RCTs ist die zufällige Aufteilung der Probanden in Interventions- und Kontrollgruppe(n). Es muss nachvollziehbar beschrieben sein, wie die Randomisierung abgelaufen ist, und welche Person die Zuteilung der Probanden vorgenommen hat. Es muss ersichtlich sein, wie der Zufall generiert wurde (z.B. Blocklisten), und wie die verdeckte Zuteilung (allocation concealment) gewährleistet wurde. Forschende werden immer (un)bewusst versuchen, die Ergebnisse in ihrem Sinne zu beeinflussen - das ist nur menschlich ("führe mich nicht in Versuchung"). Wenn Forschende (in welcher Weise auch immer) Einfluss auf die Zuteilung nehmen können, werden sie die Einteilung in ihrem Sinne beeinflussen, z.B., indem sie einen Probanden, der aus ihrer Sicht "besser" in die Kontrollgruppe passen würde, eben jener zuteilen.

Die Erfahrung lehrt uns, dass Studien mit fragwürdiger Randomisierung oder ungenügendem Allocation Concealment dazu neigen, den Behandlungseffekt um bis zu 40% zu überschätzen

[2] zur Fallzahlkalkulation in klinischen Studien siehe Röhrig et al. (2010)

[3] https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

[4] https://r-project.org, siehe hierzu auch große Schlarmann (2024)

(Schulz & Grimes (2002a)), und dass deren Daten breiter um den "wahren Wert" streuen (also ungenauer sind).

2.8 Waren die gebildeten Gruppen ähnlich/vergleichbar?

Beim RCT werden die Probanden per Zufall in Kontroll- und Interventionsgruppe(n) eingeteilt. Durch die zufällige Einteilung entstehen *vergleichbare* Gruppen. Das bedeutet, dass die jeweiligen Merkmale der Probanden (Alter, Geschlecht, Bildungsstand, usw.) in allen Gruppen vergleichbar ist. Wenn die Gruppen vergleichbar *sind*, dann unterscheiden sie sich theoretisch nur in der Art der Interventionen, die sie erhalten werden. Wenn das aber so ist, dann lässt sich das gemessene Ergebnis alleine auf die Interventionen zurückführen. Dies ist der größte Vorteil des RCTs gegenüber anderen Studiendesigns.

Wenn die Gruppen von Anfang an ("zur Baseline") schon unterschiedlich sind, dann kann das gemessene Ergebnis eben nicht mehr allein der Intervention zugeschrieben werden. Darum ist es so wichtig zu prüfen, ob vergleichbare Gruppen zur Baseline vorliegen.

Um zu überprüfen, ob die Randomisierung erfolgreich war, werden in Studien so genannte Baseline-Tabellen präsentiert. In ihnen sind die Charakteristika der Probanden aller Gruppen gegenübergestellt.

Table 1 | Baseline characteristics of patients with acute chest pain randomised to receive verbal advice followed by an information sheet (intervention) or verbal advice alone. Values are numbers (percentages) unless stated otherwise

Variables	Intervention group (n=349)	Control group (n=351)	Total (n=700)
Mean (SD) age (years)	48.3 (11.8)	48.9 (11.2)	48.6 (11.5)
Men	214 (61)	217 (62)	431 (61.6)
Diagnostic group receiving information sheet:			
Benign non-cardiac chest pain	81 (23)	81 (23)	162 (23)
Chest pain uncertain, no follow-up	228 (65)	230 (66)	458 (65)
Chest pain uncertain, referred to cardiology	30 (9)	31 (9)	61 (9)
Angina	10 (3)	9 (3)	19 (3)

Abbildung 1: Baseline-Tabelle ohne p-Werte (aus Arnold et al. (2009))

In Abbildung 1 ist ein Ausschnitt der Baseline-Tabelle aus Arnold et al. (2009) zu sehen. Lesende werden es schwer haben zu entscheiden, ob die Werte aus Kontroll- und Interventionsgruppe vergleichbar sind. Müssen die Werte exakt gleich sein, oder ist eine gewisse Abweichung tolerierbar? Aus dem Bauch lässt sich dies schwer entscheiden, aber es gibt einen Weg, statistisch zu klären, ob es sich um vergleichbare Gruppen handelt. Hierfür hätten die Autoren Signifikanztests[5] durchführen, und deren Ergebnisse über die Modell-p-Werte als eigene Spalte der Tabelle hinzufügen müssen. Anhand der p-Werte lässt sich leicht ablesen, ob Unterschiede vorliegen.

