

BACHELORARBEIT

von Roman Hornung
Sommersemester 2009

Metaanalysen

Betreuung:
PD Dr. Christian Heumann

Institut für Statistik
Ludwig-Maximilians-Universität München

In Zusammenarbeit mit

Dr. Susanne Rösner

Inhaltsverzeichnis

1	Einführung	1
1.1	Einleitung	1
1.2	Begriff	2
1.3	Geschichte	2
1.4	Gründe für die Durchführung	2
1.5	Anforderungen an die einzuschließenden Studien	3
1.6	Typen	3
2	Ablauf und Methodiken	5
2.1	Formulierung des Themas	5
2.2	Ein- und Ausschlusskriterien	5
2.3	Suche nach relevanten Studien	6
2.4	Datenaufbereitung	6
2.5	Publikationsbias	8
2.5.1	Diagnose	8
2.5.2	Ansätze zur Korrektur	14
2.6	Homogenität / Heterogenität	17
2.6.1	Formaler Test	17
2.6.2	Informelle Analyse	18
2.6.3	Einschätzung des Ausmaßes	21
2.7	Modelle	22
2.7.1	Modell mit festen Effekten:	22
2.7.2	Modell mit zufälligen Effekten:	23
2.7.3	Modellwahl:	24
2.7.4	Schätzung der Zwischenstudienvarianz	25
2.8	Berechnung der Gesamteffektgröße	28
2.9	Sensitivitätsanalyse	30
2.10	Metaregression	31
3	Vor- und Nachteile	35

4	Neuere Entwicklungen und Tendenzen	38
4.1	Mixed Treatment Comparison (MTC) Metaanalysen	38
4.2	Individual Patient Data (IPD) Metaanalysen:	41
5	Anwendungsbeispiel	43
5.1	Einleitung	43
5.2	Effektgrößen	43
5.2.1	Gewinnung der Effektgrößen und der Standardfehler	44
5.2.2	Untersuchung der Heterogenität	45
5.3	Struktur der Kovariablen	46
5.4	Methodik zur Schätzung der Modelle	48
5.5	Bemerkungen zur Interpretation	49
5.6	Einfachregressionen getrennt nach Substanz	50
5.7	Regressionen über beide Substanzen	53
5.8	Backward- und Forward-Selection	54
5.9	Placebo-Response als Kovariable	57
6	Zusammenfassung	59
A	Tabellen	62
A.1	Einfachregressionen getrennt nach Substanz	62
A.2	Regressionen über beide Substanzen	67
A.3	Backward- und Forward-Selection	69
A.4	Placebo-Response als Kovariable	75
B	Grafiken	76
B.1	Zwischenstudienvarianzschätzer und Forest Plots	76
B.2	Verteilungen der Werte der Kovariablen	78
B.3	Meta-Regression	81
B.3.1	Streudiagramme	81
B.3.2	Normal-Quantil-Plots	90
C	Konfidenzintervall der I^2-Statistik	99
D	R-Code	101

Kapitel 1

Einführung

1.1 Einleitung

Mit der wachsenden Menge an Information und deren immer besseren Verfügbarkeit werden Metaanalysen in ihrer Eigenschaft, beliebig viele Einzelergebnisse gemeinsam zu analysieren, immer wichtiger. So werden sie in verschiedensten Wissenschaften, wie z.B. Medizin, Psychologie, Wirtschaft oder in gewissem Umfang in den Sozialwissenschaften durchgeführt. Im Zuge dessen wird auch die Methodik stetig immer weiter entwickelt.

Der erste Teil dieser Arbeit befasst sich allgemein mit Metaanalysen. Dabei wird besonderer Wert auf deren konkreten Ablauf gelegt. Da Metaanalysen am häufigsten auf den auf Makroebene zusammengefassten Ergebnissen der verwendeten Studien beruhen, konzentriert sich die Darstellung hierbei vor allem auf die diesbezügliche Form der Durchführung. Allerdings werden später auch die Besonderheiten von IPD Metaanalysen, bei denen die einzelnen Mikrodaten verfügbar sind, behandelt. In der Diskussion wird auf die Vor- und auch Nachteile eingegangen. So sind Metaanalysen trotz ihrer offenkundigen Vorteile auch oft berechtigter Kritik ausgesetzt. Es gibt neben einigen Fehlerquellen auch Probleme, die schwer bis gar nicht überwindbar sind. Zum Abschluss dieses Teils werden zwei neuere Entwicklungen im Bereich der Metaanalysen beschrieben: Mixed Treatment Comparison (MTC) Metaanalysen und Individual Patient Data (IPD) Metaanalysen.

Im zweiten Teil wird in explorativer Weise die Technik der Meta-Regression auf einen Meta-Datensatz angewendet, der Studien über die Wirksamkeit von Acamprosat und Naltrexon, zweier Medikamente zur Alkoholentwöhnung enthält. Dabei wird angestrebt, die beobachtete Heterogenität in der Studienpopulation durch ausgewählte Kovariablen weitestgehend aufzuklären.

1.2 Begriff

Die griechische Silbe "meta" bedeutet u.a. "über" oder "höhere Ebene" (Tintelnot & Bausch (2003); Bachim (2008)). Eine Metaanalyse ist demnach eine Analyse über Analysen. Der Metaanalytiker erreicht auch eine "höhere Ebene", in dem Sinne, dass er mehr Einblick in größere Zusammenhänge und den aktuellen Forschungsstand hat als die Betreiber der einzelnen Studien.

1.3 Geschichte

Die Grundlage der Methodik der Metaanalyse wurde 1904 von Karl Pearson gelegt. Pearson wurde die Aufgabe gegeben, die Wirksamkeit von Typhusimpfungen bei Soldaten zu untersuchen. Dazu analysierte er den Zusammenhang zwischen Impfung und Erkrankung bzw. Mortalität. Er betrachtete dabei post hoc die Ergebnisse aus einzelnen Studien zu diesem Thema, wobei auch die unterschiedlichen Messgenauigkeiten berücksichtigt wurden.

In den darauf folgenden Jahren forschten andere Statistiker gemäß Pearson an Methoden, um mehrere Einzelergebnisse entsprechend zusammenzufassen (Cebulla (2008)).

Der Begriff Metaanalyse wurde allerdings erst 1977 von Glass geprägt. Aufgrund einer heftigen Debatte über die Wirksamkeit von Psychotherapie waren in den vorigen Jahrzehnten hunderte von Studien im psychotherapeutischen Bereich durchgeführt worden. Trotzdem war die Frage aufgrund der Diversität der Ergebnisse nicht geklärt. Um dies zu erreichen, berechnete Glass aus den standardisierten Behandlungs-Kontrollgruppendifferenzen von 375 Studien den Mittelwert und erhielt damit ein Gesamtergebnis, aus dem er folgern konnte, dass Psychotherapie in der Tat wirksam ist.

Seit dieser Zeit wurden tausende Metaanalysen durchgeführt und die Methoden immer weiter verbessert (Lipsey & Wilson (2001)).

1.4 Gründe für die Durchführung

Die einzelnen Studien werden immer unter bestimmten Bedingungen durchgeführt, sodass sie häufig keine generalisierbaren oder untereinander eindeutigen Resultate liefern. Wenn die Schwankungen in den Ergebnissen nur so gering sind, dass sie durch Zufall erklärbar sind oder die Heterogenität zumindest sehr schwach ist, können die Ergebnisse gemittelt und so das Gesamtergebnis von der Abhängigkeit von den einzelnen Studienbedingungen

weitgehend bereinigt werden. Ist die Heterogenität dagegen stärker, fließen die Bedingungen als Kovariablen in die Analyse ein. Dies geschieht mittels Sensitivitätsanalyse (Abschnitt 2.9) bzw., wenn möglich, allgemeiner mit Hilfe von Metaregression (Abschnitt 2.10).

Dazu kommt, dass die Studien häufig zu klein sind, um einzeln verlässliche Ergebnisse zu liefern. Metaanalysen sind dagegen, dadurch dass sie mit (den Ergebnissen) einer viel größeren Datenmenge arbeiten, präziser (Blettner & Schlattmann (2005)).

Die mit Metaanalysen nachgewiesenen Effekte sind meist sehr gering, aber dennoch von großer Bedeutung. Auf der einen Seite könnten diese Effekte zwar für den Einzelnen weniger wichtig sein, aber sich dennoch stark auf die Gesamtheit auswirken, beispielsweise die Folgen von Passivrauchen. Umgekehrt könnte bereits eine geringe Chance auf ein positives Ergebnis für den Einzelnen wichtig sein, z.B. die Wirksamkeit eines Medikaments gegen eine schwere Krankheit (Sauerbrei & Blettner (2003)).

1.5 Anforderungen an die einzuschließenden Studien

Studien, die in Metaanalysen eingeschlossen werden, müssen allgemein bestimmte Bedingungen erfüllen. Es können nur Studien mit der gleichen Thematik zusammen analysiert werden. Dazu müssen sie natürlich empirisch sein, z.B. können also keine theoretischen Papers eingeschlossen werden, und die Ergebnisse müssen quantitativ und vergleichbar sein, d.h. sie müssen vom gleichen Typ sein oder zumindest alle in einen einheitlichen Typ umgewandelt werden können (Lipsey & Wilson (2001)).

1.6 Typen

Blettner et al. (2003) unterscheiden vier verschiedene Typen, Studien zusammen zu fassen.

Der erste ist der narrative Review, welcher nicht zu den Metaanalysen zählt, da die Ergebnisse der Studien rein qualitativ in einem Bericht zusammengefasst werden und keine quantitativen Ergebnisse geliefert werden.

Zweitens gibt es Metaanalysen, die nur auf veröffentlichten Daten beruhen, MAL. Die Informationen, die bei der Metaanalyse verwandt werden, ergeben sich also nur aus den Berichten der Studien. Wie bereits erwähnt, ist das der

in der Praxis am häufigsten verwendete Typ.

Als dritter Typ werden Metaanalysen aufgeführt, in denen alle Einzeldaten zur Verfügung stehen, MAP oder auch IPD Metaanalysen. Diese stellen den Goldstandard aller Metaanalysen dar, werden aber (noch) selten durchgeführt, da sie sehr aufwendig und kostenspielig sind.

Der letzte Typ ist die prospektive Metaanalyse. Bei dieser ist zur Zeit der Durchführung der einzelnen Studien schon geplant, dass sie später einmal gemeinsam analysiert werden. Dadurch können sie hinsichtlich ungewünschter Heterogenitätsfaktoren standardisiert werden. Was sie von Multicenter-Studien unterscheidet ist, dass die Studien auf der ganzen Welt durchgeführt werden. Dadurch ergibt sich zusätzliche wichtige Heterogenität. Z.B. gibt es in Deutschland keine gesamtstaatliche Erfassung von Krebsfällen, in anderen Ländern hingegen schon (Sauerbrei & Blettner (2003)).

Kapitel 2

Ablauf und Methodiken

2.1 Formulierung des Themas

Zu Beginn einer Metaanalyse muss, wie in jeder Wissenschaft, das Thema unmissverständlich und umfassend formuliert werden. Dieser Teil ist sehr wichtig, da das gesamte weitere Vorgehen darauf aufbaut. Je exakter die Formulierung des Themas ist, desto genauer sind später die Ergebnisse. Allerdings muss die Formulierung nicht sehr detailliert sein (Lipsey & Wilson (2001)).

2.2 Ein- und Ausschlusskriterien

Die genaue Festlegung der Ein- und Ausschlusskriterien ist von zentraler Bedeutung. Durch sie wird sichergestellt, dass die Selektion nicht durch den Selektierenden verfälscht wird, indem beispielsweise nur Studien mit positiven Ergebnissen ausgewählt werden, um subjektive Überzeugungen zu stützen. Beispiele sind bestimmte Schlüsselvariablen, die alle Studien beinhalten müssen, das Design, so könnten z.B. Beobachtungsstudien ausgeschlossen werden, das Veröffentlichungsjahr, z.B. Ausschließen von älteren Studien, oder der kultureller Bereich, man könnte z.B. nur Studien aus Europa oder den USA wählen.

Sehr wichtig ist, dass die Kriterien vor Beginn der Selektion festgelegt werden, da sonst die bereits gewonnen Studien Einfluss auf die Kriterien haben könnten (Lipsey & Wilson (2001)).

Die Ein- und Ausschlusskriterien machen den Selektionsprozess reproduzierbar, d.h., jemand anderer, der die Metaanalyse nochmal durchführt, sollte

anhand dieser Kriterien wieder auf die gleiche Studienpopulation kommen (Petitti (1994)).

2.3 Suche nach relevanten Studien

Dazu ist es allerdings auch nötig, dass die Suche nach potentiell wählbaren Studien so gründlich wie möglich durchgeführt wird. Noch wichtiger: der Publikationsbias wird dadurch so weit wie möglich eingeschränkt, der in dieser Arbeit später genauer behandelt wird. Das Ziel muss sein, alle durchgeführten Studien, die für die Metaanalyse wählbar sind, zu lokalisieren, obgleich dies unerreichbar sein könnte, weil die Daten mancher Studien noch nicht einmal analysiert worden sein könnten. Die Suche erfolgt mit Hilfe verschiedener Quellen. Computergestützte Datenbanken müssen durchsucht werden. In diesen sind sehr viele Studien mit Titel und Abstract zu finden. Für einige Beispiele im medizinischen Bereich, wurde gezeigt, dass circa 50 Prozent der relevanten Studien auf diese Art lokalisiert werden können. Weiter sollten Experten in dem entsprechenden Bereich kontaktiert werden. Diese haben häufig auch Informationen über nicht-veröffentlichte Studien. Relevante Zeitschriften, Doktorarbeiten und Bücher sollten durchgesehen werden. Schließlich finden sich in bereits gewonnen Studien häufig Verweise auf weitere interessierende Studien. Wenn noch weitere Quellen zur Verfügung stehen, sollten diese natürlich genutzt werden (Blettner & Schlattmann (2005)).

2.4 Datenaufbereitung

Nach der Studiensammlung müssen die interessierenden Informationen aus den Studienberichten abgeleitet werden. Das sind zum einen die Ergebnisse der Studien in Form von Effektgrößen mit den zugehörigen Standardfehlern und zum anderen mögliche Kovariablen für die Sensitivitätsanalyse bzw. Metaregression.

Die Gewinnung der Daten aus den Studienberichten stellt den zeitintensivsten Teil von Metaanalysen dar (Blettner & Schlattmann (2005)). Die abgeleiteten Informationen können aus mehreren Gründen nicht verlässlich sein. So könnten Zahlen aus der Publikation falsch entnommen werden. Außerdem werden Zahlen häufig in den Berichten mehrmals genannt und Fehler, die an einer Stelle korrigiert wurden verbleiben an einer anderen. Drittens könnten die Informationen nicht genau in der Form präsentiert werden, in der sie gewünscht sind und daher unterschiedlich interpretiert werden.

Aus diesen Gründen ist es wichtig, dass zwei verschiedene Personen die Datengenerierung unabhängig voneinander parallel durchführen und die Interrater-Verlässlichkeit überprüft wird. Die Personen erhalten genaue Anweisung, wie die Datengenerierung durchzuführen ist. Vor der Gewinnung aller notwendigen Informationen muss in einer Testphase überprüft werden, ob wirklich die interessierenden Größen gemessen werden.

Auch hier könnte es das Problem eines Bias geben, wenn bestimmte Informationen die Codierer in ihren Einschätzungen beeinflussen. Zum Beispiel könnten sie Studien aus etablierten Zeitschriften prinzipiell bessere Werte in den Qualitätsmerkmalen zuweisen. Um dies zu verhindern könnten die Codierer hinsichtlich bestimmter Merkmale verblindet werden. Am ehesten beeinflussbar könnten die Codierer von dem Wissen über die Autoren der Studien und ihren Instituten, den Zeitschriften, in denen sie veröffentlicht wurden und den Geldgebern sein. Problematisch ist auch, wenn die Codierer erfahrene Forscher in dem betreffenden Gebiet sind, da solche beteiligte Studien häufig schon kennen und die Verblindung nur teilweise durchgesetzt werden kann. Daher wäre es besser, unabhängige Personen mit der Datengenerierung zu beauftragen. Wissen über die Ergebnisse der Studien könnte außerdem auch Einfluss auf das Urteil über die verwendete Methodik haben. Deshalb sollte für alle Studien jeweils eine Person die Codierung der Ergebnisse und eine die der Methoden übernehmen (Petitti (1994)).

Im Folgenden soll die Gewinnung der Effektgrößen und ihrer Standardfehlern behandelt werden.

Die in den Studienberichten angegebenen Effektgrößen sind häufig unterschiedlichen Typs. Beispielsweise könnten sowohl Typen für stetige, als auch für dichotome Outcomes, die künstlich aus einer stetigen Variable dichotomisiert wurden, vorkommen oder wenn der Outcome binär ist, könnten Odds Ratios, relative Risiken und Risikodifferenzen angegeben sein. Man muss sich aber für einen Typ entscheiden, damit die Effektgrößen später gemeinsam analysiert werden können. Es existieren verschieden einfache Methoden, um die Effektgrößen in den einzelnen Fällen zu standardisieren, siehe Lipsey & Wilson (2001).

Als zweites müssen noch die Standardfehler bestimmt werden. Für den Fall, dass sie nicht gegeben sind und auch keine Information gegeben ist, aus der sich diese analytisch berechnen ließen, existieren Methoden zur Imputation.

2.5 Publikationsbias

Im Zusammenhang mit der Studiensuche wurde bereits auf den Publikationsbias eingegangen. Er entsteht, wenn negative und nicht signifikante Studienergebnisse seltener veröffentlicht werden als positive und somit die veröffentlichten Studien nicht repräsentativ für alle durchgeführten Studien sind. Natürlich sind Wissenschaftler gewillter, Studien zur Veröffentlichung zu geben, die ein signifikant positives Ergebnis liefern, da in ihren Augen nur diese einen Erkenntnisgewinn darstellen. Viele Fachzeitschriften unterstützen den Publikationsbias zusätzlich, da sie hauptsächlich signifikant positive Ergebnisse veröffentlichen (Petitti (1994)). Allerdings machen die führenden medizinischen Zeitschriften heute die Registrierung von Studien vor ihrer Durchführung zur Bedingung einer späteren Veröffentlichung (Stanley (2005)).

Durch die mangelnde Veröffentlichung von nicht-signifikanten und noch wichtiger negativen Studien wird der Gesamteffekt in nicht vernachlässigbarer Weise überschätzt (Blettner & Schlattmann (2005)).

Wenn zu einem Thema sehr viele Studien existieren, ist ein gutes Mittel gegen Publikationsbias, nur die größten für die Metaanalyse zu verwenden, da große Studien in der Regel veröffentlicht werden (Stanley (2005)). Außerdem könnten ältere Studien ausgeschlossen werden. In früheren Jahren war es noch üblicher, hauptsächlich positive Studien zu veröffentlichen und unveröffentlichte Studien älterer Jahrgänge sind schwieriger zu lokalisieren.

Publikationsbias könnte in Forschungsbereichen, in denen konkurrierende Ansichten bestehen oder zumindest keine vorherrschende Meinung besteht, weniger ausgeprägt sein (Stanley (2005)). Also könnte es nützlich sein, inhaltlich zu diskutieren, welche Anreize bzw. Hindernisse es geben könnte, negative und nicht-signifikante Ergebnisse zu dem betreffenden Thema zu veröffentlichen, um so das Risiko eines Publikationsbias abzuschätzen.

2.5.1 Diagnose

Grafische Mittel

Das gängigste grafische Mittel um Publikationsbias festzustellen, ist der Funnel Plot. In diesem werden die Effektgrößen der einzelnen Studien gegen ein Maß für deren Präzision abgetragen. Dies könnte z.B. die inverse Varianz (Blettner & Schlattmann (2005)), der inverse Standardfehler, die Fallzahl, die Wurzel der Fallzahl (Stanley (2005)) oder der Standardfehler (Schwarzer et al. (2003)) sein. Die inversen Varianzen nehmen darin eine besondere

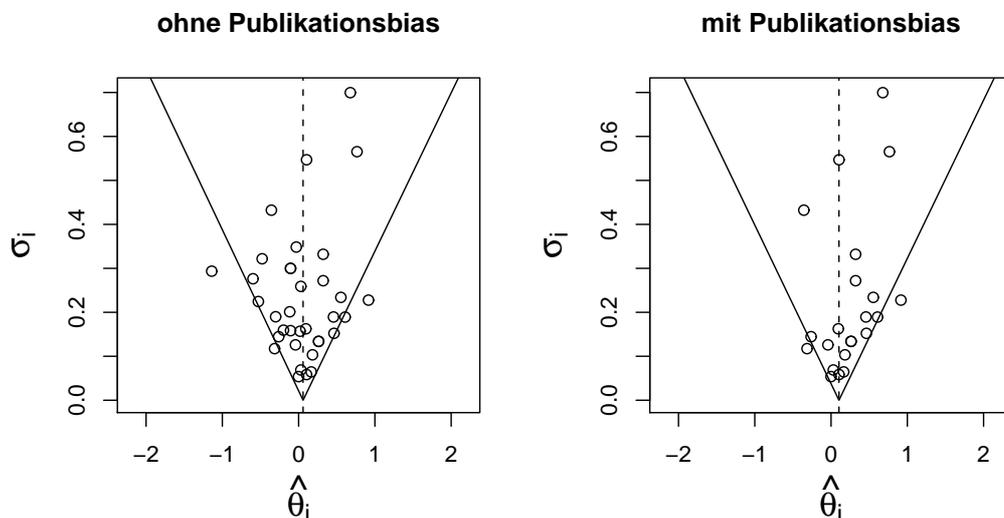


Abbildung 2.1: Funnel Plots

Stellung ein, da das die Gewichte sind, mit denen die einzelnen Effektgrößen später in die Berechnung der Gesamteffektgröße eingehen werden. Bei Verwendung dieses Maßes nimmt der Funnel Plot in Abwesenheit von Publikationsbias am klarsten die Form eines auf den Kopf gestellten Trichters an, der sein Zentrum an der Stelle der Gesamteffektgröße hat. Allerdings wurde der Standardfehler von der Arbeitsgruppe für statistische Methoden der Cochrane Collaboration als die bevorzugte Variante gewählt. Damit sollte der Plot die Form eines symmetrischen, auf den Kopf gestellten Dreiecks haben, an dessen Spitze die Gesamteffektgröße liegt (Schwarzer et al. (2003)).

Abbildung 2.1 zeigt zwei Funnel Plots. Im linken streuen die Effektgrößen zufällig und symmetrisch um die Gesamteffektgröße, es gibt also keine Anzeichen für einen Publikationsbias. Im Rechten sind links oben weniger Studien, der Funnel Plot stellt somit kein symmetrisches Dreieck dar. Negative und nicht-signifikante Studien scheinen also seltener veröffentlicht werden, was zu Publikationsbias führt.

Die Spitze kann bereits als grobe Einschätzung des Gesamteffekts dienen (Stanley (2005)).

Eine weitere nützliche Darstellungsart, die auch beim Verständnis des später behandelten Egger Test hilft, ist der Galbraith Plot. Hier werden die mit dem Standardfehler standardisierten Effektgrößen gegen die inversen Standardabweichungen geplottet, Abbildung 2.2.

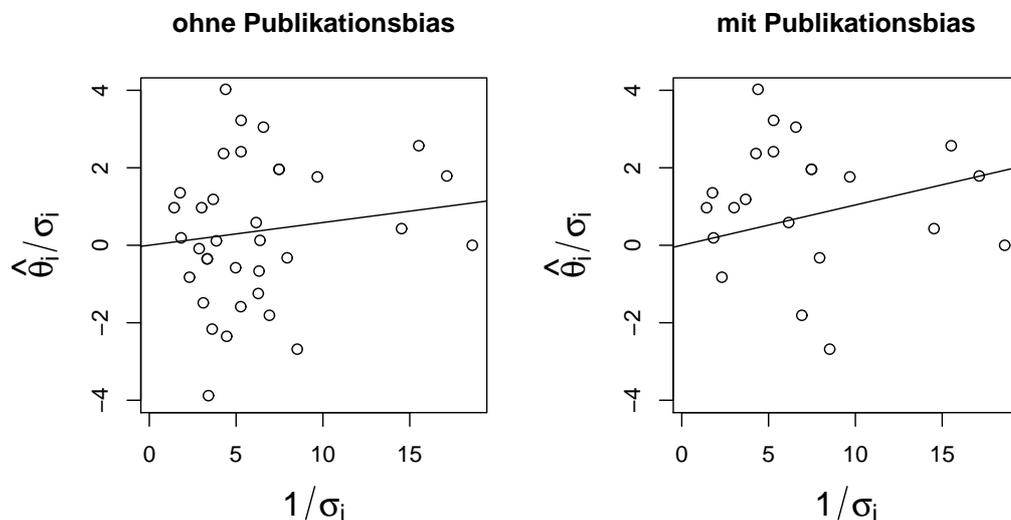


Abbildung 2.2: Galbraith Plot

Die Steigungen der Geraden, die durch den Ursprung und die Punkte gehen, entsprechen den einzelnen Effektgrößen. Der geschätzte Koeffizient $\hat{\beta}$ in einer linearen Regression

$$\frac{\hat{\theta}_i}{\sigma_i} = \beta \frac{1}{\sigma_i} + \epsilon_i, \quad i = 1, \dots, k, \quad \epsilon_i \sim N(0, 1) \quad (2.1)$$

entspricht dem geschätzten Gesamteffekt.

Ein Nachteil des Galbraith Plot gegenüber dem Funnel Plot ist, dass Publikationsbias weniger offensichtlich zu erkennen ist (Schwarzer et al. (2003)).

Bei beiden Plots kann dazu nicht sicher gesagt werden, ob eine abweichende Form wirklich von dem Fehlen unveröffentlichter Studien herrührt, oder ob eine andere, womöglich unbeobachtete und unberücksichtigte Variable, die auch mit der Varianz korreliert ist, diese Form verursacht oder verstärkt. So wurde z.B. beobachtet, dass bestimmte Qualitätsmerkmale die Schätzungen der Effektgrößen beeinflussen können. Verzerrungen dieser Art, die von abweichenden Ergebnissen kleiner Studien herrühren, werden als “small study effects” bezeichnet.

Sutton et al. (2008) schlagen eine Lösung dieses Problems für den Funnel Plot vor. In den Plot werden Konturlinien eingezeichnet, die den in der Forschung üblichen Signifikanzniveaus 1%, 5% und 10% entsprechen, wodurch den Studien in dem Plot sofort ihre Signifikanzen zugeordnet werden können.

Da davon ausgegangen wird, dass die Nicht-Veröffentlichung einer Studie auf ein unsignifikantes Ergebnis zurückzuführen ist, sollten bei Publikationsbias hauptsächlich in dem Bereich weniger Punkte sein, in dem Studien nicht signifikant sind. Ist der Grund für die Verzerrung dagegen ein anderer, werden (auch) in anderen Bereichen weniger Punkte sein. Die Linien werden dabei auf Basis des Standard-Wald-Test gewonnen (Moreno et al. (2009)).

Tests

Grafische Methoden sind generell subjektiv. Jeder Betrachter kann zu einer etwas anderen Einschätzung kommen (Sutton & Higgins (2008)). Daher können sie nur für eine grobe Abschätzung verwendet werden und es werden formale statistische Tests benötigt.

Es kann eine lineare Regression der Effektgrößen auf die inversen Varianzen gemacht werden (Blettner & Schlattmann (2005)). Herrscht kein Publikationsbias, so müssten die Effektgrößen alle zufällig um den Gesamteffekt streuen und die Regressionsgerade demnach (annähernd) parallel verlaufen, siehe Abbildung 2.3 links.

Verläuft sie hingegen nicht parallel, weist dies auf Assymetrie des Funnel Plots hin, was wahrscheinlich auf Publikationsbias zurückzuführen ist, Abbildung 2.3 rechts.

Ein weiterer Regressions-Test ist der Egger Test. Bei diesem wird durch den Galbraith Plot eine Regressionsgerade gelegt, wobei die standardisierten Effektgrößen auf die inversen Standardfehler regressiert werden (Schwarzer et al. (2003)).

$$\frac{\hat{\theta}_i}{\sigma_i} = \beta_0 + \beta_1 \frac{1}{\sigma_i} + \epsilon_i, \quad i = 1, \dots, k, \quad \epsilon_i \sim N(0, 1) \quad (2.2)$$

Dann wird ein zweiseitiger t-Test auf Verschiedenheit von Null des Intercepts durchgeführt. Es lässt sich leicht nachvollziehen, warum der Intercept in Abwesenheit von Publikationsbias Null sein müsste. Alle Effektgrößen haben, unabhängig von ihrer Standardabweichung, die Gesamteffektgröße als Erwartungswert, d.h. es ergibt sich für alle durchschnittlich der gleiche Wert. Damit und da bei den Effektgrößen weiter links durch eine größere Standardabweichung geteilt wird, ergeben sich dort durchschnittlich kleinere Werte als weiter rechts. Wenn $\frac{1}{\sigma_i}$ gegen Null geht, geht die Varianz gegen unendlich und es ergeben sich im Schnitt Nullwerte für $\frac{\hat{\theta}_i}{\sigma_i}$, Abb. 2.4 links.

Fehlen nicht-signifikante und stärker negative Studien vermehrt, so überwiegen weiter links die positiven Studien, der Intercept ist signifikant

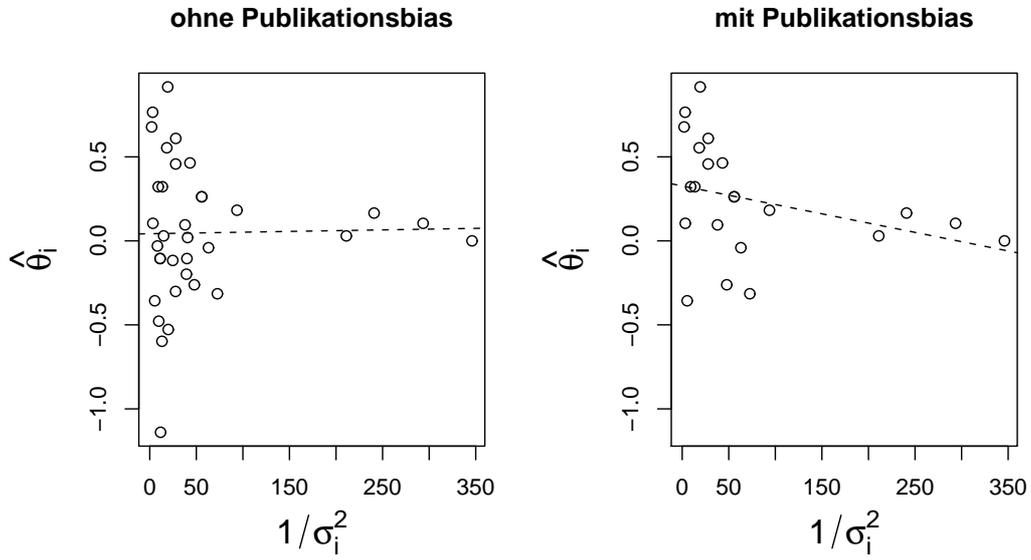


Abbildung 2.3: Regressionstest

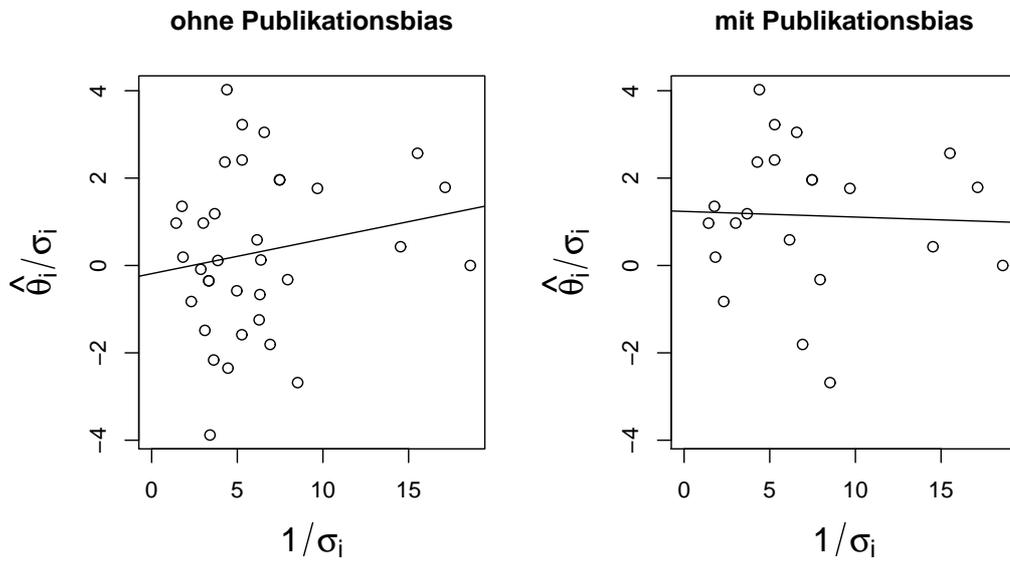


Abbildung 2.4: Egger-Test

von Null verschieden und Publikationsbias wird vermutet, Abb. 2.4 rechts. Nachteilig an diesem Test ist, dass er bei binären Outcomes durch die Korrelation zwischen der Schätzung der Effektgrößen (Harbord et al. (2006)) und ihren Standardfehlern unter gewissen Umständen eine hohe Fehler 1. Art-Quote aufweist, d.h. häufig lässt der Test vermuten, es bestünde Publikationsbias, obwohl keiner besteht (Sutton & Higgins (2008)).

Daher wurden eine Reihe von Tests speziell für binäre Outcomes entwickelt (Sutton & Higgins (2008)). Bei Verwendung von Odds Ratios als Effektgrößen empfiehlt die Cochrane Collaboration davon den Test von Harbord et al. (2006). Dieser basiert auf den Komponenten des Score-Tests, also der Score-Statistik und der Fisher-Information.

Die Daten liegen in folgender Form vor:

Tabelle 2.1: Notation für die Outcomes einer Studie

	Event	kein Event	Σ
Präparat	r_i^T	$n_i^T - r_i^T$	n_i^T
Placebo	r_i^C	$n_i^C - r_i^C$	n_i^C
Σ	r_i	$n_i - r_i$	n_i

Damit lauten die geschätzten logarithmierten Odds Ratios:

$$\hat{\theta}_{OR_i} = \log \left[\frac{r_i^T (n_i^C - r_i^C)}{r_i^C (n_i^T - r_i^T)} \right] \quad (2.3)$$

mit geschätzten Varianzen

$$\hat{V}(\hat{\theta}_{OR_i}) = \left(\frac{1}{r_i^T} + \frac{1}{n_i^C - r_i^C} + \frac{1}{r_i^C} + \frac{1}{n_i^T - r_i^T} \right) \quad (2.4)$$

Die Score-Statistiken ergeben sich zu:

$$Z_{OR_i} = r_i^T - r_i n_i^T / n_i \quad (2.5)$$

und die Fisher-Informationen, ausgewertet bei $\theta_{OR_i} = 0$, lauten:

$$V_{OR_i} = \frac{n_i^C n_i^T r_i (n_i - r_i)}{n_i^2 (n_i - 1)} \quad (2.6)$$

V_{OR_i} ergibt sich, indem die auf die Randhäufigkeiten in Tabelle 2.1 bedingte Likelihood verwendet wird.

Im Gegensatz zu den Standardschätzern der Varianzen $\hat{V}(\hat{\theta}_{OR_i})$ hängen die V_{OR_i} nur von den Randhäufigkeiten in Tabelle 2.1 ab, wodurch sie weniger stark von den Stichprobenfehlern beeinflusst werden. Dadurch wird die Korrelation zwischen den geschätzten Odds Ratios und ihren geschätzten Varianzen viel geringer.

Die $Z_{OR_i}/\sqrt{V_{OR_i}}$ werden ungewichtet auf die $\sqrt{V_{OR_i}}$ regressiert und wie beim Egger-Test die Verschiedenheit von Null des Intercepts mit einem zweiseitigen t-Test überprüft.

Dieser Test weist, wenn die Heterogenität gering bis mittel stark ausgeprägt ist, eine Fehler-1.-Art-Quote nahe dem Signifikanzniveau, also dem Sollwert, auf und hat dennoch eine dem Egger-Test vergleichbare Power.

Bei starker Heterogenität wächst die Fehler-1.-Art-Quote, obwohl viel geringer als beim Egger-Test, auch hier stark an. In diesem Fall ist es allerdings unwahrscheinlich, dass die Heterogenität nur durch Publikationsbias erklärt werden kann und es sollten andere potentielle Gründe mittels Sensitivitätsanalyse (Abschnitt 2.9) und / oder Metaregression (Abschnitt 2.10) untersucht werden.

Harbord et al. (2006) geben auch die sich bei Verwendung von logarithmierten relativen Risiken ergebenden Statistiken Z_{RR_i} und V_{RR_i} an:

$$Z_{RR_i} = (r_i^T n_i - r_i n_i^T) / (n_i - r_i) \quad (2.7)$$

$$V_{RR_i} = n_i^C n_i^T r_i / n_i (n_i - r_i) \quad (2.8)$$

Allerdings wurden die statistischen Eigenschaften dieser Form des Tests noch nicht evaluiert (Harbord et al. (2006)).

Natürlich gilt, wie für die grafischen Methoden auch hier, dass andere, unberücksichtigte Größen das Ergebnis der Tests beeinflussen könnten.

2.5.2 Ansätze zur Korrektur

Da unveröffentlichte Studien nicht beobachtet werden, kann auch nicht festgestellt werden, warum sie nicht veröffentlicht werden. Damit kann Publikationsbias nicht korrigiert werden, ohne dass Annahmen über die Mechanismen gemacht werden, die zu dem Vorbehalt von Studien führen (Bowden et al. (2006)). Deshalb werden Methoden zur Korrektur häufig im Rahmen einer Sensitivitätsanalyse vorgestellt, mit der überprüft wird, wie empfindlich das Ergebnis gegenüber Änderungen in diesen Mechanismen ist (Sutton & Higgins (2008)).

Es werden im Folgenden exemplarisch zwei solcher Methoden vorgestellt:

Regressions-Methode:

Es gibt eine Reihe verwandter Methoden, die mit Hilfe von Regression Publikationsbias adjustieren. Unter diesen hat sich in einer Simulation, neben einer Methode speziell für Odds Ratios, eine Adjustierung mit Hilfe linearer Regression der Effektgrößen auf die Varianzen in den Studien als am besten hinsichtlich der mittleren quadratischen Abweichung und des absoluten Bias erwiesen:

$$\hat{\theta}_i = \beta_0 + \beta_1 \sigma_i^2 + g_i, \quad g_i \sim N(0, V(\hat{\theta}_i)) \quad (2.9)$$

Dabei kann die Varianz der Effektgrößen $V(\hat{\theta}_i)$ gleich der beobachteten Varianz σ_i^2 in den Studien oder der Varianz in den Studien plus der zwischen den Studien τ^2 sein, siehe Abschnitt 2.7.2. Die von Publikationsbias bereinigte Schätzung des Gesamteffekts ist dann der geschätzte Intercept $\hat{\beta}_0$, da dies die Vorhersage für eine hypothetische Studie mit Varianz Null ist. Wenn man davon ausgeht, dass die Schätzung unverzerrt ist, müsste eine solche Studie genau den wahren, von Publikationsbias bereinigten Gesamteffekt liefern (Moreno et al. (2009)).

Korrektur mit Hilfe von Pseudodaten:

Diese Methode stammt aus dem Bereich der Ansätze mit Hilfe von Selektionsmodellen. Bei diesen wird davon ausgegangen, dass bei der Generierung der Daten zwei Modelle wirken: Ein Modell, das Effektgrößenmodell, bestimmt die Werte der Ergebnisstatistiken der einzelnen Studien. Das zweite, das Selektionsmodell, bestimmt, ob diese Werte veröffentlicht werden. Letzteres ist eine Funktion bestimmter Größen, die Einfluss auf die Publikation haben, wie den Effektgrößen oder den p -Werten der Studien. Selektionsmodelle fallen in zwei Lager, in denen unterschiedliche Annahmen über die Parameter der Modelle gemacht werden: Bei den einen wird angenommen, dass die Parameter unbekannt sind, aber geschätzt werden können, bei den anderen wird vorausgesetzt, dass die Parameter bekannt sind, es wird dann aber die Sensitivität der Ergebnisse auf Änderungen in diesen Parametern untersucht.