TABLE 1 Participants' baseline sociodemographic and clinical characteristics

Variables	Total n = 94	Usual care n = 47	Intervention n = 47	p
Sex, n (%)				0.322[a]
Female	73 (77.7)	34 (72.3)	39 (83.0)	
Male	21 (22.3)	13 (27.7)	8 (17.0)	
Age (years), Mean (SD)	49.3 (7.4)	49.2 (8.4)	49.4 (6.4)	0.776[b]
Years of education, n (%)				
<9	59 (62.8)	29 (61.7)	30 (63.8)	0.797[a]
9–12	32 (34.0)	16 (34.0)	16 (34.0)	
13–16	2 (2.1)	1 (2.1)	1 (2.1)	
>16	1 (1.1)	1 (2.1)	–	

Abbildung 2: Baseline-Tabelle mit p-Werten (aus Mattei Da Silva et al. (2020))

In Abbildung 2 ist ein Ausschnitt der Baseline-Tabelle von Mattei Da Silva et al. (2020) zu sehen. In der rechten Spalte sind die geforderten p-Werte angegeben. Alle p-Werte sind *größer* als $0,05$, somit sind die Unterschiede *nicht signifikant*. Das bedeutet, dass es keinen statistischen Unterschied zwischen den Gruppen gibt, und diese somit *vergleichbar* sind. Die Randomisierung scheint also funktioniert zu haben.

Die Baseline-Tabelle ist so gesehen die einzige Tabelle, in welcher Forschende gerne *keine* signifikanten p-Werte sehen möchten.

2.9 Wurden alle Möglichkeiten der Verblindung ausgeschöpft?

Verblindung bedeutet, dass nicht bekannt ist, in welcher Gruppe (Intervention oder Kontrolle) sich ein Proband befindet. Verblindet werden können

- die Probanden (wissen nicht, ob sie zur Kontroll- oder Interventionsgruppe gehören),
- die Applikanten (Durchführende wissen nicht, *wen* sie gerade behandeln),
- und die Forschenden (Datenauswerter wissen nicht, welche Daten zu welcher Gruppe gehören).

Menschen verhalten sich anders, wenn sie beobachtet werden, und sie verhalten sich anders, wenn etwas *Neues* an ihnen ausprobiert wird. In Abhängigkeit ihrer Präferenzen fühlen sie mehr oder weniger Schmerzen, oder sie fühlen sich generell besser oder schlechter. Wenn ein Proband nichts von Akkupunktur hält, wird er in einer Studie über die Wirksamkeit von Akkupunktur zur Rauchentwöhnung sich wahrscheinlich so *fühlen* und *verhalten*, wie es seiner Grundeinstellung

[5] zu Signifikanztests siehe Lange & Bender (2007c) und du Prel et al. (2010)

entspricht. Pflegende, die vom Bobath-Konzept überzeugt sind, werden in einer Studie zur Wirkung von Bobathpflege sich gegenüber den Patienten in der Bobath-Gruppe (Intervention) wahrscheinlich (unbewusst) anders verhalten als gegenüber solchen in der Kontrollgruppe (nach dem Motto "ich zeig jetzt mal wie viel besser das ist"). Und Forschende werden (unbewusst) alles tun, um möglichst große Effekte nachweisen zu können. Wenn bekannt ist, welche Daten zu welcher Gruppe gehören, können im Vorfeld geschickt "Ausreißer" ausgeschlossen werden, so dass sich die eigene Vorannahme bestätigt.