Die Methode der Korrektur unter Benutzung von Pseudodaten stammt aus dem zweiten Bereich. Bei dieser wird beispielsweise vorausgesetzt, dass die Veröffentlichung nur vom p -Wert abhängt und dass Studien mit $p < 0.05$ immer veröffentlicht werden und solche mit $p > 0.05$ mit einer Wahrscheinlichkeit von $P_{p > 0.05}(\text{“Veröffentlichung”}) = 0.4$. Eine solche Regel wird Selektionskriterium C genannt.

Im Folgenden wird die Methode beschrieben, für die Annahme, dass keine Varianz zwischen den Studien besteht (Modell mit festen Effekten, siehe Ab-

schnitt 2.7.1). Zunächst wird der unkorrigierte Gesamteffekt $\hat{\theta}$ aus den Daten geschätzt. Dann wird ein $\hat{\theta}_{i1}$ aus $N(\hat{\theta}, \sigma_i^2)$ gezogen und mit dem festgelegten C verglichen. Wenn es die Voraussetzung für die Veröffentlichung erfüllt, wird es angenommen, ansonsten verworfen und die Prozedur wiederholt, bis ein $\hat{\theta}_{i1}$ die Voraussetzung erfüllt. Das wird für alle $i = 1, \dots, k$ durchgeführt, sodass ein vollständiger Pseudodatensatz $\vec{\hat{\theta}}_1$ entsteht. Um die Genauigkeit der Schätzung zu verbessern, werden mehrere Datensätze $\vec{\hat{\theta}}_1, \dots, \vec{\hat{\theta}}_m$ erzeugt, wobei sich $m = 5$ als ausreichend erwiesen hat.

Insgesamt liegen nun der echte Datensatz $\vec{\theta}_0$, erzeugt unter einer unbekanntem Gesamteffektgröße θ vor und m Pseudo-Datensätze, erzeugt unter der bekannten Gesamteffektgröße $\hat{\theta}$.

Nun soll, abhängig von θ die Wahrscheinlichkeit maximiert werden, dass der Datensatz $\vec{\theta}_0$ aus der Verteilung mit Erwartungswert θ kommt, gegeben, dass genau ein Datensatz aus dieser Verteilung kommt und der Rest aus der Verteilung mit Erwartungswert $\hat{\theta}$.

Das führt zu folgender bedingten Likelihood-Funktion:

$$\prod_{i=1}^k \frac{g(\hat{\theta}_{i0} | \theta, C) \prod_{k=1}^m g(\hat{\theta}_{ik} | \hat{\theta}, C)}{g(\hat{\theta}_{i0} | \theta, C) \prod_{k=1}^m g(\hat{\theta}_{ik} | \hat{\theta}, C) + \sum_{j=1}^m g(\hat{\theta}_{ij} | \theta, C) \prod_{k=0, k \neq j}^m g(\hat{\theta}_{ik} | \hat{\theta}, C)} \quad (2.10)$$

Dabei ist g die Verteilung, die sich ergibt, wenn das Selektionskriterium auf die Normalverteilungen der Ursprungsdaten angewendet wird:

$$g(y | \theta, C) = \frac{f(y | \theta)}{p(y \in C)} \quad (2.11)$$

Da alle der Terme in der Likelihood dieselbe Selektionswahrscheinlichkeit haben, lässt diese sich kürzen, wodurch sich die Likelihood-Funktion nach Umformung wie folgt schreiben lässt:

$$\prod_{i=1}^k \frac{f(\hat{\theta}_{i0} | \theta) / f(\hat{\theta}_{i0} | \hat{\theta})}{\sum_{j=0}^m (f(\hat{\theta}_{ij} | \theta) / f(\hat{\theta}_{ij} | \hat{\theta}))} \quad (2.12)$$

Wenn Variation zwischen den Studien angenommen wird, muss zusätzlich die Zwischenstudienvarianz τ^2 (Abschnitt 2.7.2) geschätzt werden und es ergibt sich analog folgende bedingte Likelihood-Funktion:

$$\prod_{i=1}^k \frac{f(\hat{\theta}_{i0} | \theta, \tau) / f(\hat{\theta}_{i0} | \hat{\theta}, \hat{\tau})}{\sum_{j=0}^m (f(\hat{\theta}_{ij} | \theta, \tau) / f(\hat{\theta}_{ij} | \hat{\theta}, \hat{\tau}))} \quad (2.13)$$

Die Lösung der Likelihood könnte z.B. in R mit der Funktion `optim()` geschehen.

Eine Simulation hat gezeigt, dass diese Methode, wenn die Annahmen über C stimmen, unverzerrte Ergebnisse liefert. Die Varianz der Schätzung lässt sich verringern, wenn die Prozedur mit dem Ergebnis der ersten Anwendung als neues $\hat{\theta}$ und, gegeben den Fall, als neues $\hat{\tau}$ noch einmal wiederholt wird (Bowden et al. (2006)).

2.6 Homogenität / Heterogenität

Die Frage nach der Homogenität der Studienpopulation ist zentral, da sie entscheiden kann, wie bei der Analyse weiter vorgegangen wird. Homogenität bedeutet, dass alle Effektgrößen denselben Gesamteffekt schätzen.

2.6.1 Formaler Test

Der gängigste Test ist der χ^2 -Homogenitätstest (Blettner & Schlattmann (2005)):

$$\chi_{het}^2 = \sum_{i=1}^k \frac{(\hat{\theta}_i - \hat{\theta})^2}{\sigma_i^2} \sim \chi_{k-1}^2 \quad (2.14)$$

Dabei bezeichnen $\hat{\theta}_i$ die einzelnen Effektgrößen, $\hat{\theta}$ den geschätzten Gesamteffekt (Berechnung siehe Kap. 2.8), σ_i^2 die Varianzen und k die Anzahl der Effektgrößen. Die quadratischen standardisierten Abweichungen werden aufsummiert, welche bei Homogenität einer χ^2 -Verteilung mit $k - 1$ Freiheitsgraden folgen. Überschreitet die Teststatistik das $1 - \alpha$ -Quantil der χ^2 -Verteilung, wird die Nullhypothese "Es besteht keine Heterogenität" abgelehnt.

Dieser Test hat geringe Power. Die Anzahl der Studien hat dabei keinen bedeutenden Einfluss. Vielmehr ist die Power besonders gering, wenn die Zahl der zugrunde liegende Individualdaten gering ist und sich die Varianz der Effektgrößen zwischen den Studien stark unterscheidet, was meistens der Fall ist (Baujat et al. (2002)). Nichtsdestotrotz bietet die Teststatistik ein nützliches informelles Maß.

2.6.2 Informelle Analyse

Da formale Tests Heterogenität generell schlecht entdecken können, sollte zusätzlich eine informelle Analyse durchgeführt werden (Sauerbrei & Blettner (2003)).

Hierbei werden neben den bereits im Zusammenhang mit Publikationsbias behandelten Funnel Plots und Galbraith Plots weitere grafische Verfahren angewendet.

Eine erste Einschätzung liefert der Forest Plot, der in Metaanalysen standardmäßig eingesetzt wird. Dabei werden die geschätzten Effektgrößen der einzelnen Studien mit ihren Konfidenzintervallen untereinander geplottet. Zur Identifikation der einzelnen Studien werden üblicherweise die Betreiber und das Jahr der Durchführung angegeben. Häufig findet sich unten die Gesamteffektgröße mit Konfidenzintervall in Form einer Raute (Cooper & Hedges (1994)). Streuen die Effektgrößen zu stark und haben dazu noch kleine Konfidenzintervalle (Abb. 2.5 links), wird Heterogenität vermutet. Sind sie hingegen vergleichbar groß ohne auffallend abweichende Werte mit dazu kleinen Konfidenzintervallen, wird angenommen, dass die Schwankungen symmetrisch um eine Gesamteffektgröße herum und zufällig sind, was Homogenität bedeutet (Abb. 2.5 rechts).

Des weiteren zählt dazu ein Plot des Einflusses der Effektgrößen auf die χ^2 -Heterogenitätsstatistik (\hat{x}) gegen ihren Einfluss auf die Gesamteffektgröße (\hat{y}):

$$\hat{x}_i = \frac{(\hat{\theta}_i - \hat{\theta})^2}{\sigma_i^2} \quad \hat{y}_i = \frac{(\hat{\theta}_{-i} - \hat{\theta})^2}{V(\hat{\theta}_{-i})} \quad (2.15)$$

Dabei ist $\hat{\theta}_{-i}$ die geschätzte Gesamteffektgröße, wenn die i -te Studie weggelassen wird und $V(\hat{\theta}_{-i})$ die Varianz dieser Größe (Berechnung siehe Abschnitt 2.8). \hat{x}_i ist damit der i -te Summand der χ^2 -Heterogenitätsstatistik und \hat{y}_i ist die standardisierte quadratische Abweichung zwischen der alten und der neuen geschätzten Gesamteffektgröße, die sich ergeben würde, wenn in einer Metaanalyse mit $k - 1$ Studien die i -te Studie hinzugenommen würde. In den Plot werden als Zeichensymbole statt Punkten Zahlen verwendet, die der Reihenfolge der Studien im Datensatz entsprechen, damit die Studien zugeordnet werden können. Anders als z.B. beim Galbraith Plot wird hier die Heterogenität und der Einfluss der einzelnen Studien nur in Bezug zu den anderen betrachtet. Besonders einflussreiche Studien, die zudem viel zur Heterogenität beitragen, finden sich rechts oben. Somit lassen sich abweichende und das Gesamtergebnis stark beeinflussende Studien identifizieren, ohne notwendigerweise die Gründe für die Heterogenität zu betrachten.

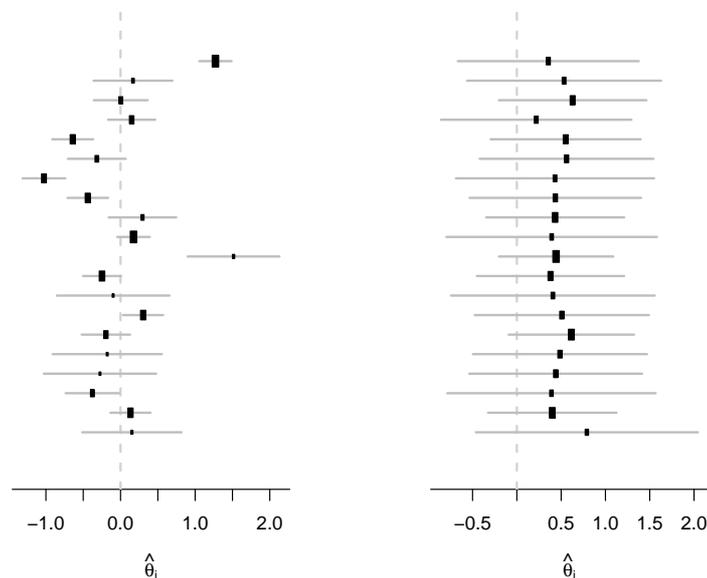


Abbildung 2.5: Forest Plots bei unterschiedlich starker Heterogenität

Es können aber auch Subgruppen untersucht werden, indem, beispielsweise durch unterschiedliche Farben, die Zugehörigkeit zu den unterschiedlichen Gruppen gekennzeichnet wird, siehe Abb. 2.6. Subgruppen, die besonders heterogen und einflussreich sind, werden entsprechend oben rechts oben in der Graphik sein.

Mittels Sensitivitätsanalyse können die Auswirkungen auf das Gesamtergebnis untersucht werden, wenn besonders heterogene Studien oder Subgruppen entfernt werden.

Dabei muss allerdings beachtet werden, dass derartige Untersuchungen *post hoc* sind, da keine vor Beginn der Analyse aufgestellten Hypothesen überprüft, sondern Zusammenhänge, die durch die Daten suggeriert werden, weiter untersucht werden. Dadurch sind die Ergebnisse weniger verlässlich. Möglicherweise gibt es aber vor der Analyse keine klaren Hypothesen, sondern man möchte Hinweise auf unbekannte Zusammenhänge erhalten. Ansonsten könnte, wenn die interessierenden Hypothesen nicht zu kompliziert sind, Metaregression angewendet werden, mit der es auch möglich ist, stetige Einflussgrößen einzuschließen (siehe Abschnitt 2.10). Die Methode kann aber auch zusammen mit Metaregression angewendet werden, wobei mit dem Plot, anders als bei der Metaregression untersucht wird, wie stark sich einzel-

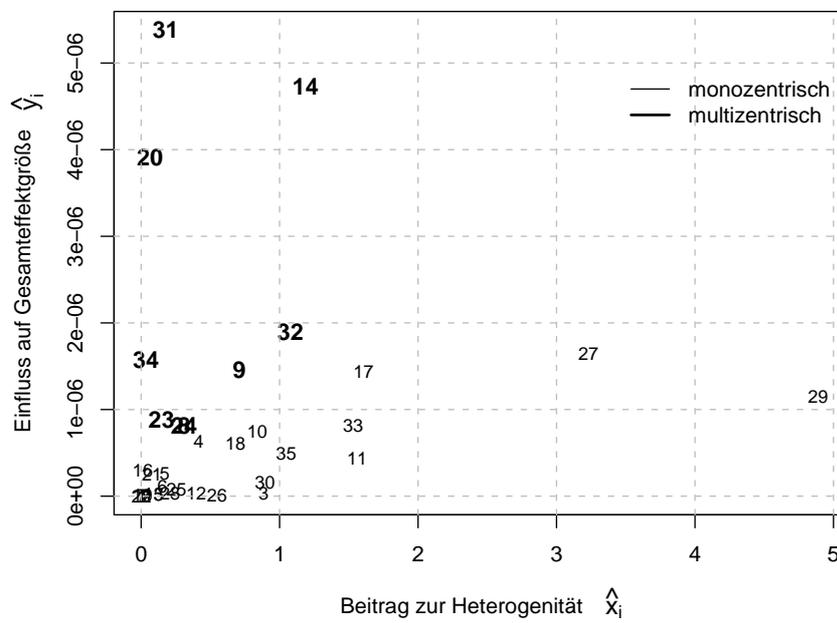


Abbildung 2.6: graphische Methode zur Untersuchung der Heterogenität

ne Subgruppen auf das Gesamtergebnis auswirken. Bei Metaregression wird hingegen nur untersucht, wie stark sich die Subgruppen unterscheiden (Baujat et al. (2002)).

Wenn die Daten in Form von 2x2-Tabellen vorliegen, kann der L'Abbé-Plot angewendet werden (Sutton & Higgins (2008)). Dabei werden die beobachteten Risiken oder Chancen der Kontrollgruppen gegen die der Interventionsgruppen gegeneinander geplottet. Die Größen der Punkte sind proportional zu den Gewichten der Studien. Häufig wird auch eine Regressionsgerade und die Diagonale, welche gleichen Risiken bzw. Chancen in den beiden Gruppen entspricht, eingezeichnet (Sharp & Thompson (2000)). Bei stärkerer Heterogenität sollten die Abweichungen von der Regressionsgerade groß sein.

Außerdem könnte die Studienpopulation hinsichtlich einzelner Merkmale, die die Werte der Effektgrößen möglicherweise stark beeinflussen, aufgesplittet und dann geprüft werden, ob sich die Werte in den entstehenden Subgruppen wirklich unterscheiden (Sauerbrei & Blettner (2003)).

2.6.3 Einschätzung des Ausmaßes

In letzter Zeit wird oft von vornherein davon ausgegangen, dass Heterogenität besteht, da es unrealistisch wäre, anzunehmen, dass die Studien, welche sich faktisch immer in gewissen Punkten unterscheiden, alle denselben Effekt schätzen (Sutton & Higgins (2008)). Aufgrund dessen wird mehr die Stärke der Heterogenität abgeschätzt, als dass ihr Vorhandensein überprüft wird. Ein Maß hierzu ist die I^2 -Statistik:

$$I^2 = \frac{\chi_{het}^2 - (k - 1)}{\chi_{het}^2} \quad (2.16)$$

Dabei ist χ_{het}^2 die Statistik des obigen Heterogenitätstests und k die Anzahl der Studien in der Metaanalyse. I^2 ist der Anteil der Zwischenstudienvarianz an der Gesamtvarianz. Wenn der Wert der Statistik negativ ist, wird sie auf Null gesetzt. So können Werte von Null bis eins angenommen werden. Null bedeutet, dass gar keine Heterogenität beobachtet wurde, Werte nahe bei eins bedeuten, dass die Studienpopulation sehr heterogen ist. Higgins et al. (2003) sprechen von schwacher Heterogenität bei Werten von 0,25 bis 0,5, mittlerer bei Werten von 0,5 bis 0,75 und starker ab 0,75.

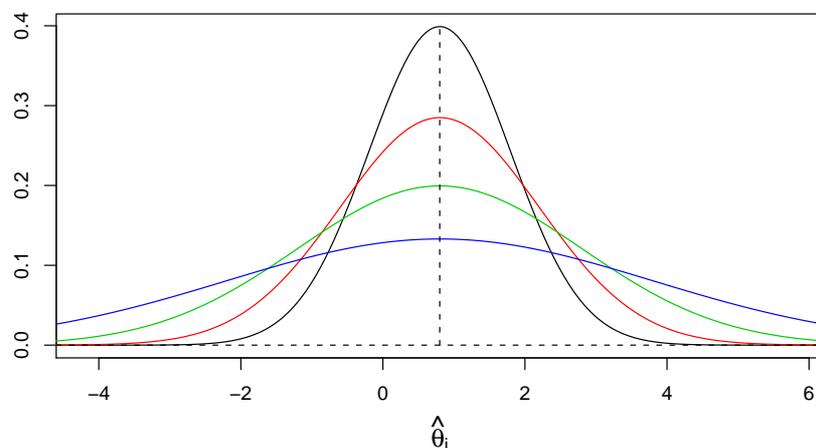


Abbildung 2.7: Modell mit festen Effekten

2.7 Modelle

Um die Effektgrößen gemeinsam analysieren zu können, müssen Verteilungen für sie angenommen werden. Im Folgenden werden zunächst die wichtigen Verteilungen und die daraus resultierenden Modelle vorgestellt. Anschließend wird die Wahl zwischen diesen Modellen diskutiert und schließlich die Schätzung der Varianz zwischen den Studien genauer behandelt.

2.7.1 Modell mit festen Effekten:

Hier wird vorausgesetzt, dass alle Studien denselben Gesamteffekt schätzen. Demnach streuen die Schätzer alle um diesen und die Variabilität ist nur den unterschiedlichen Unsicherheiten zuzuschreiben, mit denen dieser immer gleiche Effekt geschätzt wird, siehe Abb. 2.7. Es ergibt sich folgendes Modell (Blettner & Schlattmann (2005)):

$$\hat{\theta}_i = \theta + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_i^2) \quad (2.17)$$

$$\text{Also gilt: } V(\hat{\theta}_i) = \sigma_i^2. \quad (2.18)$$

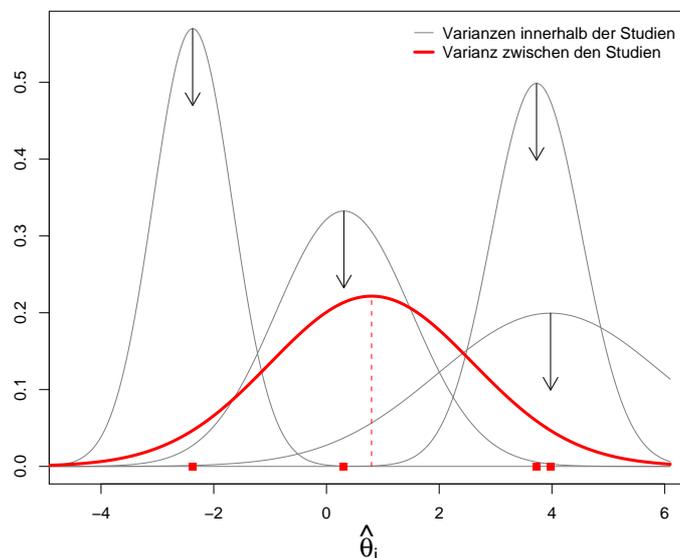


Abbildung 2.8: Modell mit zufälligen Effekten

2.7.2 Modell mit zufälligen Effekten:

In diesem Modell schätzen die Studien alle unterschiedliche Effekte. Diese haben eine eigene Verteilung, deren Erwartungswert geschätzt werden soll. Im Unterschied zum Modell mit festen Effekten existiert also kein Gesamteffekt, sondern nur der gemeinsame Erwartungswert der unterschiedlichen Effektgrößen (Blettner & Schlattmann (2005)). Die vorhandenen Effektgrößen werden demnach als zufällige Auswahl einer hypothetischen Effektgrößenpopulation betrachtet (Petitti (1994)). Anders als im Modell mit festen Effekten kommt hier zusätzlich zu den von Unsicherheiten in der Schätzung der Effekte verursachten Varianzen eine Varianz der wahren Effekte, oder auch Varianz zwischen den Studien hinzu, siehe Abb. 2.8. Das Modell lässt sich wie folgt notieren:

$$\hat{\theta}_i = \theta + b_i + \epsilon_i, \quad b_i \sim N(0, \tau^2), \quad \epsilon_i \sim N(0, \sigma_i^2) \quad (2.19)$$

$$\text{Also gilt: } V(\hat{\theta}_i) = \tau^2 + \sigma_i^2 \quad (2.20)$$

Die Varianz zwischen den Studien kann auch als Maß für das Ausmaß an Heterogenität interpretiert werden.

Die Verteilung der Effekte ist hier die Normalverteilung. Es ist jedoch auch eine diskrete Verteilung möglich, wenn angenommen wird, dass die geschätzten Effektgrößen zufällig aus m verschiedenen Subpopulationen stammen, innerhalb welcher jeweils derselbe Effekt geschätzt wird. Die Verteilung der geschätzten Effektgrößen ist also eine Mischung von Verteilungen, deren Dichte lautet:

$$f(\hat{\theta}_i, P) = \sum_{j=1}^m f(\hat{\theta}_i, \theta_j, \sigma_i^2) p_j. \quad (2.21)$$

Dabei ist

$$f(\hat{\theta}_i, \theta_j, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2} \frac{(\hat{\theta}_i - \theta_j)^2}{\sigma_i^2}\right). \quad (2.22)$$

Für die einzelnen Effektgrößen gegeben ihre Populationen wird also wieder eine Normalverteilung angenommen, mit Erwartungswert θ_j und Varianz σ_i^2 . p_j bezeichnen die a priori-Wahrscheinlichkeiten, dass eine Effektgröße zu Population j mit Parameter θ_j gehört.

Die Parameter P der Verteilung sind dann:

$$P \equiv \begin{bmatrix} \theta_1, \dots, \theta_m \\ p_1, \dots, p_m \end{bmatrix} \quad \text{mit} \quad p_j \geq 0 \quad j = 1, \dots, m. \quad (2.23)$$

$$\sum_{j=1}^m p_j = 1. \quad (2.24)$$

Dabei muss zusätzlich die Anzahl der Subgruppen k geschätzt werden.

Die Schätzung könnte mit dem kostenlosen Software-Paket C.A.MAN geschehen (Blettner & Schlattmann (2005)).

2.7.3 Modellwahl:

Im Fall von Homogenität sind das Modell mit festen und das mit zufälligen Effekten identisch, weil dann der Parameter τ^2 bei letzterem Null ist.

Bei Heterogenität ist das Modell mit zufälligen Effekten besser, da es die Varianz zwischen den Studien berücksichtigt und so größere Konfidenzintervalle liefert. Es ist daher auch zu bevorzugen, wenn man eher vorsichtig sein will und die Schätzung deswegen konservativ sein soll (Petitti (1994)). Die

Berücksichtigung der Heterogenität geschieht aber nur in der Form, dass die Varianz zwischen den Studien geschätzt wird, die Unterschiede werden aber nicht erklärt. Daher sollten zusätzlich mit Hilfe von Sensitivitätsanalyse und Metaregression Kovariablen in die Analyse aufgenommen werden. Dadurch ergeben sich gemischte Modelle, d.h., dass sowohl feste, in diesem Fall die Kovariablen, als auch zufällige Effekte, hier die Variabilität zwischen den Studien, im Modell sind.

Wenn die Parameter mit Likelihood-Methoden geschätzt wurden, können die Modelle mit formalen Anpassungstest, wie dem Likelihood Ratio Test oder dem bayesianischen Informationskriterium verglichen werden (Blettner & Schlattmann (2005)).

2.7.4 Schätzung der Zwischenstudienvarianz

Die Zwischenstudienvarianz ist wahrscheinlich die beste Größe, um das Ausmaß der Heterogenität zu beurteilen (Sutton & Higgins (2008)). Jedoch ist sie häufig aufgrund geringer Studienzahl schwierig zu schätzen. Es existieren dazu verschiedene Methoden, die abhängig von der Stärke der Heterogenität und der Studienzahl unterschiedlich gute Schätzungen liefern. Daher werden im Folgenden mehrere Schätzer vorgestellt und anschließend für die verschiedenen Umstände diskutiert, welcher von diesen jeweils am geeignetsten ist.

Die gebräuchlichste ist eine Momenten-Methode von DerSimonian und Laird. Bei dieser wird die Statistik χ_{het}^2 des χ^2 -Heterogenitätstest ihrem erwarteten Wert gleichgesetzt und nach τ^2 aufgelöst. Der Erwartungswert wird dabei unter der Annahme eines Modells mit zufälligen Effekten gebildet (Blettner & Schlattmann (2005)). Es gilt also $E(\hat{\theta}_i) = \theta$ und insbesondere $V(\hat{\theta}_i) = \sigma_i^2 + \tau^2$.

Der Erwartungswert ergibt sich damit zu:

$$E(\chi_{het}^2) = k - 1 + \tau^2 \left(\sum_{i=1}^k w_i - \sum_{i=1}^k w_i^2 / \sum_{i=1}^k w_i \right) \quad (2.25)$$

$$\text{mit } w_i = \frac{1}{\sigma_i^2}$$

Gleichsetzen mit χ_{het}^2 und auflösen nach τ^2 liefert:

$$\hat{\tau}_{DSL}^2 = \frac{\chi_{het}^2 - (k - 1)}{\sum_{i=1}^k w_i - \sum_{i=1}^k w_i^2 / \sum_{i=1}^k w_i} \quad (2.26)$$

Dabei kann $\hat{\tau}_{DSL}^2$ auch negative Werte annehmen. In diesem Fall wird die Schätzung auf Null gesetzt.

Der Varianz-Komponenten-Schätzer (VC) von Hedges basiert auf der Varianz-Komponenten-Schätzung in der Varianzanalyse mit zufälligen Effekten.

Man setzt dabei die quadratische Statistik

$$S_{\theta}^2 = \frac{1}{k-1} \sum_{i=1}^k (\hat{\theta}_i - \bar{\theta})^2 \quad (2.27)$$

ihrem erwarteten Wert gleich und löst wieder nach τ^2 auf. $\bar{\theta}$ bezeichnet dabei das ungewichtete Mittel der $\hat{\theta}_i$ (Hedges (1983)).

Der Schätzer ergibt sich damit zu:

$$\hat{\tau}_{VC}^2 = \frac{1}{k-1} \sum_{i=1}^k (\hat{\theta}_i - \bar{\theta})^2 - \frac{1}{k} \sum_{i=1}^k \sigma_i^2 \quad (2.28)$$

Bei negativen Werten wird die Schätzung wieder auf Null gesetzt.

Dieser Ansatz wurde allerdings in der Literatur bis auf eine spezielle Form für Mittelwertsdifferenzen selten verwendet.

Ein relativ neuer Schätzer von Sidik und Jonkman ist der Modell-Fehler-Varianz-Schätzer (MV). Er basiert auf der Reparametrisierung der totalen Varianz $V(\hat{\theta}_i) = \sigma_i^2 + \tau^2$ in $\tau^2(r_i + 1)$, mit $r_i = \sigma_i^2/\tau^2$. Die $\nu_i := r_i + 1$ lassen sich als Gewichte in der Regression ohne Kovariablen

$$\hat{\theta}_i = \theta + g_i, \quad g_i \sim N(0, \tau^2 \nu_i) \quad (2.29)$$

auffassen, die dem Modell mit zufälligen Effekten entspricht.

Die Schätzung für τ^2 ergibt sich damit gemäß der üblichen Methode zur Schätzung der Modell-Fehler-Varianz:

$$\hat{\tau}_{MV}^2 = \frac{1}{k-1} \sum_{i=1}^k \frac{1}{\hat{\nu}_i} (\hat{\theta}_i - \hat{\theta}_{\hat{\nu}})^2 \quad (2.30)$$

$$\text{mit } \hat{\theta}_{\hat{\nu}} = \sum_{i=1}^k \frac{1}{\hat{\nu}_i} \hat{\theta}_i / \sum_{i=1}^k \frac{1}{\hat{\nu}_i} \quad (2.31)$$

Wenn r_i bekannt wäre, so wäre der Schätzer unverzerrt. Es muss aber a priori geschätzt werden. Eine sehr raue Schätzung ist $\hat{r}_i = \sigma_i^2 / ((\sum_{i=1}^k (\hat{\theta}_i - \bar{\theta})^2) / k)$.

In den meisten Situationen liefert aber $\hat{r}_i = \sigma_i^2 / \hat{\tau}_{VC}^2$ ein besseres Ergebnis, wobei $\hat{\tau}_{VC}^2$ die Schätzung aus der Varianz-Komponenten-Methode ist. Für

den Fall, dass diese Null ist, müssen kleine Werte hinzu addiert werden, damit $\hat{\tau}_i$ nicht gegen unendlich geht, z.B. 0,01. Wenn diese a priori-Schätzung verwendet wird, wird die Methode MVvc genannt, um sie von der vorigen abzugrenzen.

Anzumerken ist, dass sich bei diesen beiden Methoden, im Gegensatz zu den bisherigen keine negativen Werte ergeben können. Die Schätzer müssen also nicht künstlich auf Null gesetzt werden.

Der Maximum-Likelihood-Schätzer maximiert die Likelihood-Funktion, die sich aus den Dichten der Effektgrößen ergibt. Er liefert aber in beinahe jedem Fall schlechtere Schätzungen als der approximative restringierte Maximum-Likelihood(REML)-Schätzer (ARE) und wird daher nicht weiter behandelt. Im Gegensatz zum ML-Schätzer berücksichtigt der ARE-Schätzer in seiner zugehörigen Likelihood-Funktion, dass außer τ^2 auch θ aus den Daten geschätzt wird. Mit $\hat{\theta}_i \sim N(0, \sigma_i^2 + \tau^2)$ lautet die restringierte log-Likelihoodfunktion:

$$l_R(\theta, \tau^2) = -\frac{k}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^k \log(\sigma_i^2 + \tau^2) - \frac{1}{2} \log \left(\sum_{i=1}^k \frac{1}{\sigma_i^2 + \tau^2} \right) - \frac{1}{2} \sum_{i=1}^k \frac{(\hat{\theta}_i - \hat{\theta})^2}{\sigma_i^2 + \tau^2}, \quad (2.32)$$

$$\text{mit } w_i = 1/(\sigma_i^2 + \tau^2) \quad \text{und} \quad \hat{\theta} = \sum_{i=1}^k w_i \hat{\theta}_i / \sum_{i=1}^k w_i$$

Der folgende Schätzer maximiert diese log-Likelihoodfunktion approximativ:

$$\hat{\tau}_{ARE}^2 = \frac{\sum_{i=1}^k \hat{w}_{ARE_i}^2 \{(k/(k-1))(\hat{\theta}_i - \hat{\theta}_{ARE})^2 - \sigma_i^2\}}{\sum_{i=1}^k \hat{w}_{ARE_i}^2}$$

$$\text{mit } \hat{w}_{ARE_i} = 1/(\sigma_i^2 + \hat{\tau}_{ARE}^2) \quad \text{und} \quad \hat{\theta}_{ARE} = \sum_{i=1}^k \hat{w}_{ARE_i} \hat{\theta}_i / \sum_{i=1}^k \hat{w}_{ARE_i} \quad (2.33)$$

In jeder Iteration muss der Schätzer bei Negativität auf Null gesetzt werden. Eine entsprechende Abwandlung dieses Schätzers zur Einbeziehung von Kovariablen hat sich als besonders effizient für die Metaregression erwiesen (Thompson, Sharp 1999).

Laut einer Simulationsstudie von Sidik & Jonkman (2007) ist der beste

Schätzer für die Zwischenstudienvarianz in den meisten Situationen der empirische Bayes-Schätzer (EB). Auch hier ist eine iterative Berechnung nötig:

$$\hat{\tau}_{EB}^2 = \frac{\sum_{i=1}^k \hat{w}_{EB_i} \{ (k/(k-1)) (\hat{\theta}_i - \hat{\theta}_{EB})^2 - \sigma_i^2 \}}{\sum_{i=1}^k \hat{w}_{EB_i}}$$

mit $\hat{w}_{EB_i} = 1/(\sigma_i^2 + \hat{\tau}_{EB}^2)$ und $\hat{\theta}_{EB} = \sum_{i=1}^k \hat{w}_{EB_i} \hat{\theta}_i / \sum_{i=1}^k \hat{w}_{EB_i}$ (2.34)

wobei wieder in jedem Schritt die Nicht-Negativität erzwungen werden muss. Es gibt aber Situationen, in denen andere Schätzer zu bevorzugen sind. Wenn sehr geringe Heterogenität vermutet wird, ist die Methode von DerSimonian und Laird unverzerrter. Bei stärkerer Heterogenität unterschätzt sie die Zwischenstudienvarianz aber stark, wodurch die Gefahr falsch positiver Entscheidungen steigt.

VC und MV, dagegen, überschätzen die Varianz, sind aber bei starker Heterogenität geeigneter. Wenn die Studienzahl dabei gering ist, sollte MV angewendet werden, wenn sie groß ist, dagegen VC.

In allen anderen Situationen sind der EB-Schätzer oder der MVvc-Schätzer zu wählen. Letzterer ist zwar etwas verzerrter, hat aber die Vorteile, dass keine iterative Berechnung nötig ist und, dass die Varianz nie Null geschätzt wird.

Das Ausmaß der Heterogenität kann vor der Schätzung nicht genau bestimmt werden, aber es sollte möglich sein, die Stärke in die Kategorien "niedrig", "mittel" und "stark" einzuteilen, um darauf den geeignetsten Schätzer zu wählen. Dabei helfen der Forest Plot und die I^2 -Statistik (Sidik & Jonkman (2007)).

2.8 Berechnung der Gesamteffektgröße

In den vorherigen Kapiteln wurde immer wieder die Gesamteffektgröße erwähnt. Diese sollte theoretisch nur berechnet werden, wenn die Studienergebnisse homogen sind, d.h. wenn alle Effektgrößen denselben Effekt messen. Da dies allerdings, wie schon erwähnt, unrealistisch ist, es aber dennoch häufig wichtig ist, ein einheitliches Ergebnis zu erhalten, kann man sie auch bei leichter Heterogenität ermitteln.

Die Berechnung geschieht dann wie folgt:

$$\hat{\theta} = \frac{\sum_{i=1}^k w_i \hat{\theta}_i}{\sum_{i=1}^k w_i}, \quad w_i = \frac{1}{V(\hat{\theta}_i)}. \quad (2.35)$$

Die Varianz ergibt sich zu:

$$V(\hat{\theta}) = \frac{\sum_{i=1}^k \frac{1}{V(\hat{\theta}_i)^2} V(\hat{\theta}_i)}{(\sum_{i=1}^k \frac{1}{V(\hat{\theta}_i)})^2} = \frac{1}{\sum_{i=1}^k \frac{1}{V(\hat{\theta}_i)}} = \frac{1}{\sum_{i=1}^k w_i}. \quad (2.36)$$

Dabei wird vernachlässigt, dass die Varianz der Studien geschätzt und damit selbst mit einer Unsicherheit behaftet ist (Blettner & Schlattmann (2005)).

Ein offensichtlicher Nachteil von dieser Methode ist, dass die unterschiedliche Qualität der Studien nicht berücksichtigt wird.

Ein Ansatz dies zu tun, wäre, so genannte Quality Scores in die Gewichte einzubauen. Diese würden als gewichtetes Mittel aus Scores zu verschiedenen, selbst festgelegten Qualitätsmerkmalen berechnet werden. Je besser eine Studie ist, desto einen höheren Score würde sie erhalten (Petitti (1994)).

Bei dieser Methode gibt es allerdings einige Probleme. So ist sich die Forschungswelt nicht einig, welche Punkte überhaupt Qualitätsmerkmale darstellen. Es gibt hierzu keine festgesetzten Regeln. Damit würde sich je nach Forschungsteam eine (etwas) andere Gesamteffektgröße ergeben. Dazu kommt, dass sich durch die mit abstrakten Scores veränderten Gewichte, die entstehende Effektgröße nicht mehr als konkreter Effekt interpretieren ließe. Man könnte nur noch untersuchen, ob ein Effekt signifikant von Null verschieden ist. Zum Beispiel wäre es nicht mehr möglich, festzustellen, dass ein Medikament die Chance von einer Krankheit zu genesen um 0,2 erhöht, sondern nur, dass es die Chance überhaupt erhöht (Lipsey & Wilson (2001)). Weitere Probleme ergeben sich dadurch, dass die Informationen über die Merkmale aus den Studienberichten abgeleitet werden müssen. Diese sind häufig nicht verlässlich oder unvollständig. Es könnte z.B. im Bericht nicht erwähnt werden, dass die Probanden der Versuchs- und der Kontrollgruppe nicht randomisiert zugeordnet wurden. Wenn jetzt aber fälschlicherweise angenommen wird, dass die Zuteilung randomisiert vollzogen wurde, kann diese Fehleinschätzung zu einer erheblichen Verzerrung führen (Petitti (1994)).

Aus diesen Gründen wird diese Methode allgemein nicht empfohlen (Lipsey & Wilson (2001)).

Stattdessen könnte der Einfluss von Qualitätsmerkmalen auf das Ergebnis untersucht werden, in dem die Studienpopulation nach den Qualitätsmerkmalen aufgesplittet wird und dann die Subgruppen wieder getrennt untersucht werden (Zhou et al. (2003)). Wenn sich die Ergebnisse zu stark

unterscheiden, sollte wieder keine über alle Studien gemittelte Gesamteffektgröße berechnet werden. Dies ist eine Sensitivitätsanalyse, die im Folgenden behandelt wird.

2.9 Sensitivitätsanalyse

Allgemein wird in einer Sensitivitätsanalyse untersucht, wie empfindlich das mit einem Modell gewonnene Ergebnis gegenüber unterschiedlichen Annahmen über dieses Modell ist. Wenn die Schlussfolgerung einer Analyse unabhängig von den betrachteten Annahmen ist, erhöht dies das Vertrauen in die Validität des Ergebnisses und es wird als robust bezeichnet.

In Metaanalysen gibt es unterschiedliche Anwendungen:

Häufig wird untersucht, ob sich das Ergebnis ändert wenn bestimmte Studien von der Analyse ausgeschlossen werden. Das sind zum einen sehr große, die Analyse dominierende Studien. Wenn diese stark verzerrte Ergebnisse lieferten, würde auch das Gesamtergebnis stärker verzerrt werden. Zum anderen werden umstrittene Studien ausgeschlossen, die extreme Ergebnisse liefern oder bei denen nicht in Erfahrung zu bringen ist, ob sie einzelne Einschlusskriterien erfüllen, die man aber nicht ohne Weiteres ausschließen kann (Petitti (1994)).

Allgemein können die Ergebnisse von durch kategoriale Kovariablen definierte Subgruppen verglichen werden. Diese Kovariablen können z.B. das Design oder die Qualität der Studien betreffen (Zhou et al. (2003)).

Auch interessant und sehr verbreitet ist, zu untersuchen, wie stark das Ergebnis von der Wahl zwischen dem Modell mit zufälligen und dem mit festen Effekten abhängt.

In den Studienberichten werden oft mehrere Effektgrößen angegeben. In diesem Fall kann man die Analyse getrennt für die einzelnen Effektgrößen durchführen und überprüfen, ob die Aussagen einander entsprechen (Petitti (1994)).

Wenn sich bei der Sensitivitätsanalyse starke Unterschiede ergeben, sollte wieder keine über alle Studien gepoolte Gesamteffektgröße angegeben werden, sondern mehrere Effektgrößen, getrennt nach den beeinflussenden Faktoren (Blettner & Schlattmann (2005)).

2.10 Metaregression

Metaregression ist eine Regression der Effekte der Studien auf eine oder mehrere Kovariablen auf Studienebene (Simmonds & Higgins (2007)). Die Kovariablen können dabei Merkmale der Studien sein, oder Mittelwerte von Merkmalen einzelner Beobachtungen in den Studien (Thompson & Higgins (2002)).

Metaregression ist auch eine Erweiterung der Sensitivitätsanalyse. Letztere stellt eine Metaregression mit kategorialen Kovariablen dar.

Es ist sowohl ein Modell mit festen (2.37), als auch mit zufälligen Effekten (2.38) möglich (Blettner & Schlattmann (2005)):

$$\hat{\theta}_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_l x_{i(p-1)} + \epsilon_i \quad (2.37)$$

mit $\epsilon_i \sim N(0, \sigma_i^2)$

$$\hat{\theta}_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_l x_{i(p-1)} + b_i + \epsilon_i \quad (2.38)$$

mit $b_i \sim N(0, \tau^2)$ und $\epsilon_i \sim N(0, \sigma_i^2)$, $i = 1, \dots, k$

Allerdings erscheint das zweite inhaltlich sinnvoller, da das Ziel zwar ist Heterogenität zu erklären, man aber nicht davon ausgehen kann, dass diese vollkommen erklärt wird. Modelle mit zufälligen Effekten berücksichtigen, wie schon erwähnt, die restliche Heterogenität, indem sie die nicht-erklärte Varianz abweichend von den festen Termen schätzen und so größere Konfidenzintervalle für die Modellparameter liefern. Die Schätzung der Parameter muss, wie auch bisher, gewichtet erfolgen, um die unterschiedlichen Genauigkeiten der Schätzer der Effekte zu berücksichtigen.

Wenn nur eine Kovariable im Modell ist, sollte das Ergebnis unbedingt auch grafisch dargestellt werden. Dabei müssen die Größen der Punkte, wie auch schon beim L'Abbé-Plot proportional zu den inversen Varianzen – den Gewichten der Effekte – sein.

Auch wenn der Test auf Heterogenität nicht signifikant war, ist es sinnvoll Metaregression durchzuführen. Zum einen hat der Test geringe Power, zum anderen wird nur getestet, ob Overdispersion der Effektgrößen besteht und nicht, ob bestimmte Kovariablen in irgendeiner Form Einfluss auf die Effektgrößen haben. Das wichtigste Argument ist aber, dass man heute, wie bereits mehrfach angesprochen, meist davon ausgeht, dass aufgrund der Unterschiedlichkeit der Studien immer Heterogenität besteht. Nach dieser Auffassung ist es nicht nötig auf Heterogenität zu testen.

Probleme:

Während die Randomisierung innerhalb der Studien die Verzerrung der Ergebnisse durch Confounder verhindert, ist dies bei der Metaregression nicht der Fall. Es können also andere Größen, ob beobachtet oder nicht, auf die Effektgrößen und gleichzeitig die Kovariablen einwirken. Damit kann ein scheinbarer Einfluss von Kovariablen in Wahrheit nur auf eine durch Drittvariablen verursachte Korrelation zurückzuführen sein. Dazu kommt, dass die Kovariablen, die in der Selektion betrachtet werden, oft stark miteinander korreliert sind und so die Schätzung ihrer Koeffizienten sehr ungenau wird. Zum Beispiel könnten die Studien, in denen eine besonders intensive Therapie gegeben wurde, diejenigen sein, an denen die stärker betroffenen Patienten teilnahmen. Diese Probleme sind in Metaanalysen deshalb von großer Bedeutung, da es sehr viele unterschiedliche Eigenschaften gibt, die sich von Studie zu Studie unterscheiden (Thompson & Higgins (2002)). Es sollte daher sowohl inhaltlich überlegt werden, wo Zusammenhänge bestehen könnten, als auch die Korrelationen zwischen den Merkmalen berechnet werden (Deeks et al. (2008)).

Bei den Kovariablen, die Mittelwerte von Variablen sind, deren Werte sich von Patient zu Patient in den Studien unterscheiden, z.B. Alter oder Geschlecht, ergibt sich ein spezielles Problem. Der Zusammenhang zwischen den Mittelwerten und den Effektgrößen auf Studienebene, welcher in der Metaregression betrachtet wird, kann sich von dem zugrunde liegenden Zusammenhang zwischen den Patientenmerkmalen und den Outcomes innerhalb der Studien unterscheiden. Abb. 2.9 veranschaulicht dieses Problem.

Dieses Phänomen entsteht durch Konfundierung entweder auf Studienebene, wodurch der Zusammenhang zwischen den Studien verzerrt wird, oder auf Ebene der einzelnen Studien, wodurch der Zusammenhang innerhalb der Studien verzerrt wird. Eine Verzerrung dieser Art wird meist Aggregation Bias oder Ecological Bias genannt und kann ohne Einzeldaten nicht untersucht und korrigiert werden. Für kategoriale Variablen sind aber manchmal die Effekte in den einzelnen Kategorien angegeben, so dass der Zusammenhang auch innerhalb der Studien untersucht werden kann (Thompson & Higgins (2002)).

Metaregression hat eine geringe Power, da im Vergleich zu anderen Anwendungen von Regression aufgrund der geringen Studienzahl wenig Information vorliegt. Die Cochrane-Kollaboration rät, die übliche Regel für einfache Regressionen, dass pro Kovariable im Modell zehn Beobachtungen notwendig sind, auch auf Metaregression anzuwenden. Damit sollte die Methode nur eingesetzt werden, wenn mindestens zehn Studien zur Verfügung stehen. Das

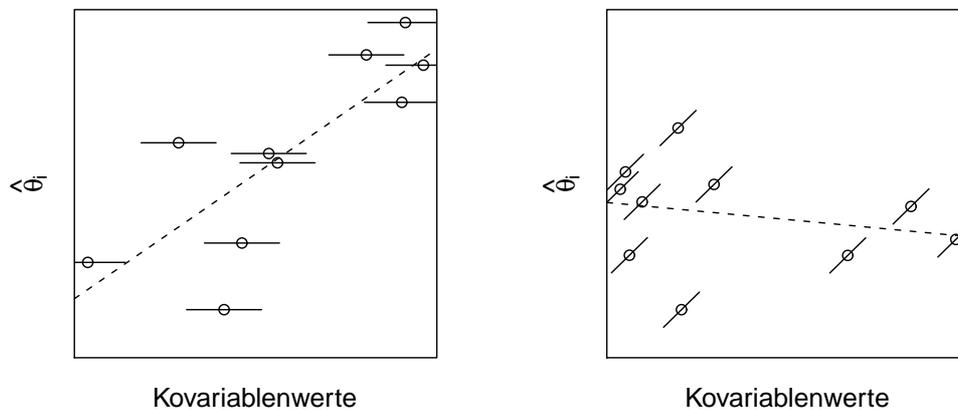


Abbildung 2.9: Verschiedene Zusammenhangstrukturen innerhalb und zwischen den Studien

Problem ist geringer, wenn die Kovariablen stark streuen und bei aggregierten Kovariablen zusätzlich die Streuung innerhalb der Studien gering ist (Deeks et al. (2008)).

Weitere Einschränkungen kann es durch fehlende Werte in den Variablen geben. Meist fehlen Angaben zu einzelnen in der Modellselektion zu berücksichtigenden Kovariablen, sodass es womöglich wenige Studien gibt, bei denen alle Werte vorliegen. Das vermindert zum einen die Power der Analyse weiter, zum anderen kann es zu einer Verzerrung kommen, wenn das Fehlen der Werte nicht zufällig ist.

Das wohl größte Problem bei Metaregression ist das “Fischen” nach signifikanten Ergebnissen. Üblicherweise stehen eine große Zahl von möglichen Einflussgrößen einer geringen Zahl an Studien gegenüber. Wenn mit einer großen Zahl an Kovariablen multiple Analysen durchgeführt werden und dabei zusätzlich häufig Zusammenhänge untersucht werden, die durch die Daten suggeriert werden, erhöht sich das Risiko falsch positiver Entscheidungen stark. Das kann nur dadurch verhindert werden, dass die Zahl der Kovariablen beschränkt wird und dass ihre Auswahl ohne Wissen über die Ergebnisse der Studien, vor Beginn der Metaanalyse, geschieht. Letzteres kann allerdings wahrscheinlich nicht erreicht werden, da die mit der Auswahl beauftragten Personen meist Experten auf dem entsprechenden Gebiet sind und vor Beginn schon über die Ergebnisse Bescheid wissen. In Metaanalysen sollte das gesammelte Wissen über das betreffende Thema enthalten sein. Daher ist

es oft schwer möglich, Einflussgrößen zu spezifizieren, ohne auf Informationen aus Studien in der Metaanalyse zurückzugreifen (Thompson & Higgins (2002)).

Die ausgewählten Kovariablen sollten im Protokoll festgehalten werden. Wenn allerdings eine Kovariable übersehen wurde, die aber von großer Bedeutung ist, was auch durch Quellen außerhalb der Metaanalyse gestützt wird, sollte sie dennoch verwendet werden. Allerdings muss dies als post hoc-Analyse gekennzeichnet werden.

Bei der Auswahl muss auch darauf geachtet werden, dass die Einflussgrößen inhaltlich wirklich mit dem Zielmerkmal in Verbindung stehen sollen und nicht nur mit dem Outcome. Zum Beispiel ist Rauchen ein guter Prädiktor für die Mortalität, nicht aber unbedingt für den Effekt einer medikamentösen Therapie auf die Mortalität (Deeks et al. (2008)).

Kapitel 3

Vor- und Nachteile

Im Folgenden werden zunächst die Vorteile von Metaanalysen, insbesondere gegenüber traditionellen Reviews beschrieben.

Metaanalysen bieten gegenüber letzteren einen viel differenzierten Umgang mit den eingeschlossenen Studien. Dadurch, dass die geschätzten Effektgrößen unter Berücksichtigung ihrer Genauigkeit eingehen, beeinflussen sie das Gesamtergebnis theoretisch genau mit der Stärke, mit der sie wirklich auch auf dieses einwirken. Es werden insbesondere auch kleine und nicht signifikante Studien mit dem jeweils nötigen Gewicht berücksichtigt. Dagegen verlassen sich narrative Reviews häufig nur auf das "Auszählen" der signifikanten Ergebnisse.

Außerdem gibt es in Metaanalysen z.B. mit Metaregression und Sensitivitätsanalyse Möglichkeiten Unterschiede in den Ergebnissen der Studien zu erklären, was mit anderen Herangehensweisen an die Zusammenfassung empirischer Forschung nur eingeschränkt möglich ist.

Ein weiterer Vorteil von Metaanalyse ist die Überprüfbarkeit und Reproduzierbarkeit der Ergebnisse. Die Ein- und Ausschlusskriterien und das Vorgehen beim Auffinden von relevanten Studien werden dokumentiert und offen dargelegt. Dazu werden die Ergebnisse und Charakteristiken der Studien formal codiert und die Methodik bei der Datenaanalyse beschrieben. Dadurch können externe Personen die Annahmen der Autoren, die Ergebnisse und Schlussfolgerungen einsehen und überprüfen und gegebenenfalls die Metaanalyse erneut durchführen.

Metaanalysen können dazu mit Informationen aus einer Vielzahl von Studien organisiert umgehen, da diese formal codiert und in einer Datenbank computergestützt gemeinsam analysiert werden. In narrativen Reviews geht der Überblick mit wachsender Studienzahl hingegen schnell verloren (Lipsey

& Wilson (2001)).

Nun wird allgemein auf Nachteile von Metaanalysen eingegangen.

Hauptkritikpunkt ist das immer die Gefahr besteht, dass das Ergebnis durch Publikationsbias, sowohl im Sinne des Fehlens ganzer Studien, als auch hinsichtlich des selektiven Veröffentlichens signifikanter Effektgrößen, verzerrt ist (Moreno et al. (2009)). Es existieren zwar schon relativ effektive Methoden, die Verzerrung durch das Fehlen von Studien zu korrigieren, diese basieren aber immer auf Annahmen, die nicht überprüft werden können, da die nicht-veröffentlichten Studien ihrer Natur nach nicht beobachtet werden. Es werden zwar häufig Sensitivitätsanalysen hinsichtlich des Einflusses dieser Annahmen durchgeführt, sollen aber, wie meist der Fall, anhand des Ergebnisses der Metaanalyse Entscheidungen getroffen werden, so sind Sensitivitätsanalysen von geringem Nutzen (Sutton & Higgins (2008)).

Ein weiteres Problem ist, dass die Studien häufig sehr unterschiedlich sind ("Äpfel & Birnen"-Problem). Insbesondere in den Sozialwissenschaften, in denen meist keine Randomisierung vorgenommen werden kann, können unzählige Konfundierungsgrößen die Ergebnisse systematisch verzerren. Auch die Outcomes in den Studien sind oft zu unterschiedlich.

Weiter sind verwendete Studien oft von geringer methodischer Qualität ("Müll rein, Müll raus"-Problem). Wenn im Extremfall nur schlechte Studien zur Verfügung stehen, wird das Ergebnis der Metaanalyse auch nicht verlässlicher sein als die Einzelstudien (Petitti (1994)).

Manchmal werden mehr als eine Effektgröße aufgenommen. Dadurch kann es zu einer erheblichen Verzerrung kommen, da Messungen an dem selben Datensatz statistisch abhängig sind. Die Unabhängigkeit der Effektgrößen ist eine fundamentale Annahme der Metaanalyse. Besondere Vorsicht ist hier geboten, da auf einen Datensatz mehrere Veröffentlichungen folgen können. Natürlich sind Metaanalysen, vor allem durch die Studiensuche und die Codierung sehr zeitintensiv und erfordern dazu Expertenwissen.

Für manche Anwendungen, vor allem in den Sozialwissenschaften, ist die Methodik der Metaanalyse zu wenig flexibel. Die starre und formale Codierung der Information, die sonst ein Vorteil ist, kann in manchen Anwendungen wichtige Aspekte nicht erfassen. Dies können z.B. das soziale Umfeld, theoretische Einflüsse und Verwicklungen, die Qualität der verwendeten Methodik, feinere oder kompliziertere Aspekte des Designs oder auch die Ergebnisse sein. In solchen Fällen wäre eine qualitative Zusammenfassung nötig. Es ist aber auch möglich Metaanalysen und narrative Reviews zusammen anzuwenden und Schlüsse aus beiden Techniken zu ziehen. Best evidence synthesis ist eine Methodik, mit der das geschehen könnte (Lipsey & Wilson (2001)).

Ein letzter, wichtiger Kritikpunkt ist, dass Metaanalysen zu einer Stagnation der Forschung führen können. Dadurch, dass so scheinbar präzise Ergebnisse geliefert werden, die aber verzerrt sein könnten, werden neue, möglicherweise bessere Studien aus ethischen und finanziellen Gründen nicht mehr unterstützt (Petitti (1994)).

Kapitel 4

Neuere Entwicklungen und Tendenzen

4.1 Mixed Treatment Comparison (MTC) Metaanalysen

In der Medizin existieren häufig verschiedene mögliche Behandlungen für die selbe Krankheit und man ist interessiert daran, welche Behandlung die effektivste ist. Oder man möchte zwei Behandlungen miteinander vergleichen, es existieren aber zu wenige Studien, die die beiden vergleichen, um über eine gewöhnliche Metaanalyse verlässliche Aussagen zu machen. In solchen Situationen kann eine MTC Metaanalyse durchgeführt werden, die laut Sutton & Higgins (2008) die möglicherweise wichtigste Entwicklung aus dem Bereich der Complex Research Synthesis ist, bei welcher zur Aggregation der Studien Modelle mit multiplen Parametern verwendet werden und /oder speziell Studien von unterschiedlichem Design zusammen analysiert werden. In einer MTC Metaanalyse können beliebig viele Behandlungen simultan analysiert und miteinander verglichen werden. Damit ist sie eine Verallgemeinerung der bisher behandelten, traditionellen paarweisen Metaanalyse, in der zwei Behandlungen oder eine Behandlung mit Placebo verglichen werden. In MTC-Metaanalysen wird Information aus direkten und indirekten Vergleichen berücksichtigt. Direkte Vergleiche ergeben sich, wenn zwei oder mehr Behandlungen innerhalb von Studien verglichen wurden. Bei indirekten Vergleichen entsteht die Verbindung über eine gemeinsame andere Behandlung, mit der die zu gegenüberstellenden Behandlungen in getrennten Studien verglichen wurden. Indirekte Vergleiche sind zwar nicht so aussagekräftig, wie direkte, liefern aber dennoch nützliche Information (Lu & Ades (2004)). In

manchen Situationen kann es aber zu Bias kommen, z.B. wenn die Studien für direkte Vergleiche systematisch unterschiedlich von denen für die indirekten Vergleiche sind (Caldwell et al. (2005)). Allerdings waren in einer Untersuchung von 44 Metaanalysen nur in drei Fällen signifikante Unterschiede zwischen den beiden Vergleichsarten zu beobachten (Song et al. (2003)).

Bei MTC-Metaanalysen gibt es zwei wichtige Annahmen:

Die wahren Odds Ratios zwischen zwei Behandlungen sind in allen Studien gleich (Modell mit festen Effekten) oder stammen zumindestens aus der gleichen Verteilung (Modell mit zufälligen Effekten), egal ob die jeweiligen Vergleiche in den einzelnen Studien wirklich durchgeführt wurden, oder nicht. Wenn also z.B. zusätzlich Behandlung B in Studien aufgenommen würde, in denen A und C verglichen werden und A in Studien, in denen B und C verglichen werden, so wären im Modell mit festen Effekten die wahren Odds Ratios zwischen A und B in diesen Studien die gleichen, wie in Studien in denen A und B direkt verglichen werden. Die Gültigkeit dieser Forderung kann kaum mit statistischen Mitteln nachgewiesen werden, sondern es sollten sachkundige Experten im klinischen Bereich herangezogen werden.

Eine weitere Forderung, die zunächst streng erscheint, ist, dass sich die Behandlungseffekte aufaddieren, dass also der relative Effekt von A gegen C durch die Effekte von A gegen B und B gegen C vorhergesagt werden kann (Caldwell et al. (2005)). Allerdings scheint diese Forderung, wie oben schon erwähnt, in den meisten Fällen erfüllt zu sein (Song et al. (2003)).

Zur Modellierung kann ein bayesianisches hierarchisches Modell verwendet werden:

Sei r_{il} die Anzahl der Ereignisse und n_{il} die Fallzahlen für Behandlung l in Studie i , wobei $i = 1, \dots, k$ und $l = 1, \dots, L$.

Dann lässt sich dieses Modell wie folgt notieren:

$$\begin{aligned}
 r_{il} &\sim \text{bin}(p_{il}, n_{il}) \\
 \text{logit}(p_{i1}) &= \mu_i - \delta_{i2}/L - \delta_{i3}/L - \dots - \delta_{iL}/L \\
 \text{logit}(p_{i2}) &= \mu_i + (L-1)\delta_{i2}/L - \delta_{i3}/L - \dots - \delta_{iL}/L \\
 &\vdots \\
 \text{logit}(p_{iL}) &= \mu_i - \delta_{i2}/L - \delta_{i3}/L - \dots + (L-1)\delta_{iL}/L
 \end{aligned} \tag{4.1}$$

$$(\delta_{i2}, \dots, \delta_{iL})^T \sim N(d, \Sigma)$$

Priori-Verteilungen für μ_i , d und Σ

Dabei ist $\text{logit}(p) = \log \{p/(1 - p)\}$ die Logit-Funktion und

$$\mu_i = \frac{1}{L} \sum_{l=1}^L \text{logit}(p_{il}) \tag{4.2}$$

der Mittelwert der Logit-Event-Raten in Studie i . Durch

$$\delta_{il} = \text{logit}(p_{il}) - \text{logit}(p_{i1}), \quad l = 2, \dots, L \tag{4.3}$$

wurde Behandlung 1 als Baseline gewählt. Abb. 4.1 veranschaulicht die hierarchische Struktur des Modells.

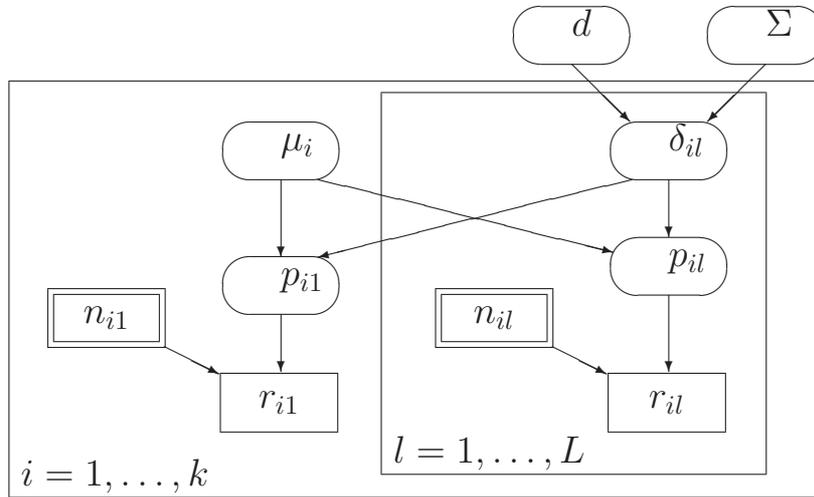


Abbildung 4.1: Hierarchische Struktur des Modells

Die Parameter d und δ_{il} sind von primärem Interesse und die μ_i sind Nuisance-Parameter. Die Priori-Verteilungen für μ_i , d und Σ sollten nicht-informativ sein. Ihre Form kann dabei sehr unterschiedlich sein und ist abhängig von der jeweiligen Anwendung. Z.B. kann für die Kovarianzen in Σ Homogenität oder Heterogenität angenommen werden, die Kovarianzmatrizen könnten aber auch für jede Studie unterschiedlich sein.

Die Annahmen des Modells, die hierarchische Struktur des Datenerzeugungsprozesses und die multivariate Normalverteilung der Logit-Event-Raten können durch sogenannte knotenbasierte Methoden überprüft werden, bei denen die Annahmen auf jeder Stufe des hierarchischen Modells anhand der vorliegenden Daten getrennt überprüft werden.

Die Posteriori-Verteilungen der Parameter können nicht analytisch bestimmt werden, sondern müssen mit Markov-Chain-Monte-Carlo- (MCMC-) Simulation approximiert werden, wozu die Software WinBUGS zur Verfügung steht. Abschließend sei noch erwähnt, dass die Analyse auch zur Planung des Designs von Studien, die Behandlungen vergleichen, die bisher noch nicht direkt miteinander verglichen wurden, herangezogen werden kann, da mit dem Modell eine prädiktive Verteilung für den Odds Ratio zwischen diesen beiden Behandlungen bestimmt werden kann (Lu & Ades (2004)).

4.2 Individual Patient Data (IPD) Metaanalysen:

IPD-Metaanalysen, bei welchen die Einzeldaten zur Verfügung stehen, gelten als Gold Standard der Metaanalysen.

Obwohl sie in letzter Zeit immer häufiger durchgeführt werden, sind sie noch weitaus seltener als Metaanalysen, die auf aggregierten Daten beruhen (Sutton & Higgins (2008)). Das hat verschiedene Gründe:

Sie sind sehr kosten- und zeitintensiv und nur mit einer Vielzahl von Mitarbeitern möglich (Stewart & Clarke (1995)). Daher muss vor der Durchführung geklärt werden, ob die Vorteile durch die Einzeldaten wirklich den erhöhten Aufwand aufwiegen (Sutton & Higgins (2008)).

Außerdem erfordern sie die enge und anhaltende Zusammenarbeit der Betreiber der Einzelstudien. Dadurch ergeben sich zahlreiche Vorteile, z.B. können durch den Austausch mit den Experten möglicherweise weitere, unveröffentlichte Studien lokalisiert werden.

IPD Metaanalysen eröffnen gegenüber Metaanalysen mit aggregierten Daten zahlreiche Möglichkeiten. So können Subgruppen untersucht werden, die sich nicht nur von Studie zu Studie unterscheiden, sondern auch in den Einzeldaten variieren. Außerdem kann die Qualität der Studien besser überprüft werden, weil sich nicht nur auf die Berichte verlassen werden muss (Stewart & Clarke (1995)). Die Daten können noch einmal neu reanalysiert werden. Dabei ist es möglich, neue, in allen Datensätzen einheitliche Variablen zu definieren und auch Regressionsmodelle über alle Daten zu rechnen. So kann mögliche Heterogenität der Studien effizienter untersucht und die Ergebnisse hinsichtlich Konfundierungsvariablen adjustiert werden. Fehler in den durchgeführten Analysen können korrigiert werden. Wenn allerdings bei der neuen Analyse Fehler gemacht werden, betreffen diese dann alle Studien (Sauerbrei & Blettner (2003)).

Dazu können Time-to-event- und Überlebenszeitanalysen durchgeführt werden, was mit aggregierten Daten nur eingeschränkt möglich ist.

Die Metaanalyse kann aktualisiert werden, wenn Follow-up-Informationen zu den Einzeldaten erhoben werden (Stewart & Clarke (1995)).

Zur Überprüfung einzelner Fragestellungen sind nicht immer alle Studien notwendig, sondern es kann sich z.B. auf neue, große und gute Studien beschränkt werden, wodurch die Verzerrung verringert wird (Sauerbrei & Blettner (2003)).

Wenn nicht zu allen Studien Einzeldaten beschafft werden können, ist es möglich, auch die aggregierten Informationen mit in die Analyse aufzunehmen.

Trotz der zahlreichen Vorteile existiert aber noch vergleichsweise wenig Literatur zu statistischen Methoden für IPD-Metaanalysen (Sutton & Higgins (2008)).

Kapitel 5

Anwendungsbeispiel

5.1 Einleitung

Acamprosate und Naltrexone sind Medikamente, die den Konsum von Alkohol reduzieren sollen. Sie wurden bereits in mehreren Metaanalysen erforscht und die Ergebnisse lassen vermuten, dass die Substanzen generell wirksam sind. Allerdings wurde auch Heterogenität in den Ergebnissen beobachtet. Durch diese könnten die Ergebnisse verzerrt sein, da Größen vernachlässigt werden, die den geschätzten oder den wahren Effekt beeinflussen (Rösner (2006)).

Daher werden nun in diesem Kapitel die Ergebnisse der Studien zu verschiedenen ausgewählten Variablen mittels Meta-Regression in Beziehung gesetzt. Die Einflussgrößen sind dabei zum größten Teil studienspezifisch, es werden jedoch auch auf Studienebene zusammengefasste Patientenmerkmale verwendet.

In der Meta-Regression werden die Kovariablen mit den geschätzten Effektgrößen verschiedener Konstrukte in Verbindung gesetzt. Dadurch ergeben sich eine Vielzahl von Regressionsmodellen. Die Analyse ist dabei allerdings hypothesengenerierend und nicht hypothesenbestätigend. Aus diesem Grund wurden die Signifikanzniveaus der Modelle auch nicht angesichts der multiplen Vergleiche adjustiert. Die Ergebnisse sind also als Trends aufzufassen, die in Zukunft noch weiter untersucht werden könnten.

5.2 Effektgrößen

Es wurden sechs verschiedene Outcomes betrachtet: Alkoholkonsum (**A**ny **D**rinking), Rückfall zu übermäßigem Alkoholkonsum (**H**eavy **D**rinking), An-

teil der Trinktage an der Gesamtzeit der Studie (**Drinking Days**), Anteil der abstinenten Tage an der Gesamtzeit der Studie (**Days Abstinent**), das Leberenzym Gamma GT, als relativ spezifischer Indikator für eine alkoholtoxische Leberschädigung (**GGT**) und Alkohol pro Trinktag in Gramm (**Grams**). AD, HD und GGT wurden als Zielvariablen bei beiden Substanzen verwendet, DrDays und Grams nur bei Naltrexon (**NAL**) und DaysAbs nur bei Acamprosat (**ACA**). Tabelle 5.1 zeigt die Anzahlen der Studien, getrennt nach Substanz, die zu den einzelnen Konstrukten zur Verfügung standen.

Tabelle 5.1: Zahl der Studien zu den Outcomes

	AD	HD	DrDays	DaysAbs	GGT	Grams
Acamprosat	23	6	2	19	7	1
Naltrexon	26	28	25	2	15	15

5.2.1 Gewinnung der Effektgrößen und der Standardfehler

Die Information zu den binären Outcomes bei AD und HD lagen in Form von Anzahlen der Events in der Behandlungs- und Kontrollgruppe und den Größen der beiden Gruppen vor:

Tabelle 5.2: Kreuztabelle der binären Outcomes

	Event	kein Event	
Präparat	r_i^T	$n_i^T - r_i^T$	n_i^T
Placebo	r_i^C	$n_i^C - r_i^C$	n_i^C

Aus diesen Tabellen wurden die logarithmierten relativen Risiken berechnet:

$$\hat{\theta}_{RR_i} = \log \left[(r_i^T/n_i^T) / (r_i^C/n_i^C) \right] \quad (5.1)$$

Der Standard-Schätzer für die Varianz des logarithmierten relativen Risikos wäre $(n_i^T - r_i^T)/r_i^T n_i^T + (n_i^C - r_i^C)/r_i^C n_i^C$. Diese Schätzung ist allerdings mit der Schätzung der Effektgrößen korreliert. In einer Simulation wurde gezeigt, dass dadurch in der Metaregression die Schätzungen der Regressionskoeffizienten und der Zwischenstudienvarianzen leicht zur Null hin verzerrt werden.

Aus diesem Grund wurde ein geglätteter Schätzer verwendet, durch den die Korrelation verringert wird:

$$V(\hat{\theta}_{RR_i}) = \left[\sum_{j=1}^k \frac{n_j^T - r_j^T}{r_j^T} \right] / kn_i^T + \left[\sum_{j=1}^k \frac{n_j^C - r_j^C}{r_j^C} \right] / kn_i^C \quad (5.2)$$

Hier geht jeweils nur der Mittelwert der $(n_i^T - r_i^T)/n_i^T$ und der $(n_i^C - r_i^C)/n_i^C$ ein. Es wird also davon ausgegangen, dass die wahren Varianzen nur von den Fallzahlen der Studien abhängen (Berkey et al. (1995)).

Zu den stetigen Outcomes lagen für die beiden Gruppen Mittelwerte der Werte pro Studie und die jeweiligen Standardabweichungen s_i^T bzw. s_i^C vor. Daraus wurden dann die Mittelwertsdifferenzen gebildet:

$$\hat{\theta}_{diff_i} = \bar{x}_i^T - \bar{x}_i^C \quad (5.3)$$

Die Varianzschätzung hierzu ist (Lipsey & Wilson (2001)):

$$\begin{aligned} V(\hat{\theta}_{diff_i}) &= \frac{1}{n_i^T} S_{pooled}^2 + \frac{1}{n_i^C} S_{pooled}^2 \\ &= \left(\frac{1}{n_i^T} + \frac{1}{n_i^C} \right) \left[\frac{(n_i^T - 1)s_i^{T^2} + (n_i^C - 1)s_i^{C^2}}{(n_i^T - 1) + (n_i^C - 1)} \right] \end{aligned} \quad (5.4)$$

5.2.2 Untersuchung der Heterogenität

In Abbildung B.2 des Anhangs finden sich Forest Plots zu den geschätzten Effektgrößen aus den einzelnen Outcomes und in Abbildung 5.1 die I^2 -Statistiken mit Konfidenzintervallen. Die Gewinnung letzterer wird im Anhang unter C beschrieben.

Die geschätzten Effektgrößen bei NAL.AD und ACA.HD sind offenbar homogen. Allerdings liegen die oberen Grenzen der Konfidenzintervalle im Bereich schwacher bzw. mittlerer Heterogenität.

ACA.AD, NAL.HD, ACA.GG, NAL.GGT und NAL.Grams scheinen sich im Bereich mittlerer Heterogenität zu befinden. Bei Betrachtung der Forest Plots fällt jedoch für NAL.HD auf, dass die Population der geschätzten Effektgrößen bis auf eine abweichende relativ homogen erscheinen. In der Tat verringert sich der Wert der I^2 -Statistik bei Weglassen dieser Studie von 0,618 auf 0,353, was schwache Heterogenität bedeuten würde.

NAL.DrDays und ACA.DaysAbs sind mit Werten nahe bei eins und sehr kleinen Konfidenzintervallen offensichtlich stark heterogen, was bei Betrachten

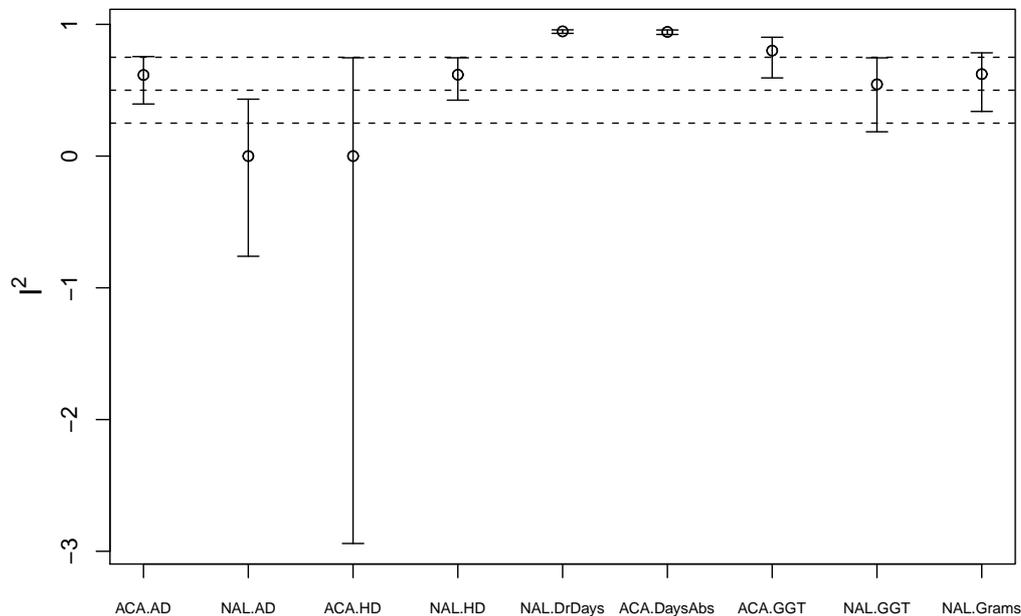


Abbildung 5.1: I^2 -Statistiken für die verschiedenen geschätzten Effektgrößen

der Forest Plots nicht unbedingt aufgefallen wäre. Diese Werte sind allerdings stark von dem Einschluss einzelner Studien abhängig. Abbildung 5.2 veranschaulicht dies. Hier wurden die I^2 -Statistiken berechnet, wenn jeweils eine Studie weggelassen wird. Die gestrichelten Linien entsprechen den I^2 -Werten, wenn keine Studie weggelassen wird.

χ^2 -Homogenitätstests waren bis auf NAL.AD und ACA.HD für alle Outcomes signifikant.

In Abb. B.1 im Anhang wurden alle in Abschnitt 2.7.4 vorgestellten Schätzer der Zwischenstudienvarianz auf die geschätzten Effektgrößen aus den verschiedenen Konstrukten angewendet. Hier wird deutlich, wie unterschiedliche Ergebnisse die verschiedenen Schätzer liefern, insbesondere bei stärkerer Heterogenität.

5.3 Struktur der Kovariablen

Es wurden zunächst zehn Kovariablen in der Regression betrachtet: Publikationsjahr, Teilnehmerzahl, Behandlungsdauer (in Wochen), Anzahl der teilnehmenden Kliniken, durchschnittliche Differenz zwischen den Drop-out-

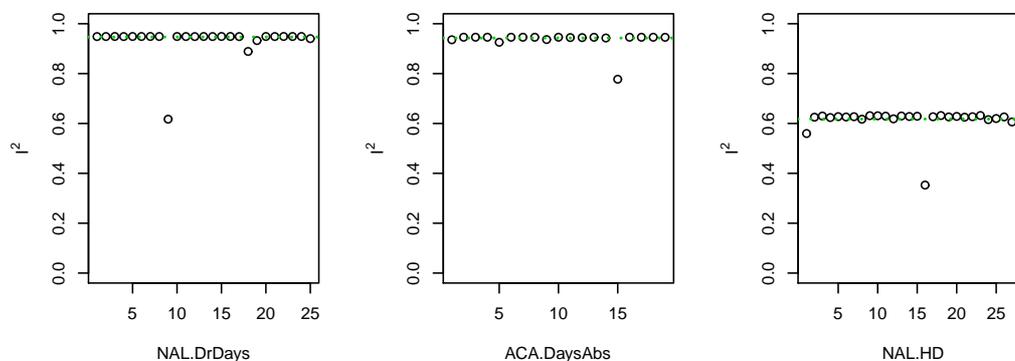


Abbildung 5.2: I^2 -Statistiken bei Weglassen der i -ten Studie, $i = 1, \dots, k$

Raten der Behandlungs- und der Kontrollgruppe (Dropout-Differenz), durchschnittliche Differenz der Raten der Dropouts aufgrund von Komplikationen zwischen der Behandlungs- und der Kontrollgruppe (Adverse-Events-Differenz), das Herkunftsland (kategorisiert in “Europa”, “USA” und “Andere”), Sponsor (von Industrie finanziert oder nicht), Art der Psychotherapie (“CBT” – kognitive Verhaltenstherapie, “MET” – Motivationale Therapie und “Andere”) und Ziel der Behandlung (“Abstinenz” und “Andere”).

Während die Werte der restlichen Kovariablen innerhalb der Studien für alle Patienten gleich sind, wurden “Drop-out-Differenz” und “Adverse-Events-Differenz” durch Zusammenfassung von einzelnen Patientendaten gebildet. Das könnte zu einer Verzerrung führen, wie weiter unten diskutiert wird.

Später wurde noch der Placebo-Response betrachtet, der allerdings aufgrund der strukturellen Abhängigkeit von den geschätzten Effektgrößen eine gesonderte Behandlung benötigte.

In Abb. B.3 des Anhangs sind die Verteilungen der vier quasi-stetigen Variablen getrennt für die beiden Substanzen visualisiert. Während “Publikationsjahr” und “Teilnehmerzahl” ausreichend über das beobachtete Spektrum verteilt sind, finden sich bei “Behandlungsdauer” und “Anzahl Kliniken”, insbesondere für Naltrexon, starke Anhäufungen bei einzelnen Werten. Naltrexon wurde beispielsweise in ca. 69 % der Fälle genau zwölf Wochen lang verabreicht und in ca. 71 % wurde die Studie nur an einer Klinik durchgeführt.

Abb. B.4 zeigt die Verteilungen der stetigen Variablen. Die Werte streuen in jedem Fall recht gleichmäßig über ihr Spektrum.

Wie man an Abb. B.5 sieht, sind die nominalen Variablen in beinahe al-

len Fällen ungleichmäßig verteilt. In den Acamprosat-Studien wurde z.B. in keinem Fall motivationale Therapie (MET) angewendet.

Allgemein ist, wie schon erwähnt, die Power geringer, wenn die Kovariablen schwach über ihr Spektrum streuen.

Die nominalen Merkmale wurden mittels Dummy-Kodierung in die Modelle aufgenommen.

Die Korrelationen nach Bravais-Pearson zwischen den (quasi-)stetigen Kovariablen waren, bis auf einen Wert von ca. 0,81 zwischen “Anzahl Kliniken” und “Teilnehmerzahl”, gering. Allerdings konnte von diesen beiden keine vernachlässigt werden, da sie aus theoretischer Sicht unterschiedlich gerechtfertigt sind.

5.4 Methodik zur Schätzung der Modelle

Es wurde eine lineare Metaregression mit zufälligen Effekten verwendet, auch um die Möglichkeit zu haben, zu messen, inwieweit die Heterogenität in jedem Fall von den Kovariablen erklärt wurde.

Zur Schätzung der Zwischenstudienvarianz wurde eine Abwandlung des iterativen approximativen REML-Schätzers zur Einbeziehung von Kovariablen verwendet. Dabei wird zwischen der Schätzung der Regressionskoeffizienten und der Kovariablen iteriert. Die Schätzung der Regressionskoeffizienten geschieht dann mit der Methode der gewichteten kleinsten Quadrate, wobei die Gewichte die Inversen der Varianzen in den Studien plus der Schätzung der Zwischenstudienvarianz des REML-Schätzers aus dem vorhergehenden Iterationsschritt sind, siehe Gleichung 5.5. Begonnen wird mit der gewichteten KQ-Schätzung, wobei $\hat{\tau}_1^2 = 0$.

l -ter Iterationsschritt:

1) Schätzung der Regressionskoeffizienten:

$$\hat{\beta}_l = (\mathbf{X}^T \hat{\mathbf{W}}_l^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{W}}_l^{-1} \hat{\theta} \quad (5.5)$$

$$\text{mit } \hat{\mathbf{W}}_l = \left(\text{diag} \left(\frac{1}{\hat{\tau}_l^2 + \sigma_i^2} \right) \right)_{i=1, \dots, k},$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1(p-1)} \\ 1 & x_{21} & \dots & \dots \\ \vdots & \vdots & \ddots & x_{k(p-1)} \end{pmatrix} \quad \text{und} \quad \hat{\theta} = \begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_k \end{pmatrix}$$

2) Schätzung der Zwischenstudienvarianz:

$$\hat{\tau}_{l+1}^2 = \frac{\sum_{i=1}^k \hat{w}_i^2 \{ [k/(k-p)] (\hat{\theta}_i - \mathbf{x}_i \hat{\beta}_l)^2 - \sigma_i^2 \}}{\sum_{i=1}^k \hat{w}_i^2} \quad (5.6)$$

$$\text{mit } \mathbf{x}_i = (1, x_{i1}, \dots, x_{i(p-1)}) \quad \text{und} \quad \hat{w}_i = \frac{1}{\hat{\tau}_l^2 + \sigma_i^2}$$

Wenn die die Zwischenvarianz in einem Schritt negativ geschätzt wird, so wird sie auf Null gesetzt.