Um diese Effekte zu schmälern, ist es ratsam, möglichst alle an der Forschung beteiligten Personen hinsichtlich der Zugehörigkeit zur Interventions- oder Kontrollgruppe zu verblinden.

Je nachdem, wie viele Parteien verblindet wurden (Probanden, Applikanten, Auswerter), spricht man von einfach-, zweifach- und dreifach-verblindeten Studien.

Während die Auswerter *immer* verblindet werden können, ist dies bei Probanden oder Applikanten nicht immer möglich. Applikanten werden wissen, ob sie einen Probanden nach dem Bobath-Konzept pflegen oder nicht, und auch Probanden werden mitbekommen, ob sie ein Nikotinpflaster erhalten haben, oder eine Akkupunktur.

Dennoch zeigt sich auch hier, dass Studien mit weniger Verblindung größere Streuungen in ihren Daten haben (Schulz & Grimes (2002b)). Die Ergebnisse werden also unpräziser und neigen zur Überschätzung des Effekts.

2.10 Ist ausreichend beschrieben, wie viele Probanden rekrutiert wurden und wie viele an welcher Stelle warum ausgeschieden sind?

Diese Informationen werden in Form eines Flow-Charts (vgl. Abbildung 3) dargestellt, in welchem der vollständige Probandenfluss - von der potentiellen Auswahl bis zum Studienende - enthalten ist.

Ein Flowchart ermöglicht es, den Verlauf der Studie zu verstehen, einschließlich der Anzahl der Probanden, die für die Studie in Frage kämen, wie viele ausgeschlossen wurden und warum, rekrutiert wurden, und wie viele an welcher Stelle ausgeschieden sind und warum.

Durch die Visualisierung des Teilnehmerflusses können potenzielle Probleme oder Bias in der

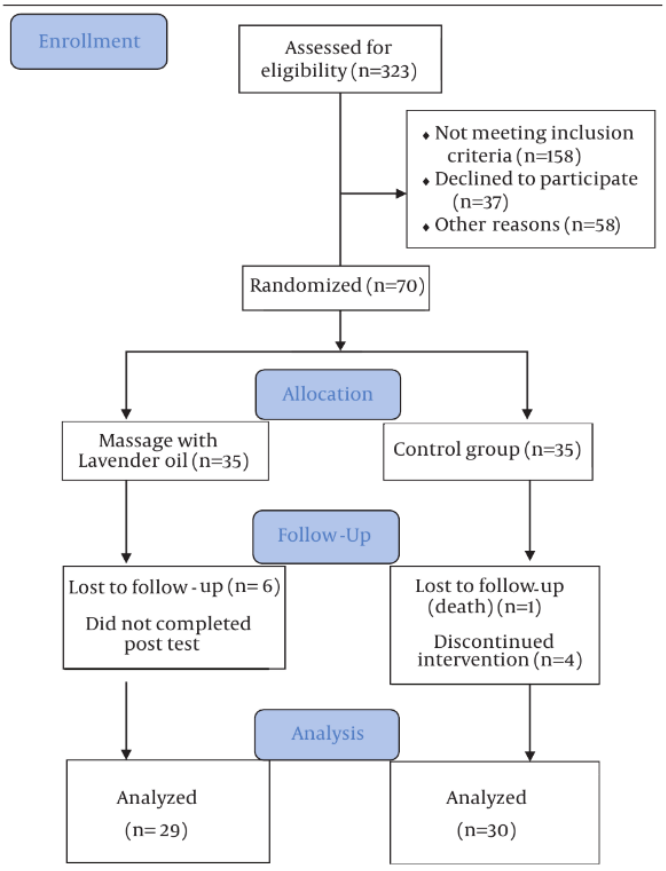

Figure 1. The Sampling Framework of the Study

Abbildung 3: FlowChart (Hashemi et al. (2015))

Studiendurchführung identifiziert werden. Zum Beispiel könnten ungleiche Verteilungen zwischen den Gruppen aufgrund von Drop-Outs oder Ausschlüssen eine Verzerrung der Ergebnisse verursachen.