Bei der Analyse wurde die Iteration abgebrochen, sobald $|\hat{\tau}_{l+1}^2 - \hat{\tau}_l^2| < 10^{-7}$ war.

Diese Methode hat sich als für Metaregression am effektivsten erwiesen (Thompson & Sharp (1999)).

5.5 Bemerkungen zur Interpretation

Die “erklärte Heterogenität” ist der Anteil der Zwischenstudienvarianz, der durch die Kovariablen erklärt wird:

$$\text{erkl. Het.} = \frac{\hat{\tau}_{Nullmodell}^2 - \hat{\tau}_{y \sim x}^2}{\hat{\tau}_{Nullmodell}^2} \quad (5.7)$$

Bei negativen Werten wurde die erklärte Heterogenität auf Null gesetzt. Um $\hat{\tau}_{Nullmodell}^2$ und $\hat{\tau}_{y \sim x}^2$ vergleichen zu können wurde die Zwischenstudienvarianz

des Nullmodells ohne Kovariablen auch mit der entsprechenden approximativen REML-Methode geschätzt, obwohl, wie in Abschnitt 2.7.4 diskutiert, für diesen Fall andere Schätzer häufig besser geeignet sind. Sie wurde auf Basis aller Studien des jeweiligen, gerade betrachteten Konstrukts berechnet, unabhängig davon, ob in den Kovariablen Werte fehlten. Wenn der Schätzwert der Zwischenstudienvarianz nach der Meta-Regression Null war, so war der erklärte Heterogenität eins. Natürlich sind diese Schätzwerte dadurch, dass hier zwei ungenaue Schätzer eingehen, nicht verlässlich und sollen daher nur als grobe Orientierung dienen.

Bei den kategorialen Kovariablen ist $\exp(\hat{\beta}_0)$ das geschätzte relative Risiko der Referenzgruppe, welche bei Herkunftsland z.B. Studien aus anderen Ländern außerhalb Europas und den USA beinhaltet. $\exp(\hat{\beta}_0 + \hat{\beta}_1)$ ist das geschätzte relative Risiko der ersten Dummyvariable entsprechenden Gruppe und, wenn vorhanden, $\exp(\hat{\beta}_0 + \hat{\beta}_2)$ das der zweiten Dummyvariable zugehörigen Gruppe. Bei Herkunftsland würden diesen die Studien aus Europa bzw. den USA entsprechen.

Betrachtet man stetige Variablen, ist allgemeiner $\exp(\hat{\beta}_0 + \hat{\beta}_1 \tilde{x})$ das geschätzte relative Risiko in Studien, bei denen die Kovariable den Wert \tilde{x} hat.

In multiplen Modellen ist $\exp(\hat{\beta}_0 + \sum_{i=1}^s \hat{\beta}_i \tilde{x}_i)$ das geschätzte relative Risiko, wenn die Kovariablen 1 bis s die Werte \tilde{x}_1 bis \tilde{x}_s haben.

Es muss auch beachtet werden, dass, mit Ausnahme von DaysAbs, also dem Anteil an trinkfreien Tagen an der Gesamtdauer der Studie, bei allen Konstrukten negative Werte der Effektgrößen einem positivem Behandlungserfolg entsprechen und umgekehrt, da die zugrundeliegenden Outcomes Schaden messen und nicht Erfolg. Daher weisen auch negative Einflüsse auf diese Effektgrößen auf einen positiven Einfluss auf den Behandlungserfolg hin.

5.6 Einfachregressionen getrennt nach Substanz

Zuerst wurden alle möglichen Modelle mit einer Kovariable bestimmt. Die Tabellen A.1 bis A.9 des Anhangs zeigen die Ergebnisse. Für die Modelle mit (quasi)stetigen Kovariablen wurden die Relationen durch Streudiagramme grafisch dargestellt (Abb. B.6 bis B.14). Die Angemessenheit der Normalverteilungsannahme der Residuen wurde mittels Normal-Quantil-Plots überprüft (Abb. B.15 bis B.23). Diese lassen vermuten, dass die Residuen zumindest approximativ normalverteilt sind.

In seltenen Fällen konnten die Schätzer für die Koeffizienten aufgrund

von Singularitäten nicht bestimmt werden. Insbesondere wurde, wie oben erwähnt, “MET” bei “Psychotherapie” kein einziges Mal angewendet.

Für NAL.AD und ACA.AD konnte die erklärte Heterogenität nicht geschätzt werden, da die geschätzte Zwischenstudienvarianz schon vor der Meta-Regression Null war. Da dies auch nach der Regression der Fall war, ergibt sich hier ein Modell mit festen Effekten.

Bei ACA.AD hatten “Publikationsjahr” und “Teilnehmerzahl” einen signifikant positiven Einfluss ($p < 0,05$). Das heißt neuere und größere Studien zeigten durchschnittlich einen geringeren Behandlungserfolg. Auch der Einfluss von “Ziel der Behandlung” war signifikant. Studien mit dem Behandlungsziel Abstinenz hatten durchschnittlich ein geringeres relatives Risiko. Schwach signifikant ($0,05 \leq p < 0,1$) positiv wirkten “Anzahl Kliniken” und “Drop-out-Differenz”. “Teilnehmerzahl” konnte am meisten Heterogenität erklären (ca. 85%).

Bei NAL.AD hatte nur “Drop-out-Differenz” einen signifikant positiven Einfluss. Allerdings war für dieses Konstrukt keine Heterogenität in den geschätzten Effektgrößen festgestellt worden, siehe Abschnitt 5.2.2.

Für ACA.HD trug keine Kovariable zur Aufklärung einer möglichen Heterogenität in signifikanter Weise bei. Allerdings standen hier zur Regression nur sechs Studien zur Verfügung.

Im Fall von NAL.HD war der Einfluss von “Herkunftsland” signifikant. In Studien aus Europa wurden durchschnittlich die geringsten Behandlungserfolge festgestellt, darauf folgen die USA und schließlich andere Länder. Allerdings konnte damit nur schätzungsweise 48 % der Heterogenität aufgeklärt werden. “Drop-out-Differenz” hatte einen schwach signifikant, positiven Einfluss. “Sponsor” scheint 100 % der Heterogenität zu erklären. Die von der Industrie gesponserten Studien scheinen einen geringeren Behandlungserfolg aufzuweisen, was inhaltlich nicht sinnvoll erscheint. Der Effekt ist auch nicht signifikant ($p \approx 0,30$).

In Abschnitt 5.2.2 wurde bemerkt, dass das Ergebnis einer Studie hier sehr von den anderen abweicht. Um die Abhängigkeit der Schlussfolgerungen von dem Einschluss dieser Studie zu untersuchen, wurde die Modelle ohne diese Studie noch einmal gerechnet. Der Einfluss von “Herkunftsland” blieb signifikant, “Drop-out-Differenz” wurde stark signifikant. Zusätzlich wurde “Publikationsjahr” stark signifikant und “Anzahl Kliniken” schwach.

Bei NAL.DrDays war nur “Drop-out-Differenz” signifikant ($p < 0,00005$). Es hatte einen stark positiven Einfluss. Bei Betrachten des zugehörigen Streudiagramms (Abb. B.10) fällt allerdings auf, dass zwei Studien die Regression stark dominieren.

Der Grund dafür soll im Folgenden erklärt werden.

Diese beiden Studien haben im Vergleich zu den anderen sehr kleine Varianzen. Die Werte von “Drop-out-Differenz” liegen dazu im Vergleich zu denen der anderen Kovariablen weit auseinander. Dadurch entsteht ein starker Hebeleffekt, durch den der Verlauf der Regressionsgerade größtenteils bestimmt wird. Wie in Abbildung B.10 zu erkennen ist, verläuft sie ziemlich genau durch die den Effektgrößen der beiden Studien zugehörigen Punkte. Also tragen diese, die vorher einen großen Teil der Heterogenität ausgemacht haben (Abb. 5.1), kaum mehr dazu bei und die Zwischenstudienvarianz wird Null geschätzt. Das führt gleichzeitig dazu, dass die Studien unterschiedlicher gewichtet werden, also insbesondere die beiden dominierenden Studien ein noch größeres Gewicht bekommen.

Um sich nicht nur auf die Ergebnisse zweier Studien verlassen zu müssen, wurden diese vorübergehend herausgenommen und die Modelle noch einmal gerechnet. Der Einfluss von “Drop-out-Differenz” war zwar immer noch positiv, aber nicht mehr signifikant ($p \approx 0,24$). Die erklärte Heterogenität war nun Null Prozent. Das Ergebnis war also nicht robust auf Entfernen der Studien. Die Schlussfolgerungen zu den anderen Kovariablen änderten sich nicht. Die Studien können aber nicht vernachlässigt werden, weil es aus qualitativer Sicht keinen Grund dazu gibt.

Bei ACA.DaysAbs hatte “Teilnehmerzahl” einen signifikant negativen Einfluss auf den Behandlungserfolg und konnte mit 45 % am meisten Heterogenität erklären. War das Ziel der Behandlung Abstinenz, so war der Behandlungserfolg durchschnittlich größer ($p \approx 0,02$), als für andere Zielkriterien. “Anzahl Kliniken” hatten einen schwach signifikant negativen Einfluss.

Für ACA.GGT standen nur sieben Studien zur Verfügung. Es war auch keine Relation signifikant. Die erklärte Heterogenität war in den meisten Fällen Null Prozent. Am größten war sie mit 59 % für Sponsor, obwohl diese Kovariable inhaltlich zu Gamma GT, einem nicht manipulierbaren Laborwert, in keinem Bezug steht. Das ist ein Beispiel dafür, dass die Schätzung der erklärten Heterogenität uns also auch der Zwischenstudienvarianz bei geringer Studienzahl sehr ungenau ist.

Auf NAL.GGT hatte nur “Publikationsjahr” einen signifikant positiven Einfluss, wobei die erklärte Heterogenität 100 % war.

“Drop-out-Differenz” und “Anzahl Kliniken” beeinflussten Nal.Grams signifikant bzw. schwach signifikant positiv. Bei motivationaler Therapie und kognitiver Verhaltenstherapie wurden zudem durchschnittlich bessere Behandlungserfolge erzielt als bei anderen Psychotherapien. Diese Relation war auch signifikant.

5.7 Regressionen über beide Substanzen

Nun wurden die Studien über Acamprosat und Naltrexon zusammen analysiert, wodurch über die Vergrößerung der Fallzahl die Genauigkeit der Schätzungen erhöht werden konnte. Die Differenzierung zwischen den beiden Medikamenten wurde durch Aufnahme von “Substanz” als Kovariable in die Modelle geleistet. Dadurch kann der Intercept zwischen Acamprosat und Naltrexon variieren, nicht aber die Steigung der Regressionsgerade. Das entspricht der Annahme, dass die Substanzen in den einzelnen Konstrukten unterschiedlich stark wirken, die Art des Einflusses der Kovariablen aber derselbe ist. Als Referenzkategorie wurde “Acamprosat” gewählt. Die Ergebnisse finden sich in den Tabellen A.1, A.2 und A.3.

Für AD standen nun 49 Studien zur Verfügung. Einen signifikant positiven Einfluss hatten “Publikationsjahr”, “Teilnehmerzahl” und “Drop-out-Differenz”. Schwach signifikant positiv wirkte “Anzahl Kliniken”. War das Ziel der Behandlung Abstinenz, so war das geschätzte relative Risiko durchschnittlich geringer ($p \approx 0,02$). Mit “Drop-out-Differenz” wurde am meisten Heterogenität aufgeklärt (ca. 93 %). Für Studien, die Naltrexon untersuchten, war das durchschnittliche relative Risiko grundsätzlich größer.

Bei HD waren 34 Studien vorhanden. Hier wirkte “Drop-out-Differenz” auch signifikant positiv auf das geschätzte relative Risiko ein. Signifikant war auch der Einfluss von “Herkunftsland”. Das war auch schon bei NAL.HD der Fall. Allerdings hat sich die Ordnung zwischen den geschätzten relativen Risiken für die Länder verändert. Studien aus Europa wiesen immer noch den geringsten Behandlungserfolg auf, als zweites kamen nun allerdings andere Länder und an dritter Stelle die USA. Die Rangordnung zwischen “USA” und “andere Länder” scheint also weniger offensichtlich zu sein. Allerdings deutete eine grafische Veranschaulichung der Effektgrößen getrennt nach “Substanz” und “Herkunftsland” darauf hin, dass die Annahme der Unabhängigkeit des Einflusses von “Substanz” und der Kovariable hier nicht gegeben sein könnte, sodass m.E. dem Modell mit “Herkunftsland” als Kovariable in NAL.HD mehr Vertrauen zu schenken ist, als diesem über beide Substanzen. Darüber hinaus ist nach Entfernen der Naltrexon-Studie mit dem extremen Ergebnis hinsichtlich HD der Einfluss von “USA” nicht mehr signifikant. “Publikationsjahr” wird dann hingegen stark signifikant und “Anzahl Kliniken” schwach. Zusätzlich würden Studien, in denen CBT-Therapie angewendet wird, durchschnittlich ein geringeres relatives Risiko aufweisen ($p \approx 0,05$).

Zu dem Konstrukt GGT gab es 22 Studien. Hier erwies sich keine Relation als signifikant. Auch der starke Einfluss von “Publikationsjahr” in NAL.GGT geht hier verloren.

5.8 Backward- und Forward-Selection

Um auch multiple Modelle betrachten zu können, aber nicht allzu viele Vergleiche durchführen zu müssen, wurde für die einzelnen Outcomes jeweils eine schrittweise Back- und Forward-Selection durchgeführt.

Bei der schrittweisen Backward-Selection wird ausgehend vom vollen Modell, das alle Kovariablen enthält, in jedem Schritt diejenige Kovariable entfernt, die zu dem günstigsten Wert eines bestimmten Informationskriteriums für die Modellgüte führt. Üblicherweise wird die Prozedur abgebrochen, wenn sich durch das Entfernen jeder Kovariable der Wert des Informationskriteriums wieder verschlechtern würde. Da im vorliegenden Fall durch die Unsicherheiten in den Schätzungen der höherdimensionalen Modelle die ersten Schritte der Selektion noch sehr unzuverlässlich sind und es daher vorkommen kann, dass sich anfangs der Wert des Informationskriteriums noch verschlechtert, sich später aber bedeutend verbessert, wurde die Selektion durchgeführt, bis nur noch eine Kovariable im Modell war. Dann wurde das Modell aus dem Schritt gewählt, in dem der Wert des Informationskriteriums am günstigsten war.

Bei der schrittweisen Forward-Selection wird analog in jedem Schritt aus den verbliebenen Kovariablen diejenige aufgenommen, die zu dem besten Wert des Informationskriteriums führt. Begonnen wird entsprechend mit dem Nullmodell. Auch hier wurde die Prozedur nicht unterbrochen, sondern so lange fortgesetzt, bis alle Kovariablen aufgenommen waren und anschließend das Modell aus dem Schritt gewählt, in dem der günstigste Wert des Informationskriteriums erzielt wurde.

Als Informationskriterium wurde das leave-one-out-Kreuzvalidierungskriterium gewählt. Bei der leave-one-out-Kreuzvalidierung wird das betrachtete Modell unter Auslassung je einer Beobachtung solange geschätzt, bis jede Beobachtung einmal ausgespart wurde. Nun werden anhand der Regressionsgeraden die ausgesparten Beobachtungen prognostiziert und die quadratischen Abweichungen dieser prognostizierten Werte von den wahren Beobachtungen gebildet. Da hier die Regression gewichtet erfolgte, mussten die quadratischen Abweichungen jeweils durch Teilen durch die ohne die ausgesparte Beobachtung geschätzte Zwischenstudienvarianz plus die Varianz innerhalb der Studie standardisiert werden. Dann wurden die standardisierten Abweichungen gemittelt.

Wenn $(\hat{\theta}_{iM})^{-i}$ die Prognose für die i -te Beobachtung bezeichnet und $(\hat{\tau}_{iM}^2)^{-i}$ die geschätzte Zwischenstudienvarianz nach Anwendung der Regression ohne Beobachtung i , lässt sich das Kreuzvalidierungskriterium wie folgt formulie-

ren:

$$CV = \frac{1}{k} \sum_{i=1}^k \frac{[\hat{\theta}_i - (\hat{\theta}_{iM})^{-i}]^2}{(\hat{\tau}_{iM})^{-i} + \sigma_i^2}$$

Da im Vergleich zur Anzahl der Studien viele Kovariablen zur Verfügung standen, konnten die vollen Modelle mit allen Kovariablen aufgrund von Singularitäten durch zu viele fehlende Werte meist nicht geschätzt werden. Daher mussten manche Kovariablen weggelassen werden. Dabei wurde abgewogen zwischen der Wichtigkeit aus theoretischer Sicht und dem Gewinn an Beobachtungen durch Herausnehmen von Kovariablen mit vielen fehlenden Werten. Bestanden nach diesem Vorgehen immer noch Singularitäten, so wurden die entsprechenden Kovariablen auch entfernt.

Die Tabellen A.4 bis A.7 zeigen die Ergebnisse für die verschiedenen Konstrukte, getrennt nach Substanz. Dabei ist auch vermerkt welche Kovariablen zu Beginn herausgenommen werden mussten, worauf im Text nun nicht mehr eingegangen wird. Für die Konstrukte ACA.HD und ACA.GGT konnten die beiden Prozeduren auch nach Herausnehmen vieler Kovariablen nicht durchgeführt werden, da zu wenig Studien zur Verfügung standen.

Zu ACA.AD lieferten Back- und Forward-Selection dasselbe Modell: Teilnehmerzahl beeinflusste das geschätzte relative Risiko signifikant positiv und es konnten ca. 85% der geschätzten Heterogenität erklärt werden. Bei den Einfachregressionen hatte diese Kovariable am meisten Heterogenität erklärt.

Für NAL.AD wurden durch beide Prozeduren Modelle mit vielen Kovariablen gefunden. Dadurch sind die Schätzungen der Koeffizienten sehr ungenau. In das aus der Backward-Selection resultierende Modell gehen “Publikationsjahr” und “Drop-out-Differenz” mit positiven und “Teilnehmerzahl” mit negativen Koeffizienten ein. Weiter mit positiven Koeffizienten sind enthalten: “Ziel der Behandlung” und “Psychotherapie”. Bei Betrachtung der Tabellen fällt auf, dass bei beiden Modellen nur “Publikationsjahr” und “Drop-out-Differenz” signifikante Koeffizienten haben. Da die gefundenen Modelle durch die Unsicherheiten in den Schätzungen einer Vielzahl von Koeffizienten keine verlässlichen Ergebnisse liefern, wurden versuchsweise alle nicht-signifikanten Koeffizienten entfernt, um ein sparsameres Modell zu erhalten. Das Ergebnis findet sich in Tabelle A.10. Die Werte der verbliebenen Koeffizienten änderten sich wenig und blieben signifikant.

Für NAL.DrDays wurde mit Backward-Selection wieder ein sehr großes Modell gefunden, in das nur die Koeffizienten von “Publikationsjahr” und “Adverse-Events-Differenz” signifikant bzw. schwach signifikant positiv eingingen. Die Forward-Selection erreichte ihren besten Wert sehr schnell, die Koeffizienten der aufgenommenen Kovariablen waren aber nicht signifikant.

Deshalb wurden wieder, wie bei NAL.AD die nicht-signifikanten Kovariablen aus dem Modell aus der Backward-Selection herausgenommen. Tabelle A.11 zeigt das resultierende Modell. Offenbar blieben die Koeffizienten signifikant und ihre Werte änderten sich auch hier nur schwach.

Während bei ACA.DaysAbs die Backward-Selection ein großes Modell fand, wurde in der Forward-Selection nur die Kovariable “Anzahl Kliniken” mit einem negativen Vorzeichen des Koeffizienten aufgenommen. Allerdings betrug die erklärte Heterogenität nur 10%. Wie in Abschnitt 5.6 gesehen, konnte jedoch unter den Einfachregressionen “Teilnehmerzahl” mit 45 % am meisten Heterogenität erklären.

Für NAL.GGT lieferten die beiden Prozeduren ähnliche Modelle. Während bei der Backward-Selection nur “Publikationsjahr” und “Drop-out-Differenz” mit einem positiven Vorzeichen im Modell verblieben, wurde bei der Forward-Selection zusätzlich “Teilnehmerzahl” aufgenommen. Der Koeffizient von letzterem ist allerdings im Gegensatz zu den anderen beiden nicht signifikant.

Zu NAL.Grams lieferten Back- und Forward-Selection schließlich wieder dasselbe Modell. Hier wurden “Anzahl Kliniken” mit positivem Vorzeichen und “Psychotherapie” mit negativen Vorzeichen in beiden Koeffizienten gewählt. “Anzahl Kliniken” war nicht signifikant. Wie in Abschnitt 5.6 bereits festgestellt, sind jedoch die Koeffizienten von “Psychotherapie” auch unter Vernachlässigung dieser Kovariable signifikant.

Nun wurden wieder alle Studien betrachtet und “Substanz” in der Selektion wie eine mit den anderen Kovariable gleichwertige Einflussgröße behandelt. Ihr Verbleiben in den Modellen wurde also nicht erzwungen. Sie war in keinem der im Folgenden diskutierten Modelle.

Im Fall von AD lieferten Back- und Forward-Selection dasselbe Ergebnis. “Publikationsjahr”, “Anzahl Kliniken” und “Drop-out-Differenz” wurden jeweils mit positivem Vorzeichen gewählt. Der Koeffizient von “Anzahl Kliniken” war schwach, die der anderen beiden stark signifikant.

In den Selektionen zu HD ergaben sich zunächst keine zufriedenstellenden Modelle. Backward-Selection lieferte ein sehr großes Modell, in dem die meisten Koeffizienten nicht signifikant waren. Das Modell aus der Forward-Selection hatte zwar weniger Einflussgrößen, die Koeffizienten waren aber auch nur teilweise signifikant. Bei Herausnehmen der nicht-signifikanten Kovariablen aus dem durch Backward-Selection gewonnenen Modell, verloren die übrigen Kovariablen ihre Signifikanz. Nach Entfernen von Teilnehmerzahl aus dem aus der Forward-Selection, ergab sich allerdings mit “Publikationsjahr” und “Herkunftsland” ein Modell, in dem alle Koeffizienten signifikant

sind, siehe Tabelle A.12.

Für GGT ergab sich über beide Prozeduren mit “Drop-out-Differenz” als einzige Kovariable das gleiche Modell. Der Koeffizient war positiv, allerdings nicht signifikant.

5.9 Placebo-Response als Kovariable

Die Untersuchung des Placebo-Response als Kovariable für den Behandlungserfolg ist nicht mit der bisher betrachteten Methode möglich. Das liegt daran, dass der Placebo-Response in die Berechnung des Behandlungserfolgs mit eingeht. Dadurch sind die beiden Größen voneinander abhängig. Genauer gesagt, ist, wenn keine echte Verbindung zwischen dem Behandlungserfolg und dem Placebo-Response besteht, eine negative Steigung der Regressionsgerade zu erwarten. Der Grund hierfür ist Regression zur Mitte. Dieser Begriff beschreibt das Phänomen, dass, wenn zwei verbundene Messungen betrachtet werden, extreme Werte der einen in der Regel mit weniger extremen Werten der anderen Messung einhergehen, diese also näher an ihrem Mittel liegen. Wenn also im vorliegenden Fall besonders kleine bzw. besonders große Werte in der Placebogruppe beobachtet werden, tendieren die Werte in der Behandlungsgruppe weniger stark in diese Richtung, wodurch sich durchschnittlich größere bzw. kleinere Messungen der Behandlungseffekte ergeben, woraufhin nun die erwartete Steigung der Regressionsgerade negativ ist.

Um diese Verzerrung zu überwinden, entwickelten Thompson et al (Thompson et al. (1997); Sharp & Thompson (2000)) ein bayesianisches, hierarchisches Modell.

Im Folgenden wird dieses Modell vorgestellt.

Es wird angenommen, dass die Zahl der Events in der Behandlungsgruppe r_i^T , bzw. in der Kontrollgruppe r_i^C einer Binomialverteilung folgen:

$$r_i^T \sim \text{Bin}(n_i^T, p_i^T) \quad (5.8)$$

$$r_i^C \sim \text{Bin}(n_i^C, p_i^C) \quad (5.9)$$

$\nu_i = \text{logit}(p_i^T)$ und $\mu_i = \text{logit}(p_i^C)$ bezeichnen die logarithmierten Chancen in den Behandlungs- bzw. in der Kontrollgruppen.

Damit ist

$$\delta_i = \nu_i - \mu_i \quad (5.10)$$

der logarithmierte Odds Ratio zwischen den Gruppen. Die Beziehung zwischen den logarithmierten Odds Ratios, also dem Behandlungseffekt und

der logarithmierten Chance in der Kontrollgruppe, einer Transformation des Placebo-Response, wird dann wie folgt angenommen:

$$\delta_i = \delta'_i + \beta(\mu_i - \bar{\mu}) \quad \text{mit} \quad \delta_i \sim N(\delta, \sigma_\delta^2) \quad (5.11)$$

δ'_i sind die hinsichtlich des Placebo-Response adjustierten Behandlungseffekte. Diese können von Studie zu Studie variieren. Es wird also nicht davon ausgegangen, dass die Heterogenität durch den Placebo-Response vollständig erklärt wird.

Es müssen nicht-informative Priori-Verteilungen für die Parameter β , δ , σ_δ^2 und μ_i angenommen werden:

$$\begin{aligned} \delta &\sim N(0, 10) \\ \sigma_\delta^2 &\sim \Gamma^{-1}(0, 001, 0, 001) \\ \beta &\sim N(0, 10000) \\ \mu_i &\sim N(0, 10), \quad i = 1, \dots, k \end{aligned} \quad (5.12)$$

Für den Parameter β wird die Varianz besonders groß gewählt, da dies der vorrangig interessierende Parameter ist. Die Empfindlichkeit der Ergebnisse auf die Wahl der Priori-Verteilungen wird in einer Sensitivitätsanalyse geprüft.

Dieses Modell wurde auf die Outcomes AD und HD getrennt nach Substanz angewendet. Diese Analyse wurde in WinBUGS 1.4.3 durchgeführt, das zum Ziehen aus den marginalen Posteriori-Verteilungen den Gibbs-Sampler verwendet. Dafür wurden zwei verschiedene Sätze von Startwerten verwendet, um die Konvergenz der Iterationen des Gibbs-Samplers gegen die interessierenden Verteilungen zu überprüfen. Nach einer Burn-in-Phase von 5000 Iterationen wurden jeweils 5000 weitere zur Approximation an die Randverteilungen durchgeführt, sodass die Approximation je gerechnetem Modell auf 10000 Zügen basiert. Eine Ausnahme stellt ACA.HD dar, bei dem aufgrund der geringen Fallzahl die Approximation langsam war, sodass hier 90000 Züge verwendet wurden.

Tabelle A.13 zeigt die Ergebnisse. Die geschätzten β -Koeffizienten waren in keinem Fall signifikant. Die Änderungen in den Parametern der Priori-Verteilungen hatten kaum einen Einfluss auf das Ergebnis, insbesondere keinen auf die Schlussfolgerungen. Wenn das Modell unter zusätzlicher Aufnahme von "Substanz" als Kovariable auf alle Studien angewendet wurde, änderten sich die Schlussfolgerungen in den beiden Konstrukten nicht.

Kapitel 6

Zusammenfassung

Im ersten Teil wurde zunächst allgemein der Begriff "Metaanalyse" erläutert. Anschließend wurde auf die wichtigsten Punkte in der Geschichte der Metaanalyse eingegangen. Außerdem wurde die Motivation für die Durchführung beschrieben und Voraussetzungen genannt, die Studien zur Aufnahme erfüllen müssen. Die vier gängigen Typen von Metaanalysen wurden vorgestellt.

Der Schwerpunkt des ersten Teils lag auf dem Ablauf von Metaanalysen und der dabei verwendeten Methodik. Dabei beschränkte sich die Darstellung größtenteils auf Metaanalysen unter Verwendung publizierter, aggregierter Daten.

Vor Beginn der statistischen Analyse muss das Thema klar statuiert und die Ein- und Ausschlusskriterien für Studien festgelegt werden, um Selektionsbias vorzubeugen. Von großer Bedeutung ist auch die erschöpfende Suche nach relevanten Studien, um den Publikationsbias zu verringern. Die strikte Codierung der gewonnenen Information ermöglicht dazu erst eine umfassende und verlässliche Analyse.

In einem größeren Abschnitt wurde der Publikationsbias behandelt, welcher entsteht wenn positive Ergebnisse in der Population der publizierten Studien überrepräsentiert sind. Dadurch wird der Gesamteffekt überschätzt. Es wurden grafische quasi-statische und formale Regressionstests zur Diagnose vorgestellt. Zur Korrektur wurden zwei repräsentative Methoden beschrieben, die allerdings auf unüberprüfbar Annahmen über die genauen Gründe für das Fehlen von Studien beruhen.

Ein zentrales Thema in Metaanalysen ist die Einschätzung der Heterogenität der Studienpopulation. Formale Tests können diese häufig nicht feststellen und es wird daher eine informelle Analyse unter Verwendung von grafischen

Mitteln benötigt. Allerdings wird heute mehr die Frage nach der Stärke als dem Vorhandensein von Heterogenität gestellt. Hierzu wurde ein gängiges Maß vorgestellt.

Darauf folgend wurden die Modelle für die Effektgrößen aus den einzelnen Studien behandelt. Hierbei werden solche mit festen und solche mit zufälligen Effekten unterschieden. In den einen streuen die geschätzten Effektgrößen zufällig um die Gesamteffektgröße, in den anderen schätzen sie unterschiedliche Effekte, die aber einer gemeinsamen Verteilung folgen. Zur Schätzung der Zwischenstudienvarianz, als zentrale Größe, um Heterogenität einzuschätzen, wurden verschiedene Methoden und deren Anwendung diskutiert. Anschließend wurde die übliche Schätzung der Gesamteffektgröße behandelt und diskutiert, warum dabei die Einbeziehung von Qualitätsmerkmalen problematisch ist.

Ein Abschnitt widmete sich der Sensitivitätsanalyse, die in Metaanalysen verschiedene Anwendungen hat.

Zum Abschluss dieses Kapitels wurde Meta-Regression genauer behandelt und auch auf die Problematiken bei der Anwendung dieser Technik zur Aufklärung von Heterogenität hingewiesen.

Das nächste Kapitel ging auf die Vorteile und insbesondere auch Nachteile ein, aufgrund derer Metaanalysen immer wieder Gegenstand von Kritik sind.

Eine neue Entwicklung im Bereich der Metaanalyse stellen Mixed Treatment Comparison-Metaanalysen dar, die den simultanen Vergleich mehrerer Behandlungen auf Grundlage von Meta-Datensätzen ermöglichen.

Schließlich wurden die Vorteile und Schwierigkeiten von Metaanalysen mit einzelnen Daten behandelt.

Im zweiten Teil wurde die beobachtete Heterogenität in Studien zu Acamprosat und Naltrexon, zweier Medikamente zur Alkoholentwöhnung mit Meta-Regression untersucht. Dabei wurde ein Modell mit zufälligen Effekten verwendet.

Zu Beginn wurde in fast allen Konstrukten starke Heterogenität festgestellt, allerdings war diese oft stark von den Ergebnissen einzelner Studien abhängig.

Wichtige Kovariablen in den Einfachregressionen waren häufig: "Publikationsjahr", "Teilnehmerzahl" und "Drop-out-Differenz". Allerdings hing dies auch deutlich von den unterschiedlichen Konstrukten ab. Auch "Herkunftsland", "Ziel der Behandlung" und "Psychotherapie" spielten eine Rolle. Meist waren für ein Konstrukt die Koeffizienten weniger oder auch nur einer Kovariablen signifikant. Bei Herausnehmen extremer Beobachtungen änderten sich die Schlussfolgerungen für manche Kovariablen.

Anschließend wurde die Regression über alle Studien durchgeführt und "Sub-

stanz" als Kovariable in die Modell aufgenommen. Dadurch änderten sich die Schlussfolgerungen über die Kovariablen kaum.

Die durch Backward- und Forward-Selection gefundenen Modelle unterschieden sich oft, was natürlich auch mit der Unsicherheit in dem Selektionsprozess zusammenhängt. Es wurden häufig sehr viele Einflussgrößen, mit großen Standardfehlern in den Schätzungen selektiert. Durch Herausnehmen von nicht-signifikanten Kovariablen, konnten allerdings geeignete Modelle gefunden werden. Auch in den multiplen Modellen spielten die bereits genannten Kovariablen die wichtigste Rolle. Es muss noch darauf hingewiesen werden, dass die Einflüsse von "Drop-out-Differenz" und "Adverse-Events-Differenz" verzerrt sein könnten. Dies sind Merkmale, die durch Aggregation von Patientenmerkmalen auf Studienebene entstanden sind. Der eigentlich interessierende Zusammenhang auf Individualebene könnte sich durch Konfundierungen von dem durch Meta-Regression gemessenen Zusammenhang zwischen den Studien unterscheiden.

Bei der Untersuchung des Placebo-Response als Kovariable für die binären Outcomes, konnte kein Zusammenhang mit dem Behandlungserfolg festgestellt werden.

Anhang A

Tabellen

A.1 Einfachregressionen getrennt nach Substanz

Abbildung A.1: Acamprosat - AD

	$\hat{\beta}_0$	$\hat{\beta}_1$	SD	p	$\hat{\beta}_2$	SD	p	erkl. Het
Publikationsjahr	-28,1619	0,0140	0,0058	0,0252	–	–	–	0,3987
Teilnehmerzahl	-0,3221	0,0005	0,0001	0,0003	–	–	–	0,8537
Behandlungsdauer	-0,1614	-0,0003	0,0021	0,8881	–	–	–	0,0000
Anzahl Kliniken	-0,2540	0,0072	0,0036	0,0576	–	–	–	0,2058
Drop-out-Differenz	-0,1413	0,0051	0,0028	0,0831	–	–	–	0,4014
Adverse-Events-Differenz	-0,1648	0,0062	0,0131	0,6426	–	–	–	0,1083
Herkunftsland	-0,1663	-0,0283	0,0958	0,7704	0,1768	0,1159	0,1423	0,4044
Sponsor	-0,1341	-0,0457	0,0873	0,6069	–	–	–	0,0426
Psychotherapie	-0,1271	-0,0320	0,0786	0,6896	–	–	–	0,0752
Ziel der Behandlung	0,0104	-0,2079	0,0775	0,0152	–	–	–	0,4407

Abbildung A.2: Naltrexon - AD

	$\hat{\beta}_0$	$\hat{\beta}_1$	SD	p	$\hat{\beta}_2$	SD	p	erkl. Het
Publikationsjahr	-31,9600	0,0159	0,0065	0,0227	-	-	-	-
Teilnehmerzahl	-0,0747	0,0000	0,0001	0,6796	-	-	-	-
Behandlungsdauer	-0,1394	0,0053	0,0048	0,2827	-	-	-	-
Anzahl Kliniken	-0,0672	0,0008	0,0025	0,7540	-	-	-	-
Drop-out-Differenz	-0,0509	0,0068	0,0027	0,0195	-	-	-	-
Adverse-Events-Differenz	-0,0341	-0,0160	0,0123	0,2139	-	-	-	-
Herkunftsland	-0,0480	-0,0003	0,0981	0,9974	-0,0168	0,0842	0,8432	-
Sponsor	-0,0738	0,0417	0,0481	0,3948	-	-	-	-
Psychotherapie	-0,0671	-0,0005	0,0666	0,9940	0,0278	0,0564	0,6272	-
Ziel der Behandlung	-0,0412	-0,0325	0,0504	0,5263	-	-	-	-

Abbildung A.3: Acamprosat - HD

	$\hat{\beta}_0$	$\hat{\beta}_1$	SD	p	$\hat{\beta}_2$	SD	p	erkl. Het
Publikationsjahr	28,7626	-0,0144	0,0125	0,3140	-	-	-	-
Teilnehmerzahl	-0,0331	0,0000	0,0002	0,9490	-	-	-	-
Behandlungsdauer	-0,0428	0,0002	0,0052	0,9783	-	-	-	-
Anzahl Kliniken	-0,0505	0,0007	0,0054	0,9038	-	-	-	-
Drop-out-Differenz	-0,0279	-0,0025	0,0043	0,5873	-	-	-	-
Adverse-Events-Differenz	-0,0644	0,0108	0,0145	0,4967	-	-	-	-
Herkunftsland	0,0114	-0,0100	0,1115	0,9328	-0,0848	0,1043	0,4619	-
Sponsor	-0,0508	0,0179	0,0679	0,8049	-	-	-	-
Psychotherapie	-0,0083	-0,0821	0,0679	0,2932	-	-	-	-
Ziel der Behandlung	-0,0734	0,0748	0,0736	0,4163	-	-	-	-

Abbildung A.4: Naltrexon - HD

	$\hat{\beta}_0$	$\hat{\beta}_1$	SD	p	$\hat{\beta}_2$	SD	p	erkl. Het
Publikationsjahr	-55,6211	0,0276	0,0184	0,1455	-	-	-	0,0942
Teilnehmerzahl	-0,4042	0,0007	0,0004	0,1240	-	-	-	0,0874
Behandlungsdauer	-0,5552	0,0211	0,0207	0,3181	-	-	-	0,0023
Anzahl Kliniken	-0,3596	0,0169	0,0104	0,1148	-	-	-	0,1036
Drop-out-Differenz	-0,2463	0,0128	0,0063	0,0522	-	-	-	0,1164
Adverse-Events-Differenz	-0,1120	-0,0185	0,0196	0,3603	-	-	-	0,9924
Herkunftsland	-0,6880	0,6338	0,1968	0,0034	0,4577	0,1690	0,0118	0,4762
Sponsor	-0,1837	0,0957	0,0896	0,2962	-	-	-	1,0000
Psychotherapie	-0,2621	-0,0630	0,1735	0,7195	-0,0276	0,2533	0,9140	0,0000
Ziel der Behandlung	-0,1502	-0,2745	0,1936	0,1733	-	-	-	0,0000

Abbildung A.5: Naltrexon - DrDays

	$\hat{\beta}_0$	$\hat{\beta}_1$	SD	p	$\hat{\beta}_2$	SD	p	erkl. Het
Publikationsjahr	-205,5809	0,1007	0,1745	0,5695	-	-	-	0,0000
Teilnehmerzahl	-4,3931	0,0025	0,0047	0,6004	-	-	-	0,0000
Behandlungsdauer	-2,7826	-0,0925	0,3757	0,8077	-	-	-	0,0000
Anzahl Kliniken	-3,7292	-0,0490	0,1360	0,7217	-	-	-	0,0000
Drop-out-Differenz	-3,8632	0,1938	0,0098	0,0000	-	-	-	1,0000
Adverse-Events-Differenz	-4,1330	0,1001	0,2959	0,7406	-	-	-	0,4205
Herkunftsland	-3,5220	-4,6430	5,4295	0,4013	-0,2369	3,6318	0,9486	0,0000
Sponsor	-3,7710	-4,2200	3,0811	0,1846	-	-	-	0,0381
Psychotherapie	-3,3463	-1,7635	1,7983	0,3370	-0,5633	3,0754	0,8563	0,0557
Ziel der Behandlung	-4,3314	1,8725	2,4468	0,4552	-	-	-	0,0000

Abbildung A.8: Naltrexon - GGT

	$\hat{\beta}_0$	$\hat{\beta}_1$	SD	p	$\hat{\beta}_2$	SD	p	erkl. Het
Publikationsjahr	-4214,8965	2,1003	0,6324	0,0055	-	-	-	1,0000
Teilnehmerzahl	-6,5192	-0,0379	0,0664	0,5780	-	-	-	0,0000
Behandlungsdauer	-12,4651	0,1748	1,4932	0,9086	-	-	-	0,0000
Anzahl Kliniken	-10,8262	0,0961	0,6152	0,8782	-	-	-	0,0000
Drop-out-Differenz	-9,3506	0,3834	0,2758	0,1898	-	-	-	0,0000
Adverse-Events-Differenz	-8,6479	-0,2820	1,4437	0,8495	-	-	-	0,0000
Herkunftsland	-2,0300	-17,3700	12,4555	0,1865	-10,0382	9,0241	0,2861	0,2668
Sponsor	-12,5619	7,2096	8,7419	0,4244	-	-	-	0,2765
Psychotherapie	-10,1235	-3,3330	8,3157	0,6951	8,0928	11,8896	0,5080	0,0000
Ziel der Behandlung	-8,7546	-9,0368	8,0995	0,3014	-	-	-	1,0000

Abbildung A.9: Naltrexon - Grams

	$\hat{\beta}_0$	$\hat{\beta}_1$	SD	p	$\hat{\beta}_2$	SD	p	erkl. Het
Publikationsjahr	-778,5607	0,3833	1,1177	0,7371	-	-	-	0,0000
Teilnehmerzahl	-15,6923	0,0325	0,0337	0,3527	-	-	-	0,0485
Behandlungsdauer	-23,1718	0,7747	0,3653	0,0538	-	-	-	1,0000
Anzahl Kliniken	-13,5726	0,9113	1,2129	0,4658	-	-	-	0,0000
Drop-out-Differenz	-7,6242	0,7097	0,3126	0,0443	-	-	-	1,0000
Adverse-Events-Differenz	-6,5857	-0,6714	1,7963	0,7214	-	-	-	0,0000
Herkunftsland	-	-	-	-	-	-	-	-
Sponsor	-12,0129	16,9129	19,1178	0,3924	-	-	-	0,0000
Psychotherapie	-3,7763	-19,5470	6,0649	0,0067	-24,3572	8,9880	0,0179	1,0000
Ziel der Behandlung	-5,6797	8,4797	12,4636	0,5155	-	-	-	1,0000

A.2 Regressionen über beide Substanzen

Tabelle A.1: Gesamt - AD

	$\hat{\beta}_0$	$\hat{\beta}_1$	SD	p	$\hat{\beta}_2$	SD	p	$\hat{\beta}$ - Naltr.	SD	p	erkl. Het
Publikationsjahr	-31,1526	0,0155	0,0037	0,0001	–	–	–	0,0111	0,0172	0,5233	0,8400
Teilnehmerzahl	-0,2854	0,0002	0,0001	0,0069	–	–	–	0,0504	0,0180	0,0075	0,5124
Behandlungsdauer	-0,1969	-0,0002	0,0017	0,9108	–	–	–	0,0456	0,0237	0,0600	0,0773
Anzahl Kliniken	-0,2608	0,0039	0,0022	0,0849	–	–	–	0,0554	0,0198	0,0074	0,2330
Drop-out-Differenz	-0,1481	0,0069	0,0016	0,0001	–	–	–	0,0327	0,0152	0,0375	0,9339
Adverse-Events-Differenz	-0,1805	-0,0035	0,0092	0,7062	–	–	–	0,0325	0,0253	0,2084	0,0000
Herkunftsland	-0,1453	-0,0468	0,0683	0,4969	0,0213	0,0229	0,3578	0,0378	0,0683	0,5833	0,3277
Sponsor	-0,2138	0,0046	0,0472	0,9225	–	–	–	0,0498	0,0232	0,0372	0,1058
Psychotherapie	-0,1547	-0,0084	0,0490	0,8648	0,0290	0,0214	0,1834	0,0241	0,0711	0,7359	0,2172
Ziel der Behandlung	-0,0938	-0,1039	0,0433	0,0212	–	–	–	0,0260	0,0213	0,2282	0,4640

Tabelle A.2: Gesamt - HD

	$\hat{\beta}_0$	$\hat{\beta}_1$	SD	p	$\hat{\beta}_2$	SD	p	$\hat{\beta}$ - Naltr.	SD	p	erkl. Het
Publikationsjahr	-48,3209	0,0241	0,0158	0,1361	-	-	-	-0,0922	0,0702	0,1987	0,1639
Teilnehmerzahl	-0,1573	0,0004	0,0003	0,1629	-	-	-	-0,0664	0,0772	0,3961	0,1489
Behandlungsdauer	-0,1310	0,0102	0,0129	0,4375	-	-	-	-0,0927	0,0778	0,2425	0,0621
Anzahl Kliniken	-0,1477	0,0131	0,0080	0,1115	-	-	-	-0,0635	0,0750	0,4032	0,1846
Drop-out-Differenz	0,0509	0,0107	0,0051	0,0458	-	-	-	-0,1001	0,0692	0,1579	0,1856
Adverse-Events-Differenz	0,0311	-0,0071	0,0143	0,6282	-	-	-	-0,0560	0,0355	0,1302	1,0000
Herkunftsland	-0,1816	0,4285	0,1598	0,0115	-0,1221	0,0633	0,0625	0,3010	0,1400	0,0392	0,3610
Sponsor	-0,0349	0,0659	0,0665	0,3297	-	-	-	-0,0470	0,0343	0,1805	1,0000
Psychotherapie	0,1164	-0,0751	0,1321	0,5735	-0,1235	0,0748	0,1085	-0,0301	0,2218	0,8930	0,0000
Ziel der Behandlung	0,1218	-0,2005	0,1544	0,2075	-	-	-	-0,0970	0,0900	0,2928	0,0000

Tabelle A.3: Gesamt - GGT

	$\hat{\beta}_0$	$\hat{\beta}_1$	SD	p	$\hat{\beta}_2$	SD	p	$\hat{\beta}$ - Naltr.	SD	p	erkl. Het
Publikationsjahr	-1194,2479	0,5919	0,7469	0,4373	-	-	-	-0,1162	3,7137	0,9754	0,0000
Teilnehmerzahl	-7,7514	-0,0170	0,0242	0,4905	-	-	-	-0,1018	3,7395	0,9785	0,0000
Behandlungsdauer	-14,1886	0,0308	0,2713	0,9107	-	-	-	1,3390	3,5552	0,7104	0,0000
Anzahl Kliniken	-8,5904	-0,2733	0,3489	0,4427	-	-	-	-0,1187	3,6060	0,9741	0,0000
Drop-out-Differenz	-9,8353	0,3962	0,2713	0,1606	-	-	-	0,4155	3,3291	0,9020	0,0000
Adverse-Events-Differenz	-18,5847	-0,1436	1,1014	0,8983	-	-	-	2,9550	3,7165	0,4408	0,0782
Herkunftsland	-17,1569	3,2562	9,7337	0,7415	2,4227	5,0349	0,6356	-0,3994	10,1782	0,9691	0,0000
Sponsor	-19,1064	0,7552	7,4871	0,9208	-	-	-	2,7890	3,8331	0,4762	0,3281
Psychotherapie	-6,4301	-8,4537	6,9108	0,2370	-0,1997	3,4959	0,9551	5,0768	12,2803	0,6842	0,0000
Ziel der Behandlung	4,7361	-8,8489	10,6350	0,4204	-	-	-	-4,3912	4,8129	0,3782	0,2260

A.3 Backward- und Forward-Selection

Tabelle A.4: Back- und Forward-Selection getrennt nach Substanz – I

ACA.AD

Herausgenommen: Adverse-Events-Differenz, Psychotherapie, Herkunftsland

BACKWARD

	$\hat{\beta}$	SD	p
(Intercept)	-0,3221	0,0461	0,0000
Teilnehmerzahl	0,0005	0,0001	0,0003

Erklärte Heterogenität: 0,85

FORWARD

	$\hat{\beta}$	SD	p
DASSELBE			

NAL.AD

Herausgenommen: Adverse-Events-Differenz

BACKWARD

	$\hat{\beta}$	SD	p
(Intercept)	-38,5395	16,6583	0,0321
Publikationsjahr	0,0192	0,0083	0,0322
Teilnehmerzahl	-0,0002	0,0001	0,1308
Drop-out-Differenz	0,0094	0,0037	0,0197
Ziel der Behandlung	0,0956	0,0683	0,1780
Psychotherapie (CBT)	0,0374	0,0772	0,6334
Psychotherapie (MET)	0,0831	0,0703	0,2514

Erklärte Heterogenität: –

FORWARD

	$\hat{\beta}$	SD	p
(Intercept)	-35,4366	16,8071	0,0485
Publikationsjahr	0,0177	0,0084	0,0484
Drop-out-Differenz	0,0093	0,0038	0,0243
Ziel der Behandlung	0,0683	0,0663	0,3161
Herkunftsland (EU)	0,0037	0,1181	0,9755
Herkunftsland (USA)	-0,0977	0,1141	0,4024
Psychotherapie (CBT)	0,0917	0,0759	0,2414
Psychotherapie (MET)	0,0856	0,0712	0,2441

Erklärte Heterogenität: –

Tabelle A.5: Back- und Forward-Selection getrennt nach Substanz – II

NAL.HD

Herausgenommen: Ziel der Behandlung

BACKWARD

	$\hat{\beta}$	SD	p
(Intercept)	-66,9947	28,4032	0,0261
Publikationsjahr	0,0331	0,0142	0,0275
Herkunftsland (EU)	0,6660	0,1741	0,0007
Herkunftsland (USA)	0,4808	0,1494	0,0034

Erklärte Heterogenität: 0,73

FORWARD

	$\hat{\beta}$	SD	p
(Intercept)	-61,2324	22,6370	0,0124
Publikationsjahr	0,0305	0,0113	0,0126
Sponsor	0,1116	0,0898	0,2263

Erklärte Heterogenität: 1,00

NAL.DrDays

Herausgenommen: Ziel der Behandlung

BACKWARD

	$\hat{\beta}$	SD	p
(Intercept)	-783,4468	46,1888	0,0000
Publikationsjahr	0,3902	0,0228	0,0000
Teilnehmerzahl	0,0046	0,0054	0,4099
Anzahl Kliniken	-0,2571	0,1587	0,1292
Adverse-Events-Differenz	0,1688	0,0818	0,0596
Herkunftsland (EU)	-4,7721	7,0973	0,5131
Herkunftsland (USA)	-0,8919	6,0542	0,8851

Erklärte Heterogenität: 1,00

FORWARD

	$\hat{\beta}$	SD	p
(Intercept)	-6,5057	3,9657	0,1151
Sponsor	-3,8567	-3,8567	0,3961
Herkunftsland (EU)	2,1982	7,0098	0,7568
Herkunftsland (USA)	2,8890	4,0541	0,4836

Erklärte Heterogenität: 0,00

Tabelle A.6: Back- und Forward-Selection getrennt nach Substanz – III

ACA.DaysAbs

Herausgenommen: Psychotherapie, Ziel der Behandlung

BACKWARD

	$\hat{\beta}$	SD	p
(Intercept)	13,7800	7,4398	0,0838
Teilnehmerzahl	-0,0264	0,0140	0,0785
Drop-out-Differenz	0,0243	0,2469	0,9230
Herkunftsland (EU)	12,8430	7,0128	0,0870
Herkunftsland (USA)	2,6659	8,9557	0,7700
Sponsor	-6,5896	6,7203	0,3424

Erklärte Heterogenität: 0,54

FORWARD

	$\hat{\beta}$	SD	p
(Intercept)	16,8885	4,0511	0,0006
Anzahl Kliniken	-0,5263	0,2726	0,0704

Erklärte Heterogenität: 0,10

NAL.GGT

Herausgenommen: Adverse-Events-Differenz, Herkunftsland, Ziel der Behandlung

BACKWARD

	$\hat{\beta}$	SD	p
(Intercept)	-6066,1000	1405,9231	0,0010
Publikationsjahr	3,0267	0,7034	0,0010
Drop-out-Differenz	0,6117	0,2138	0,0143

Erklärte Heterogenität: 1,00

FORWARD

	$\hat{\beta}$	SD	p
(Intercept)	-6347,5787	1437,9166	0,0008
Publikationsjahr	3,1706	0,7201	0,0009
Teilnehmerzahl	-0,0601	0,0644	0,3691
Drop-out-Differenz	0,6513	0,2180	0,0113

Erklärte Heterogenität: 1,00

Tabelle A.7: Back- und Forward-Selection getrennt nach Substanz – IV

NAL.Grams

Herausgenommen: Sponsor-Industrie, Adverse-Events-Differenz, Ziel der Behandlung, Herkunftsland

BACKWARD				FORWARD			
	$\hat{\beta}$	SD	p		$\hat{\beta}$	SD	p
(Intercept)	-5,0992	2,8160	0,0933				
Anzahl Kliniken	0,6864	0,7543	0,3794				
Psychotherapie (CBT)	-18,9106	6,1051	0,0085				
Psychotherapie (MET)	-26,0851	9,1863	0,0139		DASSELBE		
Erklärte Heterogenität: 1,00							

Tabelle A.8: Back- und Forward-Selection über beide Substanzen – I

AD

Herausgenommen: Ziel der Behandlung

BACKWARD				FORWARD			
	$\hat{\beta}$	SD	p		$\hat{\beta}$	SD	p
(Intercept)	-25,3155	7,1946	0,0010				
Publikationsjahr	0,0126	0,0036	0,0011				
Anzahl Kliniken	0,0028	0,0017	0,0921				
Drop-out-Differenz	0,0043	0,0018	0,0192		DASSELBE		
Erklärte Heterogenität: 1,00							

Tabelle A.9: Back- und Forward-Selection über beide Substanzen – II

HD**Herausgenommen:** Ziel der Behandlung, Adverse-Events-Differenz**BACKWARD**

	$\hat{\beta}$	SD	p
(Intercept)	-52,4113	21,6683	0,0223
Publikationsjahr	0,0261	0,0108	0,0225
Behandlungsdauer	-0,0077	0,0073	0,2949
Anzahl Kliniken	0,0106	0,0055	0,0655
Drop-out-Differenz	0,0094	0,0047	0,0538
Herkunftsland (EU)	0,2342	0,1362	0,0966
Herkunftsland (USA)	-0,1292	0,1253	0,3113
Sponsor	-0,1101	0,1069	0,3118

Erklärte Heterogenität: 1,00**FORWARD**

	$\hat{\beta}$	SD	p
(Intercept)	-61,6150	27,8819	0,0344
Publikationsjahr	0,0305	0,0139	0,0358
Herkunftsland (EU)	0,4404	0,1522	0,0068
Herkunftsland (USA)	0,2267	0,1363	0,1060
Teilnehmerzahl	0,0003	0,0002	0,2505

Erklärte Heterogenität: 0,57**GGT****Herausgenommen:** Ziel der Behandlung, Adverse-Events-Differenz**BACKWARD**

	$\hat{\beta}$	SD	p
(Intercept)	-9,0525	3,5546	0,0197
Drop-out-Differenz	0,3951	0,2642	0,1513

Erklärte Heterogenität: 0,00**FORWARD**

	$\hat{\beta}$	SD	p

DASSELBE

Tabelle A.10: NAL.AD – nach Entfernen nicht-signifikanter Kovariablen

	$\hat{\beta}$	SD	p
(Intercept)	-27,9151	13,4441	0,0492
Publikationsjahr	0,0139	0,0067	0,0496
Drop-out-Differenz	0,0057	0,0027	0,0496
erklärte Heterogenität: –			

Tabelle A.11: NAL.DrDays – nach Entfernen nicht-signifikanter Kovariablen

	$\hat{\beta}$	SD	p
(Intercept)	-771,2919	43,2366	0,0000
Publikationsjahr	0,3836	0,0216	0,0000
Adverse-Events-Differenz	0,2029	0,0792	0,0237
erklärte Heterogenität: 1,00			

Tabelle A.12: HD – nach Entfernen nicht-signifikanter Kovariablen

	$\hat{\beta}$	SD	p
(Intercept)	-71,2896	26,8754	0,0123
Publikationsjahr	0,0354	0,0134	0,0129
Herkunftsland (EU)	0,4844	0,1489	0,0027
Herkunftsland (USA)	0,2857	0,1279	0,0327
erklärte Heterogenität: 0,55			

A.4 Placebo-Response als Kovariable

Tabelle A.13: Placebo-Response - Sensitivitätsanalyse

ACA.AD						
δ	σ_δ^2	β	μ_i	$\hat{\beta}$	SD	p
$N(0, 10)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 10000)$	$N(0, 10)$	-0,1266	0,2065	0,54
$N(0, 100)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 10000)$	$N(0, 10)$	-0,1263	0,2062	0,54
$N(0, 10)$	$\Gamma^{-1}(0, 0001, 0, 001)$	$N(0, 10000)$	$N(0, 10)$	-0,1290	0,2086	0,54
$N(0, 10)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 100000)$	$N(0, 10)$	-0,1266	0,2065	0,54
$N(0, 10)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 10000)$	$N(0, 100)$	-0,1678	0,2122	0,43

NAL.AD						
δ	σ_δ^2	β	μ_i	$\hat{\beta}$	SD	p
$N(0, 10)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 10000)$	$N(0, 10)$	-0,1327	0,08933	0,14
$N(0, 100)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 10000)$	$N(0, 10)$	-0,1327	0,08933	0,14
$N(0, 10)$	$\Gamma^{-1}(0, 0001, 0, 001)$	$N(0, 10000)$	$N(0, 10)$	-0,1282	0,08718	0,14
$N(0, 10)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 100000)$	$N(0, 10)$	-0,1327	0,08933	0,14
$N(0, 10)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 10000)$	$N(0, 100)$	-0,1380	0,0889	0,12

ACA.HD						
δ	σ_δ^2	β	μ_i	$\hat{\beta}$	SD	p
$N(0, 10)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 10000)$	$N(0, 10)$	0,1016	0,2297	0,66
$N(0, 100)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 10000)$	$N(0, 10)$	0,1015	0,2301	0,66
$N(0, 10)$	$\Gamma^{-1}(0, 0001, 0, 001)$	$N(0, 10000)$	$N(0, 10)$	0,1060	0,2218	0,63
$N(0, 10)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 100000)$	$N(0, 10)$	0,1016	0,2297	0,66
$N(0, 10)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 10000)$	$N(0, 100)$	0,09854	0,2296	0,67

NAL.HD						
δ	σ_δ^2	β	μ_i	$\hat{\beta}$	SD	p
$N(0, 10)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 10000)$	$N(0, 10)$	0,0024	0,1416	0,99
$N(0, 100)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 10000)$	$N(0, 10)$	0,0023	0,1416	0,99
$N(0, 10)$	$\Gamma^{-1}(0, 0001, 0, 001)$	$N(0, 10000)$	$N(0, 10)$	0,0069	0,1409	0,96
$N(0, 10)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 100000)$	$N(0, 10)$	0,0024	0,1416	0,99
$N(0, 10)$	$\Gamma^{-1}(0, 001, 0, 001)$	$N(0, 10000)$	$N(0, 100)$	-0,0230	0,1416	0,87

Anhang B

Grafiken

B.1 Zwischenstudienvarianzschätzer und Forest Plots

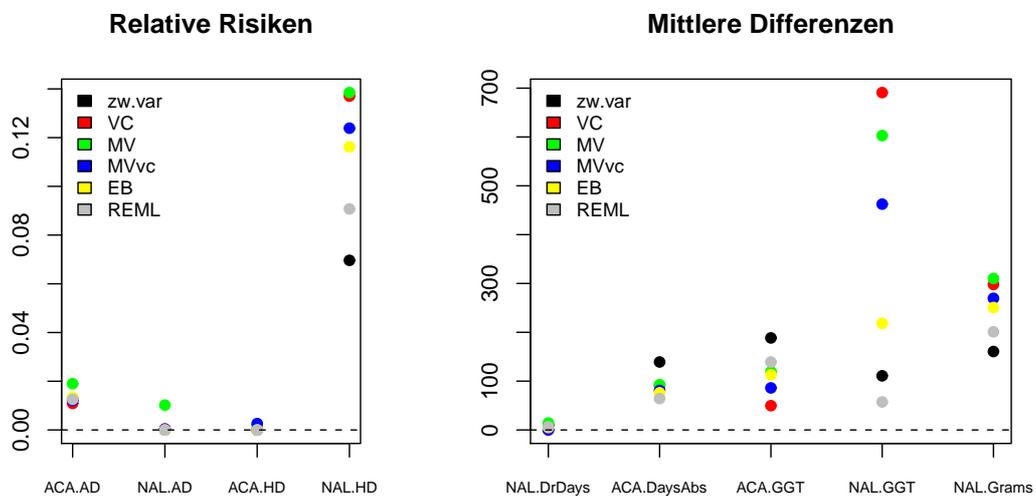


Abbildung B.1: Schätzungen der Zwischenstudienvarianzen

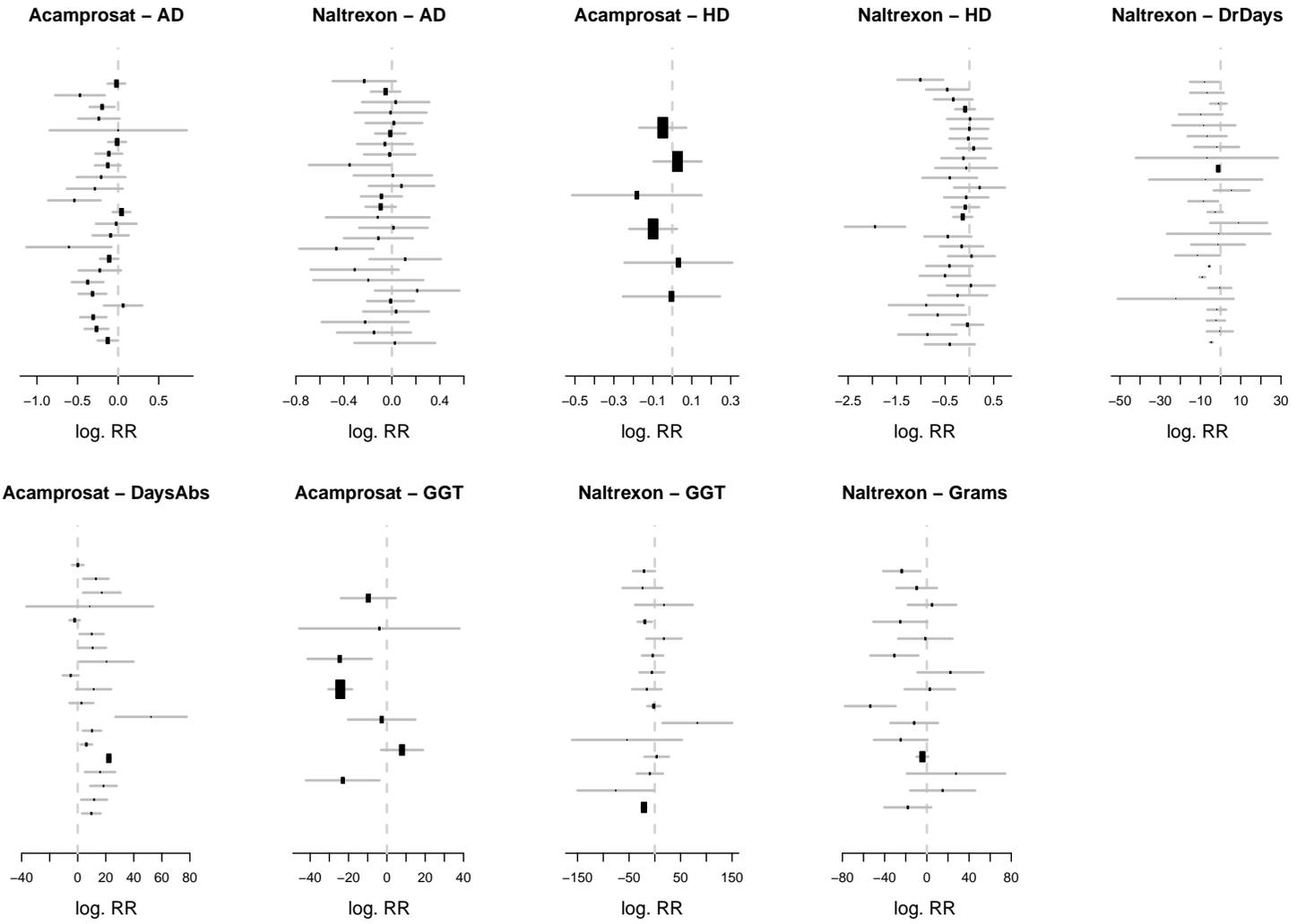


Abbildung B.2: Forest Plots

B.2 Verteilungen der Werte der Kovariablen

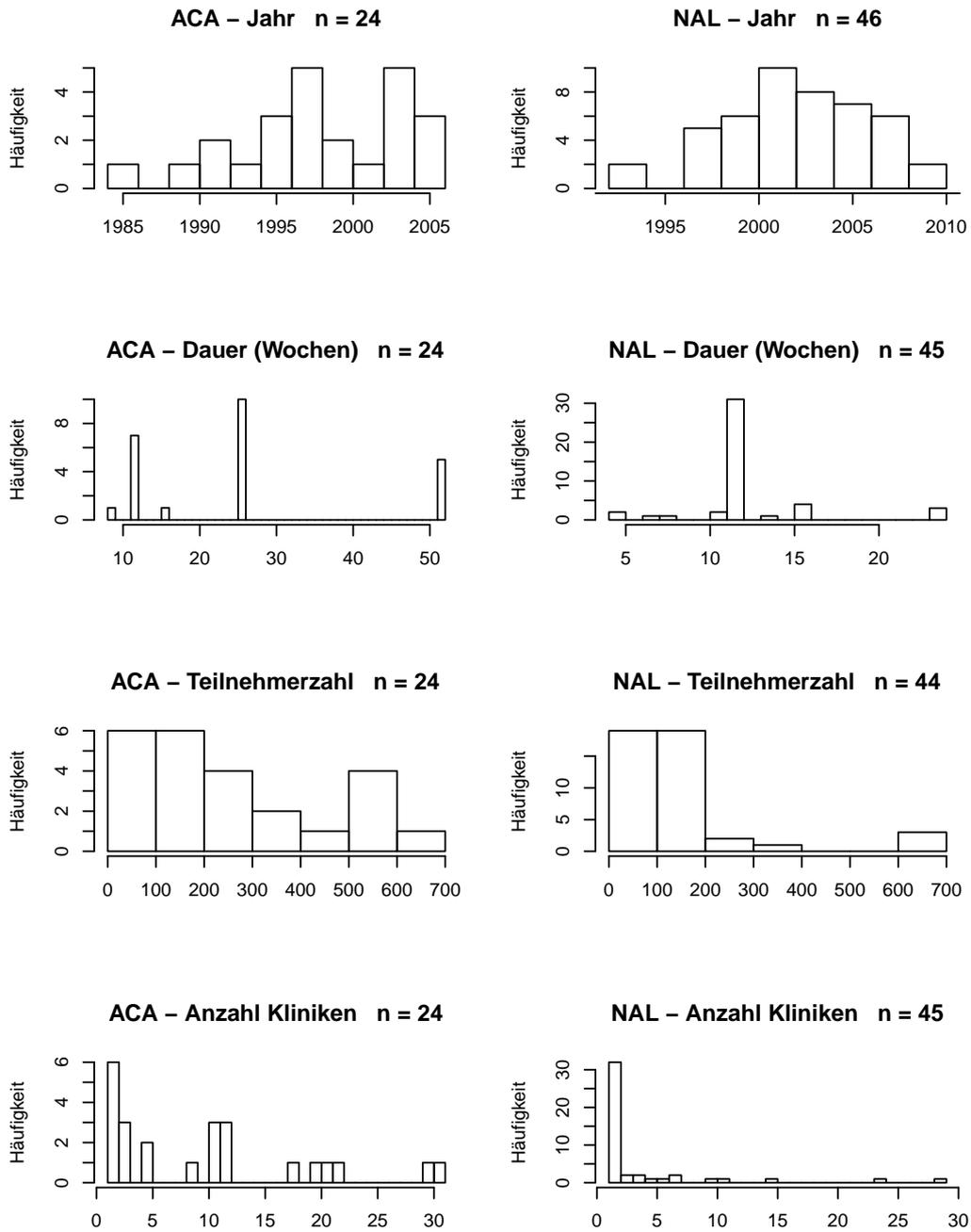


Abbildung B.3: Quasi-stetige Merkmale

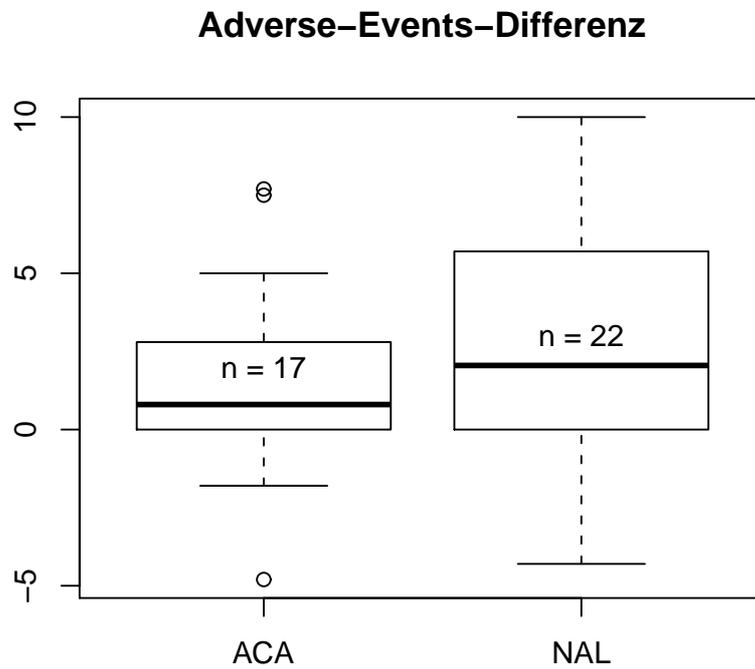
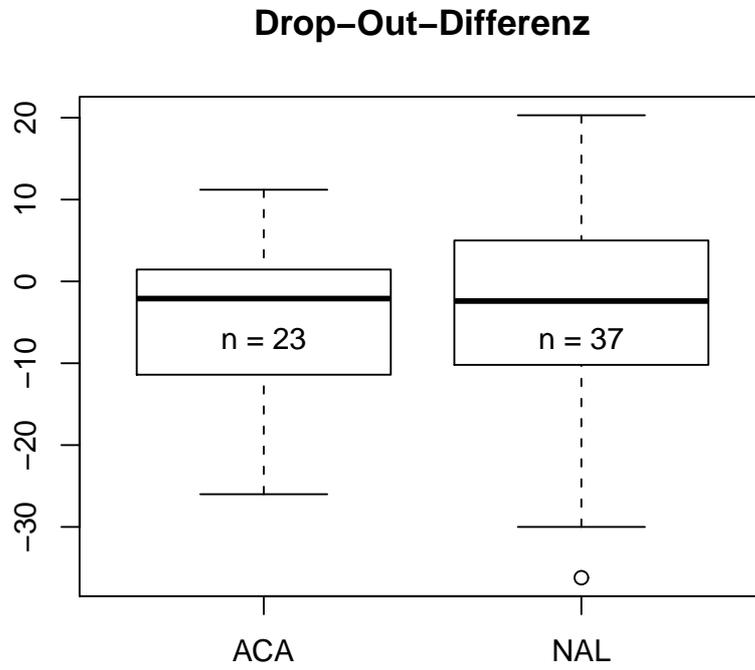