2.11 Könnte der Drop-Out die Ergebnisse verzerren?

Gerade bei Studien, die über einen längeren Zeitraum laufen, wird es immer Drop-Outs geben, also Probanden, die ihre freiwillige Teilnahme an der Studie widerrufen oder aus anderen Gründen ausscheiden. Im schlimmsten Fall ist der Drop-Out so hoch, dass die errechnete Fallzahl nicht erreicht wird. Auch wenn der Drop-Out moderat sein sollte, stellt sich die Frage, wie mit den unvollständigen Daten umgegangen werden soll. Fließen sie in die Analysen mit ein, oder werden sie vollständig entfernt?

Die Intention-to-Treat-Analyse (ITT) bezieht die Daten aller Probanden ein, die randomisiert wurden, unabhängig davon, ob sie die Behandlung (vollständig) erhalten haben oder nicht, oder ob sie die Studie abgebrochen haben. Die Drop-Out-Daten bleiben also bei ITT enthalten. Diese Methode

berücksichtigt die Verteilung von Verzerrungen, die durch Patientenabbrüche und Nichtcompliance entstehen können, und gibt ein realistischeres Bild darüber, wie die Behandlung in der Praxis wirken würde. Der Nachteil der ITT besteht darin, dass die Wirksamkeit der Intervention abgeschwächt wird, weil Probanden in die Analyse eingeschlossen werden, welche die Behandlung abgebrochen oder nicht eingehalten haben und so mit einem negativen (oder fehlendem) Outcome zum Ergebnis beitragen. Dies kann die statistische Leistungsfähigkeit (Signifikanz) beeinträchtigen.

Die Per-Protocol-Analyse (PP) hingegen bezieht nur die Daten solcher Probanden ein, die das gesamte Studienprotokoll vollständig durchlaufen haben. Daten von Drop-Outs werden für die Analysen ausgeschlossen. Diese Methode kann zu höheren Wirksamkeitsraten führen, da sie die Daten solcher Probanden ausschließt, die während der Studiendurchführung (warum auch immer) ausgeschieden sind. Die Generalisierbarkeit wird schlechter, weil sich die Studienergebnisse nur auf Probanden beziehen, die das gesamte Studienprotokoll durchlaufen haben (oder durchlaufen konnten). Dies kann dazu führen, dass die Ergebnisse optimistischer sind als in der Praxis zu erwarten. Eine Diät, die sehr harte Vorschriften macht, ist zur Gewichtsreduktion sehr wirksam. Da das Einhalten dieser Vorgaben jedoch so schwer ist, brechen viele Probanden die Diät ab. Eine solche Diät wäre zwar sehr wirksam, jedoch *für die Massen* ungeeignet.

Das Ausschließen der Drop-Out-Daten kann zudem die Effekte der Randomisierung zerstören, so dass die Gruppen zur Baseline schon unterschiedlich waren.

2.12 Ist das zentrale Ergebnis des Forschungsprojekts beschrieben?

Das Wichtigste an einer Studie sind ihre Ergebnisse. Diese müssen exakt beschrieben und quantifiziert sein, z.B. - je nach Fragestellung - in Form von statistischen Kenngrößen wie Mittelwert, Median[6] sowie deren Streuung[7], oder in Form von Odds Ratios[8].

Eine Studie über die Wirksamkeit verschiedener Diätformen sollte beispielsweise den mittleren Gewichtsverlust (Mittelwert, Median) pro Diätform inklusive der Streuung (Standardabweichung) angeben. Eine Studie über die Wirksamkeit einer Impfung sollte die Risikoveränderung in Form von Relativen Risiken oder Odds Ratios darstellen.

In vielen Studien wird zudem die *Number needed to treat* angegeben, also die Anzahl an Patienten, die mit der neuen Intervention behandelt werden müssen, um bei einem dieser Patienten einen bestimmten *Erfolg* (z.B. “1kg Gewichtsabnahme”) zu erzielen. Sie ist eine aussagekräftige und gut interpretierbare Zahl zur Beurteilung der Wirksamkeit der Intervention.

2.13 Ist die Größe des Behandlungseffekts klinisch relevant?

Es existiert ein Unterschied zwischen statistischer Signifikanz und klinischer Relevanz. Eine neue Diätform kann statistisch signifikant zu einem Gewichtsverlust beitragen. Wenn dieser Gewichtsverlust jedoch nur 1 Gramm pro Jahr beträgt, dann sind die Ergebnisse zwar statistisch signifikant (diesen Gewichtsverlut gäbe es also wirklich), aber nicht klinisch *relevant*. Forschende müssen vor Beginn der Studie angeben, welche Effektgröße sie für relevant halten. Mit diesen Informationen wird dann die oben beschriebene Fallzahlkalkulation durchgeführt.

Je nach Fragestellung ist es für Außenstehende schwer nachzuvollziehen, ab wann ein Unterschied klinisch bedeutsam ist. Daher bedienen sich viele Studien einer statistischen Hilfszahl, die *Cohen's d* genannt wird (Cohen (1992)). Die Zahl kann Werte zwischen 0 und 1 annehmen[9], wobei ein kleiner Wert für einen kleinen Effekt steht.

Abbildung 4 zeigt die Effektstärken für die jeweiligen Werte von d. Für Unterschiede von Mittelwerten gilt, dass d größer 0,8 einen großen, d größer 0,5 einen mittleren und d kleiner 0,2 einen kleinen Effekt beschreibt. Wenn für den in einer Studie gemessenen Effekt (oder Unterschied) Cohen's d angegeben ist, können fachfremde Personen auch bei ungewohnten Maßeinheiten (zB. μg pro dl) die Größe des Effekts (und somit die Wirkungsstärke) einschätzen.

[6] zu Mittelwert und Median siehe Lange & Bender (2007a)
[7] zur Variabilität siehe Lange & Bender (2007b)
[8] zur Odds Ratio siehe Shorten & Shorten (2015) und Szumilas (2010)
[9] strenggenommen auch größer als 1, aber das ist nicht mehr relevant, da der Effekt dann eh schon riesengroß ist

Table 1
ES Indexes and Their Values for Small, Medium, and Large Effects

Test	ES index	Effect size		
		Small	Medium	Large
1. m_A vs. m_B for independent means	$d = \frac{m_A - m_B}{\sigma}$	.20	.50	.80
2. Significance of product–moment r	r	.10	.30	.50
3. r_A vs. r_B for independent rs	$q = z_A - z_B$ where z = Fisher's z	.10	.30	.50
4. P = .5 and the sign test	$g = P - .50$	.05	.15	.25
5. P_A vs. P_B for independent proportions	$h = \phi_A - \phi_B$ where ϕ = arcsine transformation	.20	.50	.80
6. Chi-square for goodness of fit and contingency	$w = \sqrt{\sum_{i=1}^{k} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	.10	.30	.50
7. One-way analysis of variance	$f = \frac{\sigma_m}{\sigma}$	.10	.25	.40
8. Multiple and multiple partial correlation	$f^2 = \frac{R^2}{1 - R^2}$	.02	.15	.35

Note. ES = population effect size.

Abbildung 4: Cohen's d (aus Cohen (1992))

2.14 Ist erkennbar, ob die Ergebnisse durch Zufall entstanden sein könnten?

Nach Durchführung der Intervention(en) erfolgt der Vergleich der Daten zwischen den gebildeten Gruppen. Hierbei stellt sich die Frage, ob der gefundene Unterschied zwischen den Gruppen (also das Ergebnis der Studie) auch zufällig hätte entstehen können. Hierfür werden Signifikanztests durchgeführt, deren Ergebnisse mindestens in Form der p-Werte angegeben werden müssen. Sind die p-Werte kleiner als die festgelegte Irrtumswahrscheinlichkeit α (meistens ist $\alpha = 0,05$), so ist der gefundene Unterschied nicht zufällig entstanden.