Abbildung B.4: Stetige Merkmale

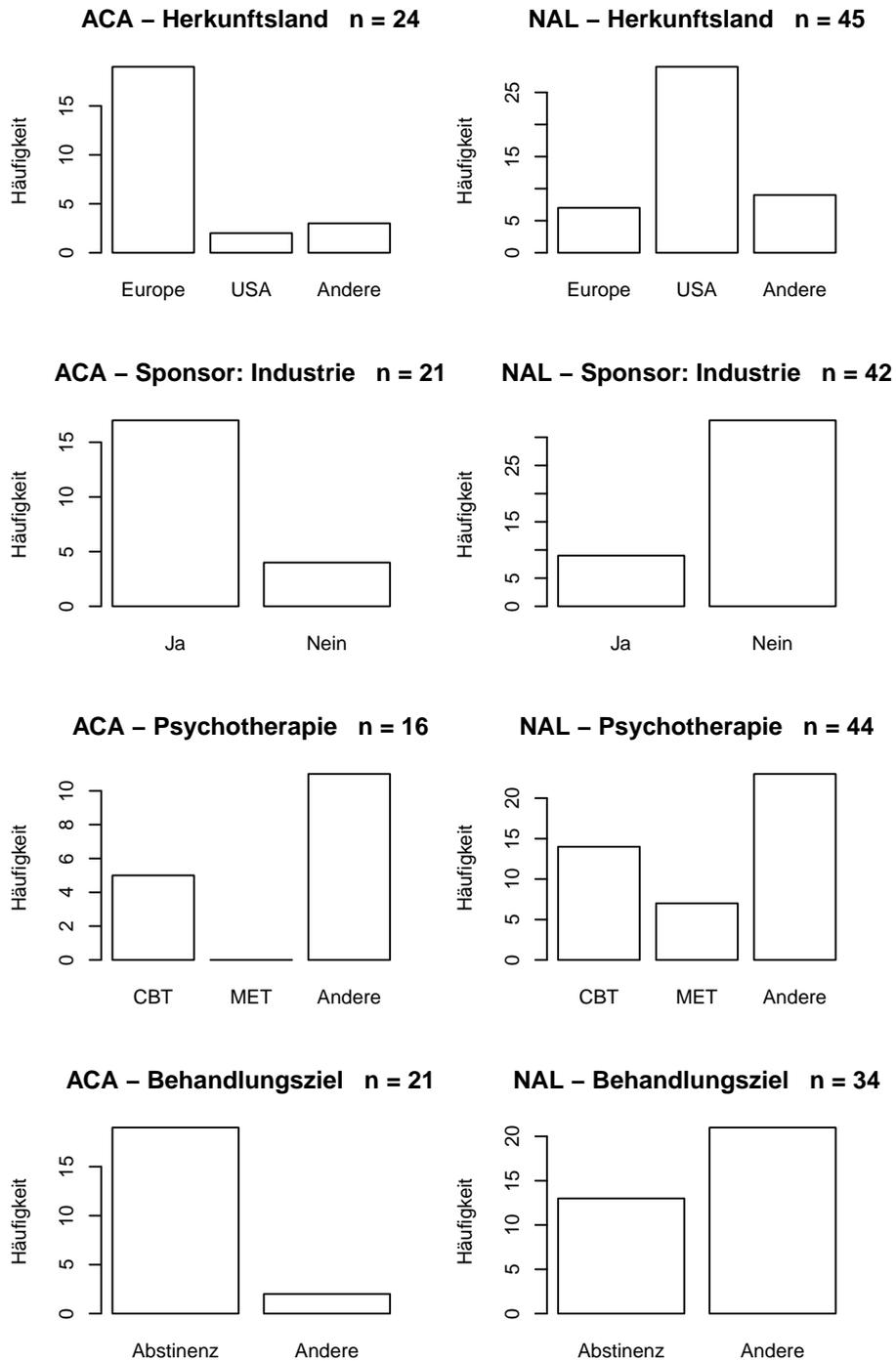


Abbildung B.5: Nominale Merkmale

B.3 Meta-Regression

B.3.1 Streudiagramme

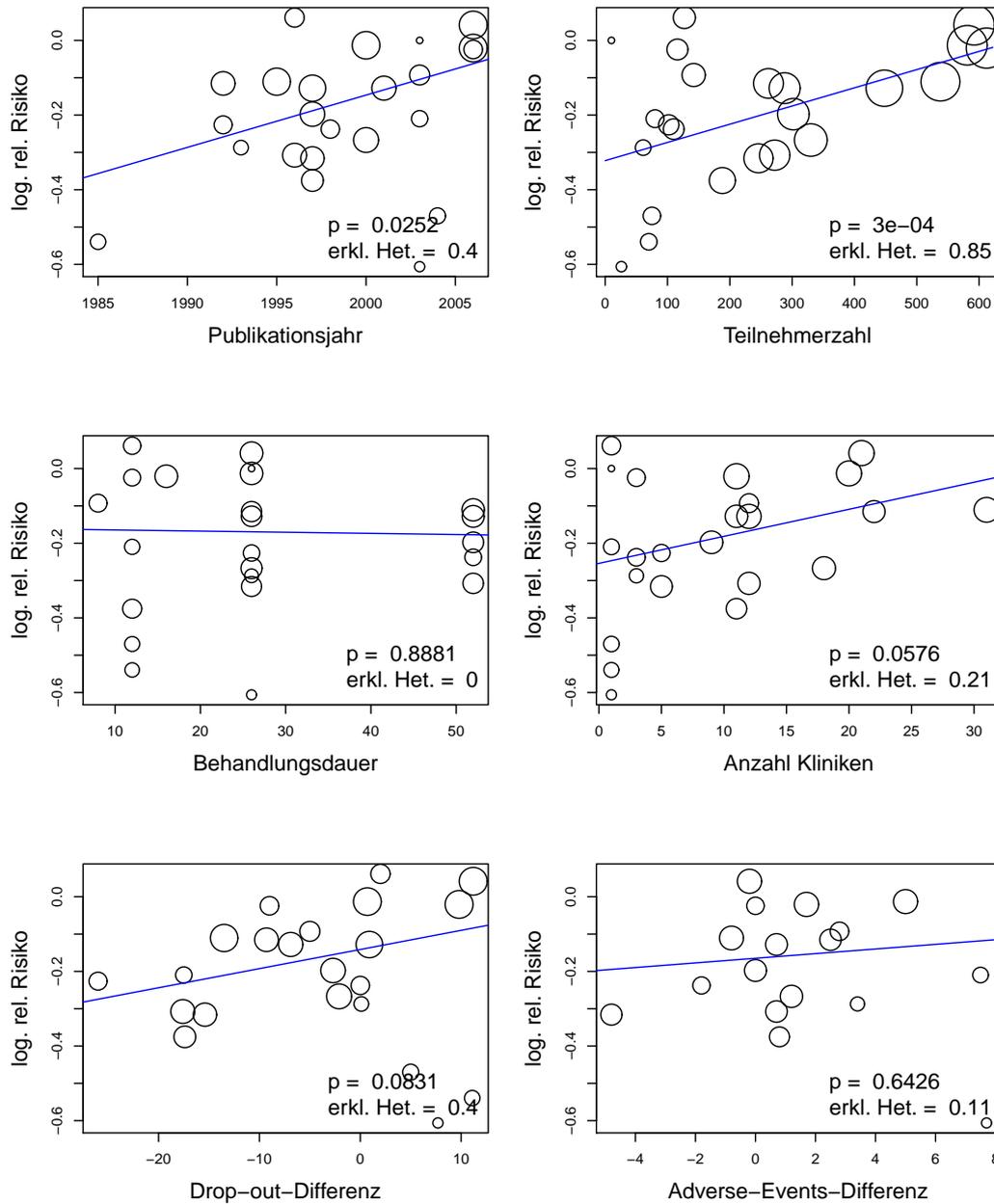


Abbildung B.6: Modelle mit (quasi)stetigen Kovariablen zu ACA.AD

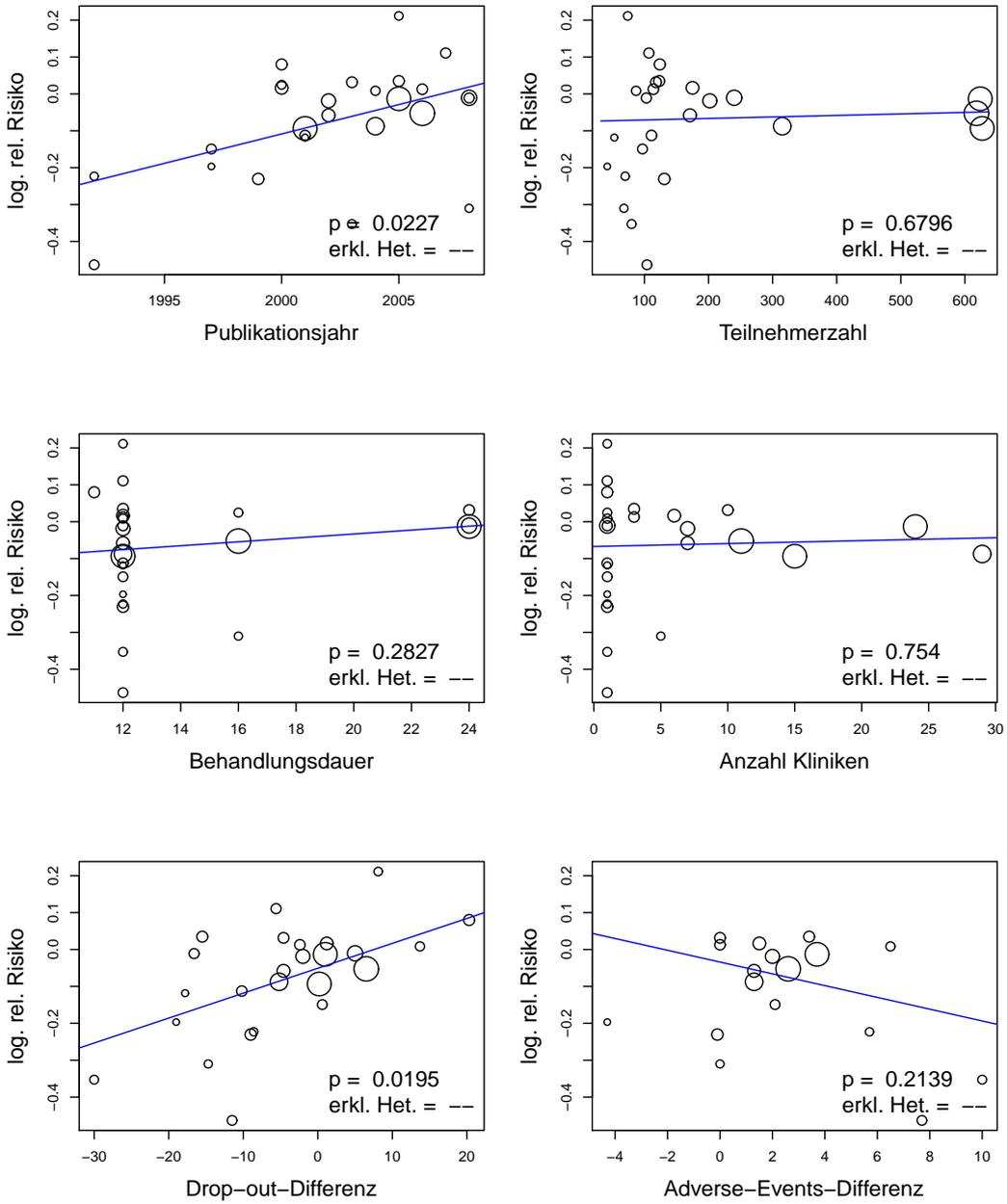


Abbildung B.7: Modelle mit (quasi)stetigen Kovariablen zu NALAD

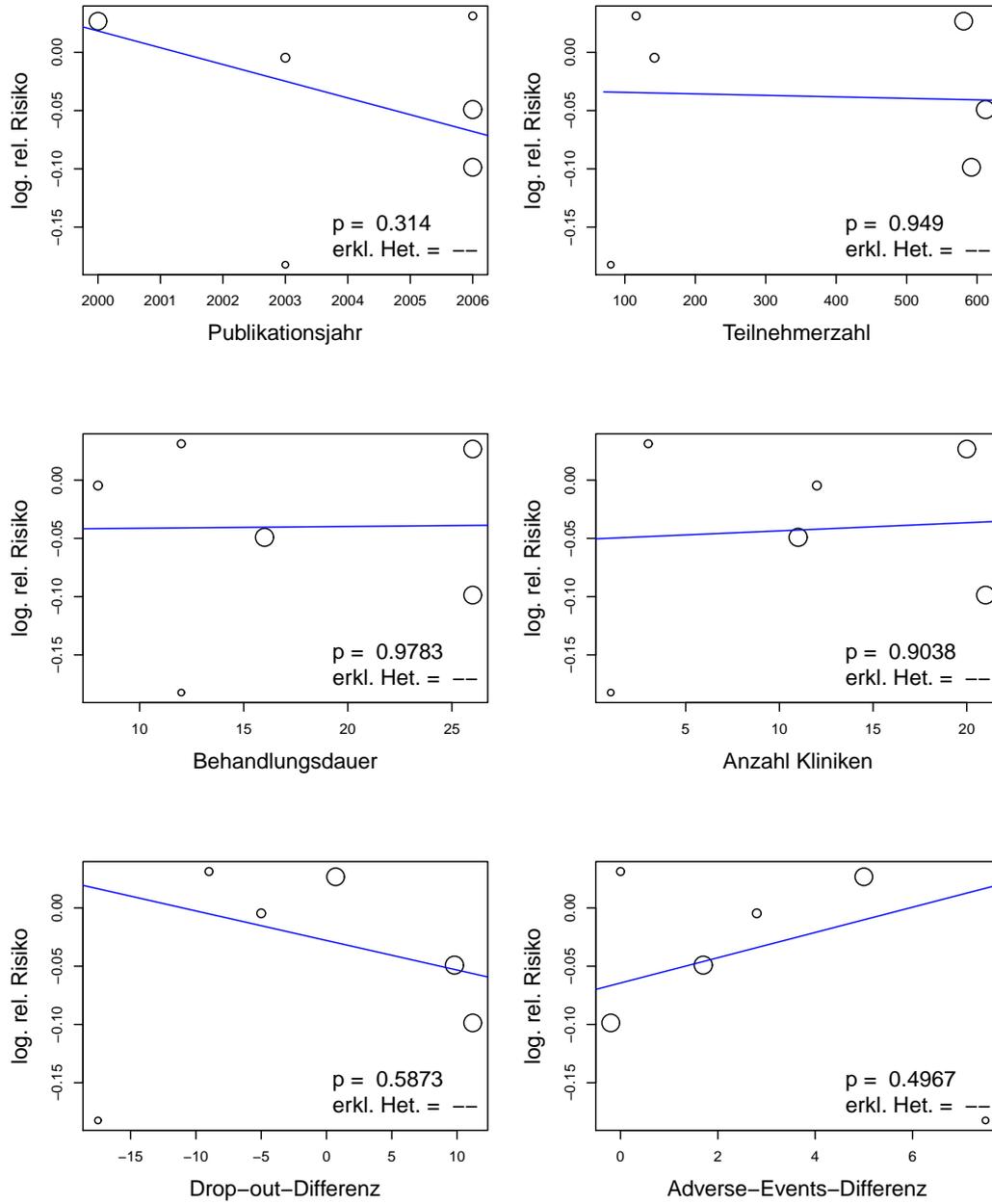


Abbildung B.8: Modelle mit (quasi)stetigen Kovariablen zu ACA.HD

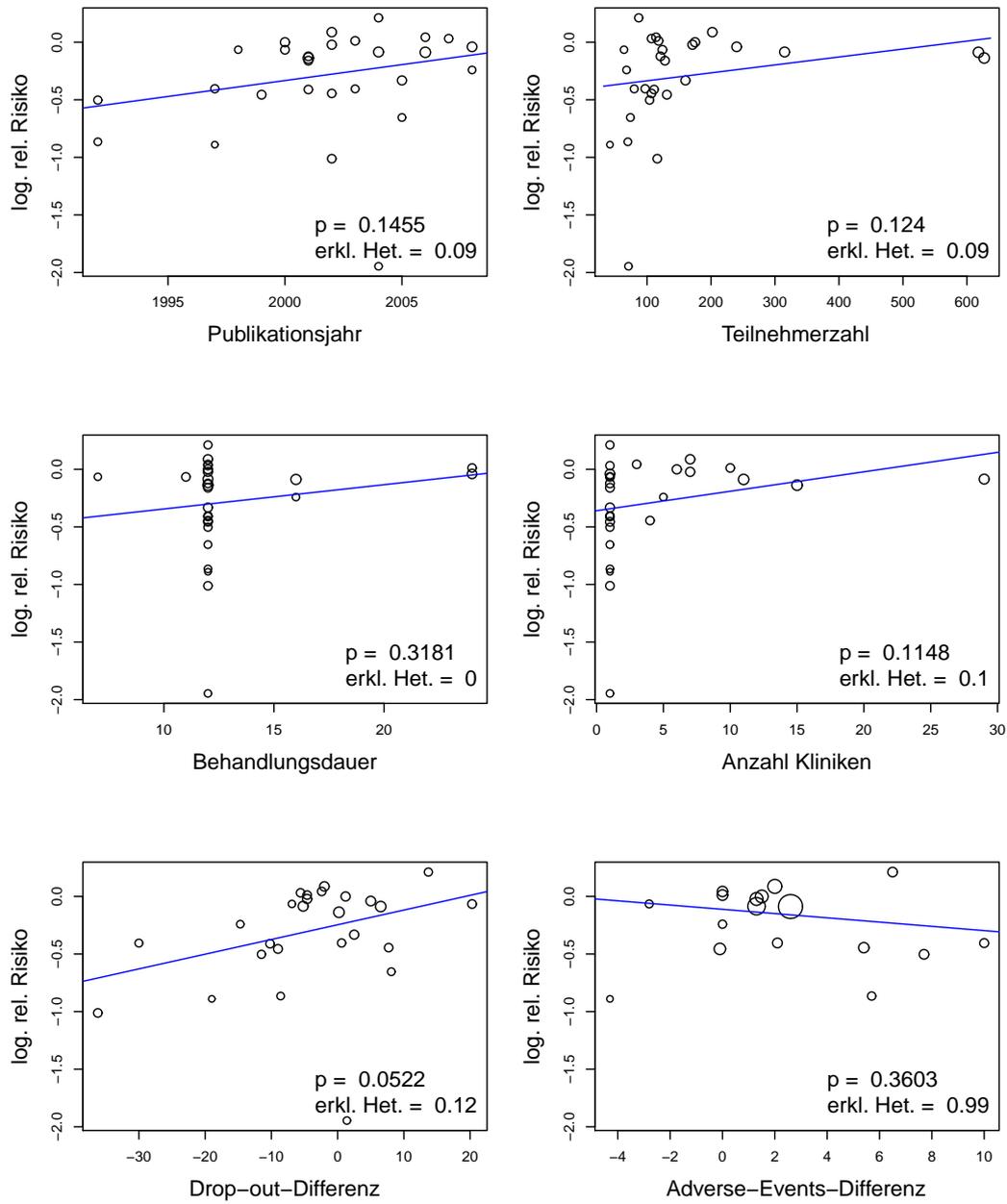


Abbildung B.9: Modelle mit (quasi)stetigen Kovariablen zu NAL.HD

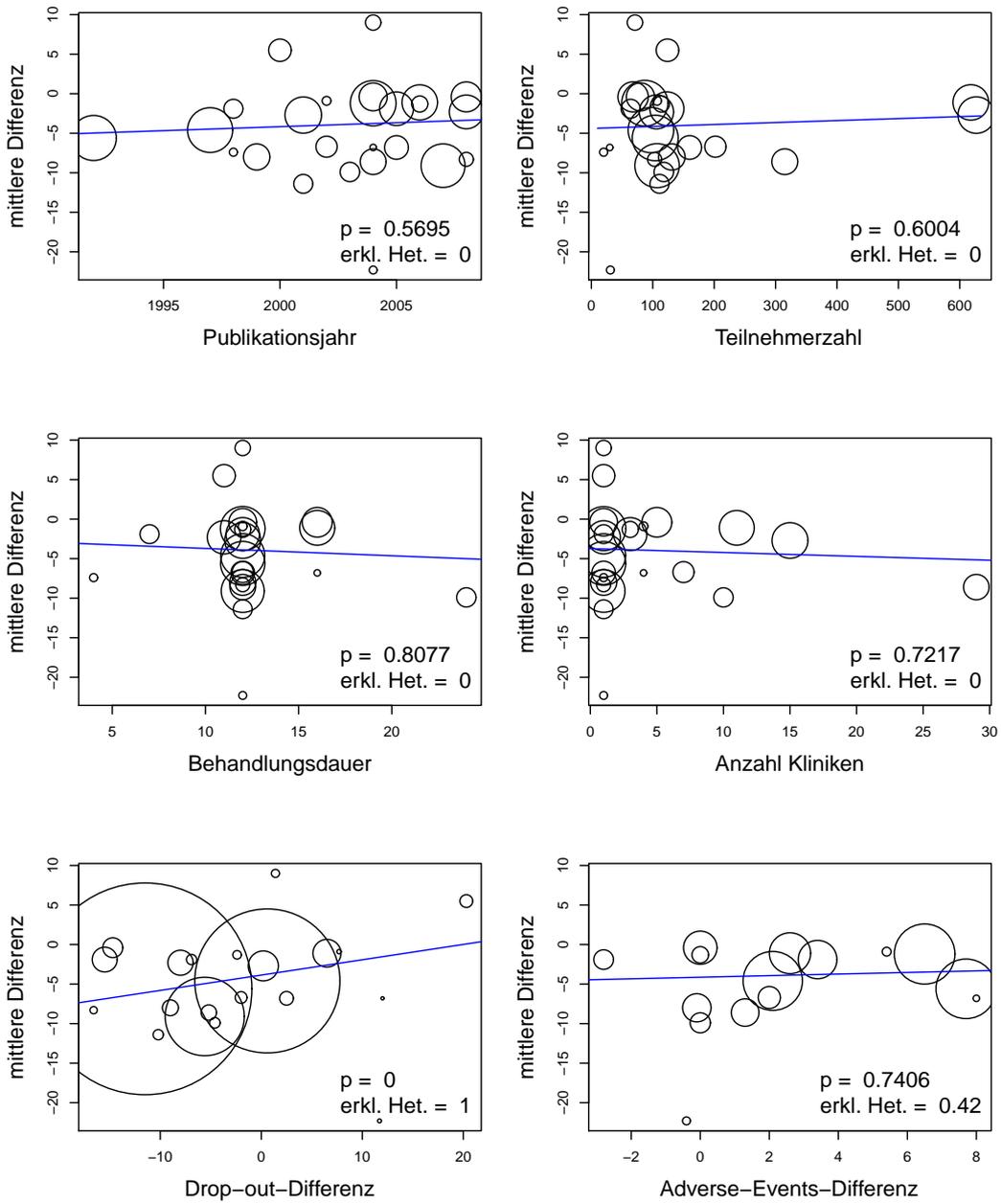


Abbildung B.10: Modelle mit (quasi)stetigen Kovariablen zu NAL.DrDays

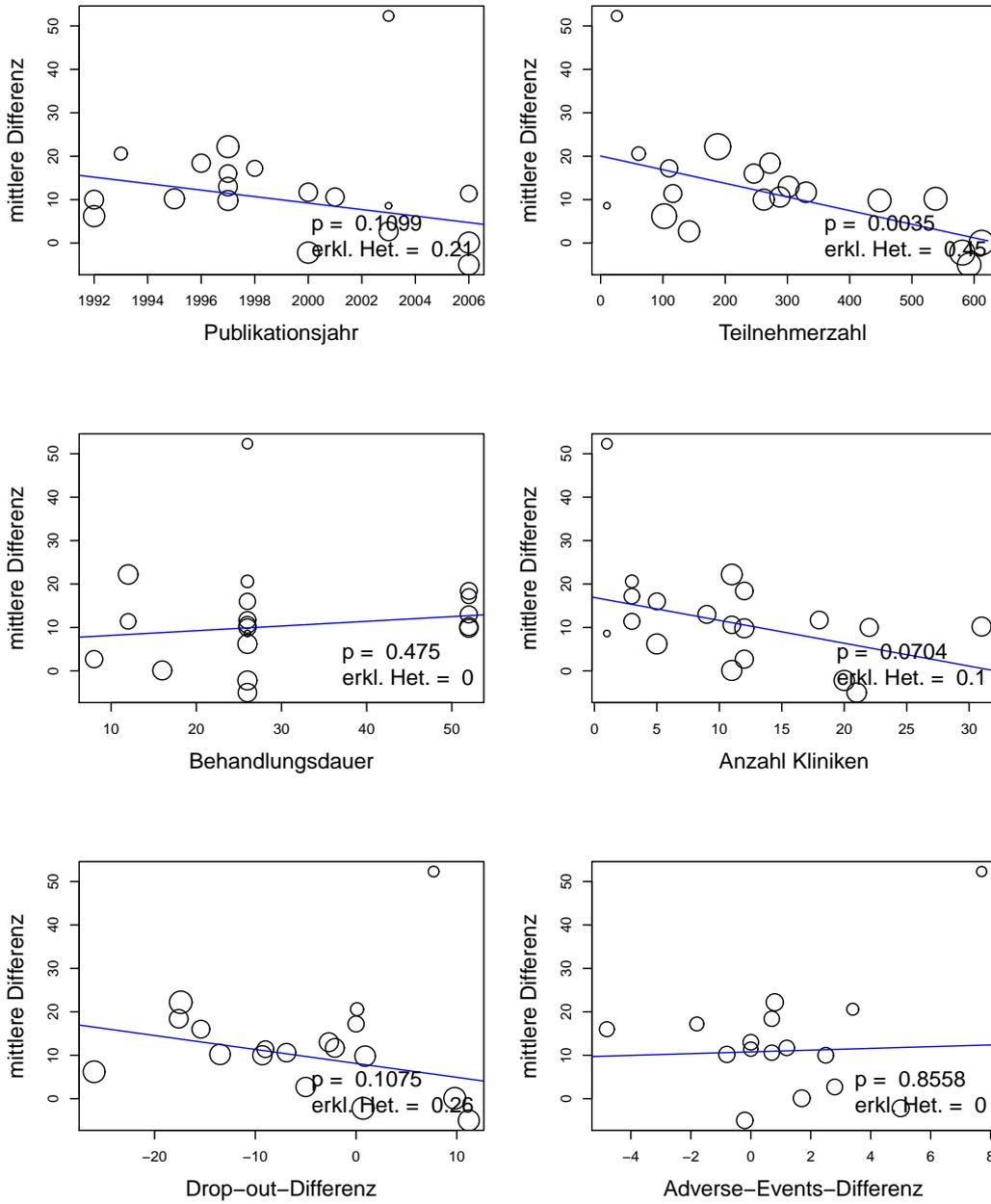


Abbildung B.11: Modelle mit (quasi)stetigen Kovariablen zu ACA.DaysAbs

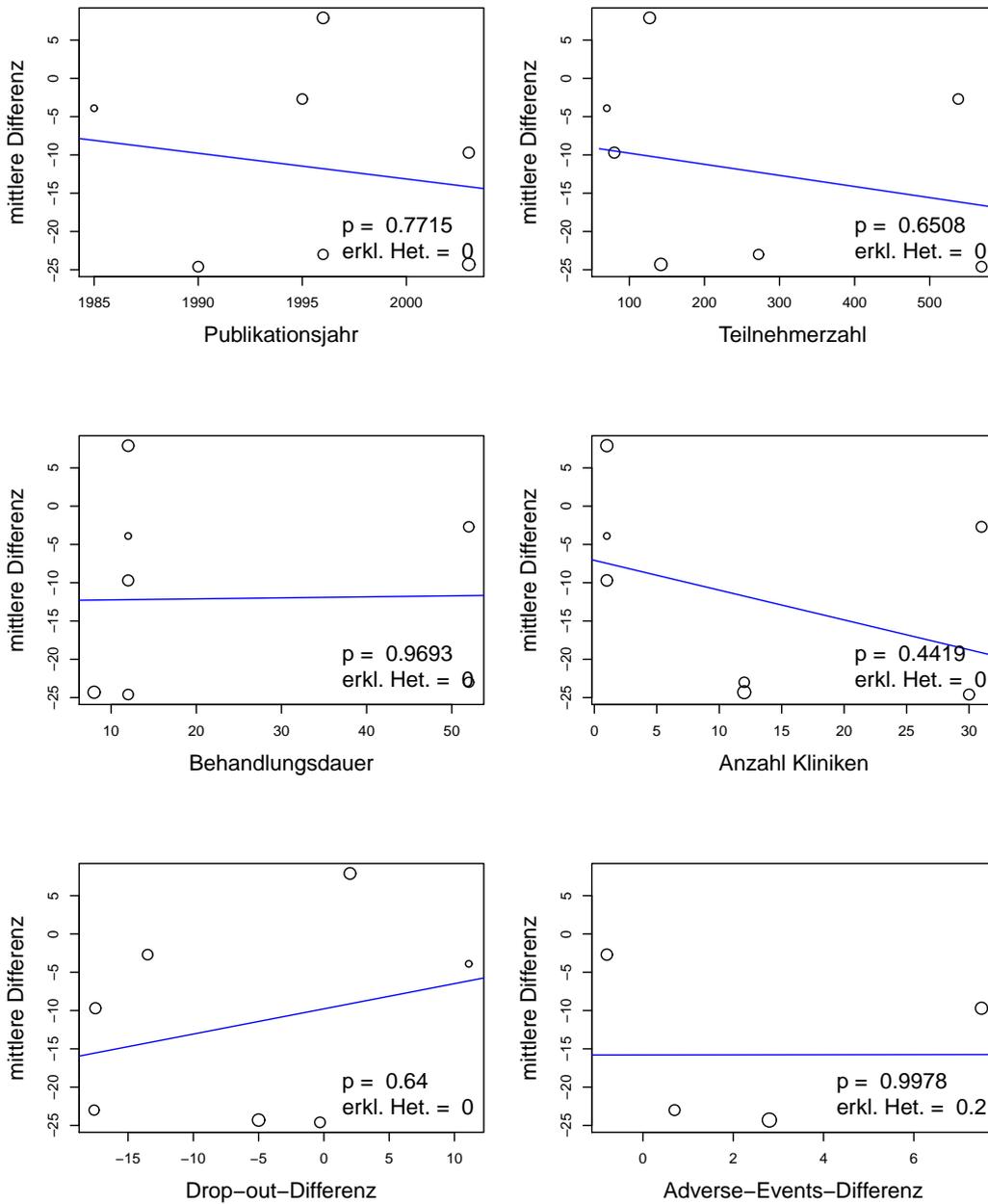


Abbildung B.12: Modelle mit (quasi)stetigen Kovariablen zu ACA.GGT

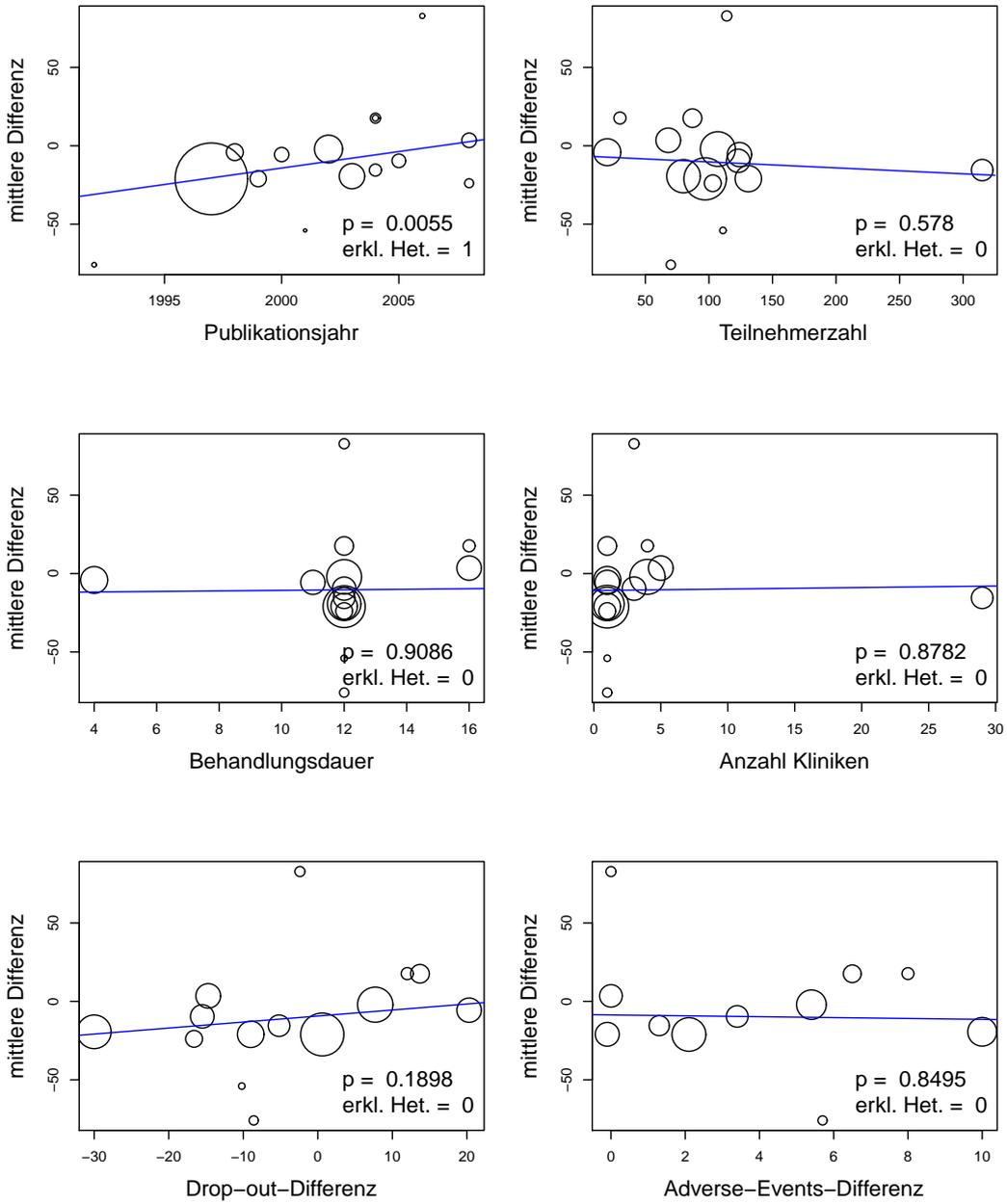


Abbildung B.13: Modelle mit (quasi)stetigen Kovariablen zu NAL.GGT

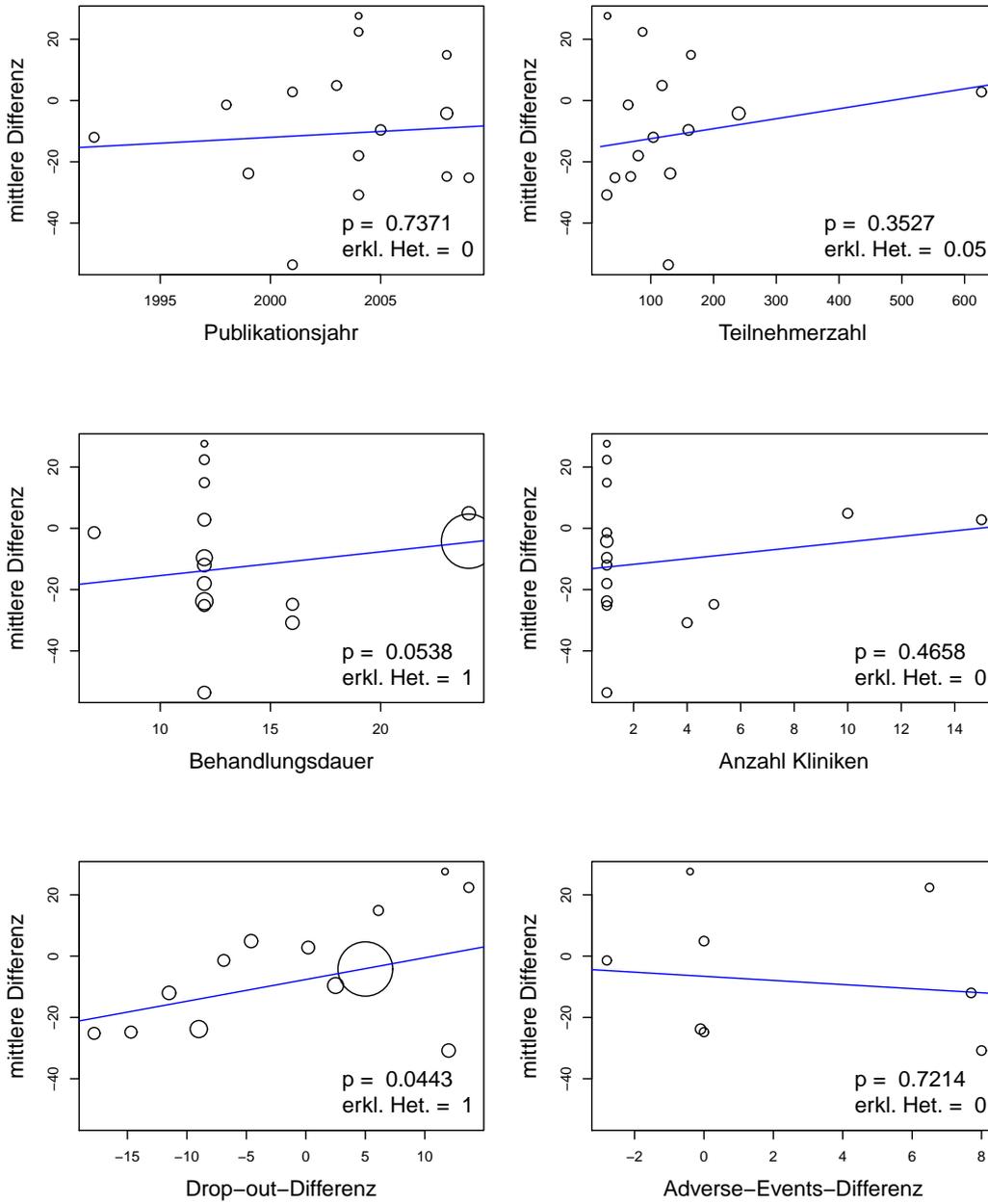


Abbildung B.14: Modelle mit (quasi)stetigen Kovariablen zu NAL.Grams

B.3.2 Normal-Quantil-Plots

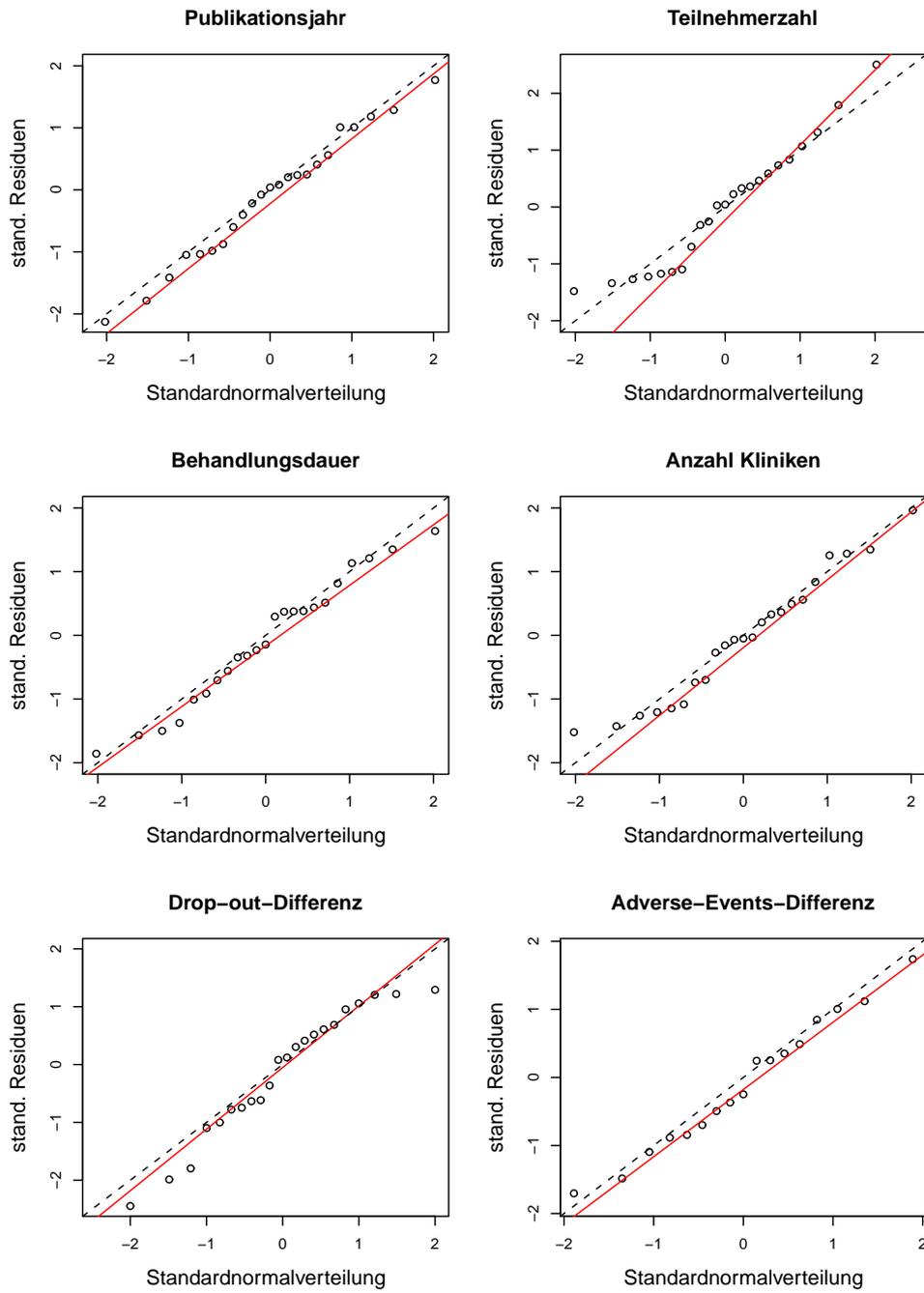


Abbildung B.15: ACA.AD

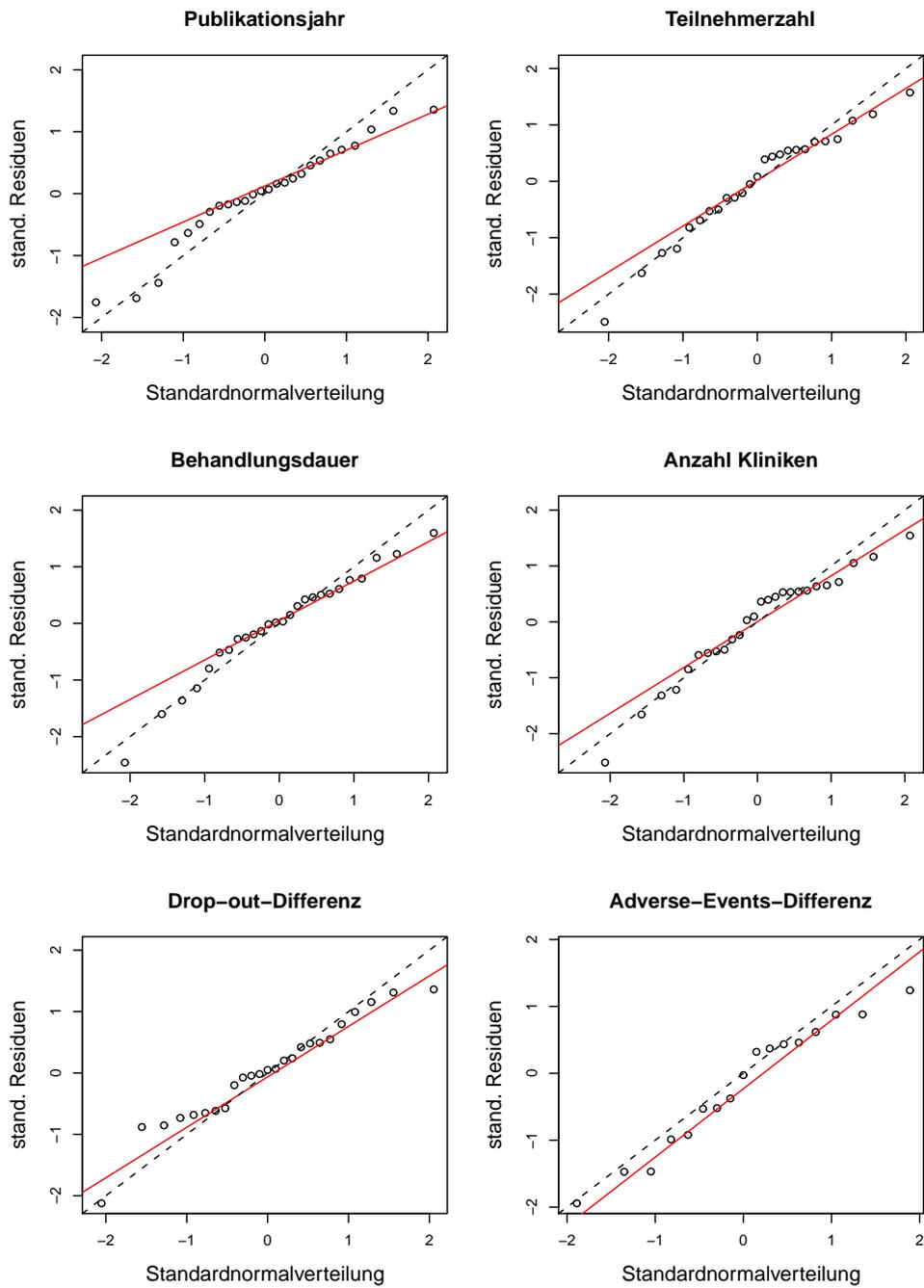


Abbildung B.16: NAL.AD

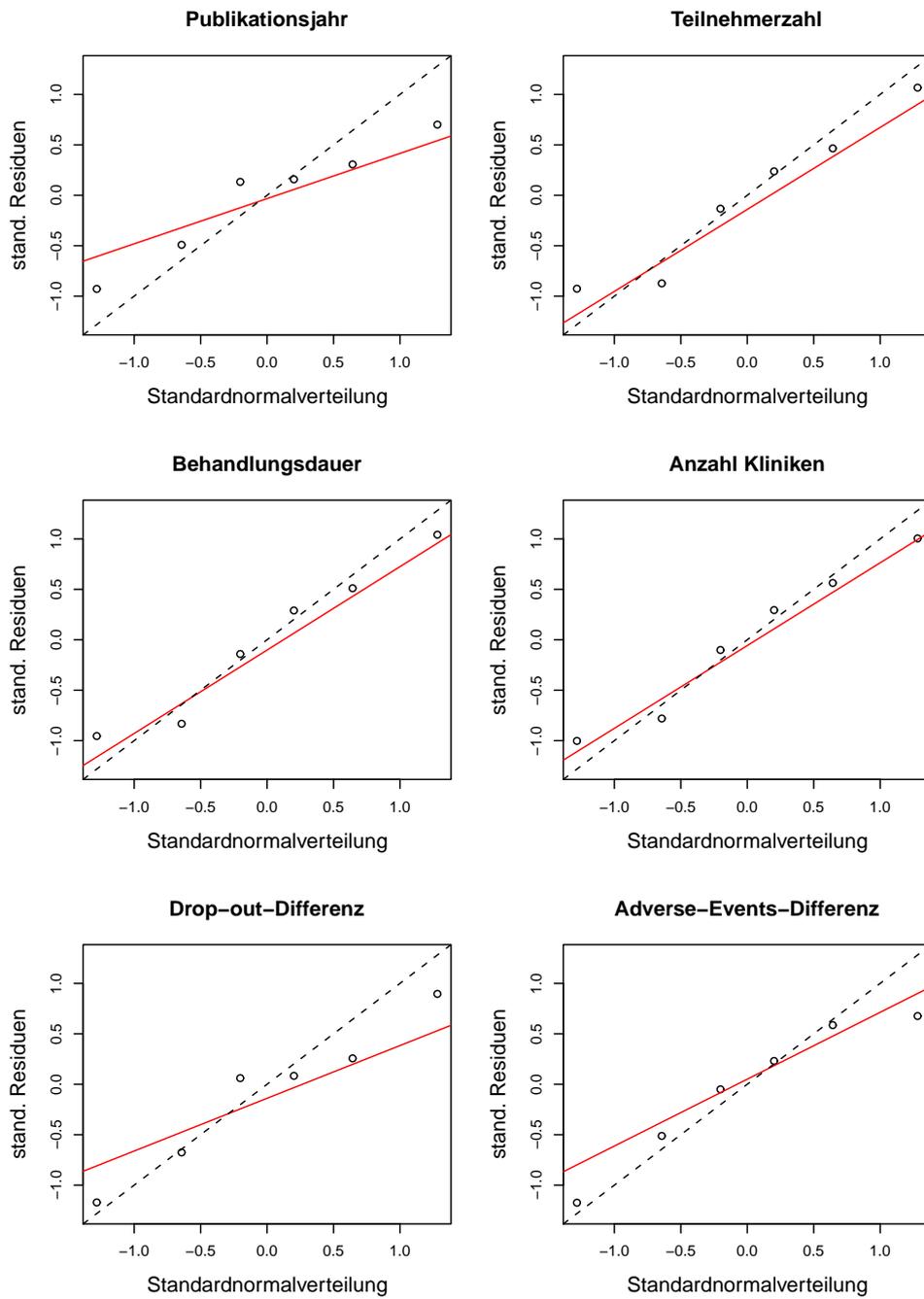


Abbildung B.17: ACA.HD

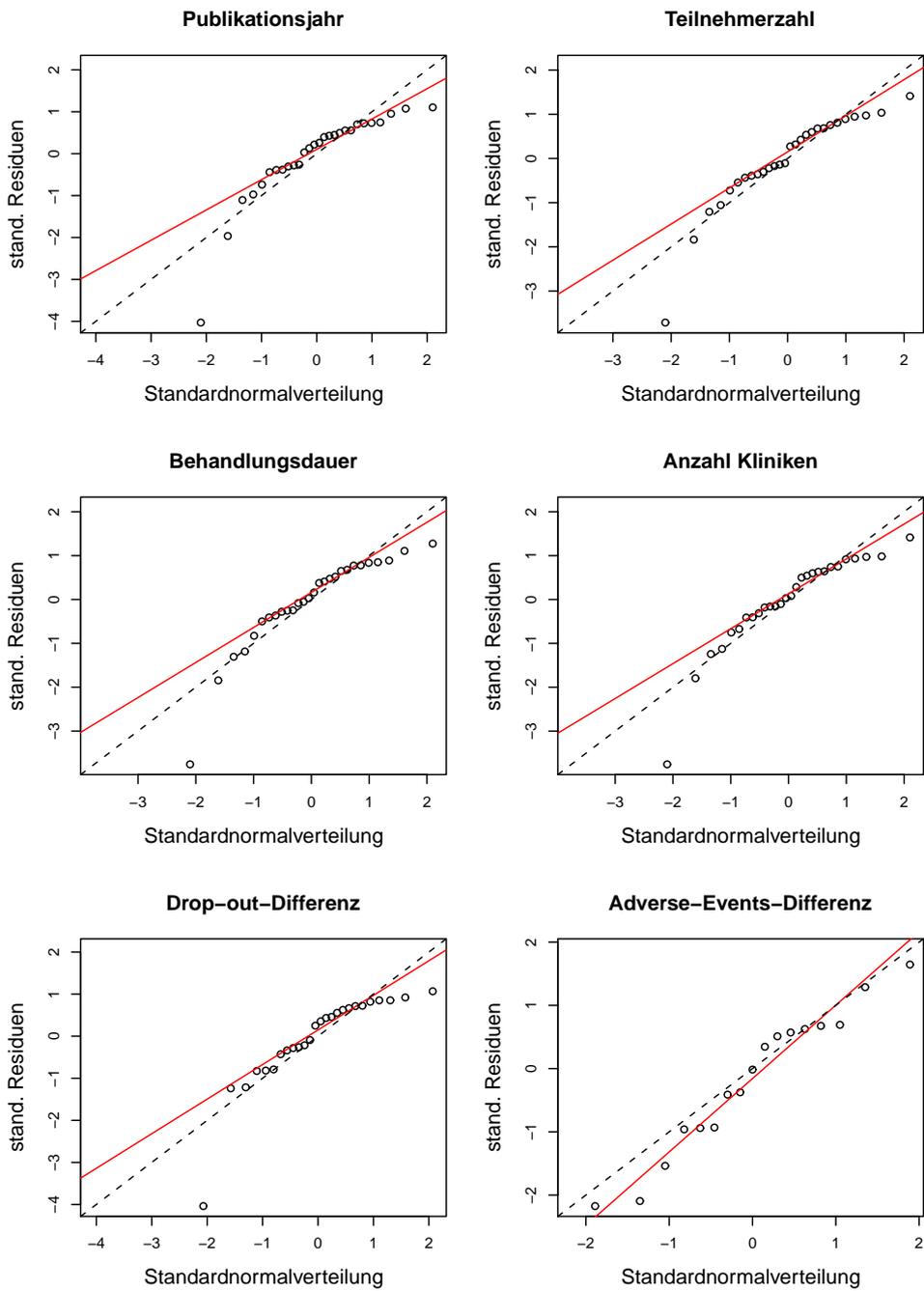


Abbildung B.18: NAL.HD

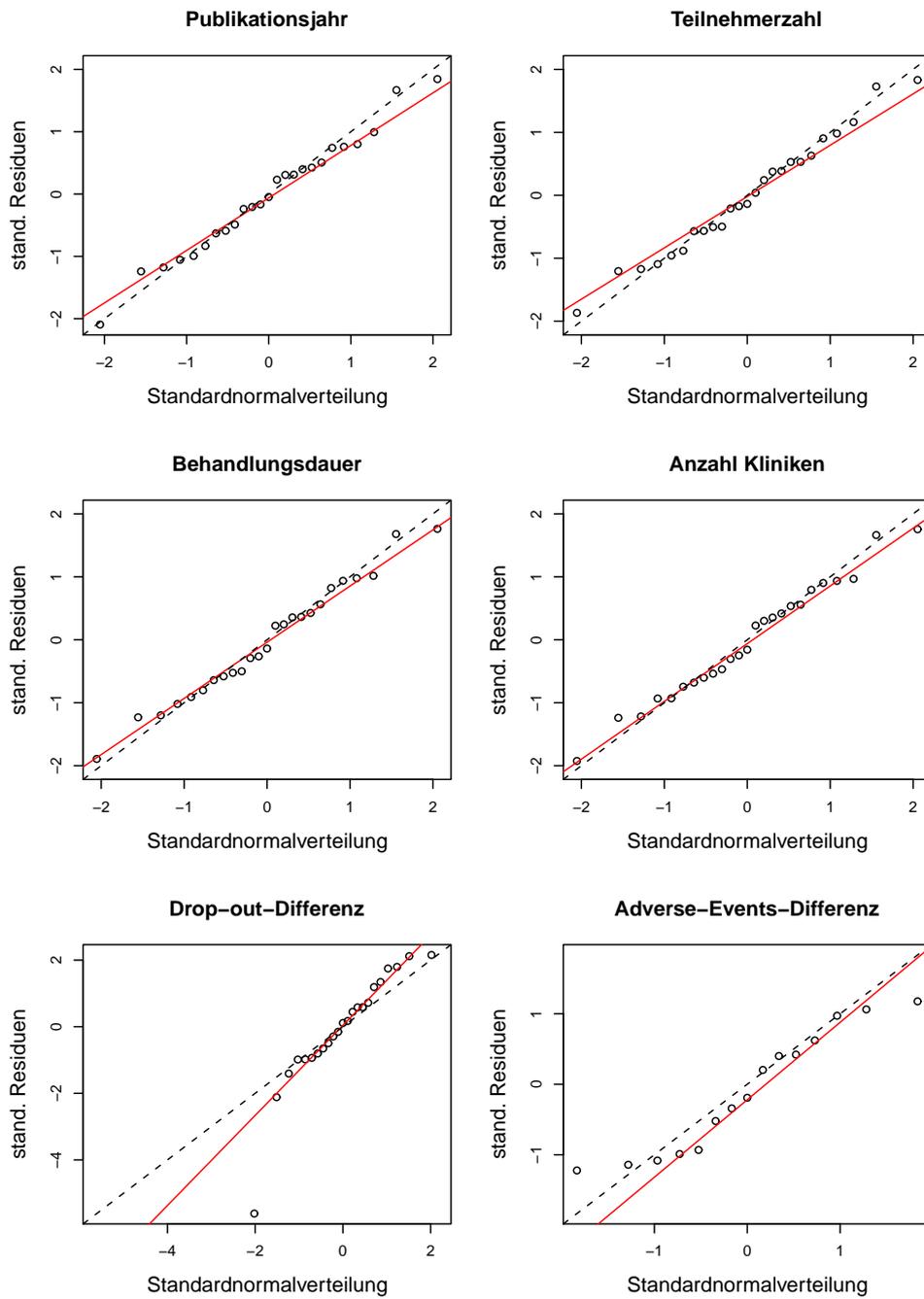


Abbildung B.19: NAL.DrDays

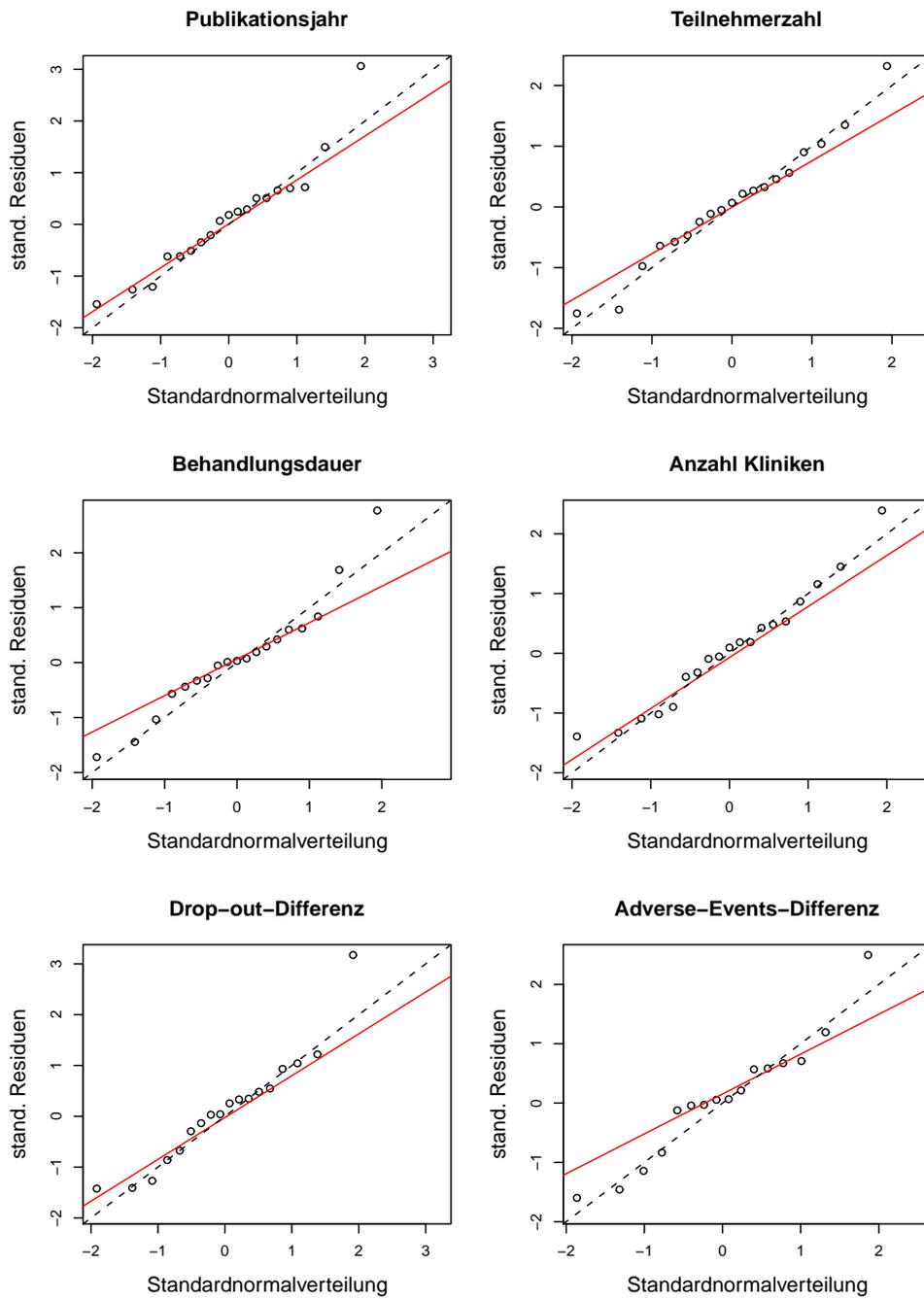


Abbildung B.20: ACA.DaysAbs

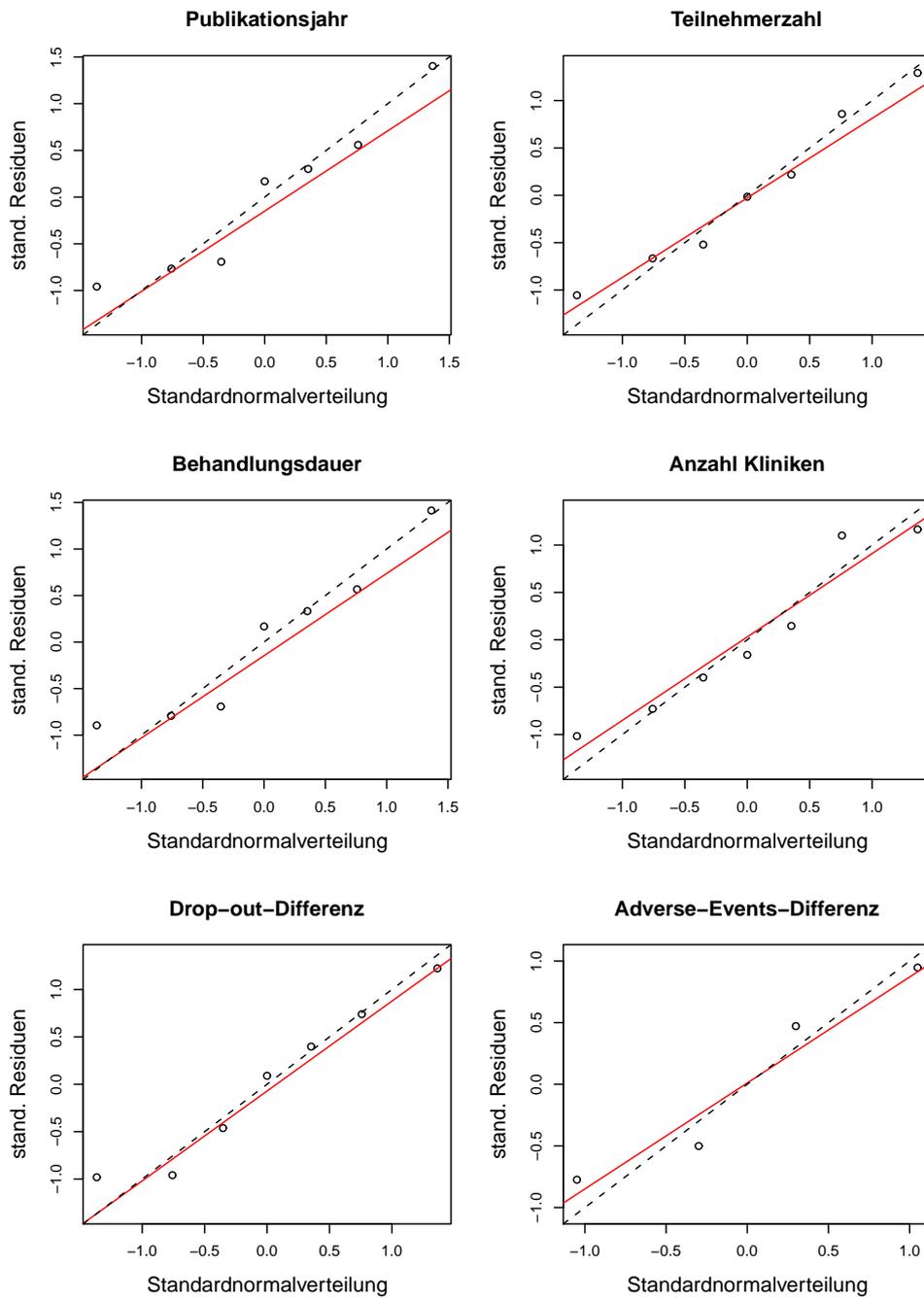


Abbildung B.21: ACA.GGT

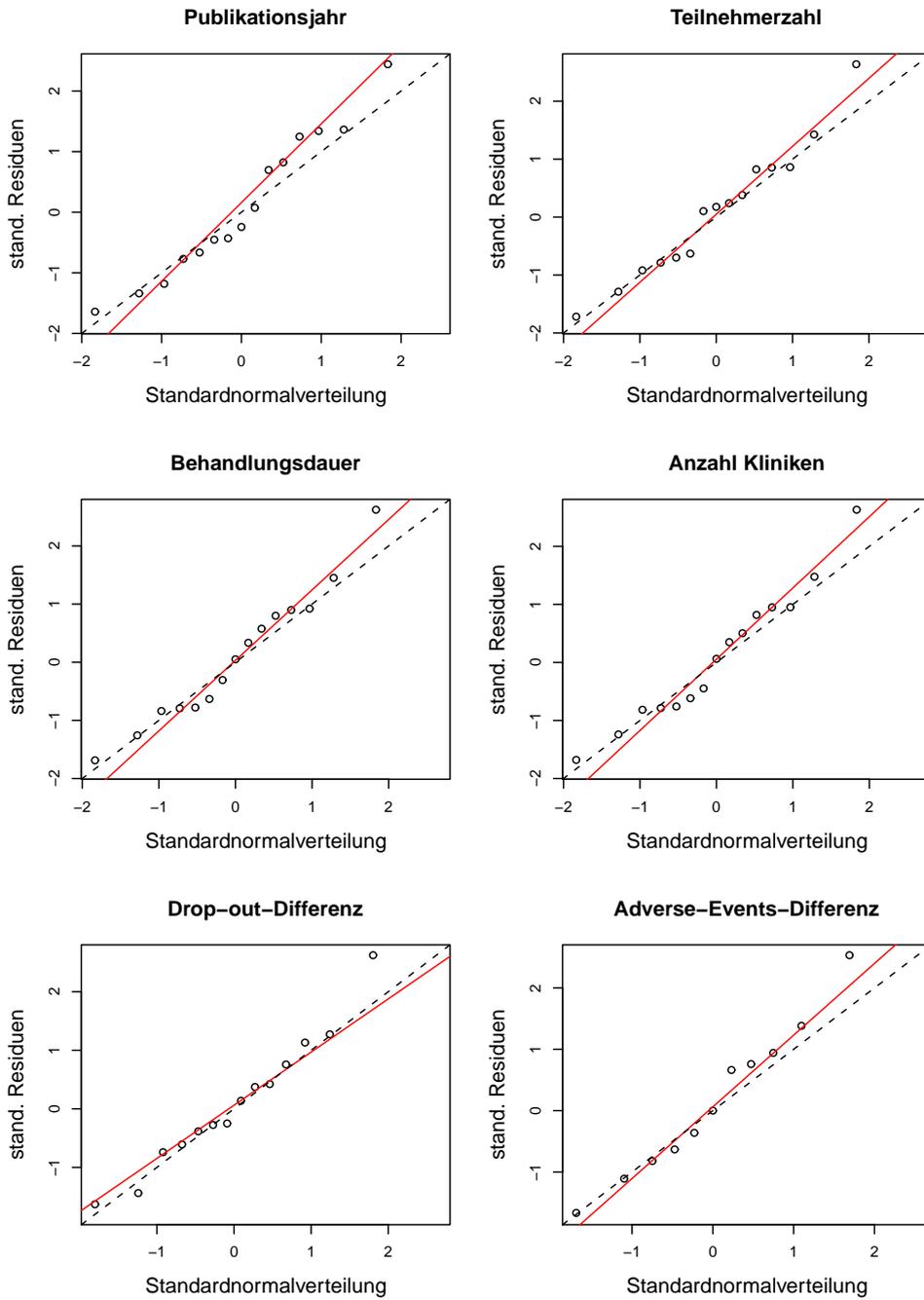


Abbildung B.22: NAL.GGT

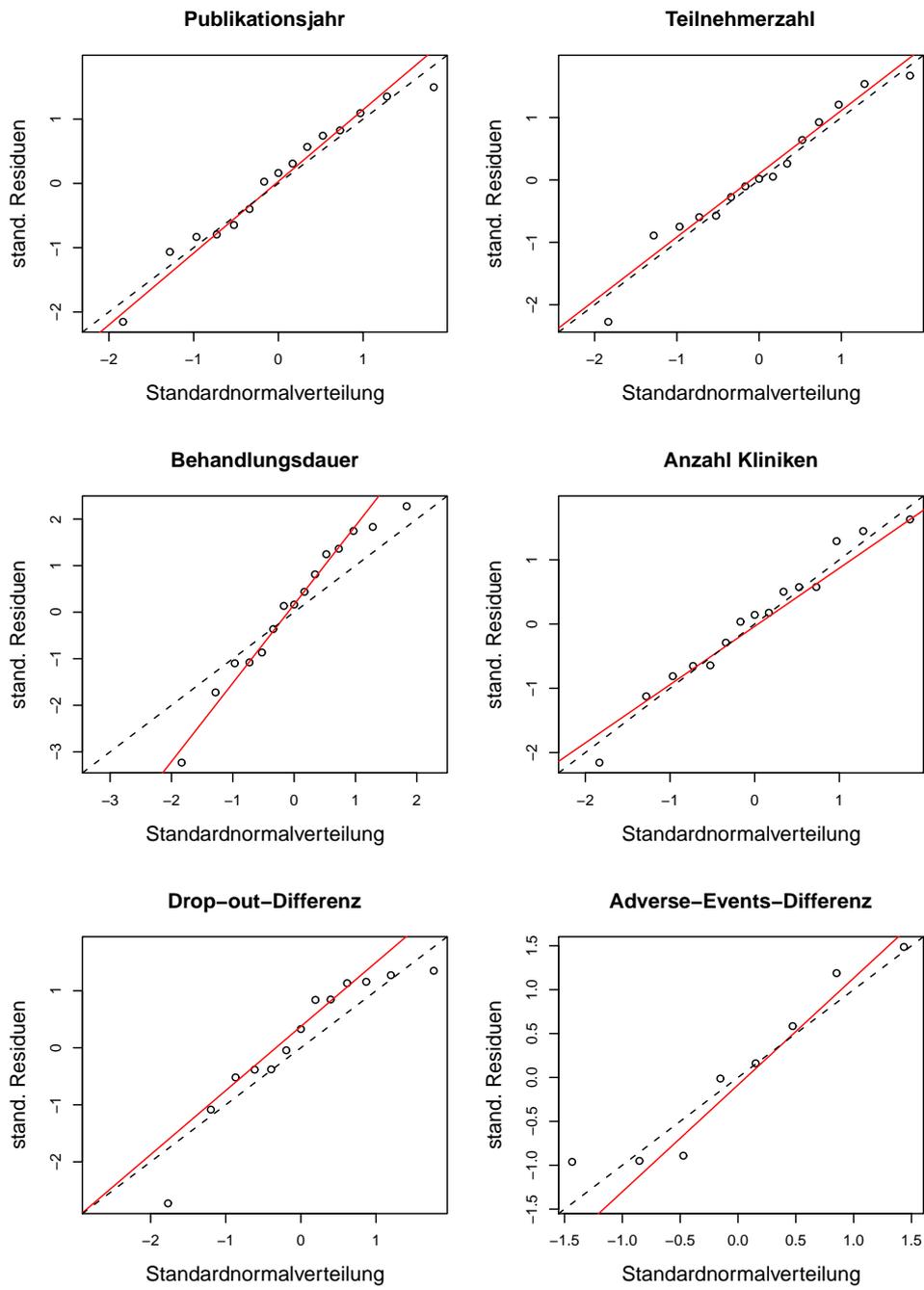


Abbildung B.23: NAL.Grams

Anhang C

Konfidenzintervall der I^2 -Statistik

Higgins & Thompson (2002) präsentieren zur Messung des Ausmaßes der Heterogenität neben I^2 noch eine andere Statistik:

$$H = \sqrt{\frac{\chi_{het}^2}{k-1}} \quad (C.1)$$

Dabei ist χ_{het}^2 die Teststatistik aus dem χ^2 -Homogenitätstest und k die Anzahl der Studien. I^2 und H haben offensichtlich folgende Beziehung:

$$I^2 = \frac{H^2 - 1}{H^2} \quad (C.2)$$

Für $\log(H)$ wird folgender Standardfehler angegeben, der auf der Approximation der χ^2 -Verteilung an die Normalverteilung für eine große Anzahl von Freiheitsgraden basiert:

$$SE_1[\log(H)] = \frac{1}{2} \frac{\log(\chi_{het}^2) - \log(k-1)}{\sqrt{2\chi_{het}^2} - \sqrt{2k-3}} \quad (C.3)$$

Wenn H von oben gegen eins geht, bzw. I^2 gegen null, so geht dieser Standardfehler gegen Null. Wie in Kapitel 2.6.3, erwähnt wird I^2 bei negativen Werten auf Null trunziert, also wird H äquivalent bei Werten kleiner eins auf eins gesetzt. Die Schätzung des Standardfehlers wäre dann aber wegen $\chi_{het}^2 \leq k-1$ kleiner gleich Null.

Daher muss für kleine χ_{het}^2 eine andere Schätzung verwendet werden, die auf der approximativen Varianz von $\log[\chi_{het}^2/(k-1)] = 2\log(H)$ basiert,

wobei χ_{het}^2 einer χ_{k-1}^2 -Verteilung folgt:

$$\text{SE}_0[\log(H)] = \sqrt{\left[\frac{1}{2(k-2)} \left(1 - \frac{1}{3(k-2)^2} \right) \right]} \quad (\text{C.4})$$

Diese Form sollte bereits angewendet werden, wenn χ_{het}^2 kleiner gleich k ist, nicht $k-1$, da sich sonst zu kleine Schätzungen für den Standardfehler ergeben würden.

Damit ergibt sich allgemein das 95%-Konfidenzintervall für H zu:

$$\exp\{ \log(H) \pm 1,96 \times \text{SE}[\log(H)] \} \quad (\text{C.5})$$

Wenn $\underline{H}_{0,95}$ bzw. $\overline{H}_{0,95}$ die obere bzw. untere Grenze des Konfidenzintervalls für H bezeichnen, lautet das Konfidenzintervall für I^2 wegen (C.2) Higgins & Thompson (2002):

$$\left[\frac{\underline{H}_{0,95}^2 - 1}{\underline{H}_{0,95}^2} ; \frac{\overline{H}_{0,95}^2 - 1}{\overline{H}_{0,95}^2} \right] \quad (\text{C.6})$$

Anhang D

R-Code

```
# Einlesen des Datensatzes:
#####

daten <- read.csv("Meta-Datensatz.csv")

#####
##   ERSTELLUNG VON DUMMY-VARIABLEN FÜR DIE KATEGORIALEN EINFLUßGRÖßEN:   ##
#####

# Herkunftsland_kategorisiert:
#####

# drei Kategorien ---> zwei Dummy-Variablen notwendig:

daten$Herkunftsland_kategorisiert_eur <- as.numeric(daten$
  Herkunftsland_kategorisiert == "Europe")
daten$Herkunftsland_kategorisiert_usa <- as.numeric(daten$
  Herkunftsland_kategorisiert == "USA")
daten$Herkunftsland_kategorisiert_eur[is.na(daten$
  Herkunftsland_kategorisiert)] <- NA
daten$Herkunftsland_kategorisiert_usa[is.na(daten$
  Herkunftsland_kategorisiert)] <- NA

# --> "andere Länder" ist Referenzkategorie.

# Sponsor_Industrie:
#####

# zwei Kategorien --> eine Dummy-Variable; Ausprägungen mit "unclear" auf
# NA setzen

daten$Sponsor_Industrie_yes <- as.numeric(daten$Sponsor_Industrie == "yes")
daten$Sponsor_Industrie_yes[daten$Sponsor_Industrie == "unclear"] <- NA

# --> "nicht gesponsort" ist Referenzkategorie.
```

```

# Psychotherapie_kategorisiert:
#####

# drei Kategorien --> zwei Dummy-Variable

daten$Psychotherapie_kategorisiert_CBT <- as.numeric(daten$
  Psychotherapie_kategorisiert == "CBT" | daten$Psychotherapie_kategorisiert
  == "CBT ")
daten$Psychotherapie_kategorisiert_MET <- as.numeric(daten$
  Psychotherapie_kategorisiert == "MET")

# --> "Andere" als Referenzkategorie.

# Ziel_der_Behandlung:
#####

# Zwei Kategorien: "Abstinenz" und "flexibel", wobei "flexibel", "harm reduction"
# und "subgroup specific" zu "flexibel" zusammengefasst wurden --> eine Dummy-Variable
# riabile notwendig:
# Ausprägungen mit "unclear" auf NA setzen

daten$Ziel_der_Behandlung_abst <- as.numeric(daten$Ziel_der_Behandlung ==
  "Abstinenz")
daten$Ziel_der_Behandlung_abst[daten$Ziel_der_Behandlung == "unclear"] <- NA

# Bei Compliance Differenz zwischen den Gruppen bilden:
#####

daten$Compliance_differenz <- daten$Compliance_Naltrexon - daten$Compliance_PBO

# Variable, die angibt, ob Multi-Center-Studie vorliegt oder nicht:
#####

daten$Multicenter <- as.numeric(daten$Anzahl_Kliniken != 1)

# Datensatz nach Substanz aufsplitten:
#####

ACA <- daten[daten$Substanz == "Acamprosat ",]
NAL <- daten[daten$Substanz == "Naltrexon",]

#####
##      ERSTELLUNG VON EINZELDATENSÄTZEN FÜR DIE VERSCHIEDENEN KONSTRUKTE:  ##
#####

# Funktion, die die Einzeldatensätze bei den binären Outcomes erstellt:
#####

```

```

outcome.daten.bin <- function(datensatz, ntrt, nctrl, ptrt, pctrl) {

  datensatz2 <- datensatz[,c(1:36,68:75)]
  datensatz2$V45 <- datensatz[,ntrt]
  datensatz2$V46 <- datensatz[,nctrl]
  datensatz2$V47 <- datensatz[,ptrt]
  datensatz2$V48 <- datensatz[,pctrl]
  names(datensatz2)[45:48] <- names(datensatz)[c(ntrt, nctrl, ptrt, pctrl)]
  a <- datensatz[,ptrt]; b <- datensatz[,ntrt] - datensatz[,ptrt]; c <-
  datensatz[,pctrl]; d <- datensatz[,nctrl] - datensatz[,pctrl]; ntr <-
  datensatz[,ntrt]; npl <- datensatz[,nctrl]
  datensatz2$logRR <- log((a/ntr)/(c/npl))