Liegt das Ergebnis in Form von Odds Ratios vor, wird das Konfidenzintervall[10] der Odds Ratio benötigt. Ist die Odds Ration größer als 1, so besteht ein positiver Effekt. Sind sie kleiner als 1, besteht ein negativer Effekt. Ist die Odds Ration exakt 1, gibt es keinen Unterschied.

Schließt dieses Konfidenzintervall der Odds Ratio die Zahl 1 ein, ist das Ergebnis *nicht* signifikant, da der wahre Wert der Odds Ratio bei exakt 1 (kein Unterschied) liegen könnte.

2.15 Sind die Ergebnisse hinreichend präzise?

Neben den zentralen Ergebniswerten sollten deren Konfidenzintervalle publiziert werden, damit die Leser sich ein besseres Bild über die Streuung der Daten (und somit über die Präzision der Ergebnisse) machen können. Das Konfidenzintervall ist ein Sicherheitsbereich, der den "wahren Wert" der Ergebnisse mit einer guten Wahrscheinlichkeit enthält. Zwar sind die angegebenen Werte (z.B. Mittelwert und Median) die besten Schätzwerte innerhalb der erhobenen Daten. Jedoch kann der wahre Wert theoretisch *überall* innerhalb des Konfidenzintervalls liegen.

In der Studie von Arnold et al. (2009) wurde beispielsweise eine Number-needed-to-treat von 9 ermittelt. Das bedeutet, dass 9 Patienten mit der Intervention behandelt werden müssen, um bei 1 Patienten eine Verbesserung zu erzielen. Das dazugehörige Konfidenzintervall für die "wahre" Number-needed-to-treat erstreckt sich von $5 - 46,1$. Das bedeutet, dass der wahre Wert theoretisch auch bei 46 Patienten liegen könnte (oder - im positiven Fall - bei 5). Beim zweiten Outcome erstreckt sich das Intervall sogar von $6,6 - \infty$. Es sollte selbsterklärend sein, dass eine Konfidenzgrenze von "unendlich" keinen großen Mehrwert darstellt.

Je enger die Intervallgrenzen, desto präziser sind die Ergebnisse.

2.16 Stehen die Ergebnisse im Einklang mit Ergebnissen anderer Forschungsarbeiten?

Die Konsistenz der Ergebnisse über verschiedene Studien hinweg stärkt die Evidenz für eine bestimmte Intervention oder ein bestimmtes Phänomen. Wenn die Ergebnisse einer RCT mit denen anderer Studien übereinstimmen, erhöht dies das Vertrauen in die Validität und Zuverlässigkeit der Ergebnisse. Es kann auch bedeuten, dass die Ergebnisse möglicherweise auf verschiedene Populationen und Kontexte übertragbar sind. Dies stärkt die Generalisierbarkeit der Ergebnisse.

Viel interessanter wird es allerdings, wenn die Ergebnisse im Widerspruch zu vorherigen Forschungsarbeiten stehen. In diesem Falle muss besonders geprüft werden, in wie weit Setting, Ort, Zeitpunkt, Maßnahmen, Datenerhebung, Studienpopulation und Stichprobe die Unterschiede erklären können. Ist die Erklärung, welche die Autoren für den Unterschied diskutieren, plausibel?

[10] zum Konfidenzintervall siehe Bender & Lange (2007) und du Prel et al. (2009)

2.17 Wurden wichtige Informationen nicht publiziert?

Mit Hilfe des CONSORT-Statments (Schulz et al. (2010)) kann überprüft werden, ob alle relevanten Informationen in der Publikation enthalten sind. Für Publikationen ab 2010 ist es fragwürdig, wenn Autoren die Statements nicht vollständig einhalten. Entweder kennen die Autoren die Statements nicht (spricht nicht für die Autoren), oder sie ignorieren diese (spricht ebenfalls nicht für die Autoren).

2.18 Sind die Ergebnisse auf Ihre Patienten / Ihr Haus übertragbar?

Diese Frage sollten Sie sich die ganze Zeit über stellen: wie relevant ist die Studie für mich, meine Patienten, mein Haus und mein Fachgebiet? Kann und sollte die vorgestellte Intervention in meinem Arbeitssetting eingeführt werden?

Die Antwort hängt u.a. stark mit dem Ort und Setting der Studie zusammen (siehe Abschnitt 2.4). Eine Studie über Ernährungsgewohnheiten von amerikanischen Großstadtmenschen wird andere Ergebnisse liefern als eine solche von bayrischen Dorfbewohnern. Ebenso werden Kinder in Düsseldorf-Oberkassel andere Angaben über ihr durchschnittliches Taschengeld geben als Kinder in Berlin-Neukölln. Die Orts- und Settinginformationen sind wichtig bei der Entscheidung, wie relevant oder übertragbar die Studienergebnisse für das eigene Fachgebiet / Haus / Studium sind, und aus welcher (oder wessen) Perspektive die Ergebnisse eingeordnet werden sollten. Ergebnisse über die Einführung ambulanter Pflegedienste im Iran sind wahrscheinlich nur schwer auf Deutschland übertragbar. Dennoch sollte man sich immer die Frage stellen: "warum sollte (oder könnte) es bei uns anders sein?"

Des Weiteren muss eingeschätzt werden, ob der Nutzen der Intervention die möglichen Risiken und Kosten wert ist, oder ob Alternativen gleich-gut oder sogar geeigneter sein könnten. Wenn eine Studie zeigt, dass Patienten weniger Angst vor der bevorstehenden Operation haben, wenn sie zuvor über eine interaktive Informations-App mittels Virtual-Reality-Brille aufgeklärt wurden, dann muss geprüft werden, ob die Anschaffung solcher (teurer) Brillen in Ihrem Haus gerechtfertigt ist, oder ob der selbe Effekt auch mit anderen Mitteln (z.B. Pflegegespräche) erzielt werden kann.

3 Anwendung der Ampel

Beim Lesen einer RCT-Studie gehen Sie die Fragen der Ampel Schritt für Schritt durch. Machen Sie sich zu jeder Frage Notizen[11], die Ihre Antworten stützen. Fällt die Antwort zu Ihrer Zufriedenheit aus, kann die Frage mit "grün" markiert werden. Ergeben sich bei einer Frage kritische Probleme, wird die Frage mit "rot" markiert. Sollten Sie "schwanken", oder sollten Ihnen Teil-Informationen oder eine gewisse Detailtiefe bei einer Frage fehlen, markieren Sie diese mit "gelb".

Schreiben Sie aussagekräftige Notizen zu jeder Frage. Dies erleichtert Ihnen den Wiedereinstieg, falls Sie die Studie (nach langer Zeit) erneut durcharbeiten müssen, oder wenn Sie die Güte der Studie jemandem vorstellen müssen.

Übung macht übrigens den Meister: je mehr Studien Sie lesen, desto tiefer wird das Verständnis des Forschungsprozesses, desto sicherer werden Sie beim Nachvollziehen der Studienverläufe, und desto leichter wird Ihnen die Anwendung der Krefelder Ampel fallen.

4 Fazit

Die Krefelder Ampel ist eine weitere Beurteilungshilfe zur Bewertung von RCT-Studien. Durch die Beantwortung der 18 Fragen erhalten Sie einen umfassenden und aussagekräftigen Überblick über die Güte der Studie.

Literatur

Arnold, J., Goodacre, S., Bath, P., & Price, J. (2009). Information sheets for patients with acute chest pain: randomised controlled trial. *BMJ*, *338*(feb26 2), b541–b541. https://doi.org/10.1136/bmj.b541

Behrens, J., & Langer, G. (2022). *Evidence based Nursing and Caring: Methoden und Ethik der Pflegepraxis und Versorgungsforschung – Vertrauensbildende Entzauberung der "Wissenschaft"* (5. vollst. überarb. u. erw. Aufl. 2022 Edition). Hogrefe AG.