  datensatz2$se.logRR <- sqrt((mean(b/a, na.rm = T)/ntr) + (mean(d/c, na.rm = T)
  /npl))
  datensatz2 <- datensatz2[!is.na(datensatz2$logRR),]
  return(datensatz2)

}

# Any Drinking:
#####

ACA.AD <- outcome.daten.bin(ACA, 38, 40, 37, 39)
NAL.AD <- outcome.daten.bin(NAL, 38, 40, 37, 39)

# Heavy Drinking:
#####

ACA.HD <- outcome.daten.bin(ACA, 43, 45, 42, 44)
NAL.HD <- outcome.daten.bin(NAL, 43, 45, 42, 44)

# Funktion, die die Einzeldatensätze bei den stetigen Outcomes erstellt:
#####

outcome.daten.diff <- function(datensatz, meanT, sdT, nT, meanP, sdP, nP) {

  datensatz2 <- datensatz[,c(1:36,68:75)]
  datensatz2$V45 <- datensatz[,meanT]
  datensatz2$V46 <- datensatz[,sdT]
  datensatz2$V47 <- datensatz[,nT]
  datensatz2$V48 <- datensatz[,meanP]
  datensatz2$V49 <- datensatz[,sdP]
  datensatz2$V50 <- datensatz[,nP]
  names(datensatz2)[45:50] <- names(datensatz)[c(meanT, sdT, nT, meanP, sdP, nP)]
  meanT <- datensatz[,meanT]; sdT <- datensatz[,sdT]; nT <- datensatz[,nT]
  meanP <- datensatz[,meanP]; sdP <- datensatz[,sdP]; nP <- datensatz[,nP]
  datensatz2$diff <- meanT - meanP
  datensatz2$se.pooled <- sqrt( ((1/nT)+(1/nP)) * ((sdT^2)*(nT-1) + (sdP^2)*
  (nP-1))/(nT+nP-2) )
  datensatz2 <- datensatz2[!is.na(datensatz2$diff),]
  return(datensatz2)

}

```

```

# Drinking Days:
#####

ACA.DrDays <- outcome.daten.diff(ACA, 47, 48, 49, 50, 51, 52)
NAL.DrDays <- outcome.daten.diff(NAL, 47, 48, 49, 50, 51, 52)

# Days Abstinent:
#####

ACA.DaysAbs <- outcome.daten.diff(ACA, 54, 55, 56, 57, 58, 59)
NAL.DaysAbs <- outcome.daten.diff(NAL, 54, 55, 56, 57, 58, 59)

# GGT:
#####

ACA.GGT <- outcome.daten.diff(ACA, 61, 62, 63, 64, 65, 66)
NAL.GGT <- outcome.daten.diff(NAL, 61, 62, 63, 64, 65, 66)
# in einer Studien fehlten Angaben zur Schätzung der Effektgröße -->
# vernachlässigen:
NAL.GGT2 <- NAL.GGT[-2,]

# Erstellung des NAL.Grams-Datensatzes:
#####

gramspddd <- read.table("GramsPerDrinkingDay.txt", header = T)

NAL.Grams <- daten[,c(1:36,68:73)]

gram_indezi <- 0
k <- 1
for(j in seq(along = daten[,1])) {
  if(any((as.character(gramspddd$Autor_grams) == as.character(
    daten$Studie)[j]) &
    (as.character(gramspddd$Jahr_grams) == as.character(
    daten$Publikationsjahr)[j]))) {
    gram_indezi[k] <- j
    k <- k+1
  }
}

NAL.Grams <- NAL.Grams[gram_indezi,]
NAL.Grams$V43 <- gramspddd$Grams_pdday_Nal_Mean
NAL.Grams$V44 <- gramspddd$Grams_pdday_Nal_SD
NAL.Grams$V45 <- gramspddd$Grams_pdday_Nal_n
NAL.Grams$V46 <- gramspddd$Grams_pdday_Placebo_Mean
NAL.Grams$V47 <- gramspddd$Grams_pdday_Placebo_SD
NAL.Grams$V48 <- gramspddd$Grams_pdday_Placebo_n
names(NAL.Grams)[43:48] <- names(gramspddd)[3:8]
meanT <- NAL.Grams[,43]; sdT <- NAL.Grams[,44]; nT <- NAL.Grams[,45]
meanP <- NAL.Grams[,46]; sdP <- NAL.Grams[,47]; nP <- NAL.Grams[,48]
NAL.Grams$diff <- meanT - meanP
NAL.Grams$se.pooled <- sqrt( ((1/nT)+(1/nP)) * ((sdT^2)*(nT-1) + (sdP^2)*(nP-1))/
  (nT+nP-2) )
# Die letzte Studie behandelte Acomprosat --> vernachlässigen:
NAL.Grams <- NAL.Grams[-16,]

```

```

# Größen der Datensätze:
#####

length(ACA.AD[,1])
length(NAL.AD[,1])
length(ACA.HD[,1])
length(NAL.HD[,1])
length(ACA.DrDays[,1])
length(NAL.DrDays[,1])
length(ACA.DaysAbs[,1])
length(NAL.DaysAbs[,1])
length(ACA.GGT[,1])
length(NAL.GGT2[,1])
length(NAL.Grams[,1])

# --> Für ACA.DrDays und NAL.DaysAbs stehen jeweils nur zwei Beobachtungen
# zur Verfügung. --> Damit kann keine Meta-Regression durchgeführt werden.
# Für ACA.HD stehen nur 6 Studien zur Verfügung, für ACA.GGT nur 7.
# ---> Auch sehr wenig. ( <--> mindestens zehn Beobachtungen für Meta-
# Regression.)
# Bei ACA.DaysAbs, NAL.GGT2 und NAL.Grams sind zwischen 10 und 20 Studien
# verfügbar.
# ---> Dort kann nur eine Kovariable im Modell sein.
# Bei ACA.AD, NAL.AD, NAL.HD und NAL.DrDays sind zwischen 20 und 30 Studien
# verfügbar.
# ---> Dort könnten auch 2 Kovariablen im Modell sein.
#
# ---> Schließe ACA.DrDays und NAL.DaysAbs im weiteren aus.

#####
##                UNTERSUCHUNG DER HETEROGENITÄT:                ##
#####

# Funktion, die den Standard-Heterogenitäts-Test durchführt und die
# I2-Heterogenitätsstatistik mit KI berechnet:
#####

het.test.I2 <- function(effekte, sds) {

  weights <- 1/(sds^2)
  ges.eff <- sum(weights*effekte)/sum(weights)
  X2 <- sum(weights*((effekte - ges.eff)^2))
  k <- length(effekte)
  p <- 1 - pchisq(X2, df = (k-1))
  I2 <- ifelse( ((X2 - (k-1))/X2)>= 0, (X2 - (k-1))/X2, 0)
  H <- ifelse(sqrt(X2/(k - 1)) > 1, sqrt(X2/(k - 1)), 1)
  if(X2 <= k) {
    selogH <- sqrt( (1/(2*(k-2)))*(1 - (1/(3*((k-2)^2)))) )
  }
  else
    selogH <- (log(X2) - log(k - 1))/(2*(sqrt(2*X2) - sqrt(2*k - 3)))
  kiH <- H*exp(c(-1,1)*qnorm(0.975)*selogH)
  kiI2 <- ((kiH^2) - 1)/(kiH^2)

  return(list(I2 = I2, KI_I2 = kiI2, X2 = X2, p = p, n = k))
}

```

```

}

# Forest Plots für die einzelnen Outcomes:
#####

library(rmeta)

par(mfrow = c(2,5))
metaplot(mn = ACA.AD$logRR, se = ACA.AD$se.logRR, xlab = "log. RR", ylab = "",
  main = "Acamprosats - AD", cex.main = 1.4, cex.lab = 1.35)
metaplot(mn = NAL.AD$logRR, se = NAL.AD$se.logRR, xlab = "log. RR", ylab = "",
  main = "Naltrexon - AD", cex.main = 1.4, cex.lab = 1.35)
metaplot(mn = ACA.HD$logRR, se = ACA.HD$se.logRR, xlab = "log. RR", ylab = "",
  main = "Acamprosats - HD", cex.main = 1.4, cex.lab = 1.35)
metaplot(mn = NAL.HD$logRR, se = NAL.HD$se.logRR, xlab = "log. RR", ylab = "",
  main = "Naltrexon - HD", cex.main = 1.4, cex.lab = 1.35)
metaplot(mn = NAL.DrDays$diff, se = NAL.DrDays$se.pooled, xlab = "log. RR",
  ylab = "", main = "Naltrexon - DrDays", cex.main = 1.4, cex.lab = 1.35)
metaplot(mn = ACA.DaysAbs$diff, se = ACA.DaysAbs$se.pooled, xlab = "log. RR",
  ylab = "", main = "Acamprosats - DaysAbs", cex.main = 1.4, cex.lab = 1.35)
metaplot(mn = ACA.GGT$diff, se = ACA.GGT$se.pooled, xlab = "log. RR", ylab = "",
  main = "Acamprosats - GGT", cex.main = 1.4, cex.lab = 1.35)
metaplot(mn = NAL.GGT2$diff, se = NAL.GGT2$se.pooled, xlab = "log. RR",
  ylab = "", main = "Naltrexon - GGT", cex.main = 1.4, cex.lab = 1.35)
metaplot(mn = NAL.Grams$diff, se = NAL.Grams$se.pooled, xlab = "log. RR",
  ylab = "", main = "Naltrexon - Grams", cex.main = 1.4, cex.lab = 1.35)
par(mfrow = c(1,1))

# "metaplot" setzt yaxt in par auf "n" --> umändern nach "s":
par(yaxt = "s")

# I2-Statistiken und Heterogenitätstests:
#####

het.test.I2(ACA.AD$logRR, ACA.AD$se.logRR)
het.test.I2(NAL.AD$logRR, NAL.AD$se.logRR)
het.test.I2(ACA.HD$logRR, ACA.HD$se.logRR)
het.test.I2(NAL.HD$logRR, NAL.HD$se.logRR)
het.test.I2(NAL.DrDays$diff, NAL.DrDays$se.pooled)
het.test.I2(ACA.DaysAbs$diff, ACA.DaysAbs$se.pooled)
het.test.I2(ACA.GGT$diff, ACA.GGT$se.pooled)
het.test.I2(NAL.GGT2$diff, NAL.GGT2$se.pooled)
het.test.I2(NAL.Grams$diff, NAL.Grams$se.pooled)

#####
##          VERSCHIEDENE SCHÄTZER FÜR DIE ZWISCHENSTUDIENVARIANZ:          ##
#####

# Die Momenten-Methode von DerSimonian & Laird:
#####