[11] ein Arbeitsblatt der Ampelfragen steht zum Download bereit, siehe https://www.schlarmann.net/Krefelder-Ampel-RCT.html

Bender, R., & Lange, S. (2007). Was ist ein Konfidenzintervall? *Deutsche Medizinische Wochenschrift (1946)*, *132 Suppl 1*, e17–18. https://doi.org/10.1055/s-2007-959031

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, *112*(1), 155–159.

du Prel, J.-B., Hommel, G., Röhrig, B., & Blettner, M. (2009). Konfidenzintervall oder p-Wert? *Deutsches Ärzteblatt*, *106*(19), 335–339. https://doi.org/10.3238/arztebl.2009.0335

du Prel, J.-B., Hommel, G., Röhrig, B., & Blettner, M. (2010). Auswahl statistischer Testverfahren. *Deutsches Ärzteblatt*, *107*(19), 343–349. https://doi.org/10.3238/arztebl.2010.0343

große Schlarmann, J. (2024). *Statistik mit R und RStudio - Ein Nachschlagewerk für Gesundheitsberufe*. Hochschule Niederrhein. https://www.produnis.de/R

Hashemi, S. H., Hajbagheri, A., & Aghajani, M. (2015). The Effect of Massage With Lavender Oil on Restless Leg Syndrome in Hemodialysis Patients: A Randomized Controlled Trial. *Nursing and Midwifery Studies*, *4*(4). https://doi.org/10.17795/nmsjournal29617

Kabisch, M., Ruckes, C., Seibert-Grafe, M., & Blettner, M. (2011). Randomisierte kontrollierte Studien. *Deutsches Ärzteblatt international*, *108*(39), 663–668. https://doi.org/10.3238/arztebl.2011.0663

Lange, S., & Bender, R. (2007a). Median oder Mittelwert? *Deutsche Medizinische Wochenschrift*, *132 Suppl 1*, e1–2. https://doi.org/10.1055/s-2007-959024

Lange, S., & Bender, R. (2007b). Variabilitätsmaße. *Deutsche Medizinische Wochenschrift (1946)*, *132 Suppl 1*, e5–6. https://doi.org/10.1055/s-2007-959026

Lange, S., & Bender, R. (2007c). Was ist ein Signifikanztest? *Deutsche Medizinische Wochenschrift*, *132*(S 01), e19–e21. https://doi.org/10.1055/s-2007-959032

Mad, P., Felder-Puig, R., & Gartlehner, G. (2008). Randomisiert kontrollierte Studien. *Wiener Medizinische Wochenschrift*, *158*(7-8), 234–239. https://doi.org/10.1007/s10354-008-0526-y

Mattei Da Silva, Â. T., De Fátima Mantovani, M., Castanho Moreira, R., Perez Arthur, J., & Molina De Souza, R. (2020). Nursing case management for people with hypertension in primary health care: A randomized controlled trial. *Research in Nursing & Health*, *43*(1), 68–78. https://doi.org/10.1002/nur.21994

Röhrig, B., Prel, J.-B. du, Wachtlin, D., Kwiecien, R., & Blettner, M. (2010). Fallzahlplanung in klinischen Studien. *Deutsches Ärzteblatt*, *107*(31-32), 552–557. https://doi.org/10.3238/arztebl.2010.0552

Schardt, C., Adams, M. B., Owens, T., Keitz, S., & Fontelo, P. (2007). Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC Medical Informatics and Decision Making*, *7*(1), 16. https://doi.org/10.1186/1472-6947-7-16

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *Lancet*, *366*. https://doi.org/10.1016/S0140-6736(10)60456-4

Schulz, K. F., & Grimes, D. A. (2002a). Allocation concealment in randomised trials: defending against deciphering. *The Lancet*, *359*(9306), 614–618. https://doi.org/10.1016/S0140-6736(02)07750-4

Schulz, K. F., & Grimes, D. A. (2002b). Blinding in randomised trials: hiding who got what. *The Lancet*, *359*(9307), 696–700. https://doi.org/10.1016/S0140-6736(02)07816-9

Shorten, A., & Shorten, B. (2015). What is an Odds Ratio? What does it mean? *Evidence Based Nursing*, *18*(4), 98–99. https://doi.org/10.1136/eb-2015-102206

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *19*(3), 227–229.