```

```

zw.var <- function(means,sds) {

  means <- means[!is.na(means)]
  sds <- sds[!is.na(sds)]
  w <- 1/(sds^2)
  Q <- sum(w*(means^2)) - (((sum(w*means))^2)/sum(w))
  zwvar <- (Q - (length(w)-1)/(sum(w) - (sum(w^2)/sum(w))))
  return(tau2 = ifelse(zwvar > 0, zwvar, 0))

}

# Die Varianz-Komponenten-Methode:
#####

VC <- function(effekte,sds) {

  k <- length(effekte)
  tau2VC <- (1/(k-1))*sum((effekte - mean(effekte))^2) - mean(sds^2)
  return(ifelse(tau2VC>0,tau2VC,0))

}

# Die Modell-Error-Varianz-Methode:
#####

MV <- function(effekte,sds) {

  k <- length(effekte)
  r <- (sds^2)/(sum((effekte - mean(effekte))^2)/k)
  meanYv <- sum((1/(r+1))*effekte)/sum(1/(r+1))
  tau2MV <- (1/(k-1))*sum((1/(r+1))*((effekte - meanYv)^2))
  return(tau2MV)

}

# Die alternative Modell-Error-Varianz-Methode:
#####

MVvc <- function(effekte,sds) {

  k <- length(effekte)
  r <- (sds^2)/ifelse(VC(effekte,sds) == 0, 0.01, VC(effekte,sds))
  meanYv <- sum((1/(r+1))*effekte)/sum(1/(r+1))
  tau2MV <- (1/(k-1))*sum((1/(r+1))*((effekte - meanYv)^2))
  return(tau2MV)

}

# Die empirische Bayes-Methode:
#####

EB <- function(thetai, sigma) {

  tau2 <- 0
  tau2_old <- 1
  k <- length(thetai)
  w <- 1/(sigma^2 + tau2)
  iter <- 0

```

```

while(abs(tau2_old-tau2)>0.0000001) {
  tau2_old <- tau2
  theta_w <- sum(w*thetai)/sum(w)
  tau2 <- sum(w*((k/(k-1))*((thetai - theta_w)^2) - (sigma^2)) ) /sum(w)
  if(tau2 < 0)
    tau2 <- 0
  w <- 1/(sigma^2 + tau2)
  iter <- iter+1
}
return(tau2 = tau2)
}

# Die approximative restringierte Maximum-Likelihood-Methode:
#####

REML <- function(thetai, sigma) {
  tau2 <- 0
  tau2_old <- 1
  k <- length(thetai)
  w <- 1/(sigma^2 + tau2)
  iter <- 0
  while(abs(tau2_old-tau2)>0.0000001) {
    tau2_old <- tau2
    theta_w <- sum(w*thetai)/sum(w)
    tau2 <- sum((w^2)*(k/(k-1))*((thetai - theta_w)^2) - (sigma^2)) ) /sum(w^2)
    if(tau2 < 0)
      tau2 <- 0
    w <- 1/(sigma^2 + tau2)
    iter <- iter+1
  }
  return(tau2 = tau2)
}

#####
##          KORRELATIONEN ZWISCHEN DEN (QUASI)STETIGEN KOVARIABLEN          ##
#####

```

```
#####

# Funktion zur Auswahl der relevanten Kovariablen:
#####

Kov <- function(datensatz, dummies = F) {
  if(dummies == F) {
    datensatz[,names(datensatz) %in% c("Publikationsjahr", "N_total",
    "Behandlungsdauer_in_Wochen", "Anzahl_Kliniken",
    "Herkunftsland_kategorisiert", "Sponsor_Industrie",
    "Psychotherapie_kategorisiert", "Ziel_der_Behandlung", "Drop.out_Differenz",
    "Adverse_events_differenz")]
  }
  else {
    datensatz[,names(datensatz) %in% c("Publikationsjahr", "N_total",
    "Behandlungsdauer_in_Wochen", "Anzahl_Kliniken", "Drop.out_Differenz",
    "Adverse_events_differenz", "Herkunftsland_kategorisiert_usa",
    "Herkunftsland_kategorisiert_eur", "Sponsor_Industrie_yes",
    "Psychotherapie_kategorisiert_CBT", "Psychotherapie_kategorisiert_MET",
    "Ziel_der_Behandlung_abst")]
  }
}

# Berechnen der Korrelationen:
#####

daten.typ.num <- Kov(daten)[,c(1,2,3,4,9,10)]
cor.dats <- cor(daten.typ.num, use = "pairwise.complete.obs")
# fix(cor.dats)

# BEMERKUNG: Zur Berechnung der Korrelationen zwischen den Kovariablen,
# wurden alle Studien, nicht getrennt nach den Konstrukten verwendet,
# da angenommen wurde, dass sich die Stärken der Korrelationen nicht
# zwischen den Konstrukten unterscheiden.

#####
##                                VORBEREITUNG DER REGRESSION                                ##
#####

# Funktion zur iterativen Schätzung der Regressionsmodelle :
#####

model.REML <- function(thetai, x, sigma, bv.estimator = "REML") {

  if(bv.estimator == "MV")
    tau2woco <- MV(thetai, sigma)
  if(bv.estimator == "MVvc")
    tau2woco <- MVvc(thetai, sigma)
  if(bv.estimator == "VC")
    tau2woco <- VC(thetai, sigma)
  if(bv.estimator == "zw.var")
    tau2woco <- zw.var(thetai, sigma)
  if(bv.estimator == "EB")
```

```

    tau2woco <- EB(thetai, sigma)
    if(bv.estimator == "REML")
      tau2woco <- REML(thetai, sigma)

    x <- as.matrix(x)
    which <- apply(x, 1, function(y) !any(is.na(y)) )
    thetai <- thetai[which]
    sigma <- sigma[which]
    x <- as.matrix(x[which,])

    tau2 <- 0
    tau2_old <- 1
    k <- length(thetai)
    w <- 1/(sigma^2 + tau2)
    iter <- 0

    while(abs(tau2_old-tau2)>0.0000001) {

      model <- lm(thetai ~ x, weights = w)
      betas <- model$coef

      if(any(is.na(betas)))
        stop("Modell nicht schätzbar!")

      tau2_old <- tau2
      tau2 <- sum((w^2)*((k/(k-(dim(x)[2]+1)))*((thetai - cbind(1,x)%*%betas)^2) -
        (sigma^2)) ) /sum(w^2)

      if(tau2 < 0)
        tau2 <- 0

      w <- 1/(sigma^2 + tau2)

      iter <- iter+1
    }

    model.final <- lm(thetai ~ x, weights = w)

    beta <- summary(model.final)$coef[1:(dim(x)[2]+1)]
    sd.beta <- summary(model.final)$coef[(dim(x)[2]+2):((dim(x)[2] + 1)*2)]/
      summary(model.final)$sigma
    p.beta <- 2*(1 - pt(abs(beta/sd.beta),df = k-2))

    model <- matrix(nrow = dim(x)[2]+1,ncol = 3,data = c(beta,sd.beta,p.beta))
    rownames(model) <- paste("beta",0:dim(x)[2],sep = "_")
    colnames(model) <- c("Estimate","Std. Error","Pr(>|t|)")

    expl.heter <- (tau2woco - tau2)/tau2woco
    expl.heter <- ifelse(expl.heter > 0, expl.heter, 0)

    return(list(Coefficients = model, tau2 = tau2, iters = iter, het.before.propexpl
      = c(tau2woco,expl.heter)))
  }

# Funktion zum Plotten der Streudiagramme mit der Regressionsgerade mit Punkten

```

```

# proportional zu den Gewichten der Studien:
#####

plot.metareg <- function(thetai,x,sigma,color.line = "blue",bv.estimator =
  "REML") {

  y.lab <- deparse(substitute(thetai))
  y.lab <- strsplit(y.lab,"")[[1]]
  wo.split <- (1:length(y.lab))[y.lab == "$"]
  main <- y.lab[1 : (wo.split-1)]
  y.lab <- y.lab[(wo.split+1) : length(y.lab)]
  main2 <- ""
  for(i in seq(along = main)) {
    main2 <- paste(main2,main[i],sep = "")
  }
  y.lab2 <- ""
  for(i in seq(along = y.lab)) {
    y.lab2 <- paste(y.lab2,y.lab[i],sep = "")
  }

  x.lab <- deparse(substitute(x))
  x.lab <- strsplit(x.lab,"")[[1]]
  x.lab <- x.lab[(wo.split+1) : length(x.lab)]
  x.lab2 <- ""
  for(i in seq(along = x.lab)) {
    x.lab2 <- paste(x.lab2,x.lab[i],sep = "")
  }

  if(x.lab2 == "N_total")
    x.lab2 <- "Teilnehmerzahl"
  if(x.lab2 == "Behandlungsdauer_in_Wochen")
    x.lab2 <- "Behandlungsdauer"
  if(x.lab2 == "Anzahl_Kliniken")
    x.lab2 <- "Anzahl Kliniken"
  if(x.lab2 == "Drop.out_Differenz")
    x.lab2 <- "Drop-out-Differenz"
  if(x.lab2 == "Adverse_events_differenz")
    x.lab2 <- "Adverse-Events-Differenz"

  if(y.lab2 == "logRR")
    y.lab2 <- "log. rel. Risiko"
  if(y.lab2 == "diff")
    y.lab2 <- "mittlere Differenz"

  limo <- model.REML(thetai, x, sigma, bv.estimator = bv.estimator)
  weights <- 1/(sigma^2 + limo$tau2)

  point.sizes <- sqrt(weights/pi)/min(sqrt(weights/pi)) # main = main2,
  if((max(point.sizes)/min(point.sizes)) > 20)
    point.sizes <- 0.5*sqrt(weights/pi)/min(sqrt(weights/pi)) # main = main2,

  plot(x, thetai, cex = point.sizes, xlab = x.lab2, ylab = y.lab2,cex.lab = 1.5)
  x1 <- c(min(x, na.rm = T)-10, max(x, na.rm = T)+10)
  x2 <- limo$Coef[1] + limo$Coef[2]*x1
  lines(x1, x2, col = color.line)
  p.value <- limo$Coef[6]

  if(is.na(limo$het[2]))
    het.exp <- "---"

```

```

else
  het.exp <- round(limo$het[2],2)
  legend("bottomright",legend = c(paste(" p = ",round(p.value,4)),
    paste(" erkl. Het. = ", het.exp)),bty = "n",cex = 1.5)
}

# Funktion, die die Normal-Quantil-Plots zeichnet:
#####

qqn.model <- function(thetai,x,sds) {

  x.lab <- deparse(substitute(x))
  x.lab <- strsplit(x.lab,"")[[1]]
  wo.split <- (1:length(x.lab))[x.lab == "$"]
  main <- x.lab[1 : (wo.split-1)]
  x.lab <- x.lab[(wo.split+1) : length(x.lab)]
  main2 <- ""
  for(i in seq(along = main)) {
    main2 <- paste(main2,main[i],sep = "")
  }
  x.lab2 <- ""
  for(i in seq(along = x.lab)) {
    x.lab2 <- paste(x.lab2,x.lab[i],sep = "")
  }

  if(x.lab2 == "N_total")
    x.lab2 <- "Teilnehmerzahl"
  if(x.lab2 == "Behandlungsdauer_in_Wochen")
    x.lab2 <- "Behandlungsdauer"
  if(x.lab2 == "Anzahl_Kliniken")
    x.lab2 <- "Anzahl Kliniken"
  if(x.lab2 == "Drop.out_Differenz")
    x.lab2 <- "Drop-out-Differenz"
  if(x.lab2 == "Adverse_events_differenz")
    x.lab2 <- "Adverse-Events-Differenz"

  model <- model.REML(thetai,x,sds)
  res.stand <- (thetai - model$Coef[1] - model$Coef[2]*x)/sqrt(model$tau2 + sds^2)

  qqnorm(res.stand,xlim = range(qqnorm(res.stand, plot = F),na.rm = T),ylim =
    range(qqnorm(res.stand, plot = F),na.rm = T), main = x.lab2,xlab =
    "Standardnormalverteilung", ylab = "stand. Residuen",cex.main = 1.4,
    cex.lab = 1.4)
  qqline(res.stand, col = 2)
  abline(c(0,1),lty = 2)
}

#####
##   EINFACHE META-REGRESSIONEN MIT STREUDIAGRAMMEN UND NORMAL-QQ-PLOTS   ##
#####

# EXEMPLARISCH FÜR ACA.AD; FÜR DIE ANDEREN KONSTRUKTE WURDE ANALOG VORGEGANGEN

```

```

# Alle Modelle mit einer Kovariablen:
#####

(ACA.AD.jahr <- model.REML(ACA.AD$logRR, ACA.AD$Publikationsjahr, ACA.AD$
  se.logRR))
(ACA.AD.N <- model.REML(ACA.AD$logRR, ACA.AD$N_total, ACA.AD$se.logRR))
(ACA.AD.dauer <- model.REML(ACA.AD$logRR, ACA.AD$Behandlungsdauer_in_Wochen,
  ACA.AD$se.logRR))
(ACA.AD.anzkllinik <- model.REML(ACA.AD$logRR, ACA.AD$Anzahl_Kliniken,
  ACA.AD$se.logRR))
(ACA.AD.dropdiff <- model.REML(ACA.AD$logRR, ACA.AD$Drop.out_Differenz,
  ACA.AD$se.logRR))
(ACA.AD.advediff <- model.REML(ACA.AD$logRR, ACA.AD$Adverse_events_differenz,
  ACA.AD$se.logRR))
(ACA.AD.landkat <- model.REML(ACA.AD$logRR, cbind(ACA.AD$
  Herkunftsland_kategorisiert_eur, ACA.AD$Herkunftsland_kategorisiert_usa),
  ACA.AD$se.logRR))
(ACA.AD.sponsor <- model.REML(ACA.AD$logRR, ACA.AD$Sponsor_Industrie_yes,
  ACA.AD$se.logRR))
(ACA.AD.psych <- model.REML(ACA.AD$logRR, ACA.AD$Psychotherapie_kategorisiert_CBT,
  ACA.AD$se.logRR))
(ACA.AD.zielbeh <- model.REML(ACA.AD$logRR, ACA.AD$Ziel_der_Behandlung_abst,
  ACA.AD$se.logRR))

# Streudiagramme für Modelle mit (quasi) stetigen Kovariablen:
#####

par(mfrow = c(3,2))
plot.metareg(ACA.AD$logRR, ACA.AD$Publikationsjahr, ACA.AD$se.logRR)
plot.metareg(ACA.AD$logRR, ACA.AD$N_total, ACA.AD$se.logRR)
plot.metareg(ACA.AD$logRR, ACA.AD$Behandlungsdauer_in_Wochen, ACA.AD$se.logRR)
plot.metareg(ACA.AD$logRR, ACA.AD$Anzahl_Kliniken, ACA.AD$se.logRR)
plot.metareg(ACA.AD$logRR, ACA.AD$Drop.out_Differenz, ACA.AD$se.logRR)
plot.metareg(ACA.AD$logRR, ACA.AD$Adverse_events_differenz, ACA.AD$se.logRR)
par(mfrow = c(1,1))

# Quantil-Plots:
#####

par(mfrow = c(3,2))
qqn.model(ACA.AD$logRR, ACA.AD$Publikationsjahr, ACA.AD$se.logRR)
qqn.model(ACA.AD$logRR, ACA.AD$N_total, ACA.AD$se.logRR)
qqn.model(ACA.AD$logRR, ACA.AD$Behandlungsdauer_in_Wochen, ACA.AD$se.logRR)
qqn.model(ACA.AD$logRR, ACA.AD$Anzahl_Kliniken, ACA.AD$se.logRR)
qqn.model(ACA.AD$logRR, ACA.AD$Drop.out_Differenz, ACA.AD$se.logRR)
qqn.model(ACA.AD$logRR, ACA.AD$Adverse_events_differenz, ACA.AD$se.logRR)
par(mfrow = c(1,1))

#####
##                               META-REGRESSIONEN ÜBER BEIDE SUBSTANZEN                               ##
#####

```

```

##          "SUBSTANZ" DABEI ALS KOVARIABLE EINBEZOGEN          ##
#####

# Funktion zur Erstellung der Gesamtdatensätze für stetige Outcomes:
#####

outcome.daten.bin.ges <- function(datensatz, ntrt, nctrl, ptrt, pctrl) {

  datensatz2 <- datensatz[,c(1:36,68:76)]
  datensatz2$V46 <- datensatz[,ntrt]
  datensatz2$V47 <- datensatz[,nctrl]
  datensatz2$V48 <- datensatz[,ptrt]
  datensatz2$V49 <- datensatz[,pctrl]
  names(datensatz2)[46:49] <- names(datensatz)[c(ntrt, nctrl, ptrt, pctrl)]
  a <- datensatz[,ptrt]; b <- datensatz[,ntrt] - datensatz[,ptrt]; c <-
    datensatz[,pctrl]
  d <- datensatz[,nctrl] - datensatz[,pctrl]; ntr <- datensatz[,ntrt]; npl <-
    datensatz[,nctrl]
  datensatz2$logRR <- log((a/ntr)/(c/npl))
  datensatz2$se.logRR <- sqrt((mean(b/a, na.rm = T)/ntr) + (mean(d/c, na.rm = T)
    /npl))
  datensatz2 <- datensatz2[!is.na(datensatz2$logRR),]
  return(datensatz2)

}

# Funktion zur Erstellung der Gesamtdatensätze für stetige Outcomes:
#####

outcome.daten.diff.ges <- function(datensatz, meanT, sdT, nT, meanP, sdP, nP) {

  datensatz2 <- datensatz[,c(1:36,68:76)]
  datensatz2$V46 <- datensatz[,meanT]
  datensatz2$V47 <- datensatz[,sdT]
  datensatz2$V48 <- datensatz[,nT]
  datensatz2$V49 <- datensatz[,meanP]
  datensatz2$V50 <- datensatz[,sdP]
  datensatz2$V51 <- datensatz[,nP]
  names(datensatz2)[46:51] <- names(datensatz)[c(meanT, sdT, nT, meanP, sdP, nP)]
  meanT <- datensatz[,meanT]; sdT <- datensatz[,sdT]; nT <- datensatz[,nT]
  meanP <- datensatz[,meanP]; sdP <- datensatz[,sdP]; nP <- datensatz[,nP]
  datensatz2$diff <- meanT - meanP
  datensatz2$se.pooled <- sqrt( ((1/nT)+(1/nP)) * ((sdT^2)*(nT-1) + (sdP^2)*
    (nP-1))/(nT+nP-2) )
  datensatz2 <- datensatz2[!is.na(datensatz2$diff),]
  return(datensatz2)

}

# Einzeldatensätze NAL und ACA verbinden:
#####

NAL.ACA <- rbind(NAL,ACA)

# Dummy-Variable zur Unterscheidungen der Substanzen:

```

```
#####
NAL.ACA$Substanz_NAL <- as.numeric(NAL.ACA$Substanz == "Naltrexon")

# Erstellung der Einzeldatensätze:
#####

GES.AD <- outcome.daten.bin.ges(NAL.ACA, 38, 40, 37, 39)

GES.HD <- outcome.daten.bin.ges(NAL.ACA, 43, 45, 42, 44)

GES.DrDays <- outcome.daten.diff.ges(NAL.ACA, 47, 48, 49, 50, 51, 52)

GES.DaysAbs <- outcome.daten.diff.ges(NAL.ACA, 54, 55, 56, 57, 58, 59)

GES.GGT <- outcome.daten.diff.ges(NAL.ACA, 61, 62, 63, 64, 65, 66)
GES.GGT2 <- GES.GGT[-2,]

# META-REGRESSION EXEMPLARISCH FÜR GES.AD; FÜR DIE ANDEREN KONSTRUKTE WURDE
# ANALOG VORGEANGEN
#####

(GES.AD.jahr <- model.REML(GES.AD$logRR, cbind(GES.AD$Publikationsjahr,GES.AD$
Substanz), GES.AD$se.logRR))
(GES.AD.N <- model.REML(GES.AD$logRR, cbind(GES.AD$N_total,GES.AD$Substanz),
GES.AD$se.logRR))
(GES.AD.dauer <- model.REML(GES.AD$logRR, cbind(GES.AD$Behandlungsdauer_in_Wochen,
GES.AD$Substanz), GES.AD$se.logRR))
(GES.AD.anzkllinik <- model.REML(GES.AD$logRR, cbind(GES.AD$Anzahl_Kliniken,
GES.AD$Substanz), GES.AD$se.logRR))
(GES.AD.dropdiff <- model.REML(GES.AD$logRR, cbind(GES.AD$Drop.out_Differenz,
GES.AD$Substanz), GES.AD$se.logRR))
(GES.AD.advevdiff <- model.REML(GES.AD$logRR, cbind(GES.AD$
Adverse_events_differenz,GES.AD$Substanz), GES.AD$se.logRR))
(GES.AD.landkat <- model.REML(GES.AD$logRR, cbind(GES.AD$
Herkunftsland_kategorisiert_eur, GES.AD$Herkunftsland_kategorisiert_usa,
GES.AD$Substanz), GES.AD$se.logRR))
(GES.AD.sponsor <- model.REML(GES.AD$logRR, cbind(GES.AD$Sponsor_Industrie_yes,
GES.AD$Substanz), GES.AD$se.logRR))
(GES.AD.psych <- model.REML(GES.AD$logRR, cbind(GES.AD$
Psychotherapie_kategorisiert_CBT, GES.AD$Psychotherapie_kategorisiert_MET,
GES.AD$Substanz), GES.AD$se.logRR))
(GES.AD.zielbeh <- model.REML(GES.AD$logRR, cbind(GES.AD$Ziel_der_Behandlung_abst,
GES.AD$Substanz), GES.AD$se.logRR))

#####
## BACKWARD- UND FORWARD-SELECTION ##
#####

#####
## IMPLEMENTIEREN DER NÖTIGEN FUNKTIONEN ##
#####
```

```

# Funktion, die anzeigt, welche Kovariablen und in welcher Reihenfolge entfernt
# werden müssen, damit der Gewinn an Fallzahl durch weniger fehlende Werte am
# größten ist - Zur besseren Einschätzung wird auch ein Plot ausgegeben:
#####

entfernen <- function(liste,kov.namen = c("Publikationsjahr", "N_total",
"Behandlungsdauer_in_Wochen", "Anzahl_Kliniken",
"Drop.out_Differenz", "Adverse_events_differenz", "Herkunftsland",
"Sponsor_Industrie_yes", "Psychotherapie", "Ziel_der_Behandlung_abst")) {

  k <- length(liste[[1]])

  kovar <- matrix(nrow = k,data = unlist(liste))
  anfang <- sum(apply(kovar,1,function(x) all(!is.na(x))))

  länge <- length(kov.namen)
  reihenfolge <- ""
  wieviel <- 0
  temp.alt <- -1000
  j <- länge

  while(j > 1) {

    temp <- 0
    for(i in seq(along = liste)) {
      kovar <- matrix(nrow = k,data = unlist(liste[-i]))
      temp[i] <- sum(apply(kovar,1,function(x) all(!is.na(x))))
    }
    entf <- (1:j)[temp == max(temp)]

    if(length(entf) == 1 & max(temp)[1]>temp.alt[1]) {
      reihenfolge[länge+1-j] <- kov.namen[entf]
      wieviel[länge+1-j] <- max(temp)
    }

    if(length(entf)>1 & max(temp)[1]>temp.alt[1]) {
      entf.index <- 1
      if(j!= länge) {
        for(l in (length(reihenfolge)+(1:length(entf)))) {
          reihenfolge[l] <- kov.namen[entf[entf.index]]
          wieviel[l] <- max(temp)[1]
          entf.index <- entf.index+1
        }
      }
      else {
        for(l in 1:length(entf)) {
          reihenfolge[l] <- kov.namen[entf[entf.index]]
          wieviel[l] <- max(temp)[1]
          entf.index <- entf.index+1
        }
      }
      j <- j-length(entf)+1
    }

    kov.namen <- kov.namen[-entf]
    liste <- liste[-entf]
    temp.alt <- max(temp)
  }
}

```

```

    j <- j-1
  }

  plot(wieviel,xlab = "",ylab = "",ylim = c(0,k))
  abline(h = c(anfang,k))
  Entfernen <- cbind(reihenfolge,wieviel)
  colnames(Entfernen) <- c("Variable","Länge")

  return(list(Entfernen = Entfernen,Ohne.Entfernung = anfang,länge.Datensatz = k))

}

# Funktion, die das Leave-One-Out-Kreuzvalidierungskriterium berechnet:
#####

CV.model <- function(thetai, x, sds) {

  x <- as.matrix(x)
  x2 <- cbind(1,x)
  abw.stand.quadr <- 0

  for(i in seq(along = thetai)) {
    model <- model.REML(thetai[-i], x[-i,], sds[-i])
    pred <- sum(x2[i,]*model$Coef[1:(dim(x2)[2])])
    abw.stand.quadr[i] <- ((thetai[i] - pred)^2)/(model$tau2 + sds[i]^2)
  }

  return(mean(abw.stand.quadr))
}

# Backward-Selection-Funktion - ausgegeben werden die Modellterme in den
# einzelnen Selektionsschritten, die herausgefallenen Kovariablen und das beste
# Modell; außerdem werden die Werte des Kreuzvalidierungskriteriums in den
# einzelnen Schritten geplottet:
#####

backward <- function(kov.liste,kov.Namen,thetai,sds) {

  welcher <- deparse(substitute(thetai))
  welcher <- strsplit(welcher,"")[[1]]
  wo.split <- (1:length(welcher))[welcher == "$"]
  splits <- welcher[1 : (wo.split)]
  datensatz <- ""
  for(i in seq(along = splits)) {
    datensatz <- paste(datensatz,splits[i],sep = "")
  }

  kov.Namen <- paste(datensatz,kov.Namen,sep = "")

  kovar <- matrix(nrow = length(thetai),data = unlist(kov.liste))
  nicht.na <- apply(kovar,1,function(x) all(!is.na(x)))

  kov.liste <- lapply(kov.liste, function(x) as.matrix(x)[nicht.na,])

```

```

thetai <- thetai[nicht.na]
sds <- sds[nicht.na]

zahl.kov <- length(kov.liste)
b <- -1000
modelterm.CV <- ""
CV <- 0
rausgefallen <- rep("",zahl.kov-1)

for(j in zahl.kov:2) {
  CVs <- 0
  for(i in 1:j) {
    temp.kov <- matrix(nrow = length(thetai),data = unlist(kov.liste[-i]))
    CVs[i] <- CV.model(thetai,temp.kov,sds)
  }

  b <- (1:j)[CVs == min(CVs)]
  which <- (1:j)[-b]

  name <- paste("cbind(",kov.Namen[which[1]])
  for(l in which[-1]){
    name <- paste(name,kov.Namen[l],sep = ", ")
  }
  name <- paste(name,")",sep = "")

  rausgefallen[zahl.kov+1-j] <- kov.Namen[b]

  modelterm.CV[zahl.kov+1-j] <- name
  CV[zahl.kov+1-j] <- min(CVs)
  kov.liste <- kov.liste[-b]
  kov.Namen <- kov.Namen[-b]

}

plot(CV,type = "o")

return(list(modellterme = modelterm.CV,rausgefallen = rausgefallen,CV = CV,
  bestes.modell = modelterm.CV[CV == min(CV)]))
}

# Forward-Selection-Funktion - analoge Ausgabe zu der bei "backward":
#####

forward <- function(kov.liste,kov.Namen,thetai,sds) {

  welcher <- deparse(substitute(thetai))
  welcher <- strsplit(welcher,"")[[1]]
  wo.split <- (1:length(welcher))[welcher == "$"]
  splits <- welcher[1 : (wo.split)]
  datensatz <- ""
  for(i in seq(along = splits)) {
    datensatz <- paste(datensatz,splits[i],sep = "")
  }

  kov.Namen <- paste(datensatz,kov.Namen,sep = "")

  kovar <- matrix(nrow = length(thetai),data = unlist(kov.liste))

```

```

nicht.na <- apply(kovar,1,function(x) all(!is.na(x)))

kov.liste <- lapply(kov.liste, function(x) as.matrix(x)[nicht.na,])
thetai <- thetai[nicht.na]
sds <- sds[nicht.na]

zahl.kov <- length(kov.liste)
b <- -1000
modellterme <- ""
CV <- 0
dazugekommen <- rep("",zahl.kov)
name.neu <- ""

for(j in zahl.kov:1) {

  CVs <- 0
  if(j == zahl.kov) {
    for(i in 1:j) {
      CVs[i] <- CV.model(thetai,kov.liste[[i]],sds)
    }
    kovar.modell <- kov.liste[[1:j][CVs == min(CVs)]]
  }
  else {
    for(i in 1:j) {
      kovars <- cbind(kovar.modell,kov.liste[[i]])
      CVs[i] <- CV.model(thetai,kovars,sds)
    }
  }
  b <- (1:j)[CVs == min(CVs)]

  if(j == zahl.kov)
    name.neu <- kov.Namen[b]
  else
    name.neu <- paste(name.neu,kov.Namen[b],sep = ", ")

  dazugekommen[zahl.kov+1-j] <- kov.Namen[b]
  modellterme[zahl.kov+1-j] <- name.neu
  CV[zahl.kov+1-j] <- min(CVs)

  if(j!= zahl.kov)
    kovar.modell <- cbind(kovar.modell,kov.liste[[b]])

  kov.liste <- kov.liste[-b]
  kov.Namen <- kov.Namen[-b]

}

modellterme <- paste("cbind(",modellterme,")",sep = "")

plot(CV,type = "o")

return(list(modellterme = modellterme,dazugekommen = dazugekommen,CV = CV,
  bestes.modell = modellterme[CV == min(CV)]))
}

```

```
#####
```

```

# BACKWARD- UND FORWARD-SELECTIONS FÜR DIE KONSTRUKTE GETRENNT NACH SUBSTANZ #
#####

# EXEMPLARISCH FÜR NAL.AD; DIE RESTLICHEN SELEKTIONEN WURDEN ANALOG DURCH-
# GEFÜHRT

# Überprüfen, welche Kovariablen aus praktischer Sicht entfernt werden müssten:
#####

entfernen(list(NAL.AD$Publikationsjahr, NAL.AD$N_total, NAL.AD$
  Behandlungsdauer_in_Wochen, NAL.AD$Anzahl_Kliniken, NAL.AD$Drop.out_Differenz,
  NAL.AD$Adverse_events_differenz, cbind(NAL.AD$Herkunftsland_kategorisiert_eur,
  NAL.AD$Herkunftsland_kategorisiert_usa), NAL.AD$Sponsor_Industrie_yes,
  cbind(NAL.AD$Psychotherapie_kategorisiert_CBT,
  NAL.AD$Psychotherapie_kategorisiert_MET), NAL.AD$Ziel_der_Behandlung_abst))

# Adverse-Events-Differenz wurde herausgenommen:

liste.alle <- list(NAL.AD$Publikationsjahr, NAL.AD$N_total, NAL.AD$
  Behandlungsdauer_in_Wochen, NAL.AD$Anzahl_Kliniken, NAL.AD$Drop.out_Differenz,
  cbind(NAL.AD$Herkunftsland_kategorisiert_eur, NAL.AD$
  Herkunftsland_kategorisiert_usa), NAL.AD$Sponsor_Industrie_yes, cbind(NAL.AD$
  Psychotherapie_kategorisiert_CBT, NAL.AD$Psychotherapie_kategorisiert_MET),
  NAL.AD$Ziel_der_Behandlung_abst)

namen.alle <- c("Publikationsjahr", "N_total", "Behandlungsdauer_in_Wochen",
  "Anzahl_Kliniken", "Drop.out_Differenz", "Herkunftsland",
  "Sponsor_Industrie_yes", "Psychotherapie", "Ziel_der_Behandlung_abst")

# Backward-Selection:
#####

backward(liste.alle,namen.alle,NAL.AD$logRR,NAL.AD$se.logRR)

# Forward-Selection:
#####

forward(liste.alle,namen.alle,NAL.AD$logRR,NAL.AD$se.logRR)

# Modell aus Backward-Selection:
#####

model.REML(NAL.AD$logRR,cbind(NAL.AD$Publikationsjahr,NAL.AD$N_total,NAL.AD$
  Drop.out_Differenz, NAL.AD$Ziel_der_Behandlung_abst, cbind(NAL.AD$
  Psychotherapie_kategorisiert_CBT, NAL.AD$Psychotherapie_kategorisiert_MET)),
  NAL.AD$se.logRR)

# Modell aus Forward-Selection:
#####

model.REML(NAL.AD$logRR,cbind(NAL.AD$Publikationsjahr, NAL.AD$Drop.out_Differenz,
  NAL.AD$Ziel_der_Behandlung_abst, cbind(NAL.AD$Herkunftsland_kategorisiert_eur,

```

```

NAL.AD$Herkunftsland_kategorisiert_usa), cbind(NAL.AD$
Psychotherapie_kategorisiert_CBT,NAL.AD$Psychotherapie_kategorisiert_MET)),
NAL.AD$se.logRR)

# ohne nicht-signifikante Kovariablen:

model.REML(NAL.AD$logRR,cbind(NAL.AD$Publikationsjahr,NAL.AD$Drop.out_Differenz),
NAL.AD$se.logRR)

#####
## FOR- UND BACKWARDSELECTIONS ÜBER ALLE STUDIEN MIT SUBSTANZ ALS KOVARIABLE ##
#####

# EXEMPLARISCH FÜR GES.AD

entfernen(list(GES.AD$Publikationsjahr, GES.AD$N_total, GES.AD$
Behandlungsdauer_in_Wochen, GES.AD$Anzahl_Kliniken, GES.AD$Drop.out_Differenz,
GES.AD$Adverse_events_differenz, cbind(GES.AD$Herkunftsland_kategorisiert_eur,
GES.AD$Herkunftsland_kategorisiert_usa), GES.AD$Sponsor_Industrie_yes,
cbind(GES.AD$Psychotherapie_kategorisiert_CBT, GES.AD$
Psychotherapie_kategorisiert_MET), GES.AD$Ziel_der_Behandlung_abst))

# Ziel der Behandlung musste rausgenommen werden:

# Neu ist "Substanz_NAL"

liste.alle <- list(GES.AD$Publikationsjahr, GES.AD$N_total, GES.AD$
Behandlungsdauer_in_Wochen, GES.AD$Anzahl_Kliniken, GES.AD$Drop.out_Differenz,
GES.AD$Adverse_events_differenz, cbind(GES.AD$Herkunftsland_kategorisiert_eur,
GES.AD$Herkunftsland_kategorisiert_usa), GES.AD$Sponsor_Industrie_yes, cbind(
GES.AD$Psychotherapie_kategorisiert_CBT, GES.AD$
Psychotherapie_kategorisiert_MET), GES.AD$Substanz_NAL)

namen.alle <- c("Publikationsjahr", "N_total", "Behandlungsdauer_in_Wochen",
"Anzahl_Kliniken", "Drop.out_Differenz", "Adverse_events_differenz",
"Herkunftsland", "Sponsor_Industrie_yes", "Psychotherapie", "Substanz_NAL")

backward(liste.alle,namen.alle,GES.AD$logRR,GES.AD$se.logRR)
forward(liste.alle,namen.alle,GES.AD$logRR,GES.AD$se.logRR)

# Modell aus Backward- und auch Forward-Selection:
#####

model.REML(GES.AD$logRR,cbind(GES.AD$Publikationsjahr, GES.AD$Anzahl_Kliniken,
GES.AD$Drop.out_Differenz),GES.AD$se.logRR)

```

Literaturverzeichnis

- S. Bachim (2008). ‘Metaanalyse zu Behandlungseffekten bei Störungen des Sozialverhaltens’. Diplomarbeit, Universität Trier, München.
- B. Baujat, et al. (2002). ‘A graphical method for exploring heterogeneity in meta-analyses: application to a meta-analysis of 65 trials’. *Statistics in Medicine* **21**:2641–2652.
- C. S. Berkey, et al. (1995). ‘A random-effects regression model for meta-analysis’. *Statistics in Medicine* **14**:395–411.
- M. Blettner & P. Schlattmann (2005). ‘Meta-Analysis in Epidemiology’. In W. Ahrens & I. Pigeot (eds.), *Handbook of Epidemiology*, pp. 829–858. Springer, Berlin.
- J. Bowden, et al. (2006). ‘Using pseudo-data to correct for publication bias in meta-analysis’. *Statistics in Medicine* **25**:3798–3813.
- D. M. Caldwell, et al. (2005). ‘Simultaneous comparison of multiple treatments: combining direct and indirect evidence’. *British Medical Journal* **331**:897–900.
- A. Cebulla (2008). ‘Meta-Analyse - Eine Einleitung’. *Sozialer Fortschritt* **10/11**:249–250.
- H. M. Cooper & L. V. Hedges (1994). *The Handbook of Research Synthesis*. Russell Sage Foundation, New York.
- J. J. Deeks, et al. (2008). ‘Analysing data and undertaking meta-analysis’. In J. P. Higgins & S. Green (eds.), *Cochrane Handbook for Systematic Reviews of Interventions*, pp. 243–296. John Wiley & Sons, Chichester.
- R. M. Harbord, et al. (2006). ‘A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints’. *Statistics in Medicine* **25**:3443–3457.

- L. V. Hedges (1983). 'A random effects model for effect sizes'. *Psychological Bulletin* **93**:388–395.
- J. P. T. Higgins & S. G. Thompson (2002). 'Quantifying heterogeneity in a meta-analysis'. *Statistics in Medicine* **21**:1539–1558.
- J. P. T. Higgins, et al. (2003). 'Measuring inconsistency in meta-analyses'. *British Medical Journal* **327(7414)**:557–560.
- M. Lipsey & D. Wilson (2001). *Practical Meta-Analysis*. Thousands Oaks: Sage, New Dehli.
- G. Lu & A. E. Ades (2004). 'Combination of direct and indirect evidence in mixed treatment comparisons'. *Statistics in Medicine* **23**:3105–3124.
- S. G. Moreno, et al. (2009). 'Novel methods to deal with publication biases: secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications'. *British Medical Journal* **339**:b2981 – b2981.
- D. B. Petitti (1994). *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis*. Oxford University Press, New York.
- R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- S. Rösner (2006). 'Meta-Analyse zur Wirksamkeit von Acamprosate und Naltrexon in der Entwöhnungsbehandlung alkoholabhängiger Patienten'. Doktorarbeit, Universität München.
- W. Sauerbrei & M. Blettner (2003). 'Issues of traditional reviews and meta-analysis of observational studies in medical research'. In R. Schulze, H. Holling, & D. Böhning (eds.), *Metaanalysis. New developments and applications in medical and social sciences*, pp. 79–98. Hogrefe & Huber, Göttingen.
- G. Schwarzer, et al. (2003). 'Statistical Tests for the Detection of Bias in Meta-Analysis'. In R. Schulze, H. Holling, & D. Böhning (eds.), *Metaanalysis. New developments and applications in medical and social sciences*, pp. 71–78. Hogrefe & Huber, Göttingen.
- S. J. Sharp & S. G. Thompson (2000). 'Analysing the relationship between treatment effect and underlying risk in meta-analysis: comparison and development of approaches'. *Statistics in Medicine* **19**:3251–3274.
- K. Sidik & J. N. Jonkman (2007). 'A comparison of heterogeneity variance

- estimators in combining results of studies'. *Statistics in Medicine* **26**:1964–1981.
- M. C. Simmonds & J. P. T. Higgins (2007). 'Covariate heterogeneity in metaanalysis: Criteria for deciding between meta-regression and individual patient data'. *Statistics in Medicine* **26**:2982–2999.
- F. Song, et al. (2003). 'Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses'. *British Medical Journal* **326**:472–476.
- D. Spiegelhalter, et al. (2007). *WinBUGS User Manual: Version 1.4.3*. MRC Biostatistics Unit, Cambridge, UK.
- T. D. Stanley (2005). 'Beyond Publication Bias'. In C. Roberts & T. D. Stanley (eds.), *Meta-Regression Analysis: Issues of Publication Bias in Economics*, pp. 15–51. Blackwell Publishing, Oxford.
- L. A. Stewart & M. J. Clarke (1995). 'Practical methodology of meta-analysis (overviews) using updated individual patient data'. *Statistics in Medicine* **14**:2057–2079.
- A. J. Sutton & J. P. T. Higgins (2008). 'Recent developments in meta-analysis'. *Statistics in Medicine* **27**:625–650.
- S. G. Thompson & J. P. T. Higgins (2002). 'How should meta-regression analyses be undertaken and interpreted?'. *Statistics in Medicine* **21**:1559–1573.
- S. G. Thompson & S. J. Sharp (1999). 'Explaining heterogeneity in meta-analysis: a comparison of methods'. *Statistics in Medicine* **18**:2693–2708.
- S. G. Thompson, et al. (1997). 'Investigating underlying risk as a source of heterogeneity in meta-analysis'. *Statistics in Medicine* **16**:2741–2758.
- J. Tintelnot & A. Bausch (2003). 'Internationalisierung und Erfolg bei kleinen und mittleren Unternehmen (KMU) - Eine Meta-Analyse'. Diplomarbeit, Universität Gießen. (Online) Stand: 05.09.09. URL: <http://ebme.uni-giessen.de>.
- X. Zhou, et al. (2003). 'Meta-Analysis'. In Y. Lu & J. Fang (eds.), *Advanced Medical Statistics*, pp. 233–317. World Scientific, River Edge, New Jersey.

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

München, den 16.09.2009

(Roman Hornung